Università di Pisa
Sant'Anna Scuola Universitaria Superiore Pisa
Consiglio Nazionale delle Ricerche
Università degli Studi di Cagliari

Book of Short Papers

SIS 2021

SIS 2021
PISA, 21-25 GIUGNO 2021

SIS Società Italiana di Statistica

Editors: Cira Perna, Nicola Salvati and Francesco Schirripa Spagnolo

BANCA DI PISA E FORNACETTE
CREDITO COOPERATIVO
Banca aderente al Gruppo Bancario Cooperativo Iccrea

TStat
Distribuzione Software | Formazione Professionale
Statistica | Economia | Finanza | Biostatistica | Epidemiologia
Sanità Pubblica | Scienze Sociali
www.tstat.it | www.tstattraining.eu

Copyright © 2021

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891927361

Contents

Preface

For more than a year, the Covid-19 pandemic has hit our most consolidated habits with serious challenges on social and economic system. Implementation of the guidelines for social distancing has led to the shifting of most of the research activities remotely. After very careful consideration, concerning the health of all conference participants and the restricted mobility of the staff of many universities and research centres, the Executive Board of the Italian Statistical Society (SIS) and the Local Organizing Committee decided to schedule the 50th Meeting of the Italian Statistical Society in remote from the 21st to the 25th of June 2021. The Conference is streamed through the Microsoft Teams platform provided by the University of Cagliari.
The conference program includes 4 plenary sessions, 15 specialized sessions, 20 solicited sessions, 37 contributed sessions and the poster exhibition. The meeting will also host three Satellite Events on 'Measuring uncertainty in key official economic statistics', 'Covid-19: the urgent call for a unified statistical and demographic challenge' and 'Evento SIS-PLS Statistica in classe: verso un insegnamento laboratoriale'. The first one, scheduled for June 17th, has been organized by prof. Tiziana Laureti and is streamed via the Zoom platform. The second satellite event, scheduled for June 18th, has been organized by the Young SIS Group and is streamed via the Zoom platform. The third satellite event is scheduled for July 8th and has been organised by prof. Laura Ventura and is hosted on Pearson's platform. A panel discussion, organized by Linda Laura Sabbadini has also been included in the program.
The conference committee had registered 323 accepted submissions, including 128 to be presented in invited plenary, specialized and solicited sessions, and 195 spontaneously submitted for oral and poster sessions.

This volume gathers most of the peer-reviewed papers submitted to the 50th Meeting of the Italian Statistical Society and presented at the virtual Conference. It is organized into 6 chapters corresponding to the plenary, specialized, the solicited sessions, satellite events, and to the general topics for contributed papers and posters. The volume covers a wide variety of subjects ranging from methodological and theoretical contributions, to applied works and case studies, giving an excellent overview of the interests of the Italian statisticians and their international collaborations.

Of course, both the SIS Conference and this volume would not be possible without the collaboration of the members of the Scientific Committee and the members of the Università di Pisa, Scuola Superiore Sant'Anna and National Research Council of Pisa. Members of these three institutions took part actively in the Local Organizing Committee. The conference also received support from sponsors, namely TStat and Banca di Pisa e Fornacette. To all of them, our thanks.
We would also like to thank the University of Cagliari for the IT support provided for the organization of the online event and in particular Prof. Francesco Mola, Rector of the University of Cagliari and our esteemed colleague, and Dr. Roberto Barreri, manager of the Systems, Infrastructure and Data Department.

Our thanks also go to all contributors for having submitted their work to the conference, the members of the Program Committee and the extra reviewers for their efforts in this difficult period. Finally, we wish to express our gratitude to the publisher Pearson Italia for all the support received.

Cira Perna, Università degli Studi di Salerno (Chair of the Program Committee)
Nicola Salvati, Università di Pisa (Chair of the Local Organizing Committee)
Francesco Schirripa Spagnolo, Università di Pisa

Program Committee: Cira Perna (Chair), Massimiliano Caporin, Paola Cerchiello, Fabio Divino, Leonardo Egidi, Raffaele Guetto, Francesco Lagona, Silvia Loriga, Francesca Lotti, Mariagiulia Matteucci, Alessio Pollice, Maria Giovanna Ranalli, Nicola Salvati, Luca Secondi, Isabella Sulis, Luca Tardella, Donatella Vicari, Domenico Vistocco

Local Organizing Committee: Nicola Salvati (Chair), Gaia Bertarelli, Bruno Cheli, Paolo Frumento, Fosca Giannotti, Caterina Giusti, Piero Manfredi, Stefano Marchetti, Lucio Masserini, Vincenzo Mauro, Barbara Pacini, Dino Pedreschi, Francesco Schirripa Spagnolo, Chiara Seghieri.

Technical Committee: Luca Frigau (Chair), Silvia Columbu, Giulia Contu, Mara Manca, Marco Ortu, Maurizio Romano, Cristian Usala, Gianpaolo Zammarchi.

Organizers of Specialized and Solicited Sessions: Giada Adelfio, Emanuele Aliverti, Michela Battauz, Marco Bee, Mauro Bernardi, Luisa Bisaglia, Claudio Ceccarelli, Claudio Conversano, Marco Corazza, Giovanni Battista Dagnino, Antonella D'Agostino, Giovanni D'Alessio, Francesco Finazzi, Michela Gnaldi, Luigi Grossi, Roberto Impicciatore, Giovanna Jona Lasinio, Brunero Liseo, Orietta Luzi, Lucio Masserini, Letizia Mencarini, Stefania Mignani, Gianna Monti, Nazareno Panichella, Pierfrancesco Perri, Lea Petrella, Alessandra Petrucci, Paolo Righi, Elena Stanghellini, Venera Tomaselli, Cristina Tortora, Rosanna Verde, Paola Vicard, Maria Prosperina Vitale, Alberto Zazzaro.

1 Plenary Sessions

Citizen Data and Citizen Science: a challenge for Official Statistics

Dati dei cittadini e Citizen Science: una sfida per la Statistica ufficiale

Monica Pratesi

Abstract Citizen Data and Citizen Science are undoubtedly a challenge and an opportunity for Official Statistics. The paper follows the evolution in the production of statistics and indicators and give some hints for using Citizen data in the production of indicators for monitoring the achievement of SDGs

Abstract *Dati dei cittadini e Citizen science sono sfide ed opportunità per la Statistica Ufficiale. Nel lavoro si segue l'evoluzione nella produzione di dati ed indicatori e si descrivono i primi passi per l'uso dei dati dei cittadini nella produzione di indicatori per il monitoraggio degli obiettivi di sviluppo sostenibile (SDG).*

Key words: Citizen Data, Citizen Science, SDGs

1 Introduction: the Next Generation data

In the last ten years official statisticians have been discussing on the impact of the Big Data in the production of Official Statistics (OS), highlighting many advantages and also disadvantages of their use. The main question was and is: "What is the future of Official Statistics in the Big data era?"

A lot has been done for the use of big data in OS by the International and National Statistical Institutes (NSIs), including the Istat. My contribution to the debate was initially on model based estimates using big data sources (Marchetti et al, 2015; 2016), then, in my capacity of President of Italian Statistical Society (SIS), I intervened on the error profile of Big Data (Pratesi, 2017; 2018). Since last year, I have been focusing on Citizen Science as the global process of digitization is so pervasive that times are mature for studying how to using and reusing Citizen Data in the production of OS (Pratesi, 2021).

As a matter of fact Official Statistics have always been evolving and the term "Trusted Smart Statistics" (TSS) was put forward by Eurostat and officially adopted by the European Statistical System (ESS) in 2018 in the so-called Bucharest memorandum to signify this evolution. But using big data, smart statistics, citizen data and citizen science in producing OS, could it be a danger? Would OS be under attack either by discussions on trust or by competition with statistics produced with lower quality? For this, the official statisticians of the future have to be more than just data engineers (Radermacher, 2019).

These data are nothing more and nothing less than Next Generation Data and following the same evolution track in data production process than NSIs have always followed we will answer to the above questions. The evolution track – presented in Section 2 - has always guaranteed trust in data collection and processing. In the sections 3 and 4, the challenges of Citizen Data (CD) and Citizen Science (CS) to OS are discussed. Finally, in section 5, a project on the use of Citizen Science and Citizen Data to estimate BES indicators (Equitable Sustainable indicators), which will be implemented by Istat, is recalled.

2 The evolution process for producing Statistics and Indicators

The mission of OS has always been to provide a quantitative representation of the society, economy, and environment for purposes of public interest, for policy design and evaluation and as basis for informing the public debate. The production of modern OS is based on a system of scientific methods, regulations, codes, practices, ethical principles, and institutional settings that was developed through the last two centuries at the national level in parallel to the developments of modern states (Ricciato et al, 2019).

The Figure 1 illustrates the evolution mechanism of the production process of a general OS system (engine), with its data sources (fuels) and User's information needs (accelerators). We see immediately that statistics and indicators are influenced both by fuels and accelerators. The rise of new data sources can give new fuels for Statistics and Indicators, but it can also act as a multiplier as it provokes new data information needs, becoming accelerators that stimulate further needs to be satisfied. Moreover, the statistical methods and the rules, for example, to guarantee the privacy and the trust on Statistics and Indicators produced, are obviously to adapt to the characteristics of the various data sources.

It is evident that this scheme of the evolution of the OS production process covers the current situation, but it is also valid for all the various breakthrough periods of data collection and statistical production lived by the NSIs.

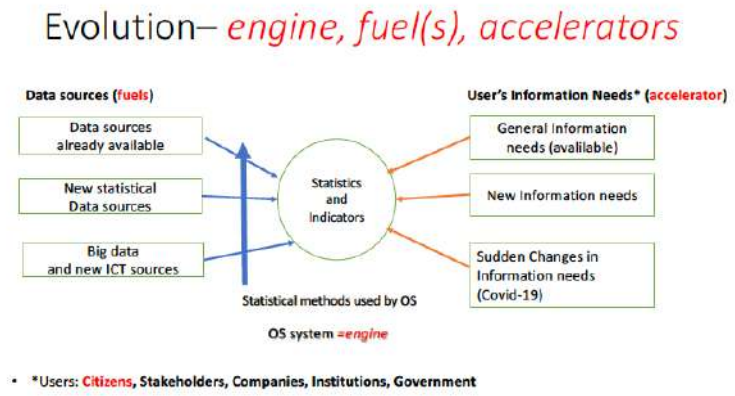


Figure 1: Evolution of the production process of a general OS system

For example, in Italy a sudden change in the information needs happened after the II world war. The government, policy makers and all the stakeholders needed new statistics to reconstruct economic situation (Marshall and Fanfani Plans) and the OS reacted designing and carrying out surveys in different domains, in particular for the construction of the National Accounts, developing new structured and standardized survey methods.

Therefore, the new scenario of data sources outlined in the introduction, moves the evolution mechanism, affecting for example the roles of the various stakeholders and their mutual relationships, to reply also to the questions by the National Recovery and Resilience Plan (NRRP). Summarizing there is a need for timely reactions, both in terms of necessary reorganizations of the NSIs and publications of the first, even provisional results of the data collection process, as Experimental Statistics.

3 The Citizen Data and the Citizen Science: a challenge for OS

As already said in the introduction, Big data, smart statistics and citizen are inseparable: from smartphones, meters, fridges and cars to internet platforms, the data of most digital technologies is Citizen Data, that is the data of the citizens and on the citizens.

In addition to raising political and ethical issues of privacy, confidentiality and data protection, the repurposing of big data call for rethinking relations between the citizens and the production of official statistics, if they are to be trusted. I am convinced that the future of Official Statistics does not depend only on the possibility to use new sources of data or new methods, but also on the possibilities that the new digital technologies offer to establish new relationships with the citizens. Their role is destined to evolve from that of respondents to that of

collaborators and co-producers of official statistics data (Ruppert, E., et al., 2018; Ruppert, E., 2019).

First, the possibility to exploit Citizen-generated data (CGD) - that are produced by non-state actors, particularly individual or civil society organizations - for official statistical purposes seems to represent a very promising avenue to collect timely and relevant new data. Privacy issues prevent citizens to fully disclosure this kind of data, while their management and storage by privately owned digital platforms generate some remarkable concerns by citizens themselves on their correct protection. In order to fully exploit this kind of data, NSIs need to develop a better understanding on the way they are generated and how can be made accessible for official purposes (Casarez-Crageda et al 2020).

The second approach, that aims at the direct collaboration of the citizens in producing OS, following the principles of the Citizen Science (CS) involves citizens along all the phases of the so-called data value chain: planning, collection, processing, analysis and use (Nascimento et al 2018). This is an important involvement that we can also link to the Post Normal Science approach (Pratesi, 2020).

The general opportunity for OS resides in gaining a new awareness of citizens in their participation to the process of official data production. Rethinking citizen involvement along the phases of the data value chain can help counter the trust deficit between citizens and governments and consequently establish a participatory data ecosystem (Misra and Schmidt, 2020)

The use phase in the data value chain requires an uptake stage that involves three activities: connecting data to users; incentivizing users to incorporate data into the decision-making process; and influencing them to value data. The active involvement of citizens in the data production is a challenge for OS to reduce the gap between users and producers. In my opinion it would also have a positive feedback on Statistical literacy, as the ability of data users to interpret and critically evaluate statistical information in a variety of contexts .

It is clear that the concept of citizen data and co-production raise practical and political questions that it is impossible to summarized here.

Moreover, CS produces data difficult to compare, the measures of precision are not clear. The challenge for OS resides in the rethinking the data collection process and of the concept of quality of the data. Traditional aspects inherent to the data production process and that are typical when NSIs conceive and govern it such as accuracy, timeliness, representativeness, completeness, etc. should be rethought and enriched introducing also other aspects. These last are important when the NSIs are not in the position to interfere in every aspect of the production process of the data as: evaluation of self-selection bias, quality checks post-production, evaluation of potential use of the data. Issues as comparability across domains, coherence and benchmarking will be even more important than in traditional data production settings. Even if there are proposals to define the quality profile of citizen-generated data, there are not yet comprehensive and meaningful empirical studies.

To fill this gap we need research in OS to generate many Experimental Statistics, producing results also by unusual tools such as inference from nonprobability

samples, data integration and data fusion of new and traditional data, model based and model assisted estimation methods.

This is true for all the thematic areas where OS is called to produce data: from economic life, as consumption expenditure, earning and usage of disposable income to the aspects of daily life, like access to public services, life-long education, participation in social and cultural life.

4 A project on the measure the quality of Citizen Data for the compilation of SDGs indicators

An important area, essential for regeneration and government in this difficult moment marked by the pandemic, is that of the Sustainable Development Goals (SDGs). SDGs are objectives included in the government programs of European countries. A recent review highlighted how data collected through CS initiatives can feed an important part of indicators for monitoring the Sustainable Development Goals (Fraisl D. et al, 2020). Among the European countries where this practice is widespread, Italy is missing.

However, the possibility for NSI to successfully use CGD data for official statistical production in general, and for the set up and maintenance of SDG indicators in particular, has to meet the essential condition that their quality for statistical purposes can be carefully assessed and their potential biases corrected using a consistent methodological and statistical data processing approach.

Table 1: Experimental settings to test the quality of CGD for the compilation of SDG indicators

Area of reference of SDG indicators	Specific topic under investigation	Characteristics of units reporting CDG data	Availability of additional information	Methodology to test for the quality of CGD
Goal 1 – Poverty	Poverty, material deprivation,	Linkable to Business Register.	EuSilc and Household Budget Survey	Record linkage, Statistical matching, Latent variables models.
Goal 4 - Education	Informal education (i.e. cinema, read books, theatre), soft skills.	Linkable to Business Register.	Labour Force Survey, educational registers and others administrative sources.	Record linkage, Statistical matching, Latent variables models.
Goal 2 – Food security improve nutrition	Food waste	Linkable to Business Register.	Household Budget Survey, Aspects of daily life.	Record linkage, Statistical matching, Latent variables models.

Table 1 highlights some possible experimental settings that can be established by ISTAT to test the quality of CGD data for the compilation of SDG indicators.
The work is in progress and the settings in the table are the initial step of a complex experiment (Pratesi et al, 2021).

Istat has a leading role in Europe for the production of Equitable and Sustainable Well-being (BES) indicators. The recent presentation of the BES Report of 10 March 2021 testifies to this. I believe that CS is a path to further improve the Institute's contribution.

References

Cázarez-Grageda, K., Schmidt, J., Ranjan, R. (2020) Reusing Citizen-Generated Data For Official Reporting A quality framework for national statistical office-civil society organisation engagement PARIS21 Working Paper

Fraisi, D., Campbell, J., See, L., When, U., Wardlaw, J., Gold, M., Moorthy, I., Arias, R., Piera, J., Oliver, J.L., Masò, J., Penker, M., Fritz, S. (2020), Mapping citizen science contributions to the UN sustainable development goals, Sustainable Science, 15, pp. 1735-1751

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinizivillo, S., Pappalardo, L., and Gabrielli, L. (2015), Small area model-based estimation using big data sources, in Journal of Official Statistics, 31, pp. 263-281

Marchetti, S., Giusti C., Pratesi M (2016), The use of Twitter data to improve small area estimates of households' share of food consumption expenditure in Italy, in AStA Wirtschafts- und Sozialstatistisches Archiv: 57 10 (2-3) 60 61 July 2016

Nascimento, S., Iglesias, J.M.R., Owen, R., Schade, S., Shanley, L. (2018), Citizen Science for policy formulation and implementation, chapter 16 in: Hecher, S,. Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (2018), Citizen Science: innovation in Open Science, Society and Policy, UCI Press London

Misra, A. and Schmidt, J. (2020), Enhancing trust in data – participatory data ecosystems for the post-COVID society, in Shaping The Covid-19 Recovery: Ideas From Oecd's Generation Y And Z © OECD 2020

Pratesi, M., (2017), Big Data: the point of view of a Statistician, Etica e Economia, 12/4

Pratesi, M. (2018), Statistica: linguaggio sovradisciplinare per comprendere e dare valore ai dati talk in the Conference on "Data to Change" held on January 15, 2018 at the Italian House of Representatives, in Statistica&Società, 2018

Pratesi, M. (2020), Parlare chiaro: statistica, dati e modelli, talk in "Parlare chiaro, i rischi della confusione dei numeri", online workshop, 30 aprile 2020, Università Politecnica delle Marche

Pratesi, M., (2021), Official Statistics and Citizen Science, Seminar held March, 18, 2021, http://www.centrodagum.it/en/seminario-scuola-dei-dottorati-delle-scienze-sociali-universita-di-firenze/

Pratesi M, Ceccarelli C, Menghinello S. (2021), Citizen generated data and Official Statistics: an application to SDGs indicators, Discussion paper n 274, Department of Economics and Management, University of Pisa.

Radermacher W. (2019), Governing-by-the-numbers/Statistical governance: Reflections on the future of official statistics in a digital and globalized society, Statistical Journal of the IAOS,

Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., Skaliotis, M. (2019), Trusted smart statistics: Motivations and principles, Statistical Journal of the IAOS, 35, pp.589-603

Ruppert, E., Grommé, F., Ustek-Spilda, F., Cakici, B. (2018), Citizen Data and Trust in Official Statistics, Economie et Statistique/Economics and Statistics, N° 505-506, pp179-193

Ruppert E. (2019), Different data futures: An experiment in citizen data, Statistical Journal of the IAOS, 35, pp. 633-641

2 Specialized Sessions

2.1 A glimpse of new data and methods for analysing a rapidly changing population

The diffusion of new family patterns in Italy: An update

Arnstein Aassve, Letizia Mencarini, Elena Pirani, Daniele Vignoli

Abstract Recent trends in cohabitation, divorce and out-of-wedlock childbearing, show that Italy has entered a stage of rapid family change. The aim of this paper is twofold. We first document those trends, holding them up against rates of tertiary education and female labor force participation, and indicators of secularization. Second, we use the recently launched Household Multipurpose Survey on Family and Social Subjects to study the underlying drivers of those macro trends. The survey provides a unique opportunity to assess the diffusion of new family behaviours among young Italians.

Abstract Le recenti dinamiche nei comportamenti familiari – divorzio, convivenza non matrimoniale, figli al di fuori del matrimonio – mostrano che anche l'Italia è entrata in una fase di rapidi cambiamenti. In questo lavoro, documentiamo questi cambiamenti, affiancandoli alle tendenze che hanno caratterizzano il contesto socio-economico italiano negli ultimi decenni, in termini di tassi di istruzione femminile a partecipazione delle donne al mercato del lavoro, e di secolarizzazione. In secondo luogo, utilizziamo l'indagine Istat su Famiglie e Soggetti Sociali per studiare i fattori alla base di queste tendenze. L'indagine rappresenta infatti un'opportunità unica per valutare la diffusione di nuovi comportamenti familiari tra i giovani italiani.

Key words: Second Demographic Transition, Italy, Multipurpose Survey, cohabitation, divorce, education, female labour force participation.

[1] Arnstein Aassve, Bocconi University; email: arnstein.aassve@unibocconi.it
Letizia Mencarini, Bocconi University; email: letizia.mencarini@unibocconi.it
Elena Pirani, University of Florence; email: elena.pirani@unifi.it
Daniele Vignoli, University of Florence; email: daniele.vignoli@unifi.it

1 Introduction

The underlying idea of the Second Demographic Transition (SDT) is that in Western societies, spearheaded by the Nordic ones, the centrality of the family is declining, being replaced by support for more liberal demographic behaviours, such as divorce, cohabitation and non-marital childbearing (Van de Kaa 1987). These new demographic behaviours are viewed as progressive independence of individuals who give growing importance to self-realization and their psychological well-being and to their personal freedom of expression (Van de Kaa 1987). The rise of individualism and secularization has, accordingly, led to shifts in the moral code, enabling major changes in family behaviour (Lesthaeghe, 2020). The source of this ideational change is, however, often elusive (Ruggles, 2012), and has generally been interpreted in terms of diffusion processes of ideas and attitudes (Casterline, 2001). Increasing female economic empowerment is also seen as an important driver for the emergence of new family behaviours (Lesthaeghe, 2020), though this view is disputed by some (Oppenheimer, 1994).

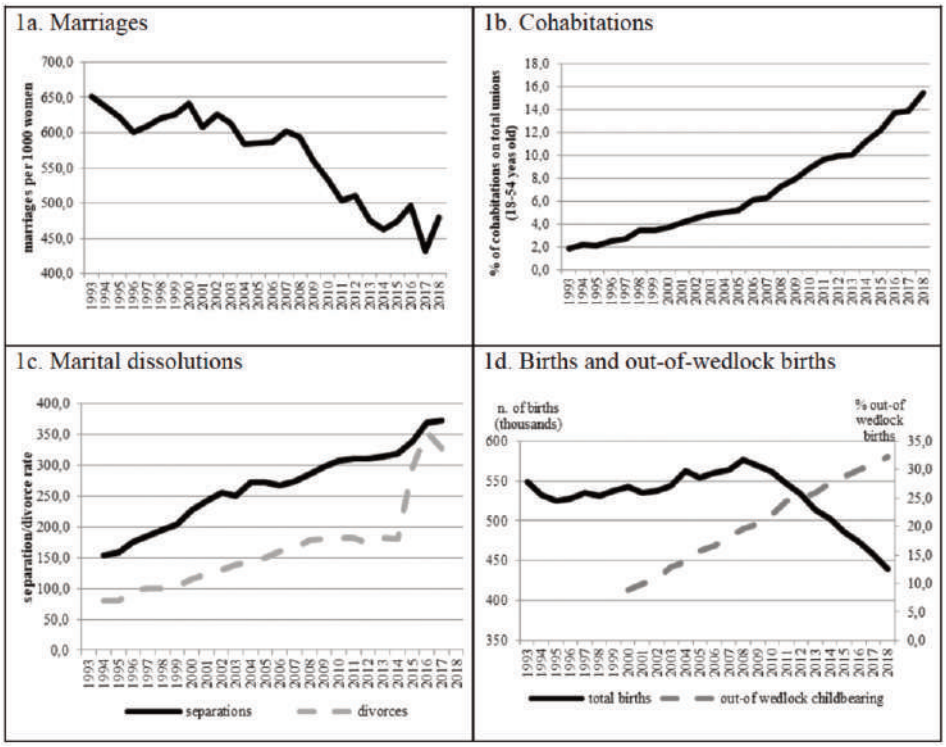
Italy has for many been viewed as the antidote to those broad demographic trends, a society where the family has remained pivotal, intergenerational co-residence remaining prevalent and where traditional attitudes towards demographic behaviour has prevailed. Italy belongs to the so-called "Southern or Mediterranean model", characterized by a very low level of social protection and by strong family ties (e.g., Reher, 1998; Viazzo, 2003). These countries are consequently classified as "traditional" in terms of value orientations, a feature not least caused by the prevalent role of the Roman Catholic Church (Caltabiano et al., 2006; Vignoli and Salvini 2012). In light of these characteristics, some have argued that the adoption of new "innovative" family behaviours, as observed in so many other countries, may not materialize in Italy - or at least - not reach the same levels as seen elsewhere (e.g., Reher, 1998; Nazio & Blossfeld, 2003).

Despite those familistic features, however, Italy did stand out as one of the first examples where fertility declined to unprecedented levels, giving rise to the term lowest-low fertility (Kohler et al., 2002), a pattern followed by a tremendous postponement of childbearing. Today the mean age of childbearing among Italian women stands at 32 years and the Total Fertility Rate is below 1.3. The contrast with the Nordic countries is startling, where new demographic behaviour is followed by "healthy" fertility rates. However, recent trends suggest that the other Second Demographic Transition indicators, such as cohabitation, out-of-wedlock childbearing and divorce, are now changing rapidly, suggesting that young Italians are moving closer to the behaviour of their Nordic counterparts (Pirani and Vignoli 2016; Vignoli et al. 2016). This article offers a general overview about the diffusion of new family-related behaviours in Italy, contesting the widely held view that Italy is a homogeneous family-oriented country.

2 Recent trends in Italy

From the Italian Statistical Office (ISTAT), we document recent trends of family behaviours. Figure 1 shows the trends in main family behaviours of the last 25 years. Although marriage continues to be central and popular among Italian couples, figures show clearly that it is no longer the unique way to form a relationship. The decline began slowly and with an irregular pace in the late 90s, but in 2008 the marriage rate started an unexpected and fast decline, likely intensified by the Great Recession (Figure 1a). From about 600 marriages every 1000 women registered in 2008, we passed to less than 500 in 2018. In addition, during the last two decades the incidence of marriages made through a civil ceremony (among all marriages) increased from less than 20% to 50% (they were only 2% in early 1970s). A non-religious marriage clearly represents a secularized choice, suggesting that traditional attitudes and norms imposed through the Roman Catholic Church are weakening. Simultaneously, non-marital unions have become popular among young Italians (Figure 1b).

Figure 1: Trends in family behaviours: Italy, 1993-2018

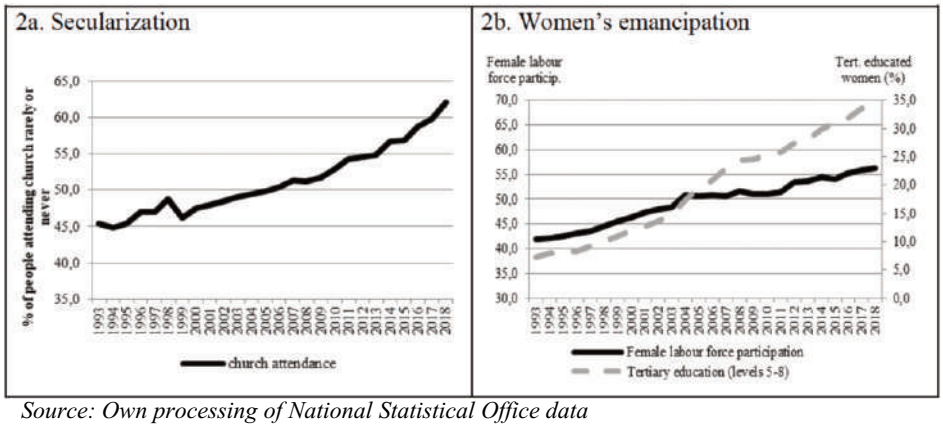


Source: Own processing of National Statistical Office data

Twenty years ago, only 2 out of 100 couples were living in a non-marital union. Today about 16% of Italians choose this form of family arrangement, at least for a part of their relationship, and this percentage is 6 points higher in the Northern regions of Italy. Whereas the level in 2018 is still modest compared to that of Nordic countries, the rising trend is remarkable. These changes are mirrored also by the percentage of out of wedlock childbearing, which has tripled since the beginning of the 21st century (Figure 1c). Currently, about one third of children are born in non-marital unions. This increase is even more dramatic considering the constant reduction in the absolute number of new born children (again Figure 1c, left-hand axis). The rising "flexibility" of union patterns are even more visible looking at marriage dissolutions. Whereas about 80 marriages out of 1000 concluded with a divorce at the beginning of 90s, the divorce rate has passed 300 in recent years (Figure 1d). This value is somewhat overestimated due to a recent change in the divorce law that has reduced the time needed to file divorce proceedings after the legal separation is made, producing an anticipation of a relevant quota of the divorces which would have been recorded in the subsequent years. Indeed, data concerning legal separation rates illustrate a clear increasing trend in marital disruption during the last 25 years.

Proponents of the Second Demographic Transition interpret these changes as a pattern of progress driven by processes such as emancipation from traditional social norms. In fact, the changes in family behaviours of Italians are occurring together with deep modifications also in the cultural and economic context. First, even in a Catholic-oriented country like Italy, it is emblematic that secularization (here approximated by the percentage of people attending church rarely or never, Figure 2a) is progressing year on year. In addition, in line with the trend of other European countries, an ever larger share of Italian women pursued higher education (one third of women aged 25-34 are currently tertiary educated, relative to the 20% of 10 years before or the 7% of the early '90s, Figure 2b). Finally, also women's labor market attachment is rising in a remarkable fashion (again Figure 2b, left-hand axis).

Figure 2: Trends in secularization and women's emancipation: Italy, 1993-2018



Source: Own processing of National Statistical Office data

3 Data and method

The analysis is based on retrospective data stemming from the 2016 Household Multipurpose Survey of Family and Social Subjects (FSS). The 2016 FSS survey was conducted by Istat, the Italian National Institute of Statistics, with a sample of about 32,000 individuals of all ages. Each individual was randomly selected from municipal registry lists, according to a sampling design aimed at constituting a statistically representative sample of the resident population. The overall response rate of the survey was greater than 80%. The 2016 FSS survey contains a wealth of information about individuals' and families' daily lives, including fertility, partnership, education and employment histories recorded with the precision of the month. This survey offers a unique – and timely – opportunity to explore trends and correlates of the diffusion of new family-related behaviours in Italy.

4 Preliminary conclusion

Given aggregate data from Istat, there is little doubt that recent trends in demographic behaviour in Italy are dramatic. With the 2016 Household Multipurpose Survey of Family and Social Subjects (FSS), we will be able to document the driving forces behind these trends. We will be able to document age differences in these trends and answer to what extent the younger cohort is the generating force. If this is the case, it is indeed possible that Italy is about to make a leap jump towards a faster and progressive process of the Second Demographic Transition. Other than facilitating insights into the extent these changes are ideational (secularization and weakening of norms), we will also be able to understand structural drivers, which would include changes in women's rates of tertiary education, labour market participation and occupations. At the same time, we will be able to hold these patterns up against the hypothesis of patterns of disadvantage (Perelli-Harris and Gerber, 2011; Perelli-Harris et al, 2010). This is an important aspect, not least because several of the rapid changes in family behaviour took place in the aftermath of the economic recession. As such, it is possible that economic hardship and increased uncertainty, which struck the younger generation particularly hard, is initializing a new path of demographic behaviour consistent with the Second Demographic Transition idea.

References

1. Caltabiano, M., Dalla-Zuanna, G., & Rosina, A. (2006). Interdependence between Sexual Debut and Church Attendance in Italy. *Demographic Research*, 14, 453-484. doi: 10.4054/DemRes.2006.14.19
2. Casterline, J. B. (2001). *Diffusion Processes and Fertility Transition: Selected Perspectives.* Washington, DC: National Research Council
3. Kohler, H.-P., E C. Billari, and J. A. Ortega. 2002. The emergence of lowest-low fertility in Europe during the 1990s, *Population and Development Review* 28(4): 641-681.
4. Lesthaeghe, R. (2020). The second demographic transition, 1986–2020: sub-replacement fertility and rising cohabitation—a global update. *Genus*, 76(1), 1-38.
5. Nazio, T., & Blossfeld, H.P. (2003). The Diffusion of Cohabitation Among Young Women in West Germany, East Germany and Italy. *European Journal of Population*, 19, 47-82. doi: 10.1023/A:1022192608963
6. Oppenheimer, V.K. (1994). Women's rising employment and the future of the family in industrial societies. *Population and Development Review* 20(2), 293-342.
7. Perelli-Harris, B., & Gerber, T. P. (2011). Nonmarital childbearing in Russia: second demographic transition or pattern of disadvantage? *Demography*, 48(1), 317–342.
8. Perelli-Harris, B., Sigle-Rushton, W., Kreyenfeld, M., Lappegård, T., Keizer, R., & Berghammer, C. (2010). The educational gradient of childbearing within cohabitation in Europe. *Population and Development Review*, 36(4), 775–801.
9. Pirani, E. and Vignoli, D. (2016). Changes in the satisfaction of cohabitors relative to spouses over time. *Journal of Marriage and Family* 78(3): 598–609. doi:10.11 11/jomf.12287.
10. Reher, D.S. (1998). Family Ties in Western Europe: Persistent Contrasts. *Population and Development Review*, 24, 203–234. doi: 10.2307/2807972
11. Ruggles, S. (2012). The Future of Historical Family Demography. Annu*al Review of Sociology*, 38, 423-431.
12. Viazzo, P.P. (2003). What's so special about the Mediterranean? Thirty years of research on household and family in Italy, *Continuity and Change*, 18, 111–137. doi: 10.1017/S0268416003004442
13. Vignoli, D., & Salvini, S. (2014). Religion and Union Formation in Italy: Catholic Precepts, Social Pressure, and Tradition. *Demographic Research*, 31, 1079-1196. doi: 10.4054/DemRes.2014.31.35
14. Vignoli, D., Tocchioni, V., & Salvini, S. (2016). Uncertain Lives. Insights into the Role of Job Precariousness in Union Formation in Italy. *Demographic Research*, 35, 253-282. doi: 10.4054/DemRes.2016.35.10.

Causes of death patterns and life expectancy: looking for warning signals

Cause di morte e speranza di vita: alla ricerca di segnali d'avvertimento

Stefano Mazzuco, Emanuele Aliverti, Daniele Durante and Stefano Campostrini

Abstract The evolution of longevity across countries is quite diverse and it still remains unclear what determined such different patterns throughout the last decades. In this paper we consider a Bayesian nonparametric mixture of B-splines for life expectancy trajectories that characterizes locally-varying similarities across functions, learning where country-specific trajectories are likely to overlap and where instead they tend to diverge. A preliminary comparison among Italy and United States indicates interesting trends in the evolution of life expectancy, with trajectories that overlap until the early 80s and then diverge substantially. We attempt to justify such differences by studying variations in the causes of premature mortality across these periods of interest, trying to justify potential driving factors for such divergences.
Abstract *Vi sono evidenti diversità nell'evoluzione della longevità tra paesi, ma non è ancora chiaro cosa abbia determinato andamenti così diversi negli ultimi anni. In questo lavoro, si considerano i dati sulle cause di morte per spiegare i diversi patterns di mortalità nel tempo, utilizzando un modello Bayesiano nonparametrico basato su misture di basi B-splines per caratterizzare similarità locali tra le diverse traiettorie di longevità nel tempo. Tale approccio consente di individuare in quali finestre temporali la longevità è simile tra paesi e in quali periodi diverge. Un primo confronto tra Italia e Stati Uniti offre dei primi risultati rilevanti che indicano l'inizio degli anni 80 come soglia da cui le differenze di longevità tra i due paesi sono divenute statisticamente rilevanti. Le motivazioni legate a tali differenze sono ricercate confrontando le traiettorie delle cause di morte.*

Key words: Bayesian nonparametrics, causes of death, life expectancy.

Stefano Mazzuco
Università degli studi di Padova, e-mail: stefano.mazzuco@unipd.it

Emanuele Aliverti
Università Ca' Foscari di Venezia, e-mail: emanuele.aliverti@unive.it

Daniele Durante
Università Bocconi di Milano, e-mail: daniele.durante@unibocconi.it

Stefano Campostrini
Università Ca' Foscari di Venezia, e-mail: stefano.campostrini@unive.it

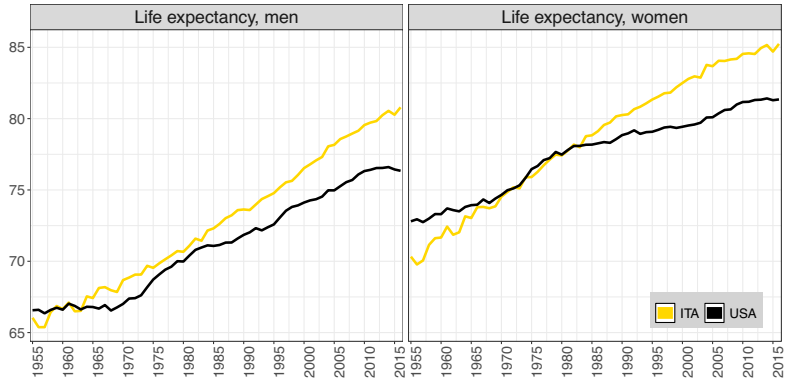


Fig. 1: Life expectancy in Italy and USA. Source: Human Mortality Database

1 Introduction

In the economically more advanced countries, longevity has steadily increased in the last decades. Comparing, for example, life expectancy at birth (e_0) in Italy and USA, we can observe in Figure 1 how these countries show comparable trends until the early 80s, resulting afterwards in substantially diverging trajectories. This result is even more striking if we consider that USA invests relatively almost twice in health of its resources (17% of GDP vs 8.7% of Italy; see [7]) and suggests the need of an improved understanding of the mechanism behind different patterns of longevity evolution, especially in relation to the mortality structure.

For example, Bergeron–Boucher et al [2] have recently shown that the extension of longevity is usually accompanied by a diversification of the causes of death. More specifically, Woolf and Schoomaker [13] analyze the trend of causes of death in USA, finding that midlife mortality caused by drug overdoses, alcohol abuses, suicides, and a list of organ system diseases have particularly increased in the last years. However, this finding has been contested (see Mehta et al, [6], who argue that cardiovascular diseases are the main responsible of US life expectancy stagnation). Such a controversy reflects the issue when dealing with cause-specific mortality, related with a competing risk setting: a cause-specific mortality rate can decline because there has been a significant improvement in treatment and/or prevention of that disease or just because other causes have grown meanwhile. Therefore, if we want to analyze the time trend of causes of death we need to take into account this feature. Stefanucci and Mazzuco [11] propose to combine Functional Data Analysis (FDA — see [10]) with Compositional Data Analysis (CDA — see [1]; [4]), limiting to a descriptive analysis of causes of death patterns.
Here, we propose a model-based analysis aimed at inferring patterns in causes of deaths that precede life expectancy stagnation. Motivated by Figure 1, we consider a flexible Bayesian nonparametric mixture of B-splines to learn local similarities

across life expectancies, assessing in which temporal blocks life-expectancy curves tend to overlap, and where instead they diverge. Subsequently, we analyze causes of death around the intervals of interest, in order to highlight what aspects of mortality have changed more considerably in those crucial years and what evidence they provide in terms of variation of life expectancy.

2 Data and methods

Data are collected from the Human Mortality Database [5], that ensures high quality data on mortality profiles of different European and non-European countries. Specifically, we focus here on Australia, Austria, Belgium, Bulgaria, Canada, Switzerland, Czech-Republic, Denmark, Spain, Finland, France, Great-Britain, Hungary, Ireland, Island, Italy, Japan, the Netherlands, Norway, Portugal, Slovakia, Sweden and United States of America. Moreover, causes of death data are taken from WHO mortality database.

We conduct analysis on these $n = 23$ countries, considering sex-specific and age-adjusted rates over a time period of $T = 62$ years ranging from 1955 to 2016. We consider 8 classes of causes of mortality, namely infections, neoplasms (all cancers with the exception of lung cancer), lung cancer, endocrines diseases, circulatory diseases, respiratory diseases, digestive diseases and external causes.

To flexibly model life expectancy patterns across countries $i = 1,\dots,n$ and years $t = 1,\dots T$, we treat the trajectory of e_0 as a function $y_i(t)$ and, following standard practice in FDA, we decompose it as

$$y_i(t) = f_i(t) + \varepsilon_i(t), \quad \varepsilon_i(t) \overset{\text{iid}}{\sim} N(0,\sigma^2), \tag{1}$$

for each country $i = 1,\dots,n$ and year $t = 1,\dots T$, where $\varepsilon_i(t)$s are independent Gaussian errors and $f_i(t)$ is an underlying smooth function. To include this smoothness, while avoiding strong assumptions on the functional form of $f_i(t)$, we model such a trajectory via B-splines [3], letting

$$f_i(t) = \sum_{k=1}^{K} \beta_{ik}\mathbf{B}_k(t), \tag{2}$$

where $[\mathbf{B}_1(t),\dots,\mathbf{B}_K(t)]$ denotes a set of fixed quadratic B-splines basis functions shared across countries, while β_{ik} denotes the country-specific coefficient referred to the k-th basis. Hence, overlapping and diverging patterns in the functions $f_i(\cdot)$, $i = 1,\dots,n$ across time are only regulated by ties among the associated coefficients β_{ik}. Motivated by our goal of learning such structures, we adapt the strategy in [8] to incorporate grouping effects for the coefficients β_{ik} via a model-based clustering induced by the following Dirichlet process prior on the coefficients

$$\beta_{1k},\dots,\beta_{nk} \mid F \overset{\text{iid}}{\sim} F, \quad F \sim \text{DP}(\alpha, F_0), \quad F_0 \sim N(0,\eta^2) \tag{3}$$

where DP(α, F_0) denotes a Dirichlet process with concentration parameter α and base measure F_0 which is assumed to be a Gaussian distribution with mean 0 and variance η^2. The discreteness of the DP is particularly appealing for our purposes, since it implies positive probabilities of ties among the coefficients β_{ik}, inducing local-clustering across curves as a byproduct. More specifically, denoting as $\beta^{\star}_{(h)k}$, $h = 1, \ldots, H_k$ the $H_k \leq n$ distinct values of the B-splines coefficients for the k-th basis, then, if $\beta_{ik} = \beta_{jk} = \beta^{\star}_{(h)k}$, countries i and j are expected to exhibit overlapping life-expectancy trajectories in the interval associated with the k-th spline basis. This local clustering can be formally characterized by $S_{(h)k} = \{i : \beta_{ik} = \beta^{\star}_{(h)k}\}$, and, thus, inference on the partitions $\rho_k = \{S_{(1)k}, \ldots, S_{(H_k)k}\}$ provides deeper insights on the trends underlying life-expectancy, flexibly learning which curves tend to overlap within specific blocks and characterizing uncertainty of this process.

Prior specification is completed by specifying a conjugate Inverse-Gamma distribution of the parameter $\sigma^2 \sim$ Inv-Gamma(a_0, b_0), while posterior inference proceeds via a collapsed back-fitted Gibbs sampler, exploiting the Polya-Urn scheme of the DP and leveraging the additive representation of the B-splines basis.

3 Preliminary results

We conduct posterior inference using 3000 Gibbs samples after a burn-in of 1000, setting $\alpha = 1$, $\eta^2 = 10$, $a_0 = b_0 = 0.01$. Effective sample size and autocorrelation plots did not provide evidence against convergence of the chains. In Figure 2 we obtain some preliminary results, focusing on the comparison between Italy and USA.

The first column of Figure 2 reports the smoothed estimated for the life expectancies and their associated credible intervals. Results indicate a common increasing trend for both states and sexes, with several important differences that are worth mention. In particular, for men we observe a period of overlapping curves in the late 50s, followed by separate trends in the late 60s and again a period of overlap during the late 70s and early 80s. Women instead report separate trends until the early 60s, followed by a long period of overlap until the late 80s and a subsequent increasing separation across the two curves.

These findings are further confirmed by the co-clustering probabilities, estimated as the proportion of MCMC sample in which Italy and United States are allocated in the same group. Such quantities are reported, as a function of time t, in the second column of Figure 2, while the third column of Figure 2 illustrates the ℓ_2 norm between the estimated life expectancy curves. Both panel indicate a common trend between men and woman after 1985, and quite different behavior in previous years.

Lastly, we compare the composition of the causes of death in the period of interest, focusing on premature mortality (≤ 70 years) in the window of interest $(1970 - 1990)$; see Figure 3. Such a composition has interestingly changed across this period of investigation, with the proportion of digestive system diseases and respiratory diseases showing markedly different trends across Italy and USA. Circulatory system diseases also report interesting trajectories, describing a notable

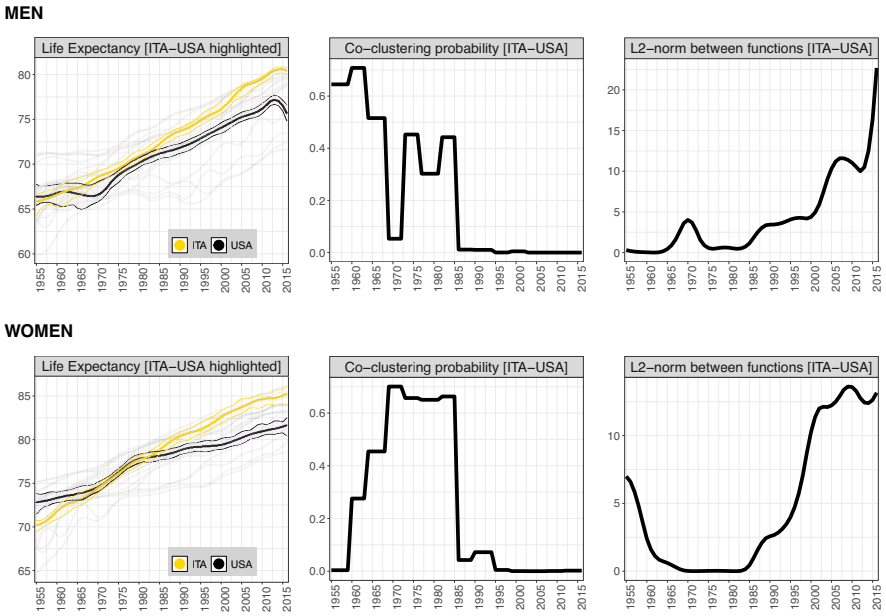


Fig. 2: First column: estimated functions $\hat{f}_i(t)$ via posterior mean and associated 95% credible intervals. Light gray curves represent all estimated countries, while Italy and United States are highlighted in yellow and black, respectively. Second column and third column focus on the comparison between Italy and United States, and depict co-clustering probability and ℓ_2 norm between estimated functions $\hat{f}(\cdot)$, respectively.

improvement for Italian women compared to Americans'. Worth to be mentioned is also the evolution of infectious diseases and the peak associated with AIDS, particularly severe in the American male population.

4 Discussion

In this article, we have provided a first step toward understanding the differences in longevity between Italy and the USA. We propose a flexible Bayesian nonparametric model to learn local-similarities across life expectancy trajectories, locating where these curves begin to diverge and exploring the composition of the causes of death around such period.

These preliminary evidences can be interpreted as *signals* of the evolving changes, more than explicit causes of the underlying processes. In fact, a proper analysis of these incredibly complex phenomena should also take into account other factors influencing the overall mortality level, along with its composition. Some potential determinants are the evolution of obesity trends and, more importantly, income in-

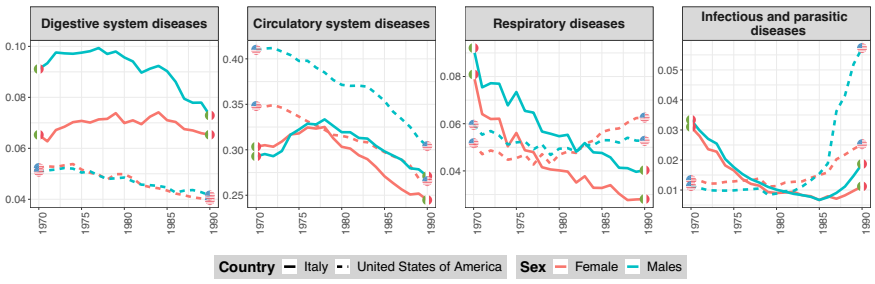


Fig. 3: Causes of death - premature mortality

equalities, which report substantial discrepancies in the period of interest [9]. Including the effects of this crucial determinants in our model will be the focus of future works, currently under investigation.

References

1. Aitchinson, J. (1986) *The Statistical Analysis of Compositional Data*. Chapman & Hall.
2. Bergeron–Boucher M-P, Aburto JM, van Raalte A. (2020) "Diversification in causes of death in low-mortality countries:emerging patterns and implications". *BMJ Global Health2020*;5:e002414. doi:10.1136/bmjgh-2020-002414
3. De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-verlag New York.
4. Egozcue, J. J. and Pawlowsky–Glahn, V. (2011) *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd.
5. Human Mortality Database (2020). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).
6. Mehta NK, Abrams LR, Myrskyla M. "US life expectancy stalls due to cardiovascular disease, not drug deaths". *Proceedings National Academy of Sciences* USA. 2020 Mar 31;**117(13)**:6998–7000. doi: 10.1073/pnas.1920391117.
7. OECD Health Statistics (2020), available at http://www.oecd.org/els/health-systems/health-data.htm
8. Paulon, G., Llanos, F., Chandrasekaran, B., and Sarkar, A. (2020). "Bayesian semiparametric longitudinal drift-diffusion mixed models for tone learning in adults"., *Journal of the American Statistical Association*, 1–14
9. Raleigh, V.S. (2019). *Trends in life expectancy in EU and other* OECD *countries: Why are improvements slowing?*
10. Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis*, vol. 2nd Ed. New York:Springer
11. Stefanucci M., Mazzuco S. (2020) "Analyzing Cause-Specific Mortality Trends using Compositional Functional Data Analysis". *Arxiv*:2007.15896.
12. Stringhini, S., Carmeli, C., Jokela, M., Avendaño, M., Muennig, P., Guida, F., Ricceri, F., d'Errico, A., Barros, H., Bochud, M. and Chadeau-Hyam, M. (2017) *Socioeconomic status and the* 25×25 *risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1.7 million men and women* The Lancet, 389(10075), pp.1229-1237.
13. Woolf, S. H. and Schoomaker, H. (2019) "Life Expectancy and Mortality Rates in the United States, 1959–2017". *Journal of the American Medical Association*,**322**, 1996–2016

2.2 Advances in ecological modelling

A Bayesian joint model for exploring zero-inflated bivariate marine litter data

Modello bayesiano congiunto per l'analisi bivariata dei dati di rifiuti marini con eccesso di zeri

S. Martino and C. Calculli and P. Maiorano

Abstract Acknowledging the spatial and spatio-temporal behavior of natural processes is crucial for management purposes. Semi-continuous datasets are common in Ecology: combining information on occurrence and conditional-to-presence abundance of species allows to improve environment effects estimates. Based on a marine litter case study, this paper proposes a two-parts model to handle 1) the zero-inflation problem and 2) the spatial correlation characterizing abundance monitoring data. In the spirit of multi-species distribution models, we propose to jointly infer different litter categories in a Hurdle-model framework. Shared spatial effects that link abundances and probabilities of occurrences of litter categories, are implemented via the SPDE approach in the computationally efficient INLA context.

Abstract *Riconoscere i trend spaziali e spazio-temporali dei processi naturali è di cruciale importanza ai fini gestionali. I dati semi-continui sono comuni in Ecologia: la combinazione di informazioni sulla presenza e sull'abbondanza (condizionata alla presenza) di specie permette di migliorare le stime degli effetti ambientali. Basato su un caso di studio riguardante i rifiuti marini, questo lavoro propone un modello in due parti per gestire 1) il problema dell'eccesso di zeri e 2) la correlazione spaziale che caratterizza i dati di monitoraggio dell'abbondanza. Ispirata ai modelli di distribuzione multi-specie, la proposta permette di inferire congiuntamente diverse categorie di litter con un modello Hurdle. Effetti spaziali comuni che mettono in relazione abbondanze e occorrenze di diverse categorie, vengono*

Sara Martino
Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway e-mail: sara.martino@ntnu.no

Crescenza Calculli
Department of Economics and Finance, University of Bari Aldo Moro, Largo Abbazia S. Scolastica - 70124 Bari, Italy e-mail: crescenza.calculli@uniba.it

Porzia Maiorano
Department of Biology, LRU CoNISMa, University of Bari Aldo Moro, 70125 Bari, Italy e-mail: porzia.maiorano@uniba.it

implementati con l'approccio SPDE nel contesto, computazionalmente efficente, di INLA.

Key words: Marine litter, Spatial joint modeling, Hurdle models, Integrated nested Laplace approximation (INLA), Stochastic Partial Differential Equation (SPDE) approach

1 Introduction

Marine environment is changing rapidly due to increasing anthropogenic activities. The continuously growing quantity of litter and debris items ending into the sea has been recognized as the prevalent form of marine pollution that impacts ecological, economic and aesthetic values of the marine and coastal environment [3]. Understanding the spatial and spatio-temporal distribution of marine litter is essential for sustainable management of ecosystems. Despite this urgency, marine litter data are scarce, poorly monitored and their analysis represents a challenging complex ecological problem. Marine litter items, mostly collected by experimental fishery surveys, can be classified as different special abiotic items (or additional species), thus the analysis of litter abundances can be treated in the context of Joint Species Distribution Models (JSDMs) [2]. This class of models explicitly acknowledges the multivariate nature of communities by assuming the joint species response to the environment and to each other species. For marine litter data a joint models approach allows a more complete understanding of the distribution and dynamics of multiple litter categories and of the effects of environmental covariates by considering spatial structures shared by different categories. The specification of these components provides a method to link together ecological data generation processes concerning different species or categories. However, the inclusion of spatial correlation structures and the presence of excesses of zeros increase model complexity and lead to high computational costs. Hurdle and zero-inflated mixture models have been used to deal with this issue (differences are mainly based on the interpretation of zero observations). While both models are widely used to model count data, hurdle semi-continuous models are especially useful to model density data [5].
A Hierarchical joint modeling approach is proposed to investigate environmental drives and the spatial distribution of waste amounts found at the sea-floor in a Central Mediterranean area. Extending our previous analysis in [1], we fit a Hurdle model to accommodate the excess of zeros resulting from the semi-continuous nature of the data and consider the spatial patterns of two litter categories: plastic and other litter categories all together. The INLA method proved to be an efficient approach to model spatially correlated data with excess of zeros including the effects of environmental covariates. Spatial structures are modeled by the stochastic partial differential equations (SPDE) approach. It consists in defining a continuously indexed Matérn Gaussian field (GF) as a discretely indexed spatial random process (GMRF) using piece-wise linear basis functions defined on a triangulation of the

domain of interest [6]. The SPDE approximation is implemented in INLA [7] via the R-INLA package (`http://www.r-inla.org`) designed to make Bayesian inference accessible for a large class of latent Gaussian models providing, at the same time, accurate and computationally efficient approximations of the posterior marginals.

2 Case study description

Monitoring litter data come from experimental trawl surveys carried out from 2013 to 2016 in the North-Western Ionian Sea as part of MEDITS (MEDiterranean International Trawl Surveys) activities. The study area represents the deepest sea in the Mediterranean basin characterized by a complex geomorphology and the presence of important fisheries and main harbors. The same 70 geo-referenced depth-stratified hauls are sampled between 10 and 800 m in depth every year, summing to 280 hauls in 4 years. Wastes caught during the trawl surveys are classified in 8 categories: plastic, rubber, metal, glass/ceramic, cloth/natual fibres, processed wood, paper/cardboard, other/unspecified. The number of collected items for each litter category was scaled to the swept surface unit (1km^2), thus obtaining density indices (N/km^2) for each litter category and survey at every haul location. Since plastic items represent the most abundant fraction of wastes collected at surveyed hauls, densities of two litter categories are considered: *plastic* and the aggregation of *all other categories*, leading to a bivariate response. Different environmental covariates are investigated as possible drivers that might affect the spatial distribution of the bivariate litter abundances over the study region. In particular, data on sea currents and fishing activities, collected with different spatial supports, were first aligned and then used to investigate environmental factors affecting the bivariate distribution of the density of the two litter categories. The effects of the superficial eastward (U) and northward (V) sea water velocities (available at `http://marine.copernicus.eu/`) and the daily average transit (MVH) and fishing time (MFH) for 3 types of vessels (Drifted Longlines, Fixed Gears and Trawler available at `https://globalfishingwatch.org/`), are investigated.

3 The proposed Bayesian joint Hurdle model

Spatial and spatio-temporal abundance processes commonly result in semi continuous non-negative datasets. These data can be conveniently modeled within the Hurdle models framework, where two independent sub-processes are considered: an occurrence process and a conditional-to-presence continuous process. For the bivariate abundance response, let p indicates *plastic* litter and $\overline{p}$ *all other* litter categories, where

$$\pi_{stp}, \pi_{st\overline{p}} \sim Ber(\pi_{stp}), Ber(\pi_{st\overline{p}})$$
$$\mu_{stp}, \mu_{st\overline{p}} \sim Gamma(a_{stp}, b_{stp}), Gamma(a_{st\overline{p}}, b_{st\overline{p}})$$

being the occurrence and the conditional-to-presence abundance sub-processes at time t ($t = 1, \ldots, T$) and at location s ($s = 1, \ldots, n_t$), respectively. Then the abundances of the two litter categories can be specified as follows:

$$\text{logit}(\pi_{stp}) = \beta_{0p}^{(\pi)} + \sum_{i=1}^{m} \beta_{1p}^{(\pi)} x_{ist} + V_s + V_{stp} \tag{1}$$

$$\log(\mu_{stp}) = \beta_{0p}^{(\mu)} + \sum_{i=1}^{m} \beta_{1p}^{(\mu)} x_{ist} + \alpha_s^{(1)} V_s + \alpha_{sp} V_{stp} \tag{2}$$

$$\text{logit}(\pi_{st\overline{p}}) = \beta_{0\overline{p}}^{(\pi)} + \sum_{i=1}^{m} \beta_{1\overline{p}}^{(\pi)} x_{ist} + \alpha_s^{(2)} V_s + V_{st\overline{p}} \tag{3}$$

$$\log(\mu_{st\overline{p}}) = \beta_{0\overline{p}}^{(\mu)} + \sum_{i=1}^{m} \beta_{1\overline{p}}^{(\mu)} x_{ist} + \alpha_s^{(3)} V_s + \alpha_{s\overline{p}} V_{st\overline{p}} \tag{4}$$

where $\pi_{st.}$ and $\mu_{st.} = a_{st.}/b_{st.}$ are modeled through the logit and logarithm links, respectively. In linear predictors, the $\beta_{0.}^{(\pi),(\mu)}$ represent the intercepts and $\beta_{1.}^{(\pi),(\mu)}$ are fixed effects of spatio-temporally varying covariates x_i (U, V, MVH and MFH). In order to account for spatially structured effect common to the two litter categories, a shared component V_s is specified in Eqs. (1)-(4). Moreover, specific spatial components $V_{st.}$ are assumed to be common to occurrence and abundance sub-processes of each category. The $\alpha_s^{(.)}$ and $\alpha_{s.}$ parameters represent the scale effects of the spatial variation common to the four processes and to the category-specific processes, respectively. Common and category-specific spatial components, V are modeled by GMRFs through the SPDE approach [6], as $V \sim N(0, Q(\kappa, \tau))$ and $(log(\kappa), log(\tau)) \sim MVN(\mu, \rho)$ where the covariance function of the spatial effect Q depends on a range effect (κ) and a total variance parameter (τ). For the hyper-parameters of spatially structured effects, Penalised-Complexity priors (PC-priors) [8] are used to design sensible hyperpriors for the precision and mixing parameter distributions.

4 Main Joint model outcomes

In Table 1 we report the estimated fixed effects for the occurrences and the positive abundances considering both litter categories in Eqs. (1)-(4). Relevant effects of MVH and MFH covariates are estimated for the plastic category occurrences: while fishing activities have a negative effect on the presence of plastic items, the

higher transit vessel time the higher the probability of presence of items of this category. Moreover, a higher presence of plastic is estimated in the presence of currents towards north-western direction. Even tough not relevant, the same signs for regression coefficients characterize the occurrences of other litter category.

None of the estimated effects are relevant in affecting positive abundances of both litter categories.

The posterior mean maps in Figure 1, show the estimated spatial fields, V_s and V_{st}.. In particular, the spatial variation common to both litter categories, allows to identify hot-spots with higher densities and higher presence probabilities of all litter items (Figure 1(a)). On the other hand, category-specific spatial effects suggest smooth differences in the relative abundance and spatial locations of plastic items with respect to common trend while, for others litter, a higher concentration of items is found on the southern-west area (Sicily and Calabria).

Table 1: Estimated fixed effects for occurrences and abundances of the two litter categories. 95% CI in brackets.

	Coeff.	Plastic		Other categories	
Occurence	β_{MVH}	0.227	(0.003,0.496)	0.097	(-0.034,0.249)
	β_{MFH}	−0.274	(-0.586,-0.011)	−0.066	(-0.248,0.095)
	β_U	−10.598	(-19.306,-2.158)	−5.097	(-11.654,1.238)
	β_V	7.696	(0.288,15.180)	9.379	(3.369,15.642)
Abundance	β_{MVH}	0.004	(-0.035,0.045)	−0.033	(-0.079, 0.014)
	β_{MFH}	0.006	(-0.043, 0.053)	0.017	(-0.038,0.070)
	β_U	−1.057	(-3.309, 1.163)	−1.024	(-4.806,2.700)
	β_V	1.747	(-0.321, 3.793)	−0.658	(-3.690,2.345)

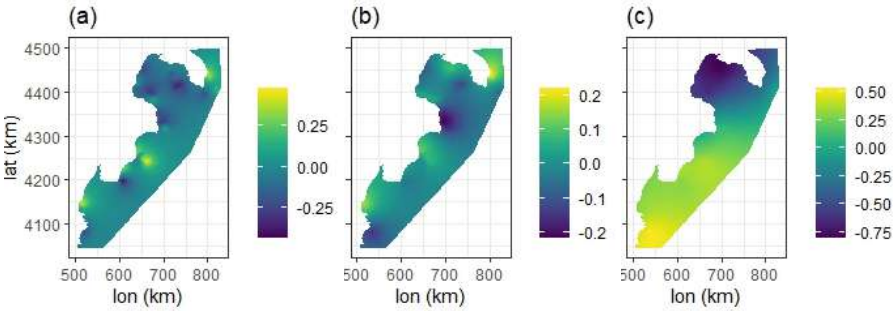


Fig. 1: Estimated spatial effects (posterior means): (1) common trend for the two litter categories; specific trend for plastic (2) and all other litter (3) abundances

5 Discussion and further developments

This study showed the use of common and shared spatial components as a more realistic approach to infer semi-continuous density data. Developments of this work include the analysis of shared spatio-temporal structured components for multiple litter categories. Further goals also concern the application of new methods for the analysis of data displaying spatial dependencies over complex domains and for addressing the change-of-support problem occurring within different levels of data aggregation.

References

1. Calculli, C., Pollice, A., Paradinas, I., Sion, L., Maiorano, P.: An INLA spatio-temporal model for zeroinflated marine plastic litter abundance. Proceedings of the GRASPA 2019 Conference, Pescara, 15-16 July 2019, 62–65 (2019). https://aisberg.unibg.it/handle/10446/146842?mode=full.5485#.XcKwZehKg2w
2. Clark, J.S., Nemergut, D., Seyednasrollah, B., Turner, P.J., Zhang, S.: Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. Ecol Monogr, **87**,34–56 (2017) doi: 10.1002/ecm.1241
3. Galgani, L., Beiras, R., Galgani, F.,Panti, C., Borja, A.: Editorial: Impacts of Marine Litter, Frontiers in Marine Science,**6**, 208 (2019) doi:10.3389/fmars.2019.00208
4. Illian, J.B., Martino, S., Sørbye, S.H., Gallego-Fernndez, J.B.,Zunzunegui, M., Esquivias, M.P., Travis, J.M.J.: Fitting complex ecological point process models with integrated nested Laplace approximation. Methods Ecol Evol, **4**, 305–315 (2017) doi: 10.1111/2041-210x.12017
5. Krainski, E.T., Lindgren, F., Simpson, D., Rue, H.: The R-INLA tutorial on SPDE models (2015). https://www.math.ntnu.no/inla/r-inla.org/tutorials/spde/spde-tutorial.pdf (2015)
6. Lindgren, F., Rue, H., Lindstrom, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach (with discussion).Journal of the Royal Statistical Society, **73**, 423–498 (2011)
7. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by usingintegrated nested Laplace approximations. Journal of the Royal Statisitcal Society, B, **71**, 319–392 (2009)
8. Simpson, D., Rue, H., Riebler, A., Martins, TG, Sørbye, S.H.: Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. Statistical Science, 32 (1), 1–28 (2017)

2.3 Advances in environmental statistics

Bayesian small area models for investigating spatial heterogeneity and factors affecting the amount of solid waste in Italy

C. Calculli and S. Arima

Abstract This study aims at investigating factors that impact solid waste generation rates at the Italian province-level scale. An approach based on Bayesian small area models is used to evaluate potential spatial dependencies and local variations on the distribution of waste generation rates and its determinants. An extension of the basic Fay-Herriot model to include spatial correlation among neighboring areas is considered (SFH). Preliminary results suggest a highest per capita production of differentiated (or recyclable) waste fraction in wealthier Italian provinces highlighting the intrinsic correlation between areas characterized by similar socio-economic scenarios.

Abstract *L'obiettivo di questo lavoro è quello di analizzare i fattori che influenzano i tassi di produzione dei rifiuti solidi urbani su scala provinciale in Italia. Un approccio modellistico basato sulla stima bayesiana per piccole aree viene utilizzato per valutare la dipendenza spaziale e le variazioni terriroriali nella distribuzione dei tassi di produzione dei rifiuti e delle loro determinanti. In particolare, viene considerata una estensione del modello di Fay-Herriot (SFH) al caso di correlazione spaziale tra dati di aree adiacenti. I risultati preliminari suggeriscono una produzione maggiore pro-capite della frazione differenziata (o riciclabile) di rifiuti urbani nelle province italiane più ricche, evidenziando una correlazione intrinseca tra aree caratterizzate da scenari socio-economici simili.*

Key words: Bayesian SAE, Spatial random effects models, Municipal solid wastes

Crescenza Calculli
Department of Economics and Finance, University of Bari Aldo Moro, Largo Abbazia S. Scolastica - 70124 Bari, Italy e-mail: crescenza.calculli@uniba.it

Serena Arima
Department of History, Society and Human Studies, University of Salento, Piazza Tancredi, n.7 - 73100 Lecce, Italy e-mail: serena.arima@unisalento.it

1 Introduction

Municipal solid waste (MSW) commonly refers to the semi-solid or solid materials disposed by residents of urban areas. Due to population growth, urbanization, economic development, inappropriate recycling and governance programmes, MSW has become a global issue that poses serious threats to the environment and human health. Several industrialized countries have raised concerns about the economic viability and environmental acceptability of waste-disposal practices. For this reason, estimating waste production and treatment are crucial to quantify impacts, plan capacities and set policy targets. Several studies showed the impacts of socio-economic and demographic factors on solid waste generation trends, although neglecting the potential spatial dependency at local level. Few studies addressed the spatial characteristics of MSW generation rates applying spatial stratification or visually mapping their distributions using spatial statistical techniques such as simultaneous spatial autoregression (SAR) and geographically weighted regression (GWR) [7, 4]. An alternative model-based approach is proposed in order to capture the local variations in the distribution of waste generation rates accounting for their determinants at the Italian province level. This approach, based on methods of Small Area Estimation (SAE), assumes area-specific random effects to account for the between-area variation. A detailed overview of this methods is reported in Section 2.

2 Small area models

In recent years, small area estimation established as an important area of statistics as private and public agencies try to extract the maximum information from sample survey data. Small area methods arise when one wants to use sample surveys, generally designed to provide estimates of total or means for large subpopulations or domains, to draw inferences about smaller domains, such as as states, provinces, or different racial and/or ethnic subgroups that are unplanned by design. These small domains are called small areas. In recent years, demand for reliable estimates for small area estimates has greatly increased due to their growing use in formulating policies and programmes, allocating government funds, regional planning and other uses. Policy makers are often interested in targeting areas with particular needs in order to conduct specific actions: as in our case study, identifying those areas with highest waste production is of fundamental importance for decision makers who should plan the opening of new filtration and recycling plant or monitoring the consequences of such a waste production on the health of the population.
The simplest approach to small area estimation is to consider *direct estimators*, that is estimating the variable of interest using the domain-specific sample data. However, it is well known that the domain sample sizes are rarely large enough to support and accurate direct estimators [5]. Small area estimation tackles the problem of providing estimates of one or several variables of interest in areas where the infor-

mation available on those variables is, on its own, not sufficient to provide accurate direct estimates. Estimates for all areas are produced using the sample and some additional auxiliary information which should be available for all small areas. *Indirect estimators* are often employed in order to increase the effective domain sample size by borrowing strength from the related areas using linking models, census, administrative data and other auxiliary variables associated with the small areas. The model-based approach is widely used to develop indirect estimates. Depending on the type of data available, small area models are classified into two types, area-level and unit-level:

- area-level models where only area-level data are available for the response variable and the covariates;
- unit-level models where data on the response variable, and possibly on covariates, are available at the unit level.

The model proposed by [2] is the most popular small area model when data are available at area-level. It borrows strength from data available from all areas by assuming a hierarchical structure and uses auxiliary information from other data sources such as administrative records or censuses. The frequentist predictor of small area means, which is also known as empirical best linear unbiased predictor (EBLUP), results in a convex combination of the direct estimator and the synthetic estimator from the model. Properties of the predictors of small area means, such as bias and mean squared error, are derived conditionally on the auxiliary information. In this paper we focus on an area-level nested error linear regression model and propose a Bayesian hierarchical model for estimating finite population small area means.

2.1 The proposed model

Let m denote the number of small areas under consideration and let θ_i be the population characteristic of interest in the $i-$th area. The quantity of interest may be a total, a mean (or a proportion). Let y_i be a direct estimator of θ_i for area i and let X_i be the p dimensional vector of auxiliary data collected at the area level. The Fay-Herriot model is defined as

$$y_i = X_i'\beta + v_i + e_i, i = 1, \ldots, m, \tag{1}$$

where the random effects $v_1, \ldots, v_m$ and sampling errors $e_1, \ldots, e_m$ are independent with $v_i \sim N(0, \sigma_v^2)$ and $e_i \sim N(0, \psi_i)$. We assume that the ψ_i's are known but, in general, different, reflecting the potential difference in sample sizes [5]. Our goal is the prediction of small area mean

$$\theta_i = X_i'\beta + v_i \tag{2}$$

The best linear predictor of θ_i for model (1) when model parameters (β, σ_v^2) are known is the EBLUP estimator $\tilde{\theta}_i = \gamma_{iv} y_i + (1 - \gamma_{iv}) X_i'\beta$, with $\gamma_{iv} = \sigma_v^2/(\sigma_v^2 + \psi_i)$.

In this paper, we consider a Bayesian extension of the Fay-Herriot model in Equation (1) and rewrite the model as the following multi-stage model:

Stage 1. $y_i = \theta_i + e_i \qquad i = 1,\ldots,m$, with $e_i \overset{\text{iid}}{\sim} N(0,\psi_i)$ and ψ_i a-priori known;
Stage 2. $\theta_i = x_i^T\beta + v_i \quad i = 1,\ldots,m$, with $v_i \overset{\text{iid}}{\sim} N(0,\sigma_v^2)$;
Stage 3. β, σ_v^2 are, loosely speaking, mutually independent with flat priors over location parameters and inverse gamma distributions over the scale parameters.

In particular, we assume a Normal flat prior for the regression coefficients and an inverse gamma with parameters equal to 0.001 for the scale parameter σ_v^2.
Following [6], we extend the aforementioned model to incorporate spatially correlated random effects in the linking model: in particular, we add a spatial random effect denoted with b_i in the linking model in Stage 2 as follows:

$$\theta_i = x_i^T\beta + b_i \tag{3}$$

where

$$\mathbf{b} \sim MVN(\mathbf{0}, \Sigma(\sigma_b^2, \lambda))$$

and

$$\Sigma(\sigma_b^2,\lambda) = \sigma_b^2 \qquad\qquad D = \lambda\mathbf{R} + (1-\lambda)\mathbf{I}$$

defining a conditional auto-regressive (CAR) model on the area specific spatial effects. In particular, σ_b^2 is a spatial dispersion parameter and λ is a spatial autocorrelation parameter ($\lambda \in [0,1]$) and $\mathbf{I}$ is an identity matrix of dimension m. The matrix $\mathbf{R}$, commonly known as the neighbourhood matrix, has ith diagonal element equal to the number of neighbours of the area i, and the off-diagonal elements in each row equal to -1 if the corresponding areas are neighbours and 0 otherwise.
As the posterior distribution cannot be obtained analytically, full conditional distributions have been derived and samples from the posterior are obtained by Gibbs sampling.

3 Data description

The Italian territory covers an area of 301,340 km^2 dived into different classification levels according to the European Nomenclature of Territorial Units for Statistics (NUTS). For this analysis, the administrative province levels disaggregation (NUTS-3) is considered for a total amount of 107 units over the whole territory. At province level, data related to the production and collection of municipal solid waste (MSW) are retrieved from the waste national registry provided by the Italian Institute for Environmental Protection and Research, ISPRA (`http://www.catastorifiuti.isprambiente.it`). For 2019 and for each unit, the amount of MSW (in *tons*) and the amount of the differentiated fraction of MSW (DSW, in *tons*) are provided. This latter information refers to the

organic and other materials (such as plastic, metal, glass, wood, paper and cardboard) fraction of generated MSW bound to recycling programmes and processed in composting facilities (including integrated anaerobic and aerobic treatment and anaerobic digestion treatment plants). In Italy as a whole for 2019, the production and collection of DSW represents the 61.35% of the total quantity of MSW with an amount generated per person of 306.29 Kg on average [3]. At province level for 2019, greater per capita productions of DSW ($\geq$ 480 Kg/inh) are observed in North-Central Italian units (*i.e.* Reggio Emilia, Rimini, Ferrara, Piacenza, Lucca). According to data availability and comparability at the Italian province level, potential determinants of per capita DSW generation are considered. In particular, factors related to demographic and socio-economic characteristics are evaluated: population (Pop), sex-ratio (SexR, as Male/Female), population density (PopDen, as pop/km^2), number of cities (Cit), average number of housing (nHous), unemployment rate (UnempR, as the % of not working people older than 15 years), graduates (Grad, as the % of people with higher educational degree), added value per capita (AddV in euros, as a proxy of the provincial Gross domestic product per capita). All indicators are provided by the the Italian Statistics Institute, ISTAT (available at `http://dati.istat.it/`). Besides, the number of plants where DSW are dumped for recycling, composting or treating by other methods is considered as potential factor affecting waste generation.

4 Preliminary results and further developments

The Hierarchical model in Eqs. 1 and 3 is fitted to evaluate the dependence relation between per capita production of DSW and the demographic and socio-economic covariates in Section 3. The joint posterior distribution of model parameters are obtained using 10,000 iterations discarding the first 1,000 using `R`'s `BayesSAE` package [1]. Model selection allows to determine relevant predictors as reported in Tab. 1. Estimated coefficients suggest positive effect of the population density and the added value covariates on the per capita DSW production; opposite signs are estimated for the number of housing and the unemployment rate implying higher production of differentiated waste fraction in wealthier provinces. The estimated mean spatial component in Fig. 1 shows the areal effect on the whole Italian territory highlighting the intrinsic correlation between similar provinces in terms of socio-economic characteristics: northern provinces are typically more productive and prosperous then the rest of the country. These preliminary results show that the spatial parametrization plays a significant role in modelling waste data. However, further discussion about the spatial component should be considered taking into account different diagnostic tools and alternative parameterizations of the spatial correlation.

Table 1: Posterior means of fixed effects and 95% Highest Posterior Density Intervals.

Parameter	mean	lower HPDI	upper HPDI
β_{PopDen}	0.144	0.058	0.217
β_{nHous}	−0.186	−0.275	−0.097
β_{UnempR}	−0.017	−0.033	−0.001
β_{AddV}	0.007	0.003	0.011

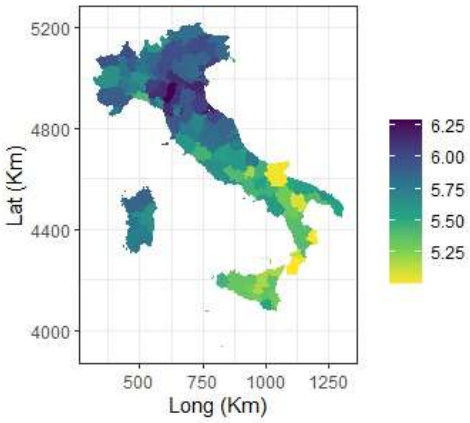


Fig. 1: Estimated mean effect of provinces on the DSW production per capita for 2019

References

1. Chengchun, S.: BayesSAE: Bayesian Analysis of Small Area Estimation. R package version 1.0-2. https://CRAN.R-project.org/package=BayesSAE, (2018)
2. Fay, R.E. and Herriot, R.A.: Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, **74**, 269–177, (1979)
3. ISPRA: Rapporti 331/2020. ISBN: 978-88-448-1030-6 (2020)
4. Keser, S., Duzgun, S., Aksoy, A.: Application of spatial and non-spatial data analysis in determination of the factors that impact municipal solid waste generation rates in Turkey. Waste Management **32**, 359–371 (2012)
5. Rao, J.N.K. and Molina, I.: Small Area Estimation, 2nd edn. Wiley, Hoboken, New Jersey (2015)
6. You, Y. and Zhou, Q.M.: Hierarchical Bayes small area estimation under a spatial model with application to health survey data. Survey Methdology, **37**, 1, 25–37, (2011)
7. Zhang, G., Lin, T., Chen, S. et al.: Spatial characteristics of municipal solid waste generation and its influential spatial factors on a city scale: a case study of Xiamen, China. J Mater Cycles Waste Manag, **17**, 399–409 (2015)

A spatial regression model for for predicting abundance of lichen functional groups

Un modello di regressione spaziale per la prevsione dell'abbondanza dei gruppi funzionali lichenici

Pasquale Valentini, Francesca Fortuna, Tonio Di Battista and Paolo Giordani

Abstract Lichen functional traits such as growth, photosynthetic and reproductive strategies, are considered indicators of changes in environmental conditions resulting especially from air pollution. However, due to the high variability of lichen flora, it is often difficult to differentiate the effects of atmospheric pollutants and those of other environmental variables. The aim of this paper is to evaluate the synergistic effect of atmospheric pollutants and environmental variables on lichen functional groups distribution through a Bayesian generalized spatial regression model.

Abstract *I tratti funzionali dei licheni, come la crescita, le strategie fotosintetiche e riproduttive, sono considerati importanti indicatori dei cambiamenti ambientali derivanti soprattutto dall'inquinamento atmosferico. Tuttavia, a causa dell'alta variabilita' della flora lichenica e' spesso difficile differenziare gli effetti degli inquinanti atmosferici da quelli di altre variabili ambientali. Questo articolo ha lo scopo di valutare l'effetto degli inquinanti atmosferici e delle variabili ambientali sulla distribuzione dei gruppi funzionali lichenici attraverso un modello di regressione bayesiano generalizzato.*

Key words: Bayesian spatial regression model, Lichen biodiversity, Air pollution, Biomonitoring techniques, Ecological functional traits

---

Pasquale Valentini
"G.d'Annunzio" University of Chieti-Pescara (Italy) e-mail: pasquale.valentini@unich.it

Francesca Fortuna
"Roma Tre" University, Rome (Italy) e-mail: francesca.fortuna@uniroma3.it

Tonio Di Battista
"G.d'Annunzio" University of Chieti-Pescara (Italy) e-mail: dibattis@unich.it

Paolo Giordani
"University of Genoa", Genoa (Italy) e-mail: giordani@dipteris.unige.it

1 Introduction

Lichens are particularly sensitive to environmental stresses due to their physiological characteristics and respond to phytotoxic gases at cellular, individual and community level [9]. However, different species react to pollutants in different ways, thus, it is often difficult to identify a direct and clear relationship between lichen biodiversity and pollution [2]. Moreover, the high variability of lichen diversity makes difficult to differentiate the effects of atmospheric pollutants and those of other environmental variables since lichen flora highly vary across geographic, climatic and ecological gradients [5].
For these reasons, approaches based on morpho-functional species traits (such as photosynthetic strategy, growth form or reproductive strategy) have been recently used to assess monitoring change in ecosystems [5, 8]. Functional traits allows to define functional groups, that is, groups of species that share some functional characteristics and, thus, react similarly to an environmental factor. However, the influence of environmental conditions on lichen functional traits is still poorly documented, hindering their use in environmental monitoring [5].
The aim of this paper is to examine the relationships between lichen functional groups and some covariates linked to environmental, habitat and air pollutant variables. At this purpose, we aim to determine the spatial pattern of lichens functional groups in response to a spatial gradient of pollution alteration. In this paper we introduce a lognormal and gamma mixed NB spatial regression model for counts [10]. In particular, the model proposed are presented in a hierarchical framework. This allows the inclusion of a "nugget" term in the spatial part of the model. The full model is estimated in a Bayesian setting and posterior inference is performed hierarchically via a Markov chain Monte Carlo scheme. The remainder of the paper proceeds as follows. In section 2 we briefly describe the data used in this study, while in section 3 we introduce the model. In section 4 we consider Bayesian inferential issues and, finally, in section 5 concludes the paper with a discussion.

2 The data

Epiphytic lichen biodiversity of Liguria region, in Northwestern Italy, has been considered. Data on lichen abundance has been collected following the standards suggested by [1]; the survey lasted from 2002 to 2003 and involved a total of 151 sampling sites of a 30 m radius plot (see Fig. 1) and 196 epiphytic lichen species i.e. lichens which live on trees bark [4].
Although the dataset does not refer to particularly up-to-date data, it has been selected as a model database to describe lichen colonisation processes under heterogeneous environmental conditions and included in The PREDICTS project (Projecting Responses of Ecological Diversity In Changing Terrestrial Systems), which has collated representative database of comparable samples of biodiversity from multiple sites that differ in the nature or intensity of human impacts relating to land use [6].

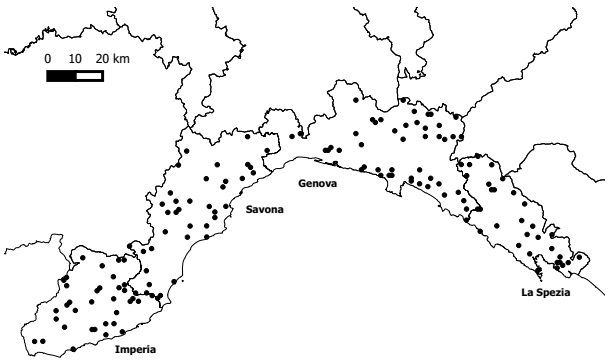


Fig. 1 Location of the sample sites in Liguria region

Twelve lichen functional groups have been identified by aggregating the species according to the growth form, distinguishing between macrolichens (macro) and microlichens (micro); the reproductive strategy , distinguishing between sexual (sex) or asexual (asex) and the nitrogen-tolerance, distinguish between oligotrophic (oligo), mesotrophic (meso) and nitrophytic (nitro) species.
Data on atmospheric pollutants have been obtained from the Regional Inventory. Spot, dispersed and linear emissions of main pollutants have been estimated by means of an ISC3 (Industrial Source Complex) long term diffusional model [3]. The model has been applied for each $1km^2$ cell on the basis of pollutant concentrations measured by automatic gauges throughout the survey area. For each sampling site, the total annual emissions of the main atmospheric pollutants (NO_x, SO_x and PM_{10}) in 1995 and 2001 have been calculated as the average of all $1km^2$ cells within a $3km$ buffer. Several studies have shown that significant changes in species diversity and composition as a consequence of substantial impact of atmospheric pollution are generally observed in a time span ranging from two up to ten years [7]. For this reason, we have considered data of pollutants emissions referred to two time lags (1995 and 2001), respectively corresponding to five and eight years before the biomonitoring survey.
As potential drivers of lichen diversity, three main habitats have been also considered in the model, including broad-leaved forest areas; conifers and beech forests; and non-forested rural and urban areas.

3 Model

The general model that we propose here is for spatial data recorded at n sites $\mathbf{s}_i$, $i = 1,\ldots,n$. Let $\mathbf{Y} = (Y(\mathbf{s}_1),\ldots,Y(\mathbf{s}_n))'$ denote the n dimensional vector. The data model is based on the conditionally independent equation for the variables under analysis, in particular

$$Y(\mathbf{s}_i)|p(\mathbf{s}_i) \overset{ind}{\sim} NB(k, p(\mathbf{s}_i)), \quad i = 1,\ldots,n \tag{1}$$

where $p(\mathbf{s}_i) = \frac{e^{\psi(\mathbf{s}_i)}}{1+e^{\psi(\mathbf{s}_i)}}$, $\psi(\mathbf{s}_i) = logit\,[p_k(\mathbf{s}_i)] = \tilde{\psi}(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i)$ and $\varepsilon(\mathbf{s}_i) \sim N(0,\sigma^2)$ for $i = 1,\ldots,n$. The spatial process $\tilde{\psi} = (\psi(\mathbf{s}_1),\ldots,\psi(\mathbf{s}_n))'$ is thought to be the sum of parametric systematic components θ and a spatial process denoted by ν. Thus we assume that $\tilde{\psi} = \theta + \nu$ where the error term ν is assumed to be zero mean Gaussian with covariance matrix Σ_ν. In this paper we consider the exponential covariance function.
The systematic component θ is assumed to be a product of a design matrix by unknown coefficient vector (i.e. $\theta = \mathbf{X}\beta$).
In the proposed model,conditional on $\tilde{\psi}$ there are two free parameters k and σ^2 to adjust both the mean and the total variance, which become the same as those of the Negative binomial model when $\sigma^2 = 0$, and the same as those of the lognormal-Poisson model when k^{-1} [10].

4 Inference and computation

4.1 Prior information

The Bayesian specification of the model is completed with the definition of the prior distributions to the model parameters.
We assume a Normal prior distribution with mean zero and variance chosen to be large to make the priors relatively non-informative. Finally, a non informative Gamma prior is used for the parameters α_{tj} and κ_j^k. In particular, we consider the following priors $\beta \sim N(\mathbf{0},\Sigma_\beta)$, $\sigma^{-2} \sim Ga(a_1,b_1)$, $k \sim Ga(a_2,b_2)$ and $b_2 \sim Ga(a_3,b_3)$.
For ψ, we choose the exponential correlation function with partial sill ρ_1 and decay parameter ρ_2. For the decay parameter, we use a discrete uniform prior distribution such that ρ_{2k} can only take values that are within a plausible interval determined by the scale for the data locations. While for the partial sill we assume $\rho_1 \sim Ga(a_4,b_4)$

4.2 Posterior inference

Posterior inference for the proposed model is facilitated by MCMC algorithms. Standard MCMC samplers are easily adapted to our model specification such that posterior analysis is readily available. All the parameters are either sampled from normal or gamma full conditional distributions or by simple Metropolis-Hastings steps.

5 Application

In this section, we consider whether the model proposed can be useful for modelling the data set introduced in Section 2. Lichens are symbiotic organisms widely used as ecological indicators to monitor the effects of environmental changes. Understanding the impact of specific variables (i.e. pollutants) on abundance of lichen species is thus of great interest for environmental agencies.
For the fitted model, the MCMC algorithm was run for 30000 iterations. Posterior inference has been based on the last 30000 draws using every 5th member of the chain to avoid autocorrelation within the sampled values. Convergence of the chains of the model has been monitored visually through trace plots.
This study offers a model for understanding functional lichen response related to the ecosystem as a whole. The results highlight differential responses for groups of species sharing different functional traits. Moreover, the model has shown that lichen communities are affected primarily by pollution emissions of past years rather than by the more recent levels of contamination. For these reasons, the model presents potential implications for the use of lichen functional groups and, thus, to characterize areas of high environmental risk.

References

1. Asta, J., Erhardt, W., Ferretti, M., Fornasier, F., Kirschbaum, U., Nimis, P.L., Purvis, O.W., Pirintos, S., Scheidegger, C., Haluwyn, C.V., Wirth, V.: Mapping lichen diversity as an indicator of environmental quality, in: Nimis, P.L., Scheidegger, C., Wolseley, P.A. (Eds.), Monitoring with Lichens-Monitoring Lichens, Nato Science Program-IV, Kluwer Academic Publisher, The Netherlands, 273–279 (2002)
2. Ammann, K., Herzig, R., Liebendoerfer, L., Urech, M.: Multivariate correlation of deposition data of 8 different air pollutants to lichen data in a small town in switzerland. Advanced Aerobiology **51**, 401–406 (1970)
3. European Environment Agency: Corine land cover update 2000. EEA, Copenhagen (2003)
4. Giordani, P.: Variables in uencing the distribution of epiphytic lichens in heterogeneous areas: a case study for Liguria, NW Italy. Journal of Vegetation Science **17**, 195–206 (2006)
5. Giordani, P., Brunialti, G., Bacaro, G., Nascimbene, J.: Functional traits of epiphytic lichens as potential indicators of environmental conditions in forest ecosystems. Ecological Indicators **18**, 413–420 (2012)

6. Hudson, L. N., Newbold, T., Contu, S., Hill, S. L. L., Lysenko, I., De Palma, A., ... Purvis, A.: The database of the PREDICTS (Projecting Responses of Ecological Diversity In Changing Terrestrial Systems) project. Ecology and Evolution **7**, 145–188 (2017)
7. Loppi, S., Frati, L., Paoli, L., Bigagli, V., Rossetti, C., Bruscoli, C., Corsini, A.: Biodiversity of epiphytic lichens and heavy metal contents of Flavoparmelia caperata thalli as indicators of temporal variations of air pollution in the town of Montecatini Terme (central Italy). Science of the Total Environment **326**, 113–122 (2004)
8. Pinho, P., Bergamini, A., Carvalho, P., Branquinho, C., Stofer, S., Scheidegger, C.: Lichen functional groups as ecological indicators of the effects of land-use in mediterranean ecosystems. Ecological Indicators **15**, 36–42 (2012)
9. Van Dobben, H. F., Wolterbeek, H.T., Wamelink, G.W., Ter Braak, C.J.: Relationship between epiphytic lichens, trace elements and gaseous atmospheric pollutants **112**, (2), 163–169 (2001)
10. Zhou, M., Carin L.: Negative Binomial Process Countand Mixture Modeling," arXiv:1209.3442, Sept. 2012.

2.4 Advances in preference and ordinal data theoretical improvements and applications

Boosting for ranking data: an extension to item weighting

Alberi decisionali per la classificazione di rankings: un'estensione ponderata del Boosting

Alessandro Albano, Mariangela Sciandra, Antonella Plaia

Abstract Decision tree learning is one of the most popular families of machine learning algorithms. These techniques are quite intuitive and interpretable but also unstable. It is necessary to use ensemble methods that combine the output of multiple trees, to make the procedure more reliable and stable. Many approaches have been proposed for ranking data, but they are not sensitive to the importance of items. For example, changing the rank of a highly-relevant element should result in a higher penalty than changing a negligible one. Likewise, swapping two similar elements should be less penalized than swapping two dissimilar ones.
This paper extends the boosting ensemble method to weighted ranking data, proposing a theoretical and computational definition of item-weighted boosting. The advantages of this procedure are shown through an example on a real data set.

Abstract *Gli alberi decisionali sono una tecnica predittiva di machine learning particolarmente diffusa, utilizzata per prevedere delle variabili discrete (classificazione) o continue (regressione). Gli algoritmi alla base di queste tecniche sono intuitivi e interpretabili, ma anche instabili. Infatti, per rendere la classificazione più affidabile si è soliti combinare l'output di più alberi. In letteratura, sono stati proposti diversi approcci per classificare ranking data attraverso gli alberi decisionali, ma nessuno di questi tiene conto né dell'importanza, né delle somiglianza dei singoli elementi di ogni ranking. L'obiettivo di questo articolo è di proporre un'estensione ponderata del metodo boosting per ranking, che tenga conto della struttura di similarità e dell'importanza dei singoli elementi. I vantaggi di questa procedura sono mostrati con un esempio su un dataset reale.*

Key words: boosting, weighted ranking data, ensemble methods, decision trees

---

Alessandro Albano, Mariangela Sciandra, Antonella Plaia
Department of Economics, Business and Statistics, University of Palermo, e-mail: alessandro.albano@unipa.it, mariangela.sciandra@unipa.it, antonella.plaia@unipa.it

1 Introduction

Breiman et al. (1984) developed Classification and Regression Trees (CART) as an alternative non-parametric approach to classification and regression parametric procedures. It is known that the decision trees from CART suffer from high variance, i.e., the decision trees learned from different data sub-samples may be quite different. For this reason, Breiman (1996) suggested improving the accuracy of decision trees by perturbing the training set, using bootstrapping, and then combining the multiple decision trees into a single predictor. These procedures belong to the class of the so-called Perturb and Combine (P&C) methods. One of the best known P&C methods, Boosting (Freund et al., 1996), aims at a fast reduction in the training set errors. The key idea is to increase the probability of being drawn in the iterations for examples misclassified in previous iterations. Many efforts have been made to define ensemble methods for ranking data (Plaia et al. 2017, Plaia and Sciandra 2019), but none considers the possibility to give different importance to items: can each element of the ranking contribute differently to the classification tree's growing?
This paper aims to define an item-weighted version of the boosting algorithm for rankings and evaluate this algorithm on real data. The paper is organized as follows: Section 2 introduces ranking data and describes an item-weighted distance for ranking. In Section 3, an item-weighted boosting algorithm is introduced, and the steps for the implementation in the `R` statistical software environment are described. Finally, in Section 4, the procedure is applied to a real dataset. Conclusions will follow.

2 Ranking data

Preference data arise when a group of n individuals express their preferences on a finite set of items (m different alternatives of objects). Preference data can be expressed in the form of rankings when alternatives are fixed in any pre-specified order, and preferences are defined by using integers to indicate the rank of each alternative. Specifically, π is a mapping function from the set of items $\{1,...,m\}$ to the set of ranks $\pi = (\pi(1), \pi(2), ..., \pi(m))$, where $\pi(i)$ is the rank given by a judge to item i. If the m items are ranked in m different ranks, a complete (full) ranking or linear ordering is achieved (Cook, 2006). In certain cases, some items could receive the same preference, then a tied ranking or a weak ordering is obtained.
Often covariates may also be considered to explain individual differences in the evaluation of choice alternatives. In this case, defining how much each covariate contributes to identify clusters of "similar" respondents is a crucial issue to be addressed.

2.1 Item-weighted distance for ranking data

In the framework of preference data, an interesting issue is to measure the spread between rankings through dissimilarity or distance measures.
In general, distances between rankings treat all items equally, and they are not sensitive to the point of disagreement. Kumar and Vassilvitskii (2010) introduced the issue of element weights. In brief, swapping important items should receive a larger penalization than swapping negligible ones.
Albano and Plaia (2021) proposed an item-weighted version of the Kemeny distance (Kemeny, 1959) by considering a weighting vector $w = (w_1, w_2, ..., w_m)$, where $w_i \geq 0$ is the importance given to the i-th item in a ranking. The item-weighted distance d_k^{ew} between two m-size rankings, π and π^* is:

$$d_k^{ew}(\pi, \pi^*) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j \left| a_{ij} - b_{ij} \right|, \tag{1}$$

where a_{ij} and b_{ij} are the generic elements of the score matrices defined by Emond and Mason (2002). The corresponding item-weighted rank correlation coefficient (defined as an extension of τ_x provided by Emond and Mason (2002)) is:

$$\tau_x^{ew}(\pi, \pi^*) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j a_{ij} b_{ij}}{Max[d_k^{ew}]} \tag{2}$$

where the denominator represents the maximum value of the weighted Kemeny distance $d_{max}^{ew} = \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j$.
In some instances, the weights could be assigned following the idea that swapping two elements that can be considered similar in some aspect, should be less penalized than swapping two dissimilar ones. In this setting, a symmetric penalization matrix **P**, reflecting the dissimilarity among the elements, is needed. Therefore, Eqs (1) (2) are modified by replacing $w_i w_j$ with p_{ij}, where p_{ij} (≥ 0) is the generic element of the **P** matrix.

3 Item-weighted Boosting for ranking data

Decision trees (Breiman et al., 1984) are a simple non-parametric statistical methodology that allows a recursive partitioning of the predictors, such that observations with similar response values will be grouped.
For item-weighted ranking data it is possible to use the element-weighted kemeny distance d_k^{ew} (Eq. (1)) as impurity function. Nevertheless, the instability remains an issue. It is well known that in decision tree learning, the classifier predictive performance can be substantially improved by aggregating many decision trees. Dery and Shmueli (2020) demonstrated that improvements of ensemble methods over a single decision tree happen with ranking data as well. Therefore, the item-weighted

boosting procedure is proposed, and its prediction accuracy performance is studied.

3.1 Boosting Algorithm

Among the different versions of boosting algorithms in the literature, our proposal is based on AdaBoost.M1 (Freund et al., 1996), opportunely adapted to item-weighted ranking data. In the following, two vectors of weights will be used:

- the set of working weights p_b to be updated at each b iteration of the algorithm. More specifically, p_b represents the probability of each record to be included in the bootstrap sample;
- the vector of weights w, representing the importance of each item in the ranking (as defined in section 2.1). The importance of each item remains fixed during the procedure.

Algorithm 1 AdaBoost.R - Item-weighted boosting for ranking data

Input: A training set T, a number of iterations B, a vector of weights w
Output: a ranker $C_f(.)$ that maps a given x to a ranking of the labels

1: initialize $p_b(i) = 1/n \; \forall i = 1,2,...n$
2: **for** $b \leftarrow 1$ to B **do**
3: take a sample T_b, drawn from the training set T using weights $p_b(i)$
4: fit a ranking tree $C_b(.)$ to T_b
5: $e_b = \sum_{i \in T_b} p_b(i) \left(1 - \frac{\tau_x^{ew}(i)+1}{2}\right)$ where $\tau_x^{ew}(i) = \tau_x^{ew}(C_b(x_i), y_i)$
6: $\alpha_b = \frac{1}{2} \ln((1-e_b)/e_b)$
7: update the weights $p_{b+1}(i) = p_b(i) exp\left(\alpha_b \left(1 - \frac{\tau_x^{ew}(i)+1}{2}\right)\right)$ and normalize them
8: **end for**
9: $C_f(x_i) = \arg\max_{y_i \in S^m} \sum_{b=1}^{B} \alpha_b \tau_x^{ew}(C_b(x_i), y_i))$

The procedure is described in details in the Algorithm 1. At each iteration b, a tree $C_b(.)$ is trained on T_b leading to a predicted ranking for each item $\tilde{y}_i^b = C_b(x_i)$ (steps 3 and 4).
The ranking error e_b of the ranking tree $C_b(.)$ is estimated employing the distance between each predicted ranking $\tilde{y}_i^b$ and its real value y_i (step 5). Then, a factor α_b is computed, as a function of e_b, for updating the weights $w_b(i)$ (step 6). Finally, the weights $w_b(i)$ are updated and normalized after each iteration (step 7).
To obtain a final prediction, the item-weigthed boosting uses rank aggregation (Amodio et al. 2016; D'Ambrosio et al. 2017) to combine the predictions of each individual trees.
The aggregated ranking for a generic i-th observation at the b-th iteration is $\hat{y}_{ib} = \arg\max_{s^m} \sum_{k=1}^{b} \alpha_b \tau_x^{ew}(\tilde{y}_i^k, y_i)$, where α_b is the weight related to the b-th tree.
The aggregated error, after each iteration b, is $err(b) = 1 - \frac{\tau_x^{ew}(b)+1}{2}$, where $\tau_x^{ew} = \frac{1}{n}\sum_{i=1}^{n} \tau_x^{ew}$ is the average of τ_x^{ew} of the b-th tree over all the units in an example set T.
Furthermore, the procedure allows to determine the overall covariates' importance by averaging over their importance resulting in each of the b trees, with weights α_b.

4 A real data example: the German Elections dataset

To investigate the performance of the item-weighted boosting method, an application to GermanElections2009 data is shown in this Section. The dataset (Dery and Shmueli, 2020), contains socio-economic information from regions of Germany and its electoral results. The 413 records correspond to the administrative districts of Germany, which are described by 39 covariates. The outcome is the set of rankings on five items: CDU (conservative), SPD (centre-left), FDP (liberal), Green (centre-left) and Left (left-wing).
The item-weighted version of Boosting follows the intuition that swapping two similar parties should be less penalized than swapping two dissimilar ones. Therefore, we introduce the penalization matrix shown in Table 1. The item-weighted boosting

Table 1 Penalization matrix **P**.

	CDU	SPD	FDP	Green	Left
CDU	0	75	25	75	100
SPD	75	0	50	0	25
FDP	25	50	0	50	75
Green	75	0	50	0	25
Left	100	25	75	25	0

algorithm was performed on a limited number of trees ($B = 100$) and considering a depth (number of the splits in the tree) equal to 4. Fig. 4 shows the error of the item-weighted boosting applied to German Elections dataset with penalization matrix **P** (Table 1). The Boosting procedure was applied, considering only 274 observations as a training set. As expected, the accuracy improves when the number of trees grows up. The error with just one tree is 0.093 in the training and 0.099 in the test set, while, it reduces to respectively to 0.072 and 0.0826, with 100 trees. Therefore, the item-weighted boosting minimizes the prediction error taking into account the similarity structure of items.

5 Conclusions

In this paper, we investigated the role of ensemble methods for item-weighted ranking decision trees. An item-weighted version of the boosting procedure for ranking data has been proposed and implemented in `R`, by incorporating some functions of `ConsRank` package (D'Ambrosio et al., 2017) within a user-written split function of `rpart` library (Therneau et al., 2015). The item-weighted boosting can be used as an interpretative method to select and measure the overall covariates' importance, instead of a "black box" that forecasts without a clear understanding of the underlying rules. The novelty of this procedure is to provide an item-oriented approach where the structure, the interpretation and the criteria underlying the derivation of the tree itself are revised. In conclusion, this method is particularly fitting when dealing with multi-level data. The performance of the proposed method has been shown through an example in Section 4, where the data matrix contains rankings of

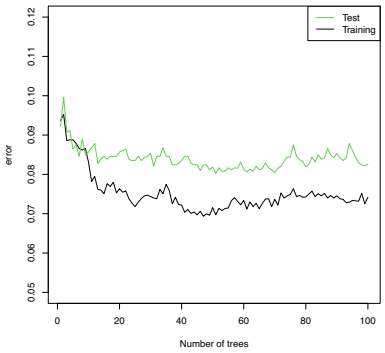


Fig. 1 Item-weighted boosting applied to German Elections dataset: $err(b)$.

political parties (level 1) who belong to political coalitions (level 2). In this case, an unweighted boosting procedure (which does not take into account the similarity of political parties) would lead to less accurate results.
Future research should consider the potential effects of weights more carefully, to improve the scalability of the algorithms with respect to the number of items.

References

Albano, A. and Plaia, A. (2021). Element weighted kemeny distance for ranking data. *In press*.

Amodio, S., D'Ambrosio, A., and Siciliano, R. (2016). Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the kemeny axiomatic approach. *European Journal of Operational Research*, 249(2):667–676.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cook, W. D. (2006). Distance-based and ad hoc consensus models in ordinal preference ranking. *European Journal of operational research*, 172(2):369–385.

D'Ambrosio, A., Amdio, S., and Mazzeo, G. (2017). Consrank-package: Median ranking approach according to the kemeny's axiomatic...

Dery, L. and Shmueli, E. (2020). Boostlr: a boosting-based learning ensemble for label ranking tasks. *IEEE Access*, 8:176023–176032.

Emond, E. J. and Mason, D. W. (2002). A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, 11(1):17–28.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer.

Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus*, 88(4):577–591.

Kumar, R. and Vassilvitskii, S. (2010). Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web*, pages 571–580.

Plaia, A. and Sciandra, M. (2019). Weighted distance-based trees for ranking data. *Advances in data analysis and classification*, 13(2):427–444.

Plaia, A., Sciandra, M., and Muro, R. (2017). Ensemble methods for ranking data. In *CLADAG 2017*, pages 1–6. Universitas Studiorum Srl Casa Editrice.

Therneau, T., Atkinson, B., Ripley, B., and Ripley, M. B. (2015). Package 'rpart'. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*.

An Extended Bradley-Terry Model For The Analysis Of Financial Data

Un'estensione del modello Bradley-Terry per l'analisi di dati finanziari

Alessio Baldassarre, Elise Dusseldorp, Mark De Rooij

Abstract The models for matched pairs allow analyzing *M* observations on *N* items. They differ from typical GLM models by permitting each item to have its probability distribution. Here, we follow the extended Bradley-Terry model with subject-specific covariates, in which observations consist of pairwise comparisons between ranked items. We applied this model to financial data, where the observations are represented by countries and the items by different types of government tax revenues. The extended Bradley-Terry model works as the basis for a new partitioning model that follows the regression trunk model. The aim is to create a partition of countries, subject to comparing their tax revenues categories and their socio-economic characteristics. The result gives a probability distribution in each terminal node for the *N* tax revenues, the main effects coefficients, and the most relevant interactions between subject-specific covariates.

Abstract *I modelli matched pairs permettono di analizzare M osservazioni su N oggetti, ognuno dei quali segue una propria distribuzione di probabilità. In questo lavoro viene utilizzato il modello Bradley-Terry esteso con subject-specific covariates, in cui le osservazioni sono rappresentate da comparazioni a coppie tra oggetti ordinati. Tale modello è stato applicato ad un campione di Paesi i cui oggetti sono rappresentati dalle rispettive entrate suddivise per tipologia di tassazione. Con il modello regression trunk si ottiene una ripartizione dei Paesi, in base alla comparazione delle proprie voci di tassazione e delle loro caratteristiche socio-economiche. I risultati mostrano la distribuzione di probabilità in ogni nodo terminale per ciascuna voce di entrata, gli effetti principali e le interazioni più significative tra le variabili considerate.*

Alessio Baldassarre, Business and Economics Department
Università degli Studi di Cagliari, e-mail: al.baldassarre1@gmail.com

Elise Dusseldorp, Methodology and Statistics, Institute of Psychology
Universiteit Leiden, e-mail: elise.dusseldorp@fsw.leidenuniv.nl

Mark De Rooij, Methodology and Statistics, Institute of Psychology
Universiteit Leiden, e-mail: rooijm@fsw.leidenuniv.nl

Key words: Bradley-Terry model, regression trunk, public finance

1 Introduction

Public finance is a branch of economics, and its goal is to study the government's role in the economic system (Gruber, 2005). One of the public authorities' goals is to find the right balance between government revenues and government expenditures to achieve desirable effects and avoid undesirable ones (Jain, 1974).

Tax revenues and government expenditures represent two components that influence each state's GDP directly or indirectly. However, it is not always clear how they interact between each other and in what direction they affect the economic wealth. Several types of research are based on the study of the causal relationship between taxation and public spending. Manage and Marlow (1986) showed that taxation causes expenditure at the state level of government, but that such causation becomes bidirectional in the short run. In contrast, Anderson et al. (1986) concluded that government expenditures Granger-cause government taxes. Several methodologies have been used in this context, including the vector autoregression model (Von Furstenberg, Green & Jeong 1986) and the conventional Granger-causality (Ram, 1988).

Our work presents an application that differs from those made previously. Specifically, the purpose of the analysis is to study the relationship between taxation, public spending, and other socio-economic components for a sample of countries worldwide. Unlike the works cited, we decided to separate government revenues based on tax categories and then order them according to their size. By doing this, it is possible to apply the Bradley-Terry for matched pairs model (Bradley & Terry, 1952), using the extended version with subject-specific covariates (Dittrich, Hatzinger & Katzenbeisser 1998). Finally, the regression trunk model is applied to search for the interactions between variables that most affect the comparisons between taxation items. The result is a small regression tree that partitions the data set and provides useful information about the main effects, the interaction effects, and the tax ordering in each terminal node.

Section 2 presents the methodology adopted with references to the log-linear Bradley-terry model with subject-specific covariates and the regression trunk model. The application on countries' tax revenues produces the results shown in Section 3. Lastly, Section 4 presents a short discussion about the achievable results with this approach.

2 Methodology

For a representative sample of $M = 100$ countries, the relationship between public revenues and government expenditures is analyzed. According to the OECD classi-

fication, we diversify revenues by $N = 4$ types of taxation (taxes on income, social security contributions, taxes on property, and taxes on goods). We applied the same procedure to the government expanses, also classified by type according to the COFOG classification (OECD data, 2018). A high number of socio-economic variables P are also considered, such as the GDP growth rate, unemployment rate, level of imports, etc. All variables refer to the year 2018, which is the last year for sufficient financial data we needed (World Bank data). They are scaled and expressed in terms of GDP to control the country's economic size.

Subsequently, we ranked the tax items by size to allow the Bradley-Terry model application with subject-specific covariates (Dittrich et al., 1998). This model belongs to the category of matched pairs models (Agresti, 2007), in which the observations are constituted by comparisons of pairs of objects, which in our case are constituted by the N tax revenues categories. For N objects it is possible to compute $n \times (n-1)/2$ paired comparisons.

Let Π_{ij} denote the probability that government revenues derived by tax category i are higher than j. The probability that another tax category j is greater than i is equal to $\Pi_{ji} = 1 - \Pi_{ij}$ (ties cannot occur in this application). The Bradley-Terry model has item parameters β_i such that (Agresti, 2007)

$$logit(\Pi_{ij}) = \log\left(\frac{\Pi_{ij}}{\Pi_{ji}}\right) = \beta_i - \beta_j. \tag{1}$$

Since taxation systems are different between M countries belonging to different regions, we have considered continuous social-economic variables, called subject-specific covariates. The basic Bradley-Terry model is a logistic model, and it can be expressed in a log-linear form (Fienberg & Larntz 1976; Sinclair, 1982), where the number of times in which i is greater than j follows a Poisson distribution. The log-linear Bradley-terry model is expressed as

$$ln(e(y_{ij;m})) = \mu_{ij;m} + y_{ij;m}(\lambda_{i;m} - \lambda_{j;m}), \tag{2}$$

where $e(y_{ij;m})$ is the expected number of comparisons in which i is greater than j for the observation m. Then, $y_{ij;m} \in \{-1, 1\}$ indicates whether for country m tax category i is more profitable than j ($y_{ij,m} = 1$) or not ($y_{ij,m} = -1$).

The parameter $\lambda_{i;m}$ can be expressed through a linear relation

$$\lambda_{i,m} = \lambda_i + \sum_{p=1}^{P} \beta_{ip} x_{p;m}, \tag{3}$$

where, $x_{p;m}$ is the pth country-specific covariate ($p = 1...P$) associated with the country m. The parameters β_{ip} express the main effects of the P covariates on the tax category i. Then, λ_i is the intercept that indicates the location of the tax category i in the overall ordering scale. This parameter is obtained through the equation $\lambda_i = \frac{1}{2} log(\pi_i)$, where π_i is the probability of the tax revenue category i being higher than the other categories.

The combination of the Bradley-Terry model with a regression tree allows creating a partition of the observations based on the comparisons and the subject-specific characteristics themselves.

Here, we apply the regression trunk model, which combines a multiple regression model and a regression tree (Dusselsorp & Meulman 2004). It solves the problem associated with additive models, whose interpretation is tricky when there are significant interactions between covariates. Furthermore, it reduces the difficulty associated with tree-based models, which capture linear effects between variables. One of the regression trunk strengths is that it does not require a priori knowledge about interaction effects. In its generic form, the final model is expressed as follows

$$\hat{Y} = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p + \sum_{t=1}^{T-1} \hat{\beta}_{P+t} I\{(X_1, ..., X_P) \in t\}, \tag{4}$$

where T is the total number of terminal nodes and $T-1$ the number of indicator variables I. The excluded region acts as a reference group, whose estimated intercept is $\hat{\beta}_0$. For the node t the estimated intercept is $\hat{\beta}_0 + \hat{\beta}_{P+t}$. The first piece of the equation estimates the linear part of the model consisting of the main effects. The second piece indicates the interactions obtained with the partitioning process.

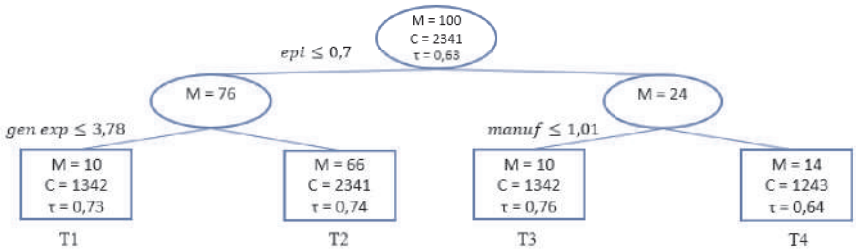
The tree is constructed by fitting logistic regressions and combining regression models with main effects with a low number of higher-order interaction effects. The split criterion used is based on reducing the deviance between one model and the next one. The final pruning is carried out by applying the V-fold cross-validation for each step of the tree construction. The method used here follows the CART procedure (Breiman & Friedman 1984) and the STIMA algorithm (Dusselsorp, Conversano & Van Os, 2010; Conversano & Dusseldorp, 2017).

3 Results

The combination of the extended Bradley-Terry model with the regression trunk model generates a tree that represents a compromise between the interpretability of the results (few interaction effects) and countries' distribution efficiency based on the relative ordering of the N tax revenue categories. Once the entire tree has been built, it is pruned to avoid overfitting problems so that the final regression trunk can be represented as in Figure 1.

In the root node and each terminal nodes $T1, ...T4$ are shown the number of countries, the consensus rankings C calculated with the minimization of the distance between rankings inside the terminal nodes, and the correlation coefficients τ within the terminal nodes. These indices are mainly used for the analysis of preference data with distance-based methods (D'Ambrosio & Heiser, 2016). Here, they are shown for the sake of completeness. It can be seen that the first split variable is represented by the environmental performance index EPI. This index has been used as a proxy for countries' environmental spending, and the fact that it is chosen as the

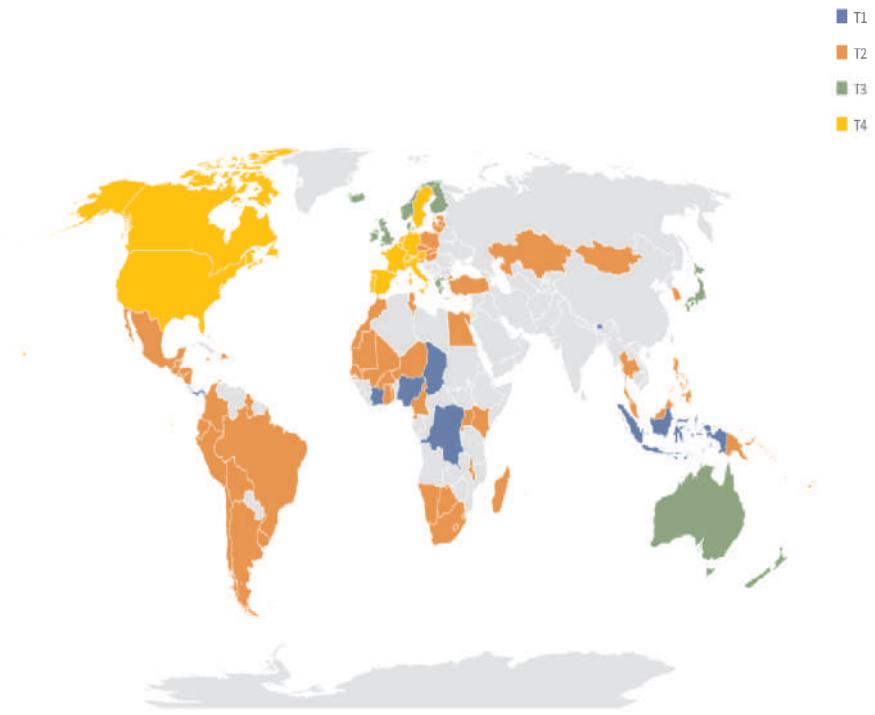
Figure 1. Pruned regression trunk



first split variable leaves room for numerous reflections on the impact of environmental performance on public finance. Furthermore, the most relevant interactions are those between EPI and the residual items of public expenditure (i.e., general exp.) as well as between EPI and the size (in terms of GDP) of the manufacturing sector (i.e., manuf.).

The countries' final partition is shown through a world map in Figure 2.

Figure 2. Countries' distribution map in each terminal node



4 Discussion

The final model allows us to find the main interaction effects and the probability distribution about the government tax revenues' ordering for each node generated by the pruned tree. For each terminal node, the model returns different coefficients for both main effects and interaction effects. In this way, the relationship between taxation and public spending emerges, and the relationship between the revenue systems and the numerous socio-economic covariates is considered.

The final result (so-called regression trunk) offers a good starting point for implementing public policies that consider the effects of socio-economic variables on the size of their tax revenues.

References

1. Agresti, A.: An introduction to categorical data analysis, Hoboken, NJ: Wiley-Interscience. (2007)
2. Anderson, W., Wallace, M.S. & Warner, J.T.: Government spending and taxation: what causes what?, Southern Economic Journal, **630**(9). (1986)
3. Bradley, R. A., & Terry, M. A.: Rank analysis of incomplete block designs. I. Biometrika, **39**, 324–345. (1952)
4. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J.: Classification and regression trees, Belmont, CA: Wadsworth International Group. (1984)
5. Conversano, C., & Dusseldorp, E.: Modeling threshold interaction effects through the logistic classification trunk. Journal of Classification, **34** (3), 399–426. (2017)
6. D'Ambrosio, A., & Heiser, W.J.: A recursive partitioning method for the prediction of preference rankings based upon Kemeny distances, Psychometrika, **81**(3), 774–794. (2016)
7. Dittrich, R., Hatzinger, R., & Katzenbeisser, W.: Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings, Journal of the Royal Statistical Society C, **47**, 511–525. (1998)
8. Dusseldorp, E., & Meulman, J. J.: The regression trunk approach to discover treatment covariate interaction, Psychometrika, **69**(3), 355–374. (2004)
9. Dusseldorp, E., Conversano, C., & Van Os, B.J.: Combining an additive and tree-based regression model simultaneously: STIMA, Journal of Computational and Graphical Statistics, **19**(3), 514-–530. (2010)
10. Fienberg, S. E. & Larntz, K.: Loglinear representation for paired and multiple comparison models, Biometrika, **63**, 245–254. (1976)
11. Gruber, J.: Public Finance and Public Policy. New York: Worth Publications. p. 2. (2005)
12. Jain, P.C.: The Economics of Public Finance. (1974)
13. Manage, & N., Marlow, M.L.: The causal relation between federal expenditures and receipts, Southern Economic Journal, 617-29. (1986)
14. Marden, J. I.: Analyzing and modelling rank data. London: Chapman & Hall (1995)
15. OECD, Revenue statistics. (2018)
16. Ram, R.: Additional evidence on causality between government revenue and government expenditure, Southern Economic Journal, **763**(9). (1988)
17. Sinclair, C. D.: GLIM for preference, Biometrika, **14**, 164–178. (1982)
18. Von Furstenberg, G.M., Jeffery Green, R. & Jeong, J. Tax and spend, or spend and tax?, The Review of Economics and Statistics, **179**(8), (1986)
19. World Bank, World Development Indicators. (2018)

2.5 Business system innovation, competitiveness, productivity and internationalization

An analysis of the dynamics of the competitiveness for some European Countries

Un'analisi delle dinamiche della competitività a livello europeo

Andrea Marletta, Mauro Mussini and Mariangela Zenga

Abstract In 2019, the European Union registered its seventh consecutive year of economic growth, but significant differences among the European Union countries are still present. In this context, competitiveness is an important element of human-centric and sustainable economic progress. Focusing on this concept, this study proposes an approach to track the dynamics of competitive growth in a set of member countries. After selecting some economic indicators for a 6-years period from 2014 to 2019, a time trajectory showing the dynamics of innovation and digitalization for each country has been plotted. This graphical analysis allows to extract some evidences about different paths to competitiveness within the European Union.

Abstract *L'Unione Europea ha registrato nel 2019 una crescita economica per il settimo anno consecutivo, ma nonostante ciò, notevoli differenze sono ancora presenti al suo interno. Possibili motivazioni di queste differenze possono essere misurate in termini di competitività. Questo studio propone un approccio utile per tracciare le dinamiche di una crescita competitiva in un gruppo di paesi membri dell'Unione Europea. Utilizzando 6 indicatori macroeconomici dal 2014 al 2019, un'analisi delle traiettorie per ogni paese ha mostrato le dinamiche legate all'innovazione ed alla digitalizzazione. Sulla base di questa analisi grafica, è stato possibile ricavare alcune evidenze sui differenti percorsi verso la competitività all'interno dell'Unione Europea.*

Key words: Competitiveness, principal component analysis, innovation, time trajectory, digitalization

---

Andrea Marletta, University of Milano-Bicocca
e-mail: andrea.marletta@unimib.it

Mauro Mussini, University of Milano-Bicocca
e-mail: mauro.mussini1@unimib.it

Mariangela Zenga, University of Milano-Bicocca
e-mail: mariangela.zenga@unimib.it

1 Introduction

In the EU economies, several common policy actions have been made to improve economic resilience for the enterprises after the economic crisis, even if significant regional differences remain. In particular, the territorial competitiveness results an important element of regional differentiation among EU countries. According to the World Economic Forum, competitiveness at the national level is the 'set of institutions, policies and factors that determine the level of productivity of a country' [12]. In this paper, the attention is posed on a pillar of competitiveness: the innovation. Theoretical and empirical studies confirm, in fact, that innovation is a key determinant of the competitiveness of enterprises and countries [11, 6, 13]. Undoubtedly, a very significant factor influencing the realization the innovation activities is the R&D expenditure from government and business sector [1]. On the other hand, considering the sectors, ICT plays a key role in enabling innovations in many technological and economic activities.

The aim of the paper is to provide an overview of the dynamics of the competitiveness for some EU countries, limiting to two dimensions of the innovation: Research & Development and digitalization.

The paper is structured as follows: the second section presents the method to analyse three-way data, the third section describes the data and the results for 10 EU countries observed for 6 years. The section four concludes the paper.

2 Methods

In this study, a set of variables for a subgroup of EU countries is observed over the 2014-2019 period. These data form a multivariate time array $\mathbf{X}$ [8, 4], the structure of which is

$$\mathbf{X} \equiv \left\{ x_{ijt} : i = 1,\ldots,I; j = 1,\ldots,J; t = 1,\ldots,T \right\} \tag{1}$$

where i is a generic unit (i.e. a country), j is one of the observed variables and t is a year within the 2014-2019 period. Such three-way data can be re-arranged to obtain the so-called multivariate time trajectories [2, 3, 4], displaying the path of each country over the years on a J-dimensional space.

The re-arrangement of the multivariate time array $\mathbf{X}$ takes place in two steps. The observed values for all countries in a given year t are selected from the multivariate time array $\mathbf{X}$, obtaining an $I \times J$ matrix which is called "slice" [8, 4].

Once a slice has been created for each year t (with $t = 1,\ldots,T$), the slices are stacked one on the top of the other until the matrix $\widetilde{\mathbf{X}}$ with $I \cdot T$ rows and J columns is achieved. The generic row of $\widetilde{\mathbf{X}}$, denoted by $\mathbf{x}_{it}$, contains the observed values for country i in year t:

$$\mathbf{x}_{it} \equiv x_{i1t},\ldots,x_{iJt}. \tag{2}$$

When a single country i is considered, the matrix displaying the time trajectory of country i is obtained by selecting the J-dimensional vectors $\mathbf{x}_{it}$, with $t = 1, \ldots, T$, from $\widetilde{\mathbf{X}}$ [5]:

$$\widetilde{\mathbf{X}}_i \equiv \{\mathbf{x}_{it} : t = 1, \ldots, T\}. \tag{3}$$

When PCA is applied to $\widetilde{\mathbf{X}}$ and only the first two PCs are held, we obtain a two-dimensional plane [7, 9, 10] in which the time trajectory of each unit is depicted in the space spanned by the first two PCs. The advantage of such an approach is that the time trajectory of a country can be displayed by connecting its PC scores, calculated for each year in the period considered, in a Cartesian plane.

3 An analysis of dynamics in competitiveness at EU level

Data are from Eurostat database, which provides high quality statistics and data on Europe. Eurostat produces European statistics in partnership with National Statistical Institutes and other national authorities in the EU Member States.

Data refers to six indicators for ten European countries from 2014 to 2019. Final dimension of the dataset is composed by 10 countries $\times$ 6 years equal to 60 rows $\times$ 6 columns representing variables.

In particular, the considered indicators are:

- Business Expenditure in R&D (BERD);
- Government Budget Appropriations or Outlays on R&D (GBAORD);
- Enterprises that provided training for ICT skills (ICTTRA);
- Enterprises that recruited ICT specialists (ICTREC);
- Employment in high- and medium-high technology (HIGEMP);
- Employed persons with ICT education by tertiary education (ICTEDU).

The BERD indicator involves data about R&D expenditure and R&D personnel broken down by the following institutional sectors: business enterprise (BES); government (GOV); higher education (HES); private non-profit (PNP); Total of all sectors. The R&D expenditure is further broken down by source of funds; type of costs; economic activity (NACE Rev.2); size class; type of R&D; fields of research and development; socio-economic objectives and by regions (NUTS 2 level). The GBAORD indicator is concerning Government Budget Allocations for R&D. GBAORD data are measuring government support to research and development activities, and thereby provide information about the priority Governments give to different public R&D funding activities. BERD and GBAORD data are available in Euro per inhabitant. The ICTTRA variable is the percentage of the enterprises that provided training to develop/upgrade ICT skills of their personnel. Data are considered for all enterprises, without financial sector (10 persons employed or more).

The ICTREC is computed as enterprise recruited/tried to recruit personnel for jobs requiring ICT specialist skills, the unit of measure is the percentage of enterprises considering all enterprises, without financial sector (10 persons employed or

more). The HIGEMP variable is represented by employment in high- and medium-high technology manufacturing sectors expressed in percentage of total employment. The ICTEDU indicator refers to employed persons with ICT education by educational attainment level, in particular, it is expressed as the percentage of employed with tertiary education (levels 5-8) over the total.

The selected countries for this study are: France and Germany belonging to the group of Central-Western Europe countries, United Kingdom and Ireland for the Anglo-Saxon area, Portugal, Italy, Greece and Spain of the Mediterranean area and Polonia and Romania for the East Europe.

Following the procedure in the previous section, a PCA has been computed on the 6 indicators obtaining 2 factors explaining 80% of the total variance. The rotated components matrix has been presented in Table 1.

Variables	PC1	PC2
BERD	**0.886**	0.401
GBAORD	**0.879**	0.383
ICTTRA	0.571	**0.672**
ICTREC	0.238	**0.906**
HIGEMP	**0.832**	-0.354
ICTEDU	-0.033	**0.701**

Table 1: Rotated components matrix for PCA

The two components are strictly related to two sub-groups of indicators: the first component is identified by BERD, GBAORD and HIGEMP and it is named "Research & Innovation", while the second one is correlated with ICTTRA, ICTREC and ICTEDU and is named "Digitalization". Beyond the coordinates of the variables, the trajectories for two countries, Germany and Greece are represented in Figure 1.

These two countries have been chosen to underline the difference in positioning for the two trajectories in the graph. The trajectory for Germany is positioned in the first quadrant of the graph, this means that Germany is a country with a high level of Research & Innovation and Digitalization. On the other hand, Greece is placed on the left side of the plot, denoting low levels of Research & Innovation, the level of digitalization is growing until a level close to the German one.

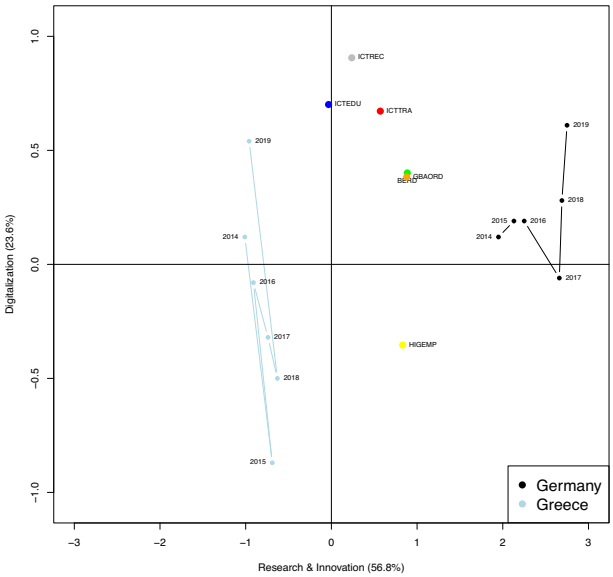


Fig. 1: Variables on cartesian plane and trajectories for Germany and Greece from 2014 to 2019

The representation of the trajectories on the cartesian plane allows to classify the countries on the basis of the quadrant occupied in the graph. For example, Germany with high level of Research & Innovation and Digitalization could be considered as a country with a high level of competitiveness. Moreover, Germany presents a quite stationary dynamic in the study because of the starting values of the considered indicators. On the other hand, Greece could be named as a country with low competitiveness because the trajectory is placed on the third quadrant for 4 years. Since the point for 2019 is positioned in the second quadrant, the competitiveness of Greece is growing thanks to indicators in digitalization. Figure 1 evidences an acceleration of innovation dynamic both for Germany and Greece in the last two years because of raising values for ICTEDU.

4 Conclusions

Reducing regional inequalities within the EU is a priority for the EU policy makers. The task of monitoring the convergence process among regional economies needs a comprehensive analysis of the dynamics in various areas related to inequality, such as per capita income, healthcare, education and job opportunities. In this scenario, disparities in competitiveness are among the main sources of inequality among the

EU member countries. Two pillars of competitiveness are innovation and the spread of digital skills, which are key drivers to create new job opportunities and to increase employability. This paper examines the convergence process in innovation and digitalization for a set of EU member countries by using time trajectories describing the paths of such countries over the 2014-2019 period.

There is evidence that the richest countries (e.g. Germany) perform better in innovation and digitalization than the poorest ones, even though some of them (e.g. Greece) show a path of improvement which may suggest the tendency to catch up with the most competitive countries. Future works could concern about the use of other dimension-reduction techniques as Dynamic Factorial Analysis to measure differences in competitiveness for European countries.

References

1. Bilbao-Osorio, B. and Rodríguez-Pose, A. (2004), From R&D to Innovation and Economic Growth in the EU. Growth and Change, 35: 434-455. https://doi.org/10.1111/j.1468-2257.2004.00256.x
2. Coppi, R., D'Urso, P. (2002). Fuzzy K-means clustering models for triangular fuzzy time trajectories. Statistical Methods & Applications, 11, p. 21-40.
3. Coppi, R., D'Urso, P. (2006). Fuzzy unsupervised classification of multivariate time trajectories with the Shannon entropy regularization. Computational Statistics & Data Analysis, 50, p. 1452-1477.
4. D'Urso, P. (2000). Dissimilarity measures for time trajectories. Journal of the Italian Statistical Society, p. 53-83.
5. D'Urso, P., De Giovanni, L., Disegna, M., Massari, R., (2019). Fuzzy clustering with spatial–temporal information. Spatial Statistics, 30, p. 71-102.
6. Edquist C. and McKelvey M. (ed.) (2000) Systems of innovation: growth, competitiveness and employment, 2 v., Cheltenham, Edward Elgar.
7. Escofier., B., Pagès J. (1994). Multiple factor analysis (afmult package). Computational Statistics and Data Analysis, 18, p. 121-140.
8. Kiers, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. Journal of Chemometrics, 14, p. 105-122.
9. Lacangellera, M., Liberati, C., Mariani, P. (2011). Banking services evaluation: A dynamic analysis. Journal of Applied Quantitative Methods, 6, p. 3-13.
10. Liberati, C., Mariani, P. (2012). Banking customer satisfaction evaluation: a three-way factor perspective. Advances in Data Analysis and Classification, 6, p. 323-336.
11. Porter M. E. (1990) The competitive advantage of nations, New York, The Free Press.
12. Schwab, K. (ed.) (2012), The Global Competitiveness Report 2012-2013, Geneva, Switzerland: World Economic Forum.
13. Solleiro J.L. and Castanon R. (2005) Competitiveness and innovation systems: the challenges for Mexico's insertion in the global context, Technovation, 25 (9), 1059-1070

National innovation system and economic performance in EU. An analysis using composite indicators

Sistema di innovazione e performance economica a livello nazionale nell'UE. Un'analisi attraverso l'uso di indicatori economici

Alessandro Zeli

Abstract In this study we propose a composite indicator to evaluate the National Innovation System for several EU countries, using similar approach proposed in literature to define composite indicators for the measure of the well-being. The analysis is carried out to compare countries across 2012-2018 years.

Abstract *In questo studio proponiamo un indicatore composito per valutare il Sistema Nazionale di Innovazione per diversi paesi dell'UE, utilizzando un approccio simile proposto in letteratura per definire indicatori compositi per la misura del benessere. L'analisi viene effettuata per confrontare i paesi dell'UE negli anni 2012-2018.*

Key words: composite indicators, National Innovation System, economic growth

1 Introduction

The relationship between scientific and technological activities and growth performance is well attested in literature and it is well-known the role played by technical change as key-driver of economic growth. The economic and statistical research attempted from several decades to provide a comprehensive reappraisal of the complex interactions between scientific and technological activities and economic growth. Countries and single economic systems can follow a variety of paths to ensure technological upgrading and catch-up. The complex mixture of the social environmental conditions and their evolution, makes very difficult to synthetize peculiarities and heterogeneities in a simple formula. However, a need of a composite indicator that includes a quantity of indexes was raised, in order to describe as broadly as possible, the social and economic conditions on which the scientific development and consequently the economic growth of a country are based (Nuvolari and Vasta, (2012)). The innovation capability of a country is related

[1] Alessandro Zeli, Istat; email: zeli@istat.it

to a set of factors linked in both market and non-market interactions which involved individuals, business firms, academic and public institutions, as well as governments. The interactions of these factors (social and economic) obtain as a result an innovation process which is characterized at national level. These considerations make the National Innovation System (NIS) notion to arise in the scientific analysis and research. Since the early 1990s, the concept of NIS had a considerable success as guideline in the processes of analysis and implementation of economic growth policies. This success is attested at national and international level (see for instance OECD and the European Commission approaches). The NIS approach is constituted by a large set of national factors that are relevant to define the national innovation environment and constitutes itself a benchmark to compare national innovation environments and explain the national differences in research innovation field and science development. The main components of NIS can be identified in private and business research for application of science, universities and public organizations, devoted to scientific and technological research, government funding of research and innovation by means of grants subsidies, R&D tax credits, interactions between private and public operators engaged in the improvement of scientific and technological capabilities (Wirkierman, Ciarli and Savona, (2018)).

2 Methodology

Several authors (Mazziotta and Pareto, (2016); Ciommi et al., (2017)) proposed a class of composite indicators for measuring the well-being and applied them to measure the BES at the regional level, in particular these composite indicators take into account the variability between and within the units. Starting from these works we replicate and apply the methodology of composite indicators to evaluate the NIS for a set of EU countries and to compare them for period covering the years from 2012 to 2018. We applied the same methodology to identify the country economic performance. Following the proposal of Mazziotta and Pareto (2016) and Ciommi et al. (2017), we use a re-scaling approach according to two 'goalposts', such as that a reference value (e.g., the indicator average) is the central value of the range. In detail, we define a matrix X_I for the NIS domain where element x_{ij} represent the value of the j-*th* elementary indicator for the i-*th* country, with $j = 1,\ldots,n$ and $i = 1,\ldots,N$. We denote by $Max_{.j}$ and $Min_{.j}$ respectively the maximum and the minimum value of the indicator j across all the countries, whereas $Rif_{.j}$ represents a reference value, in other words the average value for any indicator. As Mazziotta and Pareto (2016) work, we can introduce two 'goalposts' as follows: $Min_{I.j} = Rif_{.j} - \Delta_{.j}$ and $Max_{I.j} = Rif_{.j} + \Delta_{.j}$. Where $\Delta_{.jh} = (Max_{.j} - Min_{.j})/2$. Therefore, I_{ij} the normalized and rescaled indicator j for the i-*th* country and for NIS, denoting by variable r_{ij1}, it can be calculated as follows: $r_{ij1} = \frac{I_{ij} - Min_{I.j}}{Max_{I.j} - Min_{I.j}} 60 + 70$. (1)

Having all indicators, we considered, a positive polarity (i.e., an increase in the indicator's value corresponds to an increase in NIS domain) formula (1) can be used

for normalization, we rescale indicators into an interval of length 60 having fixed the mean (goalpost) equal to 100 (Mazziotta and Pareto, (2016)). So r_{ij1} will range, approximately, within the interval [70,130]. At same way we calculated the composite indicator (r_{ij2}) for economic variables set X_2. Literature cited above proposed several weighting and aggregation techniques, and we follow some of these approaches taking in account the distribution of the indicators among and within the territorial units (Ciommi et al., (2017)).
In particular, we use the following aggregation and weighting approaches:

Type of approach	**Formula**		
Equal weight	$EW_i = \sum_{j=1}^{n} (r_{ij} \cdot \frac{1}{n})$		
Horizontal variability approach	$AMPI_i = EW_i \pm S_i . cv_i$	$S_i = \sqrt{\frac{\sum_{j=1}^{n} (r_{ij} - EW_i)^2}{n}}$	$cv_i = \frac{S_i}{EW_i}$
Vertical variability approach	$GW_i = \frac{1}{G} \sum_{j=1}^{n} (r_{ij} \cdot G_{.j})$	$G = \sum_{j=1}^{n} G_{.j}$	
Mixed approach	$GAMPI_i = GW_i \pm S_i . cv_i$		

The equal weight approach applies a simple arithmetic average, it implies an implicit equal weight for all variables considered in the indicator. If researchers believe that not all variables could have the same weight, then two possibilities can be considered, on one hand it is possible to weight EW with a "penality" given by the variability of the indexes among the same territorial unit (horizontal variability). On the other hand, it is possible to consider a vertical variability that is an index of concentration of the single index among territorial units. Finally, we can calculate a mixed composite indicator in which both a horizontal variability and a vertical variability approaches are taken in account. As for the OECD index for compare national innovation system, they are already normalized with respect to the OECD median of each index (equal to 100), so we directly apply the aggregation method described above (EW and AMPI).

3 Results

We calculated the composite index as described above and we observed how much the indexes coming from the different approaches we used, are correlated. In Table 1, Pearson and Spearman correlations between NIS and Economic composite indicators are presented.

Table 1: *Pearson and Spearman correlations between NIS and Economic composite indicators*

NIS					**Economic**				
Spearman	**Year Average**				**Spearman**	**Year Average**			
	EW	AMPI	GW	GAMPI		EW	AMPI	GW	GAMPI

EW	1			
AMPI	0.98	1		
GW	0.81	0.84	1	
GAMPI	0.55	0.64	0.70	1

EW	1			
AMPI	0.97	1		
GW	0.97	0.91	1	
GAMPI	0.28	0.34	0.19	1

Pearson	**Year Average**			
	EW	AMPI	GW	GAMPI
EW	1			
AMPI	1.00	1		
GW	0.77	0.75	1	
GAMPI	0.60	0.64	0.72	1

Pearson	**Year Average**			
	EW	AMPI	GW	GAMPI
EW	1			
AMPI	0.99	1		
GW	0.96	0.92	1	
GAMPI	-0.22	-0.06	-0.30	1

As regards the NIS composite indicators the value we found are aligned with the outcomings attested in literature (Ciommi et al, 2017). The correlations are particularly high for the indicators EW, AMPI and GW while GAPMI presents a trend slightly different from the others. The Economic composite indicators present the same patterns of NIS composite indicators but the GAMPI behavior appears even more distant with respect to the others, presenting negative values for Spearman ranks correlations. The outcomes confirm, in general, that the composite indicators are well coordinated and represent a good synthesis of the basic index taken into account in the procedure. Now, it is important to understand if the composite indicators we found are a well representative synthesis of the phenomena we wanted to study. In particular, we had to compare the composite indicators we found with well consolidated index in the two areas we analyzed: the national innovation system and the economic performance of a country. We compared first the NIS composite indicators (EW and AMPI) with the OCSE indexes of performance of science and national innovation systems published on the OECD site (OECD 2021). Unfortunately, the OECD data are available only for 2010 and we calculated the Pearson correlation coefficients for that year between NIS composite indicators and OECD indexes. In Table 2 we present the outcomes of our elaborations and we can note that in general there is a good coincidence for the NIS composite indictors and the OECD indexes overall. AMPI shows slightly better correlations with respect to EW. As we observed above, there is in general a good approximation of the NIS composite indicators that can be considered a very good representation of the *Science base*, *Business R&D and innovation* and *Human resources* areas, while a quite good representation of *Entrepreneurship* and *Internet use for innovation* while *Knowledge flows and commercialisation* is underrepresented. As regards the Economic performance area, we compared the Economic composite indicators with the GDP index for each country. We calculated the Pearson and Spearman correlation between GDP and Economic composite indicators (Table 3). The correlations between Economic composite indicators and GDP index present, in general, a good accordance even if lesser with respect to the

NIS indicators reported in Table 2. We have to underline, again, the different behavior of GAMPI that presents a negative correlation with GDP index.

Table 2: *Pearson correlation between composite indicators NIS and composite indicators OCSE by innovation areas and total – Year 2010*

Innovation Area	EW	AMPI
Science base	0.717	0.721
Business R&D and innovation	0.790	0.801
Entrepreneurship	0.573	0.691
Internet use for innovation	0.623	0.665
Knowledge flows and commercialisation	0.354	0.348
Human resources	0.886	0.884
Total average	***0.783***	***0.806***

Note: EW_OCSE and AMPI_OCSE are based on OCSE indexes of performance of science and national innovation systems.

Table 3: *Pearson and Spearman correlations between Economic composite indicators and GDP index – Years 2012-2018*

	Pearson	Spearman
EW ec	0.67053	0.41209
AMPI	0.60582	0.41209
GW	0.66291	0.43407
GAMPI	-0.51554	-0.12637

Finally, we wanted to explore the correlation between the NIS and Economic composite indicators (i.e. if there are signals, at country level, of the influence of a strong NIS on a better economic performance). Table 4 shows the correlations between the composite indicators by country.

Table 4: *Pearson correlation coefficients between NIS composite indexes and economic composite indexes by countries – Years 2012-2018*

	EW	AMPI	GW	GAMPI	Δ GDP
Austria	-0.81	-0.79	-0.26	0.56	8.96
Belgium	-0.88	-0.89	-0.58	0.70	9.08
Finland	0.09	0.27	0.28	-0.22	6.70
France	-0.79	-0.35	-0.70	0.59	8.07
Germany	-0.30	-0.30	-0.19	0.22	10.67
Ireland	0.44	0.47	0.47	0.55	66.27
Italy	0.22	0.07	-0.32	-0.80	2.83
Netherlands	-0.54	-0.31	0.02	-0.66	11.18
Poland	0.63	0.67	-0.38	0.27	24.13
Portugal	0.80	0.79	-0.28	-0.58	10.40
Spain	0.28	0.54	-0.33	-0.71	12.76
Sweden	0.57	0.50	0.62	-0.40	15.85
United Kingdom	0.32	0.06	-0.56	-0.64	12.75

An EW and AMPI correlations outcomes analysis indicates that there is not a uniform relationship between NIS and economic performance across the EU countries. In many countries, as expected, it is confirmed a positive and quite strong correlation between NIS and economic performance. In many countries there are a counterintuitive result, for these countries there is a negative correlation between NIS composite indicators and economic performance. This is even more weird because the same countries are among those which present a stronger NIS (Figure1). On the other hand, these countries present a very weak rate of variation of the GDP index in the period 2012-2018. Hence, having these countries achieved a very high

NIS development, with no more space for further enhancement, even little GDP fluctuations may not be influenced by NIS improvements but, on the contrary, may cause the presence of negative correlations. In other words, while for less developed NIS there is a greater marginal effect of their innovative improvements on economic performance, the economic performance for more developed NIS is quite inelastic to NIS increments.

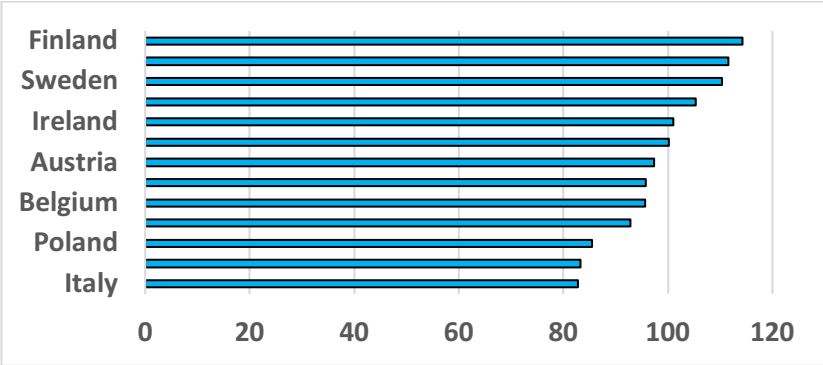


Figure 1: *AMPI_NIS composite index by country – Year 2018*

4 Concluding remarks

Preliminary results we achieved in this study can be synthetized as follows: there is a good positive concordance of composite indicators, however, it is necessary to perform a deeper analytic exploration of the meaning of GAMPI and its connections with the basic indexes. Our analyses confirm, through the comparison of our composite indicators with international benchmarks, a good representation of relevant phenomena from our composite indicators and this comfort us for the basic indexes' choice. We carried out a first exploration of the influence of NIS on economic performance, but the results are not unique, and consequently it will be necessary to carry out further analyses to highlight this relationship.

References

1. Ciommi, M., Gigliarano, C., Emili, A., Taralli, S., Chelli, F. M.: A new class of composite indicators for measuring well-being at the local level: An application to the Equitable and Sustainable Well-being (BES) of the Italian Provinces. *Ecological Indicators*, 76, 281–296 (2017)
2. Mazziotta, M., Pareto, A.: On a generalized non-compensatory composite index for measuring socio-economic phenomena. *Social Indicator Research*, 127 (3), 983–1003 (2016)
3. Nuvolari, A., Vasta, M.: The Ghost in the Attic? The Italian National Innovation System in Historical Perspective, 1861-2011. *Quaderni del Dipartimento di Economia Politica e Statistica*, n. 665 (2012)
4. OECD: *Comparative performance of national science and innovation systems*. Oecd. stat. https://stats.oecd.org/Index.aspx?DataSetCode=BENCHMARK_STIO. Last access:26 March 2021.
5. Wirkierman, A.L. Ciarli, T., Savona, M.: Varieties of European National Innovation Systems. *WP 13 ISIgrowth* (2018)/s001090000086

2.6 Challenges for observational studies in modern biomedicine

Data integration: a Statistical view

Integrazione di dati: uno sguardo statistico

Pier Luigi Conti

Abstract Data integration is a *corpus* of techniques aiming at gathering and combining together data from different sources. A short review of the main statistical problems arising from the use of data obtained from integration procedure is given.
Abstract *Il termine "integrazione di dati" indica un insieme di tecniche utilizzate per combinare dati provenienti da fonti diverse. Nella presente comunicazione viene data una breve discussione di alcuni problemi legati all'analisi statistica di dati ottenuti con tecniche di integrazione.*

Key words: Data integration, Statistical matching, Record linkage

1 Introduction

In the last twenty years, we have seen a rapid growth of different types of data sources (Official Statistics survey data, administrative data, experimental data, observational data, data from sensor, transactions data, etc.) used in different contexts (scientific, economics, commerce, official statistics, etc.). Nowadays, data are collected, at different levels, by different organizations and for different purposes. In many cases, they are in the public domain, or easily available on request. This is a part of the Big Data Revolution, that has opened new opportunities for researchers to access data potentially useful to investigate relationships among variables.

A major consequence is a significant diversification of primary data sources. This has determined a change of the traditional paradigm of a unique main source of data coming from an *ad hoc* statistical data collection process. Within this paradigm, two key points are represented by the design of a good data acquisition process (well designed experimental study, well designed sampling plan, etc.), and by good

Pier Luigi Conti
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy, e-mail: pierluigi.conti@uniroma1.it

statistical methods to analyze collected data. Although very important even now, this traditional paradigm has its limitation. In the first place, cost: *ad hoc* statistical studies are expensive, so that is it natural to resort, when possible, to inexpensive data sources. In the second place, time for data collection is dramatically reduced by taking data from already existing databases. In the third place, in several important sample surveys increasing rates of nonresponse and refusal to participate are a common problem.

Combining data from different sources, *i.e. data integration*, is becoming more and more important either to construct new databases broadening the domain of application of single databases combined together, or to make statistical inference.

The paradigm based on data integration is very promising, because of its potential advantages for scientific investigations. As a consequence, there is now an increasing interest for data to be collected *via* integration.

A frequent feature of data coming from different sources is that common variables could be observed on sample units selected with different, possibly unknown, sampling designs, and by using different measurement methodologies, that could imply, in their turn, different levels of precision.

Roughly speaking, in case of data collected from a single source, methods for assessing uncertainty due to sampling variability are rather well known, as well as methods for dealing with additional sources of errors and uncertainty, such as nonresponses, measurement errors, frame imperfections, etc..

The framework is more intricate in case of data from multiple sources combined together, because new potential sources of errors arise. They are essentially related to a basic question: "May we consider combined data as observations from the population they are supposed to represent?" As remarked in [13], this is a problem of inference under *entity ambiguity*. More concretely, with integrated data we have sample observations that do not necessarily correspond to actual population units. Such an "entity ambiguity" essentially corresponds to a source of *additional uncertainty* characterizing datasets obtained by integrating data from multiple sources.

In response to this new paradigm, the development of a framework for "analysis of integrated data" is necessary. As above remarked, in addition to the sampling and nonsampling "traditional" errors, we have to consider specific errors for integrated data, depending on the integration process. As a matter of fact, it is generally wrong to deal with integrated data as they were "fresh", single-source data, and to use standard methods for statistical analysis. The development and the use of inferential methods that take into account how data are combined is therefore necessary.

In general, the problems to be faced with data integration are essentially two.

- Development of methodologies for data integration allowing for a safe assessment of different sources of errors.
- Use of data already obtained by some integration process, as in *secondary analysis*. In this case, focus is in modeling errors and analyzing their impact on statistical inference.

In the sequel, we will mainly concentrate on two important problems of data integration, namely *statistical matching* and *record linkage* [6].

2 Statistical Matching

Consider a population composed by N units, on which three variables $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ are defined. Let A, B two independent samples of size n_A, n_B, respectively, drawn from the population according to possibly different sampling designs. In sample A, only variates $\mathbf{X}$ and $\mathbf{Y}$ are observed, and in sample B only variates $\mathbf{X}$ and $\mathbf{Z}$ are observed. Hence $\mathbf{X}$ is the set of variates *common* to the two samples, whilst $\mathbf{Y}$, $\mathbf{Z}$ are the variates *specific* of sample A, B, respectively.

Since the probability of selecting the same unit in the two samples is usually negligible, we may assume that A, B have no common units. No joint observation of $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ is available, since $\mathbf{Z}$ is missing in sample A, and $\mathbf{Y}$ is missing in sample B.

The goal of *statistical matching* is twofold. (i) At a *macro* level, statistical matching aims at estimating the joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, or at least parameters related to such a distribution. (ii) At a *micro* level, statistical matching aims at constructing, for the $n_A + n_B$ sample units, a unique, synthetic database containing $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ values.

The macro and micro approaches to statistical matching are formally different but equivalent [14]. On one hand, once an estimate of the joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is obtained via observed data, it can be used to impute missing $\mathbf{z}$-values in sample A and missing $\mathbf{y}$-values in sample B. On the other hand, once sample A is completed with $\mathbf{z}$-values and sample B is completed with $\mathbf{y}$-values, the resulting $n_A + n_B$ triples of $(\mathbf{x}, \mathbf{y}, \mathbf{z})$-values can be used to estimate the joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

Unless special assumptions are made, the statistical model for the joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is *not identifiable*; cfr. [2], [3]. The most important assumption ensuring identifiability is the *Conditional Independence Assumption* (CIA): $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}$, where $\perp\!\!\!\perp$ denotes independence. In this case, apart from considerations on the sampling designs according to which A and B are selected, the statistical matching problem reduces to statistical inference for missing data [14].

Although of interest, CIA has two main drawbacks. First of all, it is not met in several important applications. In the second place, it cannot be tested on the basis of sample data. In empirical research, a common problem is to decide which assumptions should be made. Strong assumptions, such as CIA, allow inferences that may be very powerful but not credible enough. Statistical theory, although essentially unable to resolve this problem, is useful to clarify its nature.

Statistical matching without CIA is of considerable interest. Matching methodologies and related sources of errors are studied in [2] in the case of unidimensional $\mathbf{Y}$, $\mathbf{Z}$. The consideration of discrete multivariate $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ is studied in [5]. In this field, Bayesian Networks (BNs) provide a fundamental tool to model the dependence structure of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, as well as to study error components due to matching.

In the sequel, a short account of the sources of error for statistical matching is given. The starting point is that when CIA does not longer hold the statistical model for $(\mathbf{Y}, \mathbf{Z})|\mathbf{X}$ is not identifiable. Instead of a single distribution for $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$, the data collection process is only able to identify a *class* $\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ of distributions, namely those that are compatible with the marginal distributions of $(\mathbf{Y}, \mathbf{X})$ and $(\mathbf{Z}, \mathbf{X})$ (that are identifiable, of course). Even if the marginal distributions of $(\mathbf{Y}, \mathbf{X})$

and $(\mathbf{Z}, \mathbf{X})$ were known, on the basis of the data collection process we could only say that the joint distribution of $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ lies in $\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$.

Since the marginal distributions of $\mathbf{Y}|\mathbf{X}$ and $\mathbf{Z}|\mathbf{X}$ are identifiable, they can be consistently estimated on the basis of sample data. Let $\widehat{\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}}$ be defined as $\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$, but with marginal distributions estimated on the basis of sample data. Each probability distribution in $\widehat{\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}}$ is a *matching distribution* for $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$. The *statistical matching problem* consists in choosing a matching distribution in $\widehat{\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}}$. A matching procedure is a rule to choose a distribution in the class $\widehat{\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}}$.

Let $H^*_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ be the "true" joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, and let $\widehat{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ be estimated on the basis of sample data. The *total matching error* is essentially a divergence measure between of $\widehat{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ from $H^*_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$, $d(\widehat{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}, H^*_{\mathbf{X},\mathbf{Y},\mathbf{Z}})$, say. This quantity plays a role similar to MSE in classical estimation theory, but with an important difference: due to unidentifiability, $H^*_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ cannot be consistently estimated. In many cases of interest, $\widehat{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ is actually a consistent estimator of some $\widetilde{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$, and the total matching error can be decomposed as

$$d(\widehat{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}, H^*_{\mathbf{X},\mathbf{Y},\mathbf{Z}}) = d(\widehat{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}, \widetilde{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}) + d(\widetilde{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}, H^*_{\mathbf{X},\mathbf{Y},\mathbf{Z}}). \tag{1}$$

As the sizes of samples A, B increase, $d(\widehat{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}, \widetilde{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}})$ tends to zero, whilst $d(\widetilde{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}, H^*_{\mathbf{X},\mathbf{Y},\mathbf{Z}})$ does not, as a consequence of unidentifiability. Of course, it cannot be consistently estimated on the basis of sample data. A fruitful idea consists in studying estimable upper bounds. This approach is pursued in [2], [5].

The matching error (1) can be interpreted as a measure of the quality of statistical matching: the smaller $d(\widehat{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}, H^*_{\mathbf{X},\mathbf{Y},\mathbf{Z}})$, the better the results of matching. A crucial role is played by the class $\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$. Intuitively speaking, the smaller $\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$, the smaller the matching error due to lack of identification. Hence, a special attention must be paid in constructing the class $\widehat{\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}}$. As a general principle, extra-sample information should be used to establish reasonable constraints on the dependence structure of $(\mathbf{Y}, \mathbf{Z})|\mathbf{X}$. Different approaches to make it concrete this general principle, and related applications, are given in [2], [4] and, in a multivariate BNs perspective, in [5]. In a slightly different but related perspective, in Statistical Matching the distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is only *partially specified*, and the class $\mathscr{H}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ plays a role similar to that of the *identification region* in [12].

3 Record Linkage

In record linkage [10], the aim is to identify and link together the records (with associated observations) in different files corresponding to the same statistical unit. In each data-file, records are usually identified via *key variables*, containing common identifying information. The "lucky case" of a unique identifier that allows for exact matching hardly ever occurs in practice.

In several cases of practical relevance, the key variables identifying records could not coincide, for different reasons. (i) Records could have no unique, known and accurate ID, or key variables could contain errors, or some of them could be missing. (ii) There could be differences in data captured and maintained by different databases. For instance, age vs. date of birth. (iii) Data could sometimes regularly change over time (database dynamics).

Two distinct approaches to record linkage are considered in the literature. On one hand, *deterministic* record linkage methods involve exact one-to-one character matching of the key variables. On the other hand, *probabilistic* record linkage methods involve the calculation of linkage weights estimated on the basis of likelihood ratios given all the observed agreements/disagreements of the values of the key variables. Probabilistic linkage methods can lead to better results than simple deterministic linkage.

Linkage errors are generally unavoidable. There are basically two kinds of errors.

1. *False match*: two (or more) records are matched, i.e. are referred to the same entity (unit), although they are not.
2. *False unmatch*: two (or more) records are not matched, i.e. are referred to different entities (units), although they are.

As it appears from the pioneering paper [7] by Fellegi and Sunter, the problem of linking two records is essentially a *decision problem*. One has to decide, on the basis of an appropriate statistical procedure, whether two records are: (i) *Match* (same unit); (ii) *Unmatch* (different units); (iii) *Uncertain match* (unable to decide - possible match). Uncertain matches are reviewed in the post-linkage phase, where manual / clerical review of unlinked records is performed.

The Fellegi-Sunter procedure is essentially based on an extension of the Neyman-Pearson Lemma, and consists in minimizing the uncertain match probability for fixed for fixed probabilities of false match and false unmatch.

To be used in practice, the Fellegi-Sunter procedure requires a preliminary estimation of the probability distributions of results comparisons of key variables under match and unmatch. This is a difficult problem with missing data, because for each pair of records only values of the key variables are observed, but we do not know whether they correspond to match or unmatch. An approach based on the method of moments is proposed in [7]; Maximum Likelihood estimators for incomplete data, based on the EM algorithm, are developed in [11].

The presence of (possible) errors in record linkage has a serious effect on statistical analysis of linked data. The impact of linkage error, in general, depends on the structure of the data, the distribution of linkage errors, and also on the analysis to be carried out. In some cases (for instance, fraud detection) it is of high importance to capture all matches. In other studies, for instance in the analysis of linked health care records [9], it is important that linked records are true matches, whilst false unmatches are less important. Hence, the impact of linkage error is purpose-dependent, and ideally linkage criteria should be purpose-dependent, as well. A study of bias due to linkage errors from the viewpoint of bio-medical studies is in Ch. 4 of [8].

In a more theoretical perspective, the presence of possible linkage errors is of particular relevance in studying the dependence among variables. A widely studied case is that of regression analysis, where the relationship between a response variable and a set of explanatory variables is explored by fitting a regression model to linked data. The implicit assumption is that the linked data-set is made of correctly matched records. In the presence of (possible) linkage errors, some of the records in the linked data-set will correspond to distinct units erroneously linked together. The typical effect is that the relationships between the study variable and the explanatory variables are weakened. Two options are available in this case. On one hand, attempts have bee made to reduce the probability of incorrect linkage, starting from the fundamental paper [7]. On the other hand, attempts have been made to model the bias due to incorrect linkage and then develop methods for correcting it. This approach is particularly interesting in case of secondary analyses, *i.e.* when statistical analysis is performed on already linked data-sets, with no real control on the linkage process; cfr. Ch. 5 in [8] and references therein.

References

1. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE Transactions on Knowledge and Data Engineering, **24**, 1537-1555 (2012).
2. Conti, P. L., Marella, D., Scanu, M.: Statistical Matching Analysis for Complex Survey Data With Applications. Journal of the American Statistical Association, **111**, 1715-1725 (2016)
3. Conti, P. L., Marella, D., Scanu, M.: How far from identifiability? A systematic overview of the statistical matching problem in a nonparametric framework. Communication in Statistics - Theory and Methods, **46**, 967-994 (2017)
4. Conti, P. L., Marella, D., Neri, A.: Statistical matching and uncertainty analysis in combining household income and expenditure data. Statistical Methods & Applications, **26**, 485-505 (2017)
5. Conti, P. L., Marella, D., Vicard, P., Vitale, M.: Multivariate statistical matching using graphical modeling. International Journal of Approximate Reasoning, **130**, 150-169 (2021)
6. Dong, X. L., Srivastava, D.: Big Data Integration. Morgan & Claypool Publishers, Williston (VT) (2015)
7. Fellegi, I. P., Sunter, A. B.: A theory for record linkage. Journal of the American Statistical Association, **64**, 1183-1210(1969)
8. Harron, K., Goldstein, H., Dibben, C. (Eds.): Methodological Developments in Data Linkage. Wiley, Chichester (2016)
9. German, R. R.: Sensitivity and predictive value positive measurements for public health surveillance systems. Epidemiology, **11**, 720-727 (2000)
10. Herzog, T. N., Scheuren, F. J., Winkler, W. E.: Data Quality and Record Linkage Techniques. Springer Verlag, New York (2007)
11. Jaro, M. A.: Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, **84**, 414-420 (1989)
12. Manski, C. F.: Partial identification of probability distributions. Springer Verlag, New York (2003)
13. Zhang, L.-C., Chambers, R. L. (Eds.): Analysis of Integrated Data. CRC Press, Boca Raton (FL) (2019)
14. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching - Theory and Practice, Wiley, Chichester (2006)

Exploring patients' profile from COVID-19 case series data: beyond standard statistical approaches

Studio dei profili dei pazienti con COVID-19: oltre i modelli statistici standard

Chiara Brombin, Federica Cugnata, Pietro E. Cippà, Alessandro Ceschi, Paolo Ferrari and Clelia di Serio

Abstract The statistical analysis of COVID-19 data is challenging. Data have been collected in emergency conditions, without a predefined study design, thus resulting highly heterogeneous and potentially affected by multiple sources of bias. Here we analyse a comprehensive high-quality COVID-19 dataset using different data-driven statistical techniques to disentangle effects of collected variables and gain a better insight into the role of comorbidities and medications in affecting final outcomes.
Abstract *L'analisi statistica dei dati sul COVID-19 è complessa. I dati sono stati raccolti in condizioni di emergenza, senza un disegno di studio predefinito, risultando così molto eterogenei e potenzialmente inficiati da molteplici fonti di bias. In questo lavoro analizziamo un set di dati sul COVID-19, completo e di alta qualità,*

Chiara Brombin and Federica Cugnata
University Centre for Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University e-mail: brombin.chiara@hsr.it, cugnata.federica@hsr.it

Pietro E. Cippà
Department of Medicine, Division of Nephrology, Ente Ospedaliero Cantonale, 6500 Bellinzona, Switzerland and Faculty of Medicine, University of Zurich, 8006 Zurich, Switzerland e-mail: Pietro.Cippa@eoc.ch

Alessandro Ceschi
Faculty of Medicine, University of Zurich, 8006 Zurich, Switzerland, Biomedical Faculty, Universit della Svizzera Italiana, 6900 Lugano, Switzerland, Institute of Pharmacology and Toxicology, Ente Ospedaliero Cantonale, 6500 Bellinzona, Switzerland and Department of Clinical Pharmacology and Toxicology, University Hospital Zurich, 8091 Zurich, Switzerland e-mail: Alessandro.Ceschi@eoc.ch

Paolo Ferrari
Department of Medicine, Division of Nephrology, Ente Ospedaliero Cantonale, 6500 Bellinzona, Switzerland and Biomedical Faculty, Universit della Svizzera Italiana, 6900 Lugano, Switzerland e-mail: Paolo.Ferrari@eoc.ch

Clelia Di Serio
University Centre of Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, and Biomedical Faculty, Universit della Svizzera Italiana, 6900 Lugano, Switzerland e-mail: diserio.clelia@hsr.it

utilizzando diverse tecniche statistiche, tutte in ottica data-driven, per districare gli effetti delle variabili raccolte e comprendere meglio il ruolo delle comorbidità e dei farmaci sull'esito finale della malattia.

Key words: Data-driven approaches, Classification and Regression Trees, Bayesian Networks

1 Introduction

Difficulties arising in the statistical analysis of COVID-19 data have been emphasized in the recent literature [2, 9] and have been mainly linked to the collection of the data in an emergency situation, without a planned study design. Standard statistical approaches may fail in capturing complex interrelationship among variables and collinearity issues emerge and are difficult to control and manage. Data-mining and data-driven approaches turn to be effective and powerful tools to identify risk factors associated with the outcome of interest and to unravel complex dependence structure among data. In particular we will focus on Classification and Regression Trees (CART) analysis and Bayesian Network (BN) approach to analyze a COVID-19 case-series dataset controlled for high quality standards, aimed at exploring relationships among demographic characteristics, comorbidities, treatment administration, and the final outcome.

2 Sample description

A sample of 399 hospitalized patients, admitted by Ente Ospedaliero Cantonale hospital between March 1-May 1 2020, diagnosed with COVID-19 and with complete data records has been considered for the analysis. The study was approved by the Ethical Committee of the Canton of Ticino, Switzerland. Demographic and anthropometric characteristics along with the presence of comorbid conditions and signs/symptoms of COVID-19 were recorded at admission time. Clinical and laboratory parameters have been regularly monitored every 48h during hospitalization along with the drug prescription (which was determined on clinical basis and not affected by participation in the study). Data have been electronically gathered and stored. The median age was 73 years (IQR [60.50, 81.00]) ranging from 22 to 96 years; 250 (62.7%) were male. 319 patients (80%) were discharged and 80 patients (20%) died. 69 patients (17.3%) in total were admitted to the intensive care unit (ICU).

3 Statistical methods

Multiple approaches have been implemented to identify from one side prognostic factors (among demographic/clinical characteristic and therapeutic/pharmacological treatments) that best discriminate among patients with different outcomes (death or hospital discharge) and from the other to explore data dependence structure thus uncovering complex interrelationship among collected variables. The underlying idea is to integrate results obtained from different statistical techniques to disentangle effects of demographic/clinical characteristic, the impact of comorbidities and medications for a better undestanding and management of COVID-19. Along with a standard logistic regression approach, CART analysis has been performed and BNs have been estimated. The key advantage of CART approach lies in its nonparametric nature and in its flexibility in handling mixed types of variables (continuous and/or categorical) requiring less stringent assumptions than standard modelling procedures and being based on a data-driven logic. Actually CART implements a binary recursive partitioning where parent nodes are always split into exactly two child nodes and the process is recursively repeated by treating each child node as a parent [1]. Variables and corresponding cut-off values that best differentiate observations with different outcome variable are automatically selected by the procedure. The underlying criterion used in CART is the GINI rule.
BNs implement a directed acyclic graph, defined by a set of nodes, representing random variables, and a set of arcs, implying direct dependencies among the variables. BNs allow for an effective representation and computation of a joint probability distribution over a set of random variables [7]. Actually, they represent powerful tools for examining data dependence structure, to uncover unknown and unexpected patterns and revealing possible confounding effects. Blacklists, i.e., directions, arcs, which are not allowed in the network, may be specified. All the analyses were performed using R statistical software (version 3.5.2, https://cran.r-project.org/index.html). The R package `rpart` was used to implement the classification tree analysis. The R packages `bnlearn` [8] and `gRain` [5] were used to learn the network and perform the inference required to calculate the conditional probabilities. The algorithm chosen for obtaining the BNs was the Hill-Climbing algorithm with AIC score functions. The significance level was set at 0.05.

4 Results

Based on the univariate logistic regression analysis, age, the presence of cancer, or diabetes or cardiovascular disease, having pneumonia, the use of nonsteroidal anti-inflammatory drugs (NSAIDs) and antibiotics were associated with an increasing risk of non-surviving from COVID-19. Moreover, it emerged that lower estimated glomerular filtration rate (GFR), the use of renin-angiotensin-aldosterone system inhibitors (RAASi) , and having fever were associated with a reduction of the risk of non-surviving from COVID-19 (see Table 1). In this context, estimating a multiple

regression model is meaningless since collinearity arises when jointly considering clinical symptoms, comorbidities and treatments. For this reason alternative data-driven approaches were applied. From the CART analysis (Figure 1), non-survivors had low GFR ($<$ 50 mL/min/1.73 m^2), did not use RAASi and statins, have no fever or had low GFR ($<$ 50 mL/min/1.73 m^2), did not use RAASi and statins, have fever and were overweighted (with a Body Mass Index, BMI, of at least 23). Moreover, patients with the worst outcome were also those with a high GFR ($\geq$ 50 mL/min/1.73 m^2), older ($\geq$ 77 years old), were not treated with RAASi, have fever and cardiovascular diseases. Although derived within a survival tree framework, in a previous work by Cippà *et al.* [2], GFR, BMI, Age and RAASi have emerged as best splitting variables, also with quite similar cut-off values. In the recent literature, the effect of fever at admission on COVID-19 mortality seems controversial and it has often attributed to disease stage [3, 4]. Controversy might be also attributed to the lack of a unique definition of fever cut-off. Hence, findings obtained from the CART analysis should be further investigated. Bayesian networks (BNs, Figure 2) analysis confirmed a direct link of GFR and RAAS blockers and NSAIDs with the final outcome while RAASi, gender, use of antibiotics and pneumonia showed a direct effect on the admission to the intensive care unit (ICU). Hypertension showed an indirect effect mediated by RAASi on both the final outcome and the admission to ICU.

5 Discussion

According to an information quality approach [6], the use of different statistical techniques and the adoption of a data-driven logic, relaxing stringent modelling assumptions, may shed light into complex relationships typical of complex diseases and phenomena. Data retrieved from second COVID-19 wave will be used to validate present results. Moreover, we will take advantage of a key feature of BN approach to assess alternative hypothetical scenarios.

Table 1 Univariate logistic regression model

	crude OR(95%CI)	raw p-value
Age	1.07 (1.05,1.1)	$<$ **0.001**
BMI	1.02(0.97,1.06)	0.499
GFR	0.96 (0.95,0.98)	$<$ **0.001**
Gender: M vs F	1.06 (0.64,1.76)	0.821
Cancer: yes vs no	3.56 (1.9,6.67)	$<$ **0.001**
Diabetes: yes vs no	1.88 (1.11,3.18)	**0.018**
Hypertension: yes vs no	1.29 (0.79,2.12)	0.306
Cardiovascular_disease: yes vs no	3.84 (2.29,6.43)	$<$ **0.001**
Chronic_lung_disease: yes vs no	1.46 (0.81,2.64)	0.207
Diarrhea: yes vs no	0.75 (0.37,1.51)	0.417
Resp_diff: yes vs no	0.98 (0.6,1.6)	0.941
Fever: yes vs no	0.55 (0.32,0.96)	**0.035**
Pneumonia: yes vs no	1.8 (1.09,2.98)	**0.021**
Cough: yes vs no	0.71 (0.43,1.17)	0.179
RAAS_blockers: Yes vs No	0.42 (0.23,0.75)	**0.003**
Other_antihypertensives: Yes vs No	1.39 (0.85,2.28)	0.185
NSAIDs: Yes vs No	3.24 (1.7,6.17)	$<$ **0.001**
Antidiabetics: Yes vs No	1.69 (0.99,2.87)	0.054
Statins: Yes vs No	0.76 (0.44,1.33)	0.343
Anticoagulants: Yes vs No	0.53 (0.25,1.14)	0.103
Antibiotics: Yes vs No	2.63 (1.55,4.45)	$<$ **0.001**
Immunosuppressants: Yes vs No	1.59 (0.76,3.35)	0.218
Antiviral_agents: Yes vs No	0.79 (0.47,1.35)	0.389

References

1. Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A.: Classification and regression trees. CRC press (1984)
2. Cippà, P. E., Cugnata, F., Ferrari, P., Brombin, C., Ruinelli, L., Beria, N., Bianchi, G., Schulz, L., Bernasconi, E., Merlani, P., Ceschi, A. and Di Serio, C.: A data-driven approach to identify risk profiles and protective drugs in COVID-19. Proceedings of the National Academy of Sciences, **118**, e2016877118 (2021)
3. Choron, R. L, Butts, C. A., Bargoud, C., Krumrei, N. J., Teichman, A. L., Schroeder, M. E., Bover Manderski, M. T., Cai, J., Song, C.,Rodricks, M. B., Lissauer, M. and Gupta, R.: Fever in the ICU: A Predictor of Mortality in Mechanically Ventilated COVID-19 Patients. Journal of intensive care medicine **36**, 484–493 (2021)
4. Gul, M. H. and Htun, Z. M. and Inayat, A.: Role of fever and ambient temperature in COVID-19, Expert Review of Respiratory Medicine **15**, 171–173 (2021)
5. Højsgaard, S.: Graphical independence networks with the gRain package for R. Journal of Statistical Software **46**, 1–26 (2012)
6. Kenett, R. S. and Shmueli, G.: On information quality. Journal of the Royal Statistical Society. Series A (Statistics in Society) **177**, 3–38 (2014)
7. Pearl, J.: Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press, New York (2009)
8. Scutari, M.: Learning bayesian networks with the bnlearn R package. Journal of Statistical Software **35**, 1–22 (2010)
9. Wolkewitz, M. and Puljak, L.: Methodological challenges of analysing covid-19 data during the pandemic. BMC Medical Research Methodology **20**, 1–4 (2020)

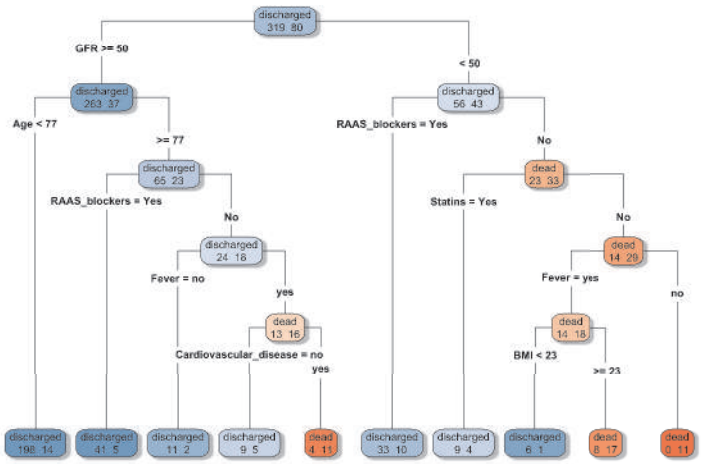


Fig. 1 Classification and Regression Trees analysis for identifying factors that best discriminate among survivors and non-survivor patients. All the covariates evaluated at the univariate level in the logistic regression were considered in the analysis.

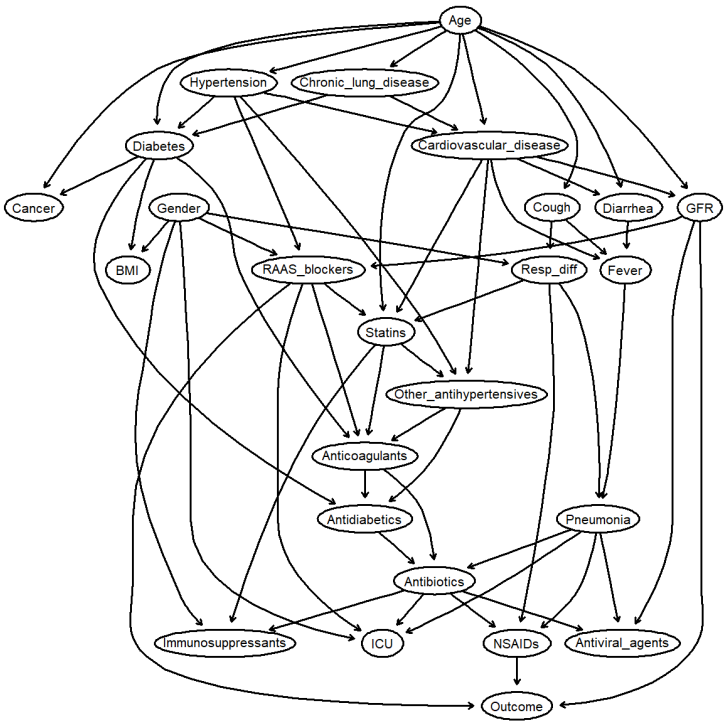


Fig. 2 Bayesian network analysis exploring the dependence structure of the data

On the statistics for some pivotal anti-COVID-19 vaccine trials

Mauro Gasparini

Abstract A small but significant part of the world-wide efforts underlying the discovery, production and distribution of anti-COVID-19 vaccines are the statistical methods informing the pivotal clinical trials leading to the approval of vaccines by Health Authorities. These methods use some fundamental concepts from Clinical Statistics and also include some epidemiological tools geared towards the analysis of semi-observational data. They will be exemplified through the comparison of the statistical sections of the two main studies about BioNTech/Pfizer and Moderna mRNA-based vaccines.

Abstract *Una parte piccola ma significativa degli sforzi che a livello mondiale soggiaciono alla scoperta, produzione e distribuzione dei vaccini anti-COVID-19 sono i metodi statistici che informano le prove cliniche pivotali per l'approvazione dei vaccini da parte delle Autorità Sanitarie. Tali metodi usano alcuni concetti fondamentali di Statistica Clinica e includono pure alcuni strumenti epidemiologici fondamentali per l'analisi di dati osservazionali. Verranno esemplificati dal confronto della statistica dei due principali studi sui vaccini di BioNTech/Pfizer e di Moderna, basati su mRNA.*

Key words: Credibility Intervals, Interim Analysis, Clopper-Pearson, BioNTech, Pfizer, Moderna

1 Introduction

Through the last pandemic year we (the human kind) have experienced grief but have also witnessed an unprecedented scientific achievement: the discovery, production and distribution of new, high-tech, life-saving vaccines in less then a year. The novel biomedical technologies used by such vaccines and the industrial effort

Mauro Gasparini
Dipartimento di Scienze Matematiche, Politecnico di Torino e-mail: mauro.gasparini@polito.it

faced to produce them in such a short time is in the headline news and this is not the right venue to discuss them once more. However, a less known part of this scientific and techological effort rests on the world-wide system of submission and approval by Health Authorities of new therapies which centers around clinical trials, and a core part of any clinical trial is its statistical methodology.

In this talk we will analyze two articles which illustrate the results of the pivotal trials of two major mRNA-based vaccines, called here for the sake of simplicity the BioNTech/Pfizer vaccine [2] and the Moderna [1] vaccine. Pivotal trials - often called Phase III trials - are conducted by the *sponsors* (BioNTtech/Pfizer and Moderna in this case) to provide evidence in favor of a new therapy to the Health Authorities, in order to obtain their market authorization. Health authorities are either national, like the US Food and Drug Administration (FDA) or supranational, like the European Medicines Agency (EMA). Needless to say, the anti-COVID-19 are so-called orphan drugs, since they have no predecessors in preventing this disease, and as such had been given a special emergency use approval track, without sacrificing any of the necessary milestones to protect people from unsafe or inefficacious drugs.

2 The three dimensions of clinical trials

The three dimensions of clinical trials are *efficacy* (we do want the therapy to fight the disease), *safety* (*primum non nocere*) and *individualization* (also called personalized medicine).

The primary efficacy variable for our vaccines was the number of confirmed symptomatic COVID-19 cases in the treatment and in the control groups (some minor differences in the definition of "confirmed symptomatic" between the two vaccines is of no importance here). Based on these counts, the vaccine treatment effect was measured in terms of incidence rates, hazard rates and finally vaccine efficacy (VE), with minor differences between the two protocols in the definition of VE which will be explored in the next section.

Safety measures center instead on *adverse events* (AEs), both solicited (i.e. predicted by the sponsors, since all vaccines cause some types of AEs) or unsolicited, in case some new unpredicted problems come up. For both efficacy and safety a variety of secondary endpoints are considered in addition to the primary efficacy endpoint.

Individualization include all efforts to tailor therapies to the specific needs of patients and include the exploration of subpopulation of patients which may differentially benefit or risk (*subgroup analysis*) as well as the study of covariates (e.g. genetic biomarkers) which may be prognostic or predictive of a differential efficacy or safety. For our vaccines, similar analyses to efficacy and safety have been conducted for some subgroups identified by BioNTech/Pfizer and Moderna based mainly on demography or comorbidity.

As usual, the not-vaccinated participants have received a sham vaccine which can not be distinguished from the vaccine neither by patients nor by medical personnel who observe the results. This is *observer-blind randomization*, an essential feature which, by avoiding biases, allows for statistical and causal modeling described in the next sections.

3 A comparison between Pfizer/BioNTech and Moderna

The main features and results of the two protocols are summarized in Table 1. The

Table 1 Main features and results of the two trials compared.

Feature	BioNTech/Pfizer	Moderna
Period	from 27 July 2020 to 14 November 2020	from 27 July 2020 to 25 November 2020
Randomized 1:1	43548	30420
Cases in active group	8	11
Cases in control group	162	185
Estimated VE (interval)	95.0 (90.3,97.6)	94.1 (89.3,96.8)
Definition of VE	% incidence rate reduction	% hazard rate reduction
Type of intervals	Bayesian	Frequentist

scenario in which vaccine trials take place makes them closer to Epidemiology than the rest of Clinical Statistics. Randomized participants are volunteers in different health and social situations, possibly with some comorbidities (the main ones are monitored in the clinical trials and they possibly identify subgroups). Volunteers keep on living their 'normal' lives during the clinical trial, with just a slightly higher level of monitoring for solicited or unsolicited adverse events. Their process of exposure, infection and recovery from the disease is therefore entangled with the process for the general population.

Now, modeling with precision the epidemic dynamics is a very challenging task, as shown by the abundance of literature and models proposed and their generalizations, starting from the basic SIR model to a variety of SEIR, SIRD, delayed SIR etc.., both in the deterministic and in the stochastic versions. But when it comes to modeling the effect of a double blindly administered vaccine on top of it all, in the Pfizer-BioNTech approach the overall model is simplified to two overlapping homogeneous Poisson processes describing the infection processes: one for vaccinated participants - with intensity λ_V - and an independent one for the not vaccinated participants - with intensity λ_C. The time dimension of the Poisson processes

is called *surveillance time* and it is measured in person-years of follow-up. It is the sum of all the durations participants have been observed in the clinical trial from 7 (BioNTech/Pfizer) or 14 (Moderna) days after the second dose up until the first of the following four endpoints happen: onset of disease, death, loss to follow up or end of study. Only onsets of disease (possibly followed by death) contributes to an event in the mentioned Poisson processes. This way, λ_V and λ_C are simple the incidence rates among the vaccinated and the not vaccinated participants, according to basic Epidemiology (see for example lesson 3 of the online course `https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section2.html`).

A common measure of comparison between two infection processes in Epidemiology is the incidence rate ratio IRR= λ_V/λ_C; specularly, in the BioNTech/Pfizer case vaccine efficacy is defined as

$$\mathrm{VE_{BP}} = 1 - \mathrm{IRR} = 1 - \frac{\lambda_V}{\lambda_C} \tag{1}$$

(or its percent equivalent), which can be interpreted as the average fraction of missed infections (fraction of vaccinated participants who are not infected but would have been infected if not vaccinated). In order to estimate VE, we can define a likelihood based on the observed quantities:

- $s_V =$ surveillance time of the vaccine group,
- $s_C =$ surveillance time of the control group,
- $x_V + x_C =$ total number of infections,
- $x_V =$ total number of infections among vaccinated participants.

If we use standard probability symbolism, the joint density of the corresponding random variables (indicated in capital letters), which is the dual way of writing the likelihood, can be expressed as

$$f_{S_V,S_C}(s_V,s_C|\lambda_V,\lambda_C) \times f_{X_V+X_C|S_V,S_C}(x_v+x_c|s_V,s_C\lambda_V,\lambda_C) \times \\ f_{X_V|X_V+X_C,S_V,S_C}(x_V|x_v+x_c,s_V,s_C\lambda_V,\lambda_C) \tag{2}$$

i.e. as the chain product of the marginal density of S_V, S_C (very difficult to derive), times the conditional density of $X_V + X_C$ (which is Poisson$(s_V\lambda_V + s_C\lambda_C)$ by the properties of overlapping independent Poisson processes), times the conditional density of X_V given $x_V + x_C$, which can be easily proved to be binomial:

$$X_V|x_v+x_c,s_V,s_C\lambda_V,\lambda_C \sim \mathrm{Binomial}\left(x_v+x_c, \frac{s_V\lambda_V}{s_V\lambda_V+s_C\lambda_C}\right). \tag{3}$$

Notice that

$$\frac{s_V\lambda_V}{s_V\lambda_V+s_C\lambda_C} = \frac{s_V(1-\mathrm{VE})}{s_V(1-\mathrm{VE})+s_C}.$$

Now, if we assume that the first two lines of Equation (2) do not depend (much) on VE, we can reduce the likelihood to the binomial likelihood (3). In other words, we perform a statistical analysis conditional on the observed surveillance times and the

total number of cases. This is what, according to my reverse engineering of [2], the BioNTech/Pfizer authors mean by "adjusting for surveillance time".

To complete the BioNTech/Pfizer analysis, a Bayesian approach is used and a prior Beta with hyperparameters 0.700102 and 1 is assumed on $s_V(1-\mathrm{VE})/(s_V(1-\mathrm{VE})+s_C)$. The second hyperparameter is set to 1 to make the prior only slightly informative, whereas the first parameter is chosen so that, approximately, VE is centered at 0.3, which is a null value for vaccine efficacy set by the company and agreed with the Health Authorities. To be precise, since the prior has to be fixed before seeing any data, the authors assumed $S_V = S_C$, so that the probability of success in the equation 3 could be written (1-VE)/(2-VE) (an equation mentioned in the protocol). By imposing VE=0.3, one could have therefore derived an approximate expectation for the probability of case equal to 0.4117647 and set to 0.7 the firsr hyperparameter (in other words, there was no need to hassle with 6 decimal places in the prior specification Beta(0.700102,1)). Finally, by standard conjugate Bayesian analysis the credible interval for VE contained in Table 1 is obtained and the following criterion for final success becomes computable: in order to declare a successful trial, the posterior probability of vaccine efficacy greater than 0.30 has to be higher than 98.6% (which actually happened in the trial).

By contrast, the definition of VE given in the Moderna protocol is

$$\mathrm{VE_M} = 1 - \frac{h_V}{h_C} \tag{4}$$

where h_V and h_C are the hazard rates for time to infection of the vaccinated and of the not vaccinated group, respectively. That is due to the fact that a stratified Cox proportional hazard model is used, with group label used as covariate and stratification factors coming from randomization. In the case a parametric specification of a constant baseline hazard (exponentiality) and no stratification, the BioNTech/Pfizer and Moderna approaches would be equivalent. The authors of [1] state in the Supplementary material that "Analyses based on the exact method conditional on the total number of cases have also been performed and the results are consistent with that from the stratified Cox model". The approach of Moderna to the final analysis is more traditionally cast into hypothesis testing theory, with the null hypothesis set at VE=0.3 and the alternative hypothesis at VE=0.6.

Another interesting aspect of both trials is the planning of a few *interim analyses*, justified by the novelty of the vaccines and the urgency to make fast decisions in front of the pandemic. Setting boundaries for the interim analysis has been done by Moderna according to the O'Brien-Fleming approach, whereas for BioNTech/PFizer the presence of interim analysis has motivated some adjustments to the success criteria.

The statistical methods used for safety are simpler: basically, Clopper-Pearson confidence intervals are used to estimate the percentage of participants with solicited adverse events, with no adjustment for multiplicity, whereas descriptive statistics are used for unsolicited adverse events. This is because safety measures are viewed differently from efficacy, the interest being in alarm signs which could compromise the use of a new therapy.

References

1. Baden, L.R. *et al.*: Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. The New England Journal of Medicine **384**, 403–416. (2021).
2. Polack, F.P. *et al.*: Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. The New England Journal of Medicine **383**, 2603–2615. (2020).

2.7 Data Science for Industry 4.0 (ENBIS)

Sample selection from a given dataset to validate machine learning models

Campionamento da una base di dati assegnata per validare modelli di apprendimento automatico

Bertrand Iooss

Abstract The selection of a validation basis from a full dataset is often required in industrial use of supervised machine learning algorithm. This validation basis will serve to realize an independent evaluation of the machine learning model. To select this basis, we propose to adopt a "design of experiments" point of view, by using statistical criteria. We show that the "support points" concept, based on Maximum Mean Discrepancy criteria, is particularly relevant. An industrial test case from the company EDF illustrates the practical interest of the methodology.

Abstract *La scelta di una base di validazione da una base di dati completa è spesso richiesta nell'uso industriale di algoritmi di apprendimento automatico supervisionati. La base di validazione è utile per produrre una valutazione indipendente del modello di apprendimento automatico. Noi proponiamo di adottare, per selezionare questa base, il punto di vista della "programmazione degli esperimenti" fatta con criteri statistici. Il concetto di "punti di supporto" basato sul criterio Maximum Mean Discrepancy è particolarmente rilevante. Un test fatto in un caso industriale presso EDF viene illustrato per evidenziare l'interesse della metodologia.*

Key words: Supervised learning, Validation, Discrepancy, Space filling design

1 Introduction

With the development of automatic diagnostics based on statistical predictive models, coming from any supervised machine learning (ML) algorithms, important issues about model validation have been raised. For example in the industrial non-destructive testing field (e.g. for aeronautic or nuclear industry), generalized automated inspection (that will allow large gain in terms of efficiency and economy) has

Bertrand Iooss
EDF R&D, 6 Quai Watier, 78401 Chatou, France, e-mail: bertrand.iooss@edf.fr
SINCLAIR AI Lab., Saclay, France

to provide high guarantees in terms of performance. In this case, it is necessary to be able to select a validation data basis that will not be used for the training nor the selection of the ML model [3, 7]. This validation data basis (also referred as verification data in the literature) has not to be communicated to the ML developers because it will serve to realize an independent evaluation of the provided ML model (applying a cross validation method is then not possible). This validation sample is typically used to provide prediction residuals (which can be finely analyzed), as well as average ML model quality measures (as the mean square error in a regression problem or the misclassification rate in a classification problem).

In this paper, we address the particular question about the way to select a "good" validation basis from a dataset useful to specify a ML model. We use indifferently the term "validation" and "test" for the basis (also called sample) because we restrict our problem to the distinction between a learning sample (which includes the ML fitting and selection phases) and a test sample. An important question is the number and the location of these test points. For the size of the test sample, no general theoretical rule can be given while the classical ML handbooks [6, 5] provide different heuristic rules (as, e.g., 80%/20% between the learning and test samples and 50%/25%/25% between the learning, model selection and test samples).

In our validation basis selection problem, the dataset already exists so the problem turns to selecting a certain number of points in a finite collection of points. For simplicity, our work is limited to a supervised classification problem with two clusters: a validation sample is extracted in each sub-dataset (corresponding to each cluster). For the test sample location issue, simple selection algorithms are sometimes insufficient to ensure the representativity of the validation basis, in particular for small-size and highly unbalanced datasets. Indeed, the simplest and usual practice to build a test sample is to randomly extract an independent Monte Carlo sample [6]. If the sample size is small, as for the space-filling design issues [4], it is well known that the proposed points can be badly localized (test samples too close from learning points or leaving large input space subdomain unsampled). Therefore, a supervised selection based on statistical criteria is necessary.

A review of classical methods for solving this issue is given in [1]. For example, CADEX [8] is a sequential selection algorithm of points inside a database to put in a validation basis, via inter-points distance computations. From chemometrics, [2] complements this literature with cluster-based selection methods. Several ideas have also been recently introduced in order to help interpreting the ML models [10]. It consists in identifying (in the dataset) the so-called prototypes (data instance representative of all the data) and criticisms (data instance not well represented by the set of prototypes). To extract prototypes and criticisms, [10] explains the principle of a greedy algorithm based on the Maximum Mean Discrepancy (MMD, see [13]).

Our work hybridizes the latter approach with the concepts of support points recently introduced by [9], and which can be used to provide a representative sample of a desired distribution, or a representative reduction of a big dataset. In Section 2, the support points based algorithm is presented, with a simple application case. Section 3 illustrates the practical interest of the methodology on an industrial test case. Section 4 concludes with some perspectives of this work.

2 Use of support points

In this section, we use the recent work of [9] about a method to compact a continuous probability distribution F into a set of representative points, called support points. With respect to more heuristic methods for solving this problem, support points have theoretical guarantees in terms of the asymptotic convergence of their empirical distribution to F. Moreover, the extraction algorithm is efficient in terms of computational cost, even for large-size test sample N (up to $N = 10^4$) and in high input space dimension d (as large as $d = 500$).

The construction of the support points is based on the optimization of the energy distance which is a particular case of the MMD criterion [14]. The MMD provides a distance between F and a uniform distribution (via a kernel metric) and can be used with a relative good computational efficiency in high dimension (thanks to the kernel trick). Let denote $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The discrete distribution of N_v support points $\mathbf{x}^{N_v} = (\mathbf{x}^{(i)})_{i=1\ldots N_v}$ is denoted F_{N_v} and the energy distance between F and F_{N_v} writes:

$$d_E^2(F, F_{N_v}) = \frac{2}{N_v} \sum_{i=1}^{N_v} \mathbb{E}\|\mathbf{x}^{(i)} - \zeta\| - \frac{1}{N_v^2} \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \mathbb{E}\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\| - \mathbb{E}\|\zeta - \zeta'\| \tag{1}$$

with $\zeta, \zeta' \sim F$ and by using the Euclidean norm. The energy distance is always non-negative and equals zero if the two distributions are the same. The support points $(\xi^{(i)})_{i=1\ldots N_v}$ are then defined by minimizing $d_E^2(F, F_{N_v})$. Finding the support points corresponds to solving an optimization problem of large complexity, where F is empirically known by the sample points (the dataset). [9] provides an efficient algorithm to solve it. The objective function is approximated by a Monte Carlo estimate, giving

$$(\xi^{(i)})_{i=1\ldots N_v} = \arg\min_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N_v)}} \left(\frac{2}{N_v n} \sum_{i=1}^{N_v} \sum_{k=1}^{n} \|\mathbf{x}^{(i)} - \mathbf{x}'^{(k)}\| - \frac{1}{n^2} \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \mathbb{E}\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\| \right) \tag{2}$$

where $(\mathbf{x}'^{(k)})_{k=1\ldots n}$ is the n-size sample from F. This cost function can be written as a difference of convex functions in $\mathbf{x}^{N_v}$ and then can be minimized thanks to a formulation as a difference-of-convex program. This procedure being quite slow, a combination of the convex-concave procedure (CCP) with resampling is used (see [9] and references therein for details) in order to obtain an efficient algorithm. The examples given by [9] clearly show that support points distribution are more uniform than the ones of Monte Carlo and quasi-Monte Carlo samples [4].

In the CCP procedure, the selected points are not extracted from the dataset but are the "best" points representative of the full dataset distribution. Therefore, for our points selection problem, an additional step is required in order to find the N_v representative points inside the dataset. For each support point, we select the nearest dataset point and call this new algorithm SPNN ("support points nearest neighbor").

To illustrate SPNN on a toy example, we build a two-class two-dimensional ($d = 2$) dataset of size $N = 100$. The classification model is the following:

$$Y = \mathbf{1}_{X_1^2 - X_1 X_2 - X_1 - 3 > 0} \tag{3}$$

with $\mathbf{1}_{(.)}$ the indicator function, $X_1 \sim \mathscr{U}(-10, 10)$ and $X_2 \sim \mathscr{U}(-10, 10)$. The goal is to extract 20% of points for the test sample, respecting the proportion of points in each class. Applying the SPNN algorithm on the two sub-datasets (corresponding to each class) gives Fig. 1 which shows that the test points distribution in each class is quite satisfactory.

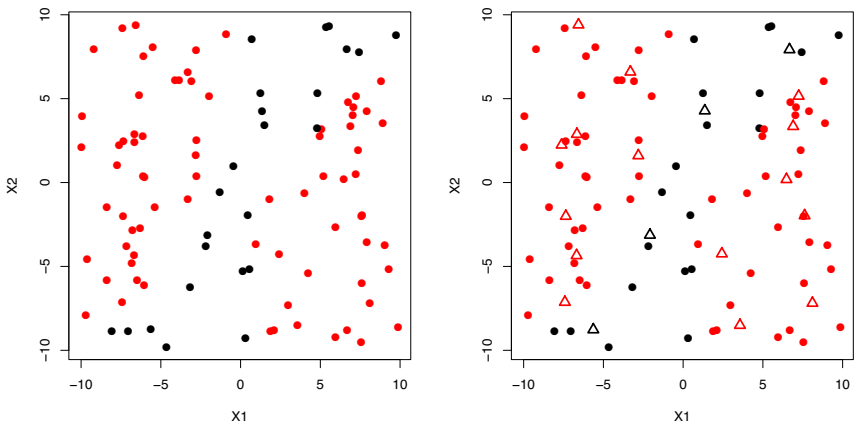


Fig. 1 Indicator function example. Left: dataset points corresponding to the two clusters (black: $Y = 0$, red: $Y = 1$). Right: test points selected by the SPNN algorithm (triangle symbol).

3 Application on an industrial use-case

This industrial problem aims at studying the fission products released in the primary circuit's water of the EDF nuclear reactors, during the load drop phase of the reactor cold shutdown. The available full dataset allows for $N = 90$ observations containing $d = 25$ covariates (describing the operation conditions of the reactor just before the shutdown) and the iodine activity level [12]. The goal is to model the event that this iodine activity level exceeds a specific threshold, that can have large impact on the scheduled planning, so on operational costs. Our classification dataset is well balanced as 48.9% of the observations (called "positive") are above the threshold and 51.1% of the observations (called "negative") are below the threshold.

For simplicity and, as it is not the subject of this work, we consider a naive logistic linear regression model (which predicts the probability for an individual to be positive) as the ML model (the probability threshold value of 0.5 is used to assign each individual to one of the two classes). To measure the quality of this ML model, we use the two main classification metrics: the error rate ε (number of missclassified observations on total number of observations) and the sensitivity τ

(number of well classified positive on the total number of positive observations). Due to the large number of covariates relatively to the observations number, the ML model applied on the full dataset (or on any sub-sample) gives unsurprisingly an error rate of zero and a sensitivity of 100%. By using a leave-one-out (LOO) procedure [6], we are able to evaluate these metrics in prediction: $\varepsilon = 18\%$ and $\tau = 82\%$. This LOO procedure is also used in the following tests on each learning sample (resulting from the extraction of the validation sample from the full sample).

Our goal is to study the capabilities of the SPNN algorithm in evaluating these metrics for different sizes of the validation sample (between 10% and 66% of the full sample size). Figure 2 provides the results that are compared to those obtained from a random sampling strategy. Error rates and sensitivities seem adequately predicted from the SPNN-based validation samples, from ratio N_v/N between 0.1 and 0.35. Of course, this result is specific to our small-size use-case. For such studies, the results also clearly show the inadequacy of the random validation samples to predict the ML model predictive capabilities. Indeed, their confidence-intervals (CI) are huge and far from reference values.

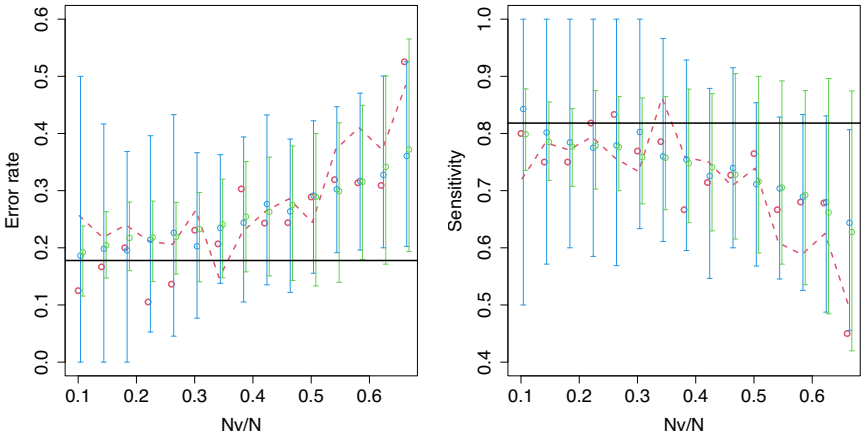


Fig. 2 Classification metrics (error rate ε at left and sensitivity τ at right) on the fission products dataset. Black line: reference values (LOO on full dataset). Red points (resp. dotted line): values from SPNN-based validation sample (resp. LOO-learning sample). Blue (resp. green) CI: 95%-CI from random validation samples (resp. LOO-learning samples).

4 Conclusion

In this work, the SPNN algorithm has been proposed for the selection of a test sample representative of a dataset. It is not restricted to an hypercubic domain (no need to transform each input to $\mathscr{U}(0,1)$) as the classical space-filling criteria in the computer experiments literature [4]. Moreover, compared to classical algorithms (as CADEX [8]), its computational cost does not depend on the dataset size and the data dimension. Its main practical limitation is that it becomes prohibitive for a test sample size N_v too large ($> 10^4$).

Further improvements of this work would be interesting to study in a near future. First, the approach gives equal importance to all the d inputs. It seems however useless to consider the inputs whose influence is negligible on the output. A preliminary step would be useful to identify important inputs and to apply the test sample selection algorithm only on these components. Second, new ideas for the support points definition can be developped, as for instance the use of the kernel Wasserstein distance [11] instead of the energy distance. Finally, this algorithm will also be useful for more complex classification problems where the inputs are temporal signals or images. Specific kernels on the input space should be adapted to these cases.

Acknowledgements This work has been funded by the international ANR project INDEX (ANR-18-CE91-0007) devoted to researches on incremental design of experiments. The author is grateful to Emmanuel Remy, Sébastien da Veiga, Luc Pronzato and Werner Müller for giving ideas during this work, as well as Emilie Dautrême, Vanessa Vergès and Marouane Il Idrissi for their help on the EDF dataset. Thanks to Grazia Vicario for the invitation to communicate this work.

References

1. T. Borovicka, M. Jr. Jirina, P. Kordik, and M. Jirina. Selecting representative data sets. In A. Karahoca, editor, *Advances in data mining, knowledge discovery and applications*, pages 43–70. INTECH, 2012.
2. M. Daszykowski, B. Walczak, and D.L. Massart. Representative subset selection. *Analytica Chimica Acta*, 468:91–103, 2002.
3. ENIQ. *Qualification of an AI / ML NDT system - Technical basis*. NUGENIA, ENIQ Technical Report, 2019.
4. K-T. Fang, R. Li, and A. Sudjianto. *Design and modeling for computer experiments*. Chapman & Hall/CRC, 2006.
5. I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. The MIT Press, 2016.
6. T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, second edition, 2009.
7. R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli. *Guidance on the assurance of machine learning in autonomous systems (AMLAS)*. Assuring Autonomy International Programme (AAIP), University of York, 2021.
8. R.W. Kennard and L.A. Stone. Computer aided design of experiments. *Technometrics*, 11:137–148, 1969.
9. S. Mak and V.R. Joseph. Support points. *The Annals of Statistics*, 46:2562–2592, 2018.
10. C. Molnar. *Interpretable machine learning*. github, 2019.
11. Jung Hun Oh, Maryam Pouryahya, Aditi Iyer, Aditya P. Apte, Joseph O. Deasy, and Allen Tannenbaum. A novel kernel Wasserstein distance on Gaussian measures: An application of identifying dental artifacts in head and neck computed tomography. *Computers in Biology and Medicine*, 120:103731, 2020.
12. E. Remy, E. Dautrême, C. Talon, Y. Dirat, and C. Dinse Le Strat. Comparison of machine learning algorithms on data from the nuclear industry. In S. Haugen, A. Barros, C. van Gulijk, T. Kongsvik, and J.E. Vinnem, editors, *Safety and Reliability – Safe Societies in a Changing World: Proceedings of ESREL 2018*, pages 825–832, Trondheim, Norway, June 2018. CRC Press.
13. A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.
14. G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013.

Reliable data-drive modelling and optimisation of a batch reactor using bootstrap aggregated deep belief networks

Changhao Zhu and Jie Zhang

Abstract To enhance the generalisation performance of DBN models, instead of building just one DBN model, several DBN models are developed from bootstrap re-sampling replication of the original modelling data and these DBN models are combined together to form a bootstrap aggregated DBN model (BAGDBN). BAGDBN is used for the modelling of a batch reactor and its generalisation performance is significantly better that that of a DBN model. Furthermore, model prediction confidence bounds can be readily obtained from the individual DBN model predictions and can be incorporated into the batch reactor optimisation framework to enhance the reliability of the resulting optimal control policy. Wide model prediction confidence bound is penalised to enhance the reliability of optimisation.

Key words: Deep belief network, process modelling, optimisation, reliability

1 Introduction

Batch reactors are suitable for the agile manufacturing of high value added products such as pharmaceuticals and specialty chemicals as the same reactors can be used to produce different products or different grades of products [2]. Batch chemical

[1] Changhao Zhu, School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK; email: zhu5@newcastle.ac.uk

Jie Zhang, School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK; email: jie.zhang@newcastle.ac.uk

reaction processes are typically highly nonlinear and batch to batch variations commonly exist in practice. Optimisation of batch process operation is essential for the enhanced production efficiency and product quality. Batch process optimisation usually requires an accurate process model that can accurately predict the end of batch product quality variables. Developing accurate mechanistic models for batch processes is typical very time consuming and effort demanding. This is because a chemical reaction network usually involves a large number of reactions and some reaction pathways and/or kinetic parameters are not readily available. To overcome this difficulty, data-driven models developed from process operation and plant testing data should be capitalised. As batch chemical reaction processes are typically very nonlinear, nonlinear data-driven modelling techniques should be utilised. Machine learning techniques including neural networks and more recently deep learning, e.g. deep belief network (DBN), are effective techniques for data-driven modelling of batch processes [6,7].

This paper presents a reliable data-driven modelling and optimisation strategy for a batch chemical reactor using bootstrap aggregated DBN (BAGDBN). BAGDBN has enhanced modelling accuracy and reliability due to the combination of multiple models. Through incorporating model prediction confidence bound from BAGDBN into the optimisation objective function, optimisation reliability can be enhanced.

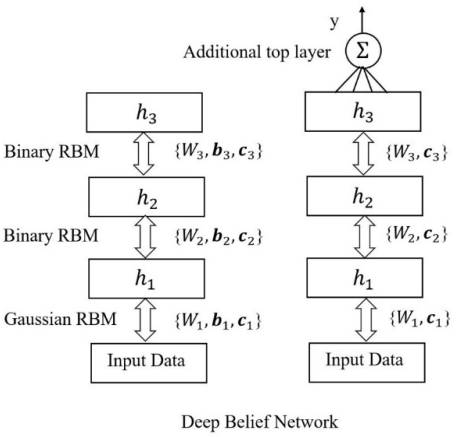


Figure 1: A DBN

2 Bootstrap Aggregated Deep Belief Networks

It is suggested that networks with a deep structure can achieve reliable learning results in recent research [4]. In a DBN model, several restricted Boltzmann machines (RBMs) can be stacked and combined as one learning network. Figure 1 shows the structure of DBN. This network contains three hidden layers, an input layer, and an output layer. Each hidden layer of DBN is regarded as one single RBM. The first phase of training is unsupervised training and the process

operational data are used to train the DBN model without any target variables involved. The unsupervised training helps the DBN to mine more correlations than feed-forward neural network. The weights are adjusted in a desired region before the supervised training phase. After unsupervised training, DBN is fine-tuned by the backpropagation algorithm in the supervised training phase.

A perfect DBN is very difficult to build due to limitation in the data. The main idea of BAGDBN is to develop multiple DBN models and then combine them to improve model prediction reliability and accuracy. To increase the diversity of these individual DBN models, each DBN model is developed from a replication of the original modelling data set generated through bootstrap resampling [3]. A diagram of BAGDBN is shown in Figure 2. These individual DBN models in a BAGDBN are trained to find the relationship between process input and output variables. Predictions from these individual DBN models are then combined to obtain the final prediction of the BAGDBN model. The output of a BAGDBN can be formulated as,

$$f(X) = \sum_{i=1}^{n} w_i \, f_i(X) \quad (1)$$

where $f(X)$ is the output of BAGDBN, $f_i(X)$ is the output of the ith DBN, w_i is the aggregating weight of the ith BAGDBN, n is the number of DBN models in the BAGDBN model, and X is a vector of model inputs. A further benefit of BAGDBN is that model prediction confidence bound can be obtained [5].

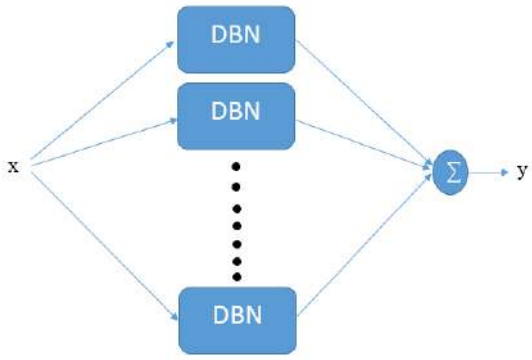


Figure 2: A BAGDBN

3 Reliable Modelling and Optimisation of a Batch Chemical Reactor

The batch chemical reactor presented in [1] is taken as a case study. The following two parallel exothermic reactions occur: A+B→C, A+C→D, where A and B are raw materials, C is the desired product and D is undesired by-product. The process operation objective is to produce the most amount of product by adjusting the reactor temperature. A simulation programme is developed using the mechanistic

model in [1] to represent the process and to test the developed modelling and optimisation strategy.

3.1 Modelling Using BAG-DBN

The batch duration is 200 hours and is divided into 10 equal stages. The reactor temperature setpoint is kept constant within each stage. The control policy contains the reactor temperature setpoints at these 10 stages. To develop a BAGDBN model for the batch reactor, 120 batches were simulated and 95 batches were used as the training and testing data while the remaining 25 batches were used as unseen validation data. 30 DBN models were developed from bootstrap re-sampling replications of the training and testing data. For each DBN, its inputs are the control policy (i.e. the 10 temperature setpoints during a batch) and its output is the final amount of product. Figure 3 shows the mean square errors (MSE) of the individual DBN models. It can be seen that the individual DBNs give various performance and the performance on training and testing data is not consistent with that on the unseen validation data. In contrast, BAGDBNs give much improved and consistent performance as shown in Figure 4, where the *x*-axis represents the number of DBNs included in the BAGDBNs. It can be seen from Figure 4 that the MSE values on training and testing data, as well as on the unseen validation data, decrease as more DBNs are aggregated and reach stable levels after sufficient number of DBNs are included.

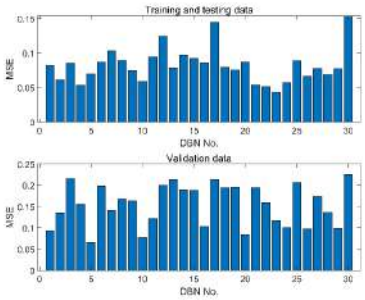


Figure 3: MSEs of individual DBNs

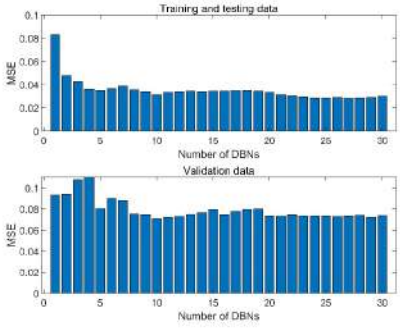


Figure 4: MSEs of BAGDBNs

3.2 Reliable Optimisation Control

Figures 5 and 6 show, respectively, the optimal control policies and the predicted and actual final product when a single DBN is used in the optimisation. It can be seen that optimisation using a single DBN gives a wide range of control policies (Figure 5) and unreliable performance as DNB predicted values can be quite different from the actual value when the "optimal" control policies are implemented (Figure 6).

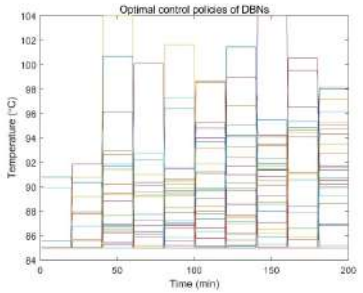


Figure 5: Optimal control policies (single DBNs)

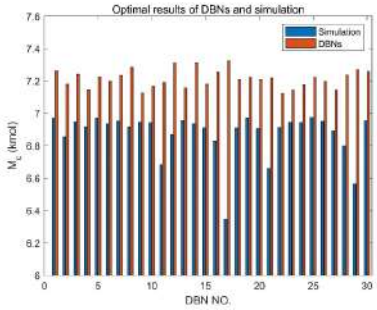


Figure 6: Optimal results of DBNs and simulation

To address this problem, a reliable optimisation strategy based on BAGDBN is studied here. The model prediction confidence bound is incorporated in the optimisation objective function so that wide confidence bounds is penalised [6]. The modified objective function is:

$$\min_{U} J = -M_c(t_f) + \lambda\sigma_e \quad (2)$$

$$\text{s.t. } M_c(t_f)=f_{BAGDBN}(U),\ T_L \le U \le T_U$$

where U is a vector of control actions (the 10 reactor temperature setpoints), $M_c(t_f)$ is the amount of product, T_L and T_U are the lower and upper bounds for U, t_f is the final batch time, λ is the weight for model prediction standard error σ_e, which indicates the width of the model prediction confidence bound.

Figure 7 shows the optimisation results under different weights (λ) for the confidence bounds. It can be seen that when λ =0, the model prediction confidence bound is quite wide and the difference between the BAGDBN prediction and the true value is quite large. As λ increases, the confidence bound gets narrower and the differences between the BAGDBN predictions and the true values get smaller and,

hence, optimisation reliability improves. The lower confidence bound indicates the worst case result and the appropriate value of λ can be selected as the value beyond which the worst case result starts to drop (around 3.75 in Figure 7).

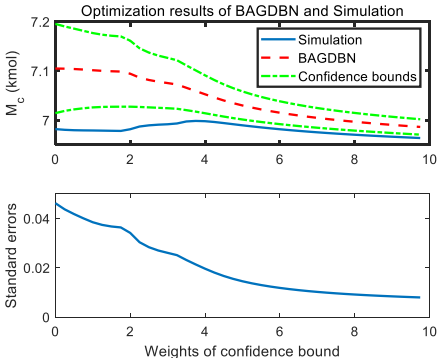


Figure 7: Optimization results of BAGDBN

4 Conclusions

A BAGDBN model is shown to give more accurate and reliable predictions than a single DBN model. Confidence bounds for BAGDBN predictions can be calculated from individual DBN predictions. Through incorporating model prediction confidence bound in the optimisation objective function, reliable optimal control policy can be obtained by penalising wide model prediction confidence bounds. Application to a batch chemical reactor demonstrate the advantages of the proposed BAGDBN based modelling and optimisation strategy.

References

1. Arpornwichanop, A., Kittisupakorn, P., Mujtaba, I.M.: On-line dynamic optimization and control strategy for improving the performance of batch reactors, Chemical Engineering and Processing: Process Intensification 44-1:101-114 (2005)
2. Bonvin, D.: Optimal operation of batch reactors: a personal view. Journal of Process Control 8: 355-368 (1998)
3. Efron, B., Tibshirani, R.: An Introduction to Bootstrap, Chapman & Hall, London, 1993.
4. Hinton, G.E., Osindero S., Teh Y.W. A fast learning algorithm for deep belief nets. Neural Comput 18: 1527–1554 (2006)
5. Zhang, J.: Developing robust non-linear models through bootstrap aggregated neural networks. Neurocomputing 25: 93–113 (1999)
6. Zhang, J.: A reliable neural network model based optimal control strategy for a batch polymerisation reactor. Ind. Eng. Chem. Res. 43: 1030-1038 (2004).
7. Zhu, C., Zhang, J.: Developing soft sensors for polymer melt index in an industrial polymerization process using deep belief networks. International Journal of Automation and Computing 17: 44-54 (2020)

2.8 Integration of survey with alternative sources of data

A parametric empirical likelihood approach to data matching under nonignorable sampling and nonresponse

Un approccio parametrico al matching per disegni di campionamento e mancate risposte non ignorabili

Daniela Marella and Danny Pfeffermann

Abstract Statistical matching attempts to combine the information obtained from different non-overlapping samples. The samples selected are often non representative of the finite population from which they are taken and not all the sampled units respond. The aim of this paper is to illustrate how informative sampling and not missing at random (NMAR) nonresponse can be handled in the statistical matching context.

Abstract *Il matching statistico combina l'informazione contenuta in due diversi campioni estratti dalla stessa popolazione. Lo scopo del presente lavoro è proporre un approccio al matching statistico quando i campioni sono informativi e la mancata risposta è non ignorabile.*

Key words: Calibration, Empirical likelihood**,** Informative sampling, NMAR nonresponse, Sample likelihood.

1 Introduction

Statistical matching attempts to combine the data obtained from different, non overlapping samples drawn from the same target population. The problem can be

[1] Daniela Marella, Università Roma Tre, Italy; email: daniela.marella@uniroma3.it

Danny Pfeffermann, Central Bureau of Statistics and Hebrew University of Jerusalem, Israel; University of Southampton, UK; email: D.Pfeffermann@soton.ac.uk

described as follows. Let A and B be two independent samples of size n_A and n_B respectively, selected from a population of N independent and identically distributed (*i.i.d.*) records, generated from some joint probability distribution function (*pdf*), $f_p(x, y, z; \theta)$ of variables (X, Y, Z) indexed by a vector parameter θ. The statistical matching problem is that (X, Y, Z) are not jointly observed in the two samples: only (X, Y) are observed for the units in sample A, and only (X, Z) are observed for the units in sample B, see D'Orazio *et al*., (2006). The sample selection in survey sampling involves complex sampling designs. The sample selection probabilities in at least some stages of the sample selection are often unequal and when the probabilities are correlated with the survey variables of interest, the sampling process becomes informative. Hence, the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population. See Pfeffermann and Sverchkov, (2009) for discussion of the notion of informative sampling and review of methods to deal with this problem. Marella and Pfeffermann, (2019) considered statistical matching under informative sampling designs with complete response, utilizing the sample likelihood (the likelihood of the observed sampled data) for parameter estimation. They illustrate that ignoring the sampling process in statistical matching can result in severely biased estimators and distort other aspects of the inference process. In what follows we consider the dual problem of informative sampling and NMAR nonresponse. Survey data are almost inevitably subject to nonresponse. Most of the approaches dealing with nonresponse assume that the missing data are missing at random (MAR, Little and Rubin, 1987). By this assumption, the probability of response does not depend on the unobserved data after conditioning on the observed data. When the response probabilities are correlated with the missing target outcomes even after conditioning on the observed data, the missing data are not missing at random (NMAR nonresponse). This type of nonresponse is the most difficult type of nonresponse to handle. No simple solutions exist in this case and serious bias may occur when it is ignored. The aim of the present paper is to analyze the statistical matching problem for the case where the sampling processes used to select the samples A and B are informative for the target variables of interest and the nonresponse in the two samples is NMAR.

2 Data matching under nonignorable sampling

Under the assumption that Y and Z are conditionally independent given X (CIA, for short), the joint population *pdf*, $f_p(x, y, z; \theta)$, factorizes as,

$$f_p(x_i, y_i, z_i; \theta) = f_p(x_i; \theta_X) f_p(y_i \mid x_i; \theta_{Y|X}) f_p(z_i \mid x_i; \theta_{Z|X}), \qquad (2.1)$$

where the parameters $\theta_X, \theta_{Y|X}, \theta_{Z|X}$ indexing the three *pdf*s are assumed to be distinct. If Z was observed in A, then following Pfeffermann *et al*., (1998), the sample *pdf* of (x_i, y_i, z_i) for $i \in A$ is defined as,

$$f_A(x_i, y_i, z_i; \theta, \kappa_A) = \frac{P(I_i^A = 1 \mid x_i, y_i, z_i; \kappa_A)}{P(I_i^A = 1; \theta, \kappa_A)} f_p(x_i, y_i, z_i; \theta) \\ = f_A(x_i) f_A(y_i \mid x_i) f_A(z_i \mid x_i, y_i) \quad (2.2)$$

where I_i^A is the sampling indicator taking the value 1 if the i th population unit is in sample A and 0 otherwise, and κ_A represents any additional parameters defining the sample distribution, resulting from the sample process. Under an informative sampling design, the observed outcomes are no longer representative of the population outcomes and the joint sample *pdf* $f_A(x, y, z)$ is different from the corresponding population *pdf*, $f_p(x, y, z)$. Under independence between observations corresponding to different sampling units, the *observed sample likelihood* of A is thus given by,

$$L_{Obs}^A(\theta_X, \theta_{Y|X}, \kappa_A) = \prod_{i=1}^{n_A} f_A(x_i, y_i; \theta_X, \theta_{Y|X}, \kappa_A). \quad (2.3)$$

An analogous expression to (2.3) holds for the sample likelihood operating in the sample B. Hence, the *sample likelihood* based on $A \cup B$ is,

$$L_{Obs}^{A \cup B} \propto \prod_{i=1}^{n_A} f_A(y_i \mid x_i) \prod_{i=1}^{n_B} f_B(z_i \mid x_i) \prod_{i=1}^{n_A} f_A(x_i) \prod_{i=1}^{n_B} f_B(x_i), \quad (2.4)$$

ommiting for convenience the parameters from the notation. As noted in Marella and Pfeffermann (2019), the sample likelihood has a similar structure to the likelihood under noninformative sampling defined in Rässler (2002). The big difference is that the population *pdf*s are replaced by the sample *pdf*s. Following Pfeffermann and Sverckov (2009), the sample *pdf*s featuring in (2.4) can be expressed as,

$$f_A(x_i; \theta_X, \kappa_A) = \frac{E_A(w_{i,A}; \theta_X, \kappa_A)}{E_A(w_{i,A} \mid x_i; \kappa_A)} f_p(x_i; \theta_X), \quad (2.5)$$

$$f_A(y_i \mid x_i; \theta_{Y|X}, \kappa_A) = \frac{E_A(w_{i,A} \mid x_i; \theta_{Y|X}, \kappa_A)}{E_A(w_{i,A} \mid x_i, y_i; \kappa_A)} f_p(y_i \mid x_i; \theta_{Y|X}) \quad (2.6)$$

where $f_p(x_i; \theta_X)$, $f_p(y_i \mid x_i; \theta_{Y|X}, \kappa_A)$ represent the corresponding population *pdf*s, E_A denotes the expectation under the sample *pdf*, $\pi_{i,A}$, $\pi_{i,B}$ and $w_{i,A} = 1/\pi_{i,A}$,

$w_{i,B} = 1/\pi_{i,B}$ denote the inclusion probabilities and the sampling weights, respectively. Equations (2.5), (2.6) define the relationship between the population *pdf*s and the sample *pdf*s. The expectation $E_A(w_{i,A} \mid x_i, y_i; \kappa_A)$ displayed in the sample *pdf* (2.6) can be estimated by regressing $w_{i,A}$ against (x_i, y_i) using the observed data. Analogous expressions to (2.5) and (2.6) hold for the sample *pdfs* $f_B(x_i; \theta_X, \kappa_B)$ and $f_B(z_i \mid x_i; \theta_{Z|X}, \kappa_B)$ operating in the sample B. Once the expectations are estimated, the model parameters $\theta_X, \theta_{Y|X}, \theta_{Z|X}$ can be estimated by maximization of the likelihood (2.4), with the weight expectations replaced by their respective estimates.

3 Data matching under nonignorable sampling and nonresponse

In practice, not all the sampled units respond. We therefore assume that in addition to the use of informative sampling designs, the nonresponse is NMAR, such that the response propensity is related to the study variables. Specifically, let R_i^A be the sample response indicator for the sample A, taking the value 1 if unit $i \in A$ responds and 0 otherwise. Let R_A denote the set of responding units in A and r_A, the size of R_A. The sub-sample R_A can be viewed as the result of a two phase sampling process: (i) a sample A is selected from the finite population with known inclusion probabilities $\pi_{i,A}$; (ii) the sub-sample R_A is selected from A with unknown response probabilities $P(R_i^A = 1 \mid I_i^A = 1)$. Under the CIA, the *pdf* for responding unit $i \in R_A$ (the respondents *pdf*) is obtained by Bayes rule as,

$$f_{R_A}(x_i, y_i, z_i; \theta, \kappa_A, \gamma_A) = \frac{P(R_i^A = 1 \mid x_i, y_i, z_i, I_i^A = 1; \gamma_A)}{P(R_i^A = 1 \mid I_i^A = 1; \theta, \gamma_A)} f_A(x_i, y_i, z_i; \theta, \kappa_A) = f_{R_A}(x_i) f_{R_A}(y_i \mid x_i) f_{R_A}(z_i \mid x_i, y_i), \quad (2.7)$$

where $f_A(x_i, y_i, z_i)$ represents the sample *pdf* for unit $i \in A$ defined in (2.2), and γ_A is the vector parameter governing the response mechanism in A. Note that unless $P(R_i^A = 1 \mid x_i, y_i, z_i, I_i^A = 1) = P(R_i^A = 1 \mid I_i^A = 1)$ for all $i \in A$, the respondents *pdf* $f_{R_A}(x_i, y_i, z_i)$ differs from the sample *pdf* $f_A(x, y, z)$ under complete response and from the target population *pdf* $f_p(x, y, z)$. Assuming that

the outcome, the sampling and the response are independent between units, the *respondents likelihoods* for A is given by,

$$L_{Obs}^{R_A}(\theta_X, \theta_{Y|X}, \kappa_A, \gamma_A) = \prod_{i=1}^{r_A} f_{R_A}(x_i, y_i; \theta_X, \theta_{Y|X}, \kappa_A, \gamma_A). \tag{2.8}$$

An analogous expression to (2.8) holds for the respondents likelihood operating in the sample B. The *respondents likelihood* of the sample $A \cup B$ is thus,

$$L_{Obs}^{R_A \cup R_B} \propto \prod_{i=1}^{r_A} f_{R_A}(y_i \mid x_i) \prod_{i=1}^{r_B} f_{R_B}(z_i \mid x_i) \prod_{i=1}^{r_A} f_{R_A}(x_i) \prod_{i=1}^{r_B} f_{R_B}(x_i), \tag{2.9}$$

omitting for convenience the parameters from the notation. Note that, the respondents *pdf* $f_{R_A}(x_i, y_i)$ is a function of the corresponding population *pdf*, the conditional expectations of the sampling weights, $P(I_i^A = 1 \mid x_i, y_i) = 1 / E_A(w_{i,A} \mid x_i, y_i)$, and the response probabilities $P(R_i^A = 1 \mid x_i, y_i, I_i^A = 1)$. Assuming that the response is independent of the sample selection, such that $E_A(w_{i,A} \mid x_i, y_i) = E_{R_A}(w_{i,A} \mid x_i, y_i)$, the probabilities $P(I_i^A = 1 \mid x_i, y_i)$ can be estimated as under complete response. Accounting for NMAR nonresponse is much more complicated than just accounting for informative sampling since the response probabilities $P(R_i^A = 1 \mid x_i, y_i, I_i^A = 1)$ are generally unknown and needed to be modelled. For this, a parametric model indexed by the unknown parameters $\gamma_A = (\gamma_{x,A}, \gamma_{y,A})$ is assumed,

$$P(R_i^A = 1 \mid x_i, y_i, I_i^A = 1) = g_A(\gamma_{x,A} x_i + \gamma_{y,A} y_i) \tag{2.10}$$

for some functions g_A, taking values in the range $(0,1)$. The same considerations apply to the respondents *pdf* $f_{R_B}(x_i, z_i)$, featuring in the likelihood (2.9). Modelling the response probabilities $P(R_i^A = 1 \mid x_i, y_i, I_i^A = 1)$ and $P(R_i^B = 1 \mid x_i, z_i, I_i^B = 1)$ by the logistic or probit functions is common, but notice that in our case the probabilities depend also on the study variables. The *respondents likelihood* (2.9) is maximized with respect to $\gamma_A, \gamma_B, \theta_X, \theta_{Y|X}, \theta_{Z|X}$. The maximization of (2.9) is often complicated numerically and might result in unstable estimates depending on the population model and the models assumed for the response probabilities. In order to overcome this problem, an alternative approach is to use the empirical likelihood (EL) (see Feder and Pfeffermann, 2019) which enables estimating the parameters γ_A, γ_B governing the response models acting in A and B, without specifying the population model. An additional important advantage of the EL approach is that it facilitates the use of calibration constraints. That is, auxiliary information on known

population means of some auxiliary variables can be incorporated by placing additional constraints on the maximization process. The EL approach under nonignorable sampling and nonresponse is investigated in Marella and Pfeffermann (2021). In Table 1 some simulation results regarding the EL approach are reported. The samples A and B are selected according to a Poisson sampling with selection probabilities depending on the variables of interest according to the sampling parameters $\kappa_A = (\kappa_{x,A}, \kappa_{y,A})$ and $\kappa_B = (\kappa_{x,B}, \kappa_{z,B})$. Furthermore, the samples A and B are subject to NMAR unit nonresponse by which the response probabilities depend on the study variables according to a logit model with parameters $\gamma_A = \gamma_B = (0.1, 0.02)$. With regard to X, Table 1 reports the Hellinger distance (HD) between the estimated *pdf* and the true *pdf* when the sample selection effects and a NMAR nonresponse are ignored (HD_1^X) and when both processes are taken into account (HD_2^X). In order to enhance the precision of the population parameters estimators a calibration constraint regarding the knowledge of the mean of X is introduced in the EL maximization.

Table 1: Hellinger distance for different choices of κ_A, κ_B and $\gamma_A = \gamma_B = (0.1, 0.02)$.

$\kappa_A = \kappa_B$	HD_1^X	HD_2^X
(0,0)	0.014	0.011
(1,0.5)	0.235	0.184
(1.25,0.5)	0.264	0.209

In Table 1 HD_1^X is larger than HD_2^X. Thus, ignoring the sample selection processes and a NMAR nonresponse affects negatively the quality of the estimates. Similar results (not reported) are obtained for the *pdf* of $Y \mid X$ and $Z \mid X$.

References

1. D'Orazio, M., Di Zio, M., Scanu, M.: *Statistical Matching: Theory and Practice*. Wiley, Chichester, (2006).
2. Little, R. J. A., Rubin, D. B.: *Statistical Analysis with Missing Data*, Wiley, New York, (1987).
3. Feder, M., Pfeffermann, D.: Statistical Inference Under Non-ignorable Sampling and Nonresponse. An Empirical Likelihood Approach. University of Southampton, Southampton, Highfield, UK: Southampton Statistical Sciences Research Institute; (2016). Available from http://eprints.soton.ac.uk/id/eprint/378245.
4. Marella, D., Pfeffermann, D.: Matching information from two independent informative sampling. *Journal of Statistical Planning and Inference*, 203, 70-81, (2019).
5. Marella, D. Pfeffermann, D.: Accounting for nonignorable sampling and nonresponse in statistical matching. In final preparation, (2021).
6. Pfeffermann, D., Krieger, A.M., Rinott, Y.: Parametric distribution of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114, (1998).
7. Pfeffermann, D., Sverchkov, M. (2009). Inference under informative sampling. In Handbook of Statistics 29B; Sample Surveys: Inference and Analysis (Eds, D. Pfeffermann and C.R.Rao), Amsterdam: North Holland, 455-487, (2009).

Survey data integration for regression analysis using model calibration

Jae-kwang Kim and Hang J. Kim

Abstract Data integration is an emerging area of research in survey sampling. By incorporating the partial information from external sources, we can improve the efficiency of the resulting estimator and obtain more reliable parameter analysis. In this paper, we consider regression analysis in the context of data integration. To combine partial information from external sources, we employ the idea of model-calibration which introduces an "working" reduced model based on the observed covariates. The working reduced model is not necessarily specified correctly, but can be a useful device to incorporate the partial information. The actual implementation is based on a novel application of the empirical likelihood method. The proposed method is particularly attractive for combining information from several sources with different missing patterns.

Key words: Empirical likelihood, Missing covariates, Measurement error models

1 Introduction

Data integration is an emerging area of research in survey sampling. By incorporating the partial information from external sources, we can improve the efficiency of the resulting estimator and obtain more reliable parameter analysis. Lohr and Raghunathan (2017), Yang and Kim (2020), and Rao (2020) provide a review of statistical methods of data integration for finite population inference. Many existing methods are mainly concerned with estimating the population mean or totals. Combining information for analytic inference, such as regression analysis, is not fully explored in the literature.

Jae-kwang Kim
Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. e-mail: jkim@iastate.edu

Hang J. Kim
Division of Statistics and Data Science, University of Cincinnati, e-mail: hang.kim@uc.edu

We are now interested in performing regression analysis in the context of data integration. When we combine data sources to perform combined regression analysis, we may encounter two problems. The covariates may not be fully observed. Also, some covariates are subject to measurement errors.

To combine partial information from external sources, we employ the idea of model-calibration which introduces an "working" reduced model based on the observed covariates. The working reduced model is not necessarily specified correctly, but can be a useful device to incorporate the partial information. The actual implementation is based on a novel application of the empirical likelihood method (Qin and Lawless, 1994). The proposed method is particularly attractive for combining information from several sources with different missing patterns. We have only to specify different working models for different missing patterns.

2 Basic Setup

Consider a finite population U of elements (x_i, y) which is believed to be an independent and identical realization of random vector (X,Y) with joint density $F(x,y)$ which is completely unspecified. Without loss of generality, write $U = \{1,\ldots,N\}$ and $X = (X_1, X_2)$. Suppose that we are interested in estimating the parameter in the regression model

$$E(Y \mid x_1, x_2) = m(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \tag{1}$$

where $m(\cdot)$ is known and $\beta = (\beta_0, \beta_1, \beta_2)'$ is the parameter of interest.

If we have a random sample S from the finite population, we can construct an estimating equation for β given by

$$\hat{U}(\beta) \equiv \sum_{i \in S} d_i U(\beta; \mathbf{x}_i, y_i) = 0 \tag{2}$$

where $d_i = \pi_i^{-1}$ is the sampling design weight for unit $i \in S$ and

$$U(\beta; \mathbf{x}, y) = \{y_i - m(x_1, x_2; \beta)\} h(x_1, x_2; \beta)$$

for some $h(x_{1i}, x_{2i}; \beta)$ such that the solution to (2) is unique (a.e.)

Now, suppose that we observe x_{1i} and y_i throughout the finite population. We wish to incorporate this extra information from the finite population. Chen and Chen (2000) first considered this problem in the context of measurement error model problems. To explain the idea in our setup, we can first consider a "working" model for $E(Y \mid x_1)$:

$$E(Y \mid x_1) = m(\alpha_0 + \alpha_1 x_1) \tag{3}$$

for some $\alpha = (\alpha_0, \alpha_1)'$. Under the working model (3), we can obtain $\hat{\alpha}$ by solving

$$\hat{U}_1(\alpha) \equiv \sum_{i \in S} d_i U_1(\alpha; x_{1i}, y_i) = 0,$$

where $U_1(\alpha; x_{1i}, y_i) = \{y_i - m(x_{1i}; \alpha)\} h_1(x_{1i}; \alpha)$.

Note that since (x_{1i}, y_i) are observed throughout the finite population, we can use the population observations to estimate α^*. That is, α^* satisfies $\sum_{i=1}^{N} U_1(\alpha; x_{1i}, y_i) = 0$. The optimal estimator of β given by

$$\hat{\beta}^* = \hat{\beta} + \widehat{Cov}(\hat{\beta}, \hat{\alpha}) \{\hat{V}(\hat{\alpha})\}^{-1} (\alpha^* - \hat{\alpha}), \tag{4}$$

where $\hat{V}(\cdot)$ and $\hat{Cov}(\cdot)$ denotes the design-based variance estimator and covariance estimator, respectively. The working model in (3) is not necessarily correctly specified, but a good working model can improve the efficiency of the final estimator. This is also the main idea of model calibration in Wu and Sitter (2001).

Chatterjee et al. (2016) formulated the above problem as a constrained maximum likelihood (CML) estimation problem when β is a parameter in the conditional distribution of Y given X with density $f(y \mid \mathbf{x}; \beta)$. In our setup, the constrained ML estimation can be expressed as maximizing

$$l_p(\beta) = \sum_{i \in S} d_i \log\{f(y_i \mid \mathbf{x}_i; \beta)\}$$

subject to

$$\sum_{i \in S} d_i \int U_1(\alpha^*; x_{1i}, y) f(y \mid x_{1i}, x_{2i}; \beta) dy = 0. \tag{5}$$

Constraint (5) can be understood as a constraint for the parameter β to satisfy $E\{U_1(\alpha^*; x_1, Y) \mid \mathbf{x}; \beta\} = 0$. By imposing this constraint into the ML estimation, the extra information α^* can be naturally incorporated to the ML estimation framework.

The CML method is not directly applicable to our conditional mean model in (2) as the likelihood function for β is not defined in our setup. Nonetheless, one can use an objective function such as those in Generalized Method of Moments (GMM) to apply this constrained optimization problem, which is asymptotically equivalent to the empirical likelihood method (Imbens, 2002). Chatterjee et al. (2016) also noted that the CML approach can be formulated using the empirical likelihood method of Qin and Lawless (1994) and Qin (2000). However, they did not explicitly discuss how to formulate the CML as an application of the empirical likelihood method.

3 Proposed method

Because the empirical likelihood method can be viewed as a calibration weighting problem in survey sampling (Wu and Rao, 2006), we can formulate the above problem as an application of the model-calibration problem of Wu and Sitter (2001). The classical calibration problem can be formulated as finding the calibration weights w_i that minimizes $Q(d, w)$ subject to some calibration constraints (Deville and Särndal, 1992). For the objective function, we may either use the pseudo EL function

$$Q(d,w) = \sum_{i \in S} d_i \log(w_i) \tag{6}$$

considered by Wu and Rao (2006) or the maximum entropy function $Q(d,w) = \sum_{i \in S} w_i \log(w_i/d_i)$ considered in Kim (2010). Our calibration constraint is

$$\sum_{i \in S} w_i U_1(\alpha^*; x_{1i}, y_i) = 0. \tag{7}$$

This is the same spirit of using (5) but without introducing the conditional density function $f(y \mid \mathbf{x}; \beta)$. Thus, we can use the following model-calibration method to estimate β efficiently as follows.

1. Using the working model (3) to obtain α^* from an external source (such as finite population).
2. Find the calibration weights that minimize $Q(d,w)$ subject to (7).
3. Once the solution $\hat{w}_i$ is obtained from the calibration problem, we can use

$$\sum_{i \in S} \hat{w}_i U(\beta; \mathbf{x}_i, y_i) = 0$$

to estimate β.

If the benchmark α^* is not available from the finite population, but another estimate is obtained from an independent source, we can combine the two samples to obtain the benchmark. In practical situations, we may not have access to the raw data of the independent source. Suppose that only the summary statistics ($\hat{\alpha}_2$ and $V_2 = \hat{V}(\hat{\alpha}_2)$) are available for the parameter α in the reduced model in (3) from the another data source, we can use

$$\hat{\alpha}^* = \frac{V_1^{-1}\hat{\alpha}_1 + V_2^{-1}\hat{\alpha}_2}{V_1^{-1} + V_2^{-1}} \tag{8}$$

to obtain the best estimator of α, where $\hat{\alpha}_1$ is the estimates of α from the current sample (S) and V_2 is the variance estimator of $\hat{\alpha}_1$. Once $\hat{\alpha}^*$ is obtained by (8), we can use $\hat{\alpha}^*$ to replace α^* in the calibration equation in (7).

Regarding the choice of $\hat{U}_1(\alpha)$, the estimating function for the parameters in the reduced model, it is based on the working model for $E(Y \mid X_1)$. To systematically construct a better control function $U_1(\alpha; \mathbf{x}_1, y)$, we note that the problem can be viewed as a missing covariate problem. Thus, we can use the regression calibration technique to build a predictor $\hat{x}_2 = \gamma_0 + \gamma_1 x_1$ and use

$$U_1(\beta; x_{1i}, \hat{x}_{2i}, y_i) = \{y_i - m(x_{1i}, \hat{x}_{2i}; \beta)\}\, h(x_{1i}, \hat{x}_{2i}; \beta) \tag{9}$$

for the control function for the model-calibration method. We can either estimate γ from sample S or use any fixed parameter value as long as the solution to $\sum_{i \in S} d_i U_1(\beta; x_{1i}, \hat{x}_{2i}, y_i) = 0$ is unique. A benchmark estimator of β can be obtained using external sources to apply the proposed model-calibration method.

If we use the control function in (9) then we are essentially treating a regression of y on x_1 and $\hat{x}_2$ as the "working" model for model-calibration. This is feasible only when we have direct access to sample S_B in addition to the internal sample S.

References

Chatterjee, N., Y.-H. Chen, P. Maas, and R. Carroll (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association 111*, 107–117.

Chen, Y. H. and H. Chen (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B 62*, 449–460.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association 87*(418), 376–382.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics 20*, 493–506.

Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology 36*, 145–155.

Lohr, S. L. and T. E. Raghunathan (2017). Combining survey data with other data sources. *Statistical Science 32*(2), 293–312.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika 87*, 484–490.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics 22*, 300–325.

Rao, J. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, DOI 10.1007/s13571–020–00227–w.

Wu, C. and J. Rao (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics 34*, 359–375.

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*(453), 185–193.

Yang, S. and J. K. Kim (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science 3*, 625–650.

Latent Mixed Markov Models for the Production of Population Census Data on Employment

Modelli Markoviani a Variabili Latenti per la stima dell'Occupazione del Censimento della Popolazione

Danila Filipponi, Ugo Guarnera and Roberta Varriale

Abstract Since October 2018, the Italian Statistical Institute is conducting yearly Census of Population and Housing in order to produce updated information on the main socio-demographic characteristics of Italian resident population. This project is carried out through the integration of survey and administrative data. In this work we deal with the estimation of the employment status of the Italian resident population over 15 integrating census survey data, labour force data and administrative information. We propose to use Mixture of Latent Markov Models in order to take into account the longitudinal data structure and the deficiency of both survey and administrative measurement process.

Key words: Population census, Hidden Markov Model, Measurement errors

1 Introduction

Nowadays, the accessibility to different, detailed and linkable data sources makes feasible the realisation of statistical registers with several variables of interest. Then, descriptive statistics can be obtained on those variables by a direct computation of the indicator of interest. This is generally referred to as register-based statistics [1], [20], [21] and it is an essential part of cutting-edge research in order to satisfy the increasingly demand for more granular statistics.

The validity of this approach relies on the important requirements that the target population is known and variables are precisely measured. However, these assumptions are often violated in practice. Administrative data, for example, are generally collected by organisations for their administrative aims, with units and variables definition that may not perfectly correspond with those of the official statistics ([21]). A remedy to overcome the deficiencies of a single source is to combine different sources of data and carry out a joint inference within a likelihood-based framework. This approach enables us to borrow information from each source that helps

in gaining efficiency in estimation and correcting selection bias and measurement errors affecting a single source ([5], [6], [7]). Of course, the linkage itself may introduce errors that have to be taken into account ([10], [11]). Whenever survey data are available, these are generally treated as a privileged source of information, based on the assumption that they provide correct measures of the target variables and that they are not affected by selection bias. Under this scenario, the other data sources play essentially the role of covariates within a prediction approach ([18]).

However, a different and more complete way of thinking can be based on the assumption that all the available sources could be potentially affected by measurement errors. Indeed, although surveys are designed to meet precise statistical requirements, they can be affected by errors that may seriously compromise the accuracy of the target estimates.

In this paper we focus on the use of latent variable models to predict the true target value given the observed measurements in the data sources. These models are proposed in the context where multiple source are available, all of them contain information closely related to the target variable, but none can be assumed as a corrected measure of the target variable. Moreover, if one of the sources is a survey it is often possible to state that the deficiencies in the measurement process of survey and administrative sources are independent and that the informative powers are complementary. Usually surveys help in variables and population identification, whereas administrative data helps in detection of errors associated with the survey response process (see the discussion on the paper of [9]). Under this scenario a latent model can estimate the true target measure taking in to account errors of all available sources.

Recent examples on linked surveys and administrative data that address the problem of measurement errors are given by [2], [3], [14], [15] , [16] and most of the applications focus on the field of employment research. ([4] , [8], [12], [13], [17]). [17] use the Latent Markov Models (LMM) to estimate temporary workers on a linked Labour Force Survey and administrative data; [8] use LMM to estimate employment status in the Italian employment register; [4] combine LMM and multiple imputation to evaluate the accuracy of the Italian employment rate using a register based approach.

In this work we propose to use a mixture of LMM to predict the employment status for the 2018 Italian Permanent Census of the Population and Housing. The paper is structured as follows; section (2) illustrates the available sources on employment and describes the model used for the prediction of the employment status and section (3) reports some results.

2 Mixed Latent Variable Model to estimate employment count

2.1 The Data

Since October 2018, the Italian Statistical Institute is conducting yearly, during the first week of October, the Census of Population and Housing (Permanent Census, PC) on a sampling basis in order to collect updated information on the main characteristics of the Italian resident population and its social and economic conditions at national, regional and local levels. The project consists on the production of data for the entire population through the integration of census survey data and administrative information. The Census, among others, produces information on the employment count the day of the Census. The Census survey is composed of two samples with different two-stage sampling designs and involves about 1,400,000 resident households located in 2,800 Italian municipalities. The information on the employment status refers to the first week of October and it is coherent with the social statistics' employment definition.

In order to predict the employment status for the Italian population over 15 years old, the statistical information collected during the Census operation has been integrated at unit level with other two sources of information on employment: (i) the Labour Force Survey (LFS) and (ii) administrative data (AD).

LFS is, of course, the main European survey to produce quarterly estimate on the employment status. For each quarter, the Italian LFS collected information on about 70,000 households living in 1,246 Italian municipalities for a total of 175,000 individuals (representing 1.2% of the overall Italian population). Then, despite sampling error and deficiencies in the survey response process, LFS is the key survey to measure correctly the employment status.

Administrative data relevant for the labour statistics come mainly from social security and fiscal authorities. After an harmonization process, data are integrated and organised in an information system having a linked employer-employees structure; from this structure it is possible to obtain information on administrative employment status for the complete population and for every month of the reference year. However, the quality of the administrative sources is quite different from the one of the statistical surveys since it depends on specific administrative definitions which may not perfectly align with those of the social statistics. Moreover, some statistical units are not covered by administrative sources, because of coverage errors or non-regular employments.

2.2 The Model

As discussed above none of the available sources of information can be considered as error free, and then taken as a benchmark. Thus, the *true* employment status at time t for subject k is modeled as a binary latent variable L_{kt} taking values 0 or 1

depending on whether at time t subject k is employed or not, respectively. In the following, for sake of simplicity, reference to subject k will be removed by notation whenever not necessary. The stochastic process L is analyzed at the finite collection of times $1,\ldots,T$ and $L_{1:T}$ denotes the random vector $L_1,\ldots,L_T$. In this work, $T=12$, each time $t\in(1,\ldots,12)$ corresponding to a specific month of the year. This choice, mainly due to computational constraints, implies the need of *moving* the concept of employment status from weekly level (the natural reference time according to statistical regulation) to monthly level. We are interested in the employment estimate for the month of October.

Information from Census, LFS and administrative sources are treated as imperfect measures of the target process L. Specifically, $Y^{(i)}_{1:T}$ with $i=1,2$ denote the binary vectors of (possibly missing) values of the employment status at times $1,\ldots,T$ resulting from Census (i=1) and LFS (i=2), while a third measure is a dichotomous vector $Y^{(3)}_{1:T}$, whose t^{th} components is 1 (employed) for a certain individual if he, or she, appears in at least one of the original administrative sources at time t, and 0 otherwise. The following individual covariates X are used: sex, age class (5 levels), income class (5 levels) and two binary flags associated with retirement status and being a student. These individual covariates explain different behavioural characteristics in the employment dynamics. Moreover, in order to take into account the different typologies of administrative sources, a four-category covariate S is defined to account for a different quality of the administrative information. Furthermore, population heterogeneity is explicitly modeled through another categorical latent variable (independent of time) G with categories representing the membership label corresponding to three different sub-populations G_1,G_2,G_3. Specifically, $G=1$ for the sub-population of *never working people*, $G=2$ corresponds to individuals with *stable employment dynamics* and $G=3$ to people who are likely to *change frequently* their employment status. In practice, the three sub-populations are characterized in terms of their behavior during the observation time (one year). Since the employment characteristics of the three groups are likely to be strictly related to demographic features as well as to the type of administrative source information is taken from, the structure of the employment population is modeled by specifying the distribution of the latent random variable G conditional on covariates X and S.

The key element of the current approach is the modelling of the employment dynamics for each sub-population. This is done by assuming that the evolution of the employment status L for each G_g is governed by a first order Markov chain with initial probabilities $P(L_1=j|G=g)$ and transition matrix M^g whose typical element is $P(L_t=k|L_{t-1}=j,G=g),\ \ (j,k\in\{0,1\})$. The assumption of a Markov Chain can be a valid assumption for the sub-population $G=2,3$, whereas is unrealistic for the sub-population of *never working people*. We can assume that the latent process L for $G=1$, is degenerate with $P(L_t=1|G=1)=0$ for $t=1,...,12$.

The next step is the definition of the measurement component of the model, i.e., the probability distribution of the random vectors $Y^{(1)}_{1:T}$, and $Y^{(2)}_{1:T}$, and $Y^{(3)}_{1:T}$ given the latent process and the covariates. According to a common approach, we assume that, conditionally on the latent process, the three measurement processes are in-

dependent one of each other and that conditional on L_t, measures associated with both LFS, admin sources, and PC at time t are independent with the corresponding measures at different times (serial conditional independence). Moreover, the manifest variables $Y^{(3)}_{1:T}$ are supposed to depend on the covariate S while no covariate is introduced in the specification of the distribution of $Y^{(1)}_{1:T}$ and $Y^{(2)}_{1:T}$. Dependence of administrative measures on the covariate S follows naturally from the definition itself of S. In fact, measurement errors crucially depend on the administrative source from which information is taken. Moreover the definition of the variable S implies some logical constraints on $Y^{(3)}$; specifically, for all $t \in (1,\ldots,12)$, $S = 1$ (that is no admin information available) implies $Y^{(3)}_t \equiv 0$. An additional constraint is introduced stating that "no false positive" data are present in the LFS data. This constraint rely on the empirical evidence that unemployed people are unlikely to declare they work. Although this hypothesis may fail in some cases due to time shifting of the responses with respect to the actual working period, we assumes that departures from the LFS constraint are negligible.

3 Some Results

The model described in the previous section has been applied to a person-linked combined data set containing monthly employment status measured by administrative sources and surveys. For parameter estimation, the syntax module of the Software Latent GOLD v.5.1 has been used ([19]).

The final model has been chosen between different alternatives: decisions have been taken based on well known indexes, such as the Bayesian Information Criterium (BIC) and the Akaike Information Criterium (AIC), the Likelihood Ratio test and *empirical* considerations. From the chosen model, estimated employees counts are obtained at different domain levels, together with measurement error estimates of the three data sources, and the estimates of initial and transition probabilities of the HMM. Furthermore, we computed bootstrap confidence interval for the number of estimated employees.

References

1. Bakker, B. F. M., Van Rooijen, J.: Methodological challenges of register-based research. Statist. Neerland., **66**, (2012)
2. Boeschoten, L., Oberski D., De Waal T.: Estimating Classification Errors under Edit Restricns in Composite Survey-Register Data Using Multiple Imputation Latent Class Modelling (Milc), Journal of Official Statistics, 33, 921–962 (2017)
3. Boeschoten, L., De Waal T, Vermunt J.K.: Estimating the Number of Serious Road Injuries per Vehicle Type in the Netherlands Using Multiple Imputation of Latent Classes, Journal of the Royal Statistical Society. Series A, 182, 1463–1486 (2019)

4. Boeschoten L., Filipponi D., Varriale R., Combining Multiple Imputation and Hidden Markov Modeling to Obtain Consistent Estimates of Employment Status, Journal of Survey Statistics and Methodology, smz052 (2020)
5. Citro, C. F. : From Multiple Modes for Surveys to Multiple Data Sources for Estimates,Survey Methodology, 40, 137–161 (2014)
6. De Waal, T., Van Delden, A., Scholtus, S.: Multi-source statistics: basic situations and methods. Discussion Paper. Statistics Netherlands, The Haguey, (2017)
7. Di Zio, M., Zhang, L.C., De Waal, T.: Statistical methods for combining multiple sources of administrative and survey data. Surv. Statistn, 76, 17–26. (2017)
8. Filipponi, D., Guarnera, U., Varriale, R.: Hidden markov models to estimate italian employment status. NTTS 2019 Bruxelles (2019)
9. Hand, D. J.: Statistical challenges of administrative and transaction data, Journal of the Royal Statistical Society Series A, 181, issue 3, p. 555-605 (2018).
10. Lahiri, P., Larsen, M. D.: Regression analysis with linked data. J. Am. Statist. Ass., 100, 222–230, (2005)
11. Kim, G. Chambers, R.: Regression analysis under incomplete linkage. Computnl Statist. Data Anal., 56, 2756–2770. (2012)
12. Magidson, J., Vermunt J. K., Tran B.: Using a Mixture Latent Markov Model to Analyze Longitudinal us Employment Data Involving Measurement Error, in New Trends in Psychometrics, eds. K. Shigemasu, A. Okada, T. Imaizumi, and T. Hoshino, pp. 235–242, Universal Academy Press, Inc. 235–242 (2009)
13. Manzoni, A., Vermunt J. K., Luijkx R. , Muffels R.: Memory Bias in Retrospectively Collected Employment Careers: A Model-Based Approach to Correct for Measurement Error, Sociological Methodology, 40, 39–73 (2010)
14. Oberski, D. L.: Estimating Error Rates in an Administrative Register and Survey Questions Using a Latent Class Model, in Total survey error in practice: improving quality in the era of big data, eds. P. P. Biemer, E. D. D. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, C. Tucker, and B. T. West, New York: Wiley (2016)
15. Oberski, D., Kirchner, A., Eckman, S., Kreuter, F.: Evaluating the quality of survey and administrative data with generalized multitrait-multimethod model Journal of the American Statistical Association, 112 (520), 1477-1489 (2017)
16. Pankowska, P., Bakker B., Oberski D. L. , Pavlopoulos D.: Reconciliation of Inconsistent Data Sources by Correction for Measurement Error: The Feasibility of Parameter Re-Use," Statistical Journal of the IAOS(Preprint), 1–13.(2017)
17. Pavlopoulos, D., Vermunt J.: Measuring Temporary Employment. Do Survey or Register Tell the Truth?, Survey Methodology, 41, 197–214 (2015)
18. Valliant R., Dorfman A. H., Royall R. M. : Finite population sampling and inference: A prediction approach. New York: John Wiley (2000)
19. Vermunt, J. K., Magidson, J.: Technical guide for Latent GOLD 5.1:Basic, advanced, and syntax. Statistical Innovations Inc., Belmont, MA V (2016)
20. Wallgren, A. , Wallgren, B.: Register-based Statistics: Statistical Methods for Administrative Data, 2nd edn. Chichester: Wiley (2014)
21. Zhang, L.C.: Topics of statistical theory for register-based statistics and data integration. Statist. Neerland., **66**, 41-–63 (2012)

2.9 Media, social media and demographic behaviours

Monitoring the Numbers of European Migrants in the United Kingdom using Facebook Data

Monitorare i Numeri di Migranti Europei nel Regno Unito attraverso i Dati di Facebook

Francesco Rampazzo, Jakub Bijak, Agnese Vitali, Ingmar Weber, and Emilio Zagheni

Abstract In June 2016, the United Kingdom voted to leave the European Union. Given the uncertainty surrounding Brexit, this paper attempts to ascertain if the number of European migrants in the United Kingdom (UK) is decreasing by using weekly estimates of the numbers of migrants obtained from the Facebook Advertising Platform. The period of analysis is from March 2019 to March 2020. The anonymised count data are disaggregated by age, education, and country of origin. We use a simple Bayesian trend model with indicator variables for age, education, and country, to analyse the changes in the numbers of migrants. The Facebook data suggests a decreasing number of EU migrants in the UK.

Abstract *Nel Giugno 2016, il Regno Unito votò per uscire dall'Unione Europea. Data l'incertezza legata alla Brexit, questo articolo ha lo scopo di investigare se il numero di migranti Europei nel Regno Unito stia diminuendo utilizzando dati settimanali da Facebook Advertising Platform. I dati sono anonimi e disaggregati per età, istruzione e paese di origine. Il modello utilizzato è un Bayesian trend con variabili dicotomiche per età, istruzione e paese per analizzare i cambiamenti nel*

Francesco Rampazzo
Saïd Business School, Leverhulme Centre for Demographic Science, and Nuffield College at the University of Oxford, Park End St, Oxford OX1 1HP, United Kingdom e-mail: francesco.rampazzo@sbs.ox.ac.uk

Jakub Bijak
Centre for Population Change, University of Southampton, Building 58, SO17 1BJ, United Kingdom e-mail: J.Bijak@soton.ac.uk

Agnese Vitali
University of Trento, Via Verdi, 26 - 38122 Trento, Italy e-mail: agnese.vitali@unitn.it

Ingmar Weber
Qatar Computing Research Institute, A157, HBKU Research Complex, B1, Doha, Qatar e-mail: iweber@hbku.edu.qa

Emilio Zagheni
Max Planck Institute for Demographic Research, Konrad-Zuse-Straße 1 18057, Rostock, Germany e-mail: zagheni@demogr.mpg.de

numero dei migranti. I dati da Facebook suggeriscono che il numero di migranti Europei nel Regno Unito sia in diminuzione.

Key words: Migration, Brexit, Facebook.

1 Introduction

The aim of this paper is to analyse the effect of the uncertainity and threat related to the departure of the United Kingdom (UK) from the European Union (EU) on the stocks of European migrants present in the UK. Since 2016 the Office for National Statistics (ONS) has reported a positive but declining net migration of EU nationals to the UK [4]. This paper focuses on investigating whether there is a declining trend in the numbers of Europeans living in the UK. To evaluate this, weekly time series data of EU migrants in the UK were collected from the Facebook Advertising Platform. The period from March 2019 to March 2020 is analysed with weekly estimates of European migrants. This period has been chosen so as not to be contaminated by the effects of the COVID-19 pandemic. In the study period, between three and four years have passed since the EU Referendum (known as the Brexit Referendum), allowing for a lag in the decision-making process of migrants regarding the decision to stay or leave the UK. Furthermore, disaggregation by age, education, and country might inform us of the differences in change of trends across these groups.

2 Data

Digital traces have been used to study migration [9, 2, 7, 6], but the capacity of this data to study migration change at a detailed timely granularity has not yet been explored. For example, [1] collected Facebook advertising data *"every two to three months"* to study the impact of Hurricane Maria on out-migration from Puerto Rico. Moreover, [6] combined data from the Labour Force Survey (LFS) with Facebook Advertising data in a Bayesian hierarchical model describing an undercount of the LFS estimates of EU migrants. The data in this paper was downloaded from the Facebook Advertising Platform. Facebook Advertising Platform provides two usage metrics: Daily Active Users (DAUs), and Monthly Active Users (MAUs). In this research, we are using the MAU, which is the "estimated number of people that have been active on your selected platforms and satisfy your targeting spec in the past month" [1]. The package pySocialWatcher was used to query the Facebook Marketing API [3]. The interest was in downloading the number of migrants disaggregated by age, education, and country of origin. As a consequence, the data is disaggregated by:

[1] https://developers.facebook.com/docs/marketing-api/reference/ad-campaign-delivery-estimate/

- **age groups**: 15-19, 20-29, 30-39, 40-49, and 50+ years old;
- **education levels**: Secondary (No Degree, In High School, High School), Tertiary (In College, In Grad School, Graduated), and Unspecified;
- **countries**: France, Germany, Ireland, Italy, Latvia, Portugal, Poland, Romania, and Spain.

3 Methodology

A simple Bayesian trend model with indicator variables for age (a), education (e), and country (i) was used to analyse the changes in the number of migrants. The index t stays for time. The trend equation m_{aeit} is the mean of y_{aeit}, a log-normal distribution with precision parameter τ. The precision parameter τ is *a priori* distributed as a Gamma with both shape and rate parameters equal to 0.01.

$$y_{aeit} \sim \text{Log-Normal}(m_{aeit}, \tau) \tag{1}$$

$$\begin{aligned} m_{aeit} = c + c_1^T d_a + c_2^T d_e + c_3^T d_i + c_{13}^T (d_a \times d_i) + c_{23}^T (d_e \times d_i) + \\ (b + b_1^T d_a + b_2^T d_e + b_3^T d_i + b_{13}^T (d_a \times d_i) + b_{23}^T (d_e \times d_i)) \times t \end{aligned} \tag{2}$$

$$\tau \sim \text{Gamma}(0.01, 0.01) \tag{3}$$

The trend m_{aeit} is divided into two parts: the c component, which is the intercept of the trend that describes the initial magnitude, and the b component, the slope of the model that describes the gradient of the decline. The parameters c and b are the overall effects, c_1^T and b_1^T are vectors of the parameters for age, c_2^T and b_2^T are for education, and c_3^T and b_3^T are for country of origin. The vectors d_a, d_e, and d_i contain the dummy indicator for the variables of age, education, and country. The reference category group for age is 15-19 years old, for education it is Secondary education, and for country it is Italy. Interactions between age and country, c_{13}^T and b_{13}^T, and education and country, c_{23}^T and b_{23}^T, are included in the model. The choice of the interactions in the model was driven by an initial descriptive analysis of the residuals from a model with all the main effects included. After analysing the residuals, the following interactions were included: Unspecified Education, Tertiary Education, and age groups 20-29, 30-39 and 40-49 for Romania and Poland. The parameters $\mathbf{c} = (c, \ldots, c_3)$ and $\mathbf{b} = (b, \ldots, b_3)$ are assumed to be normally distributed $N(0, 0.0001)$, with mean 0 and precision 0, 0.0001.

4 Analysis

Figure 1 shows the total number of estimated migrants from each country of origin studied on a log scale, using data from the Facebook Advertising Platform from

March 2019 until March 2020. The descriptive statistics clearly document a declining trend.

The model was estimated in R using JAGS [8]. The model is based on 11,880 observations of the number Monthly Active Users by country, age, and education. Table 1 reports the posterior estimates of the main effects of c and b. The parameter c indicates the initial magnitude level of migrants (i.e. the model intercept) in comparison to the reference category, while the parameter b indicates the direction of the regression slope in comparison to the reference category. The estimated median of b is negative, indicating a decreasing slope over time of the log-transformed stocks of migrants.

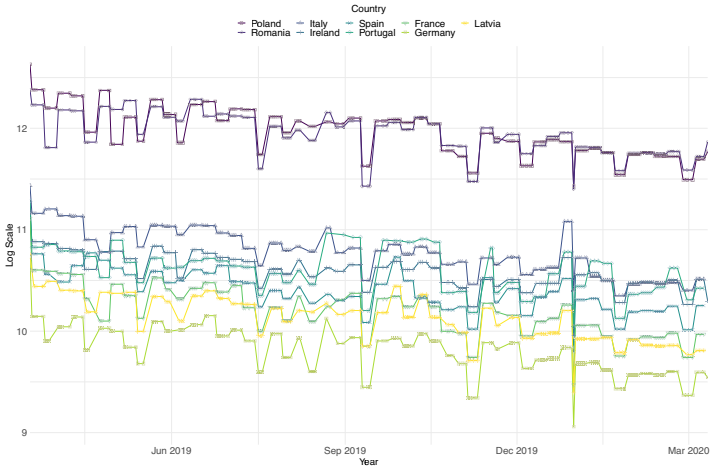


Fig. 1 Country time series with weekly data from the Facebook Advertising Platform from March 2019 to March 2020.

Table 1 Distribution of the main effect of c and b.

Parameter	2.5%	25%	50%	75%	97.5%	$\hat{R}$	$\hat{n}_{eff}$
c	8.00	8.03	8.04	8.06	8.08	1.01	97
b	-5.45×10^{-6}	-1.92×10^{-6}	-1.32×10^{-6}	1.82×10^{-6}	5.80×10^{-6}	1.02	107

In Figure 2, the respective estimated distributions of the different age groups are shown in comparison to the reference category of 15-19 years old. Looking at the b effect, the size of the age group that is decreasing fastest in comparison to the reference group is the 20-29 age group, followed by the 30-39 age group. The two groups decreasing fastest are also the two groups with a higher initial level in comparison to the 15-19 age group. The 50+ age group has a similar slope to the 15-19 age group, but a higher initial level.

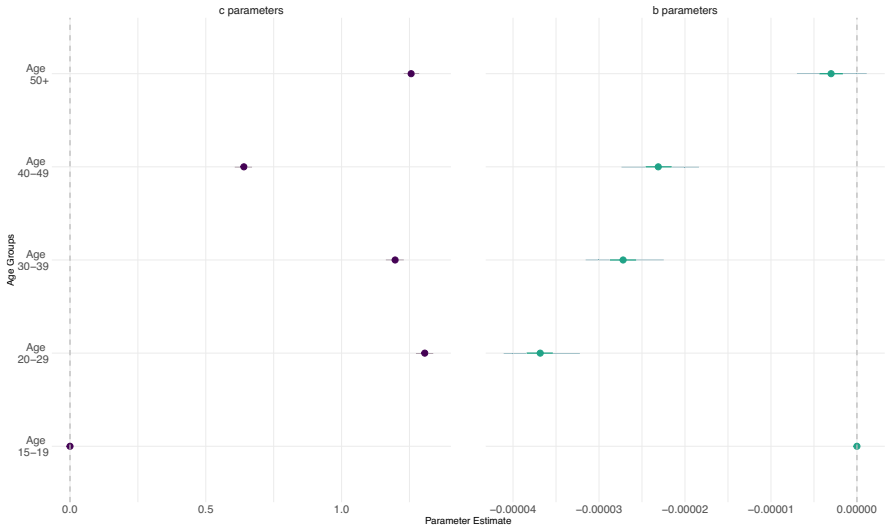


Fig. 2 Values of the *c* and *b* parameters estimated from the model for age.

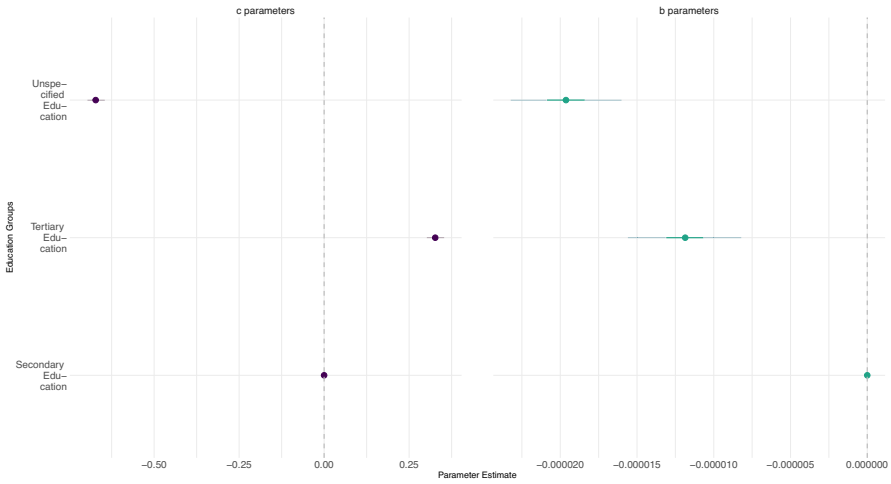


Fig. 3 Values of the *c* and *b* parameters estimated from the model for education.

Figure 3 shows the results for the education effects. The trend is decreasing fastest for the Unspecified category, followed by the Tertiary Education and then Secondary Education level. Figure 3 shows that the category with the highest number of migrants is Tertiary Education, followed by Secondary and Unspecified. In terms of countries of origin, the numbers of migrants from the largest Central and Eastern European countries, Poland and Romania, are decreasing faster than Italy and the other European countries in the analysis. The interactions effects on Poland

and Romania slightly slow down the decrease for both age groups 30-39 and 40-49, as well as for Unspecified and Tertiary Education groups.

5 Conclusions

It does seem that there is a declining trend in migrants coming to and living in the UK. The decline started after the expected Brexit date in March 2019, however as this coincided with an algorithm change in how the Facebook estimates were produced, it is difficult to establish the cause of this declining. The UK is clearly losing its attractiveness for the migrants living in the UK as well as new migrants coming to the UK, and this might be linked to the ongoing uncertainty surrounding Brexit. Although alternative data sources, such as the LFS, also show a decline in their estimates [4, 5], it is not as pronounced as the decline shown by the digital traces. This might be linked to the intrinsic timely nature of digital traces, but a more thorough enquiry into that aspect would be needed to corroborate this finding.

References

1. Alexander, M., Polimis, K., and Zagheni, E. (2019). The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45(3):617–630.
2. Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States. Population Research and Policy Review.
3. Araujo, M., Mejova, Y., Weber, I., and Benevenuto, F. (2017). Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations. In Proceedings of the 2017 ACM on Web Science Conference, WebSci '17, pages 253–257, New York, NY, USA. ACM.
4. ONS (2017). Migration Statistics Quarterly Report - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/migrationstatisticsquarterlyreport/november2017.
5. ONS (2020). Migration Statistics Quarterly Report - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/migrationstatisticsquarterlyreport/may2020.
6. Rampazzo, F., Bijak, J., Vitali, A., Weber, I., and Zagheni, E. (*Forthcoming*). A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: an Application in the United Kingdom. Demography.
7. Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Herranz, M. G., Al-Asad, M., and Weber, I. (2020). Monitoring of the Venezuelan exodus through Facebook's advertising platform. PLOS ONE, 15(2):e0229175.
8. Plummer, M., Stukalov, A., Denwood, M., and Plummer, M. M. (2016). Package 'rjags'. https://cran.r-project.org/web/packages/rjags/index.html.
9. Zagheni, E., Weber, I., and Gummadi, K. (2017). Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. Population and Development Review, 43(4):721–734.

2.10 New developments in ensemble methods for classification

An alternative approach for nowcasting economic activity during COVID-19 times

Un approccio alternativo per il nowcasting dell'attività economica in tempi di COVID-19

Alessandro Spelta and Paolo Pagnottoni

Abstract In this paper we propose a real time monitoring framework for tracking the economic consequences of various forms of mobility reductions across European countries. We adopt a granular representation of mobility patterns and we provide an analytical characterization of the rate of losses of industrial production using a nowcasting methodology. Our approach exploits the information encoded in massive dataset of human mobility provided by Facebook and Google, which are published at higher frequencies than the target economic variables, in order to obtain an early estimate before the official data becomes available. Our results show how industrial production strictly follows the dynamics of population commuting patterns an of human mobility trends which thus provide information on day-by-day variation of countries' economic activities.

Abstract *In questo documento proponiamo un quadro di monitoraggio in tempo reale per studiare le conseguenze economiche di varie forme di riduzione della mobilità nei paesi europei. Adottiamo una rappresentazione granulare della mobilità durante la prima e la seconda ondata di SARS-COV2 e forniamo una caratterizzazione analitica del tasso di perdite della produzione industriale utilizzando una metodologia di nowcasting. Il nostro approccio sfrutta le informazioni di dati sulla mobilità umana forniti da Facebook e Google, che vengono pubblicati a frequenze più elevate rispetto alle variabili economiche target, al fine di ottenere una stima anticipata prima che i dati ufficiali diventino disponibili. I nostri risultati mostrano come la produzione industriale segua rigorosamente le dinamiche dei modelli di pendolarismo. Le tendenze di mobilità umana forniscono quindi informazioni sulla variazione giorno per giorno delle attività economiche dei paesi.*

Alessandro Spelta
University of Pavia, Via S.Felice 5, 27100, Pavia (PV), e-mail: alessandro.spelta[unipv.it

Paolo Pagnottoni
University of Pavia, Via S.Felice 5, 27100, Pavia (PV), e-mail: paolo.pagnottoni@unipv.it

Key words: Covid-19; Human Mobility; Industrial Production; Nowcast

1 Introduction

Mobility restrictions has been identified as a key element to effectively limit the spread of the virus (see (1)), thus prompting the adoption of lockdown policies as ways to limit the contagion. On the other hand, such policy interventions induce severe disruptions on mobility patterns and determine relevant economic consequences because disposable workers are prevented from keeping up their activities. For this reason, a big effort is put in understanding the appropriate balance between the effect of mobility restrictions on the spreading of contagion and the direct and indirect consequences on economic outcomes.

This work aims to investigate the interplay between the SARS-COV-2 diffusion and mobility restriction measures, and how it affects productive system of some representative European countries. The economic assessment of mobility restriction measures is indeed of great interest for policy makers and motivates a growing literature related to the investigation and measurement of trade-offs between the need to limit the spread of contagion and the provision of adequate levels of economic output. Our work relies on a massive dataset of near real-time observations provided by Facebook through its Data for Good program and Google though Community Mobility Reports to build a model able to track the day-by-day economic activity of a country. We analyzed data based on the "Coronavirus Disease Prevention Maps" made available by Facebook as a part of its "Data For Good" program, a collection of unique dynamic spatial-temporal datasets illustrating worldwide populations commuting patterns over the COVID-19 pandemic period and data provided by Google within its Community Mobility Reports, an unprecedented phone-tracking based source of mobility data which aggregates anonymized information from users who have turned on their location history setting.

Our approach is grounded on nowcasting methodology, a technique previously employed to monitoring GDP dynamics in real-time (see (2)). The basic principle of nowcasting is the exploitation of the information which is published at higher frequencies than the target variable of interest in order to obtain an early estimate before the official figure becomes available. In other words, within the nowcasting framework, we are able to build a dynamic process for making a day-by-day short-term estimates of industrial production statistics that are announced at a monthly frequently and with long delays, thus providing policymakers a valuable tool for computing trade-off between the effect of mobility restrictions on the epidemic spreading and on the economy.

2 Methodology

For estimating the day-by-day impact that mobility reduction has on industrial production, we exploit the fact that these data series co-move quite strongly, so that their behaviour can be captured by few factors. In particular, through a dynamic factor model (DFM), we assume that the information of both mobility patterns and of industrial production, despite being released with different time frequency, i.e. "high" and "low" frequency, can be described by employing a small number of latent factors which follow a time series autoregressive process (see Methods).

For estimating the DFM we cast the model in a state space representation. Let $\hat{x} = (x_t', y_t^M)'$ and $\hat{\mu} = (\mu', \mu_M')$ state space representation results in:

$$\begin{aligned} \hat{x}_t &= \hat{\mu} + Z(\theta)\alpha_t \\ \alpha_t &= T(\theta)\alpha_{t-1} + \eta_t \end{aligned} \tag{1}$$

with $\eta_t \sim i.i.d.\mathcal{N}(0, \Sigma_\eta^2(\theta))$ where the vector of states includes the common factors and the idiosyncratic components and all model parameters are collected in θ. The details of the state space representation is provided in Supplementary Information. We estimate θ by maximum likelihood implemented by the Expectation Maximisation (EM) algorithm by (3). Maximum likelihood allows us to easily deal with such features of the model as substantial fraction of missing data. Given an estimate of θ, the nowcasts as well as the estimates of the factors or of any missing observations in $\hat{x}_t$, can be obtained from the Kalman filter or smoother.

Let us now denote Ψ_v the information set at time v. The nowcasting of y_t^M, i.e. the daily value assumed by the industrial production, is the orthogonal projection of y_t^M on Ψ_v given parameter estimates θ. Under the assumption that the data generating process is given by eq. 1 with θ equal to its quasi-maximum likelihood estimate, the Kalman filter and smoother can be used to obtain, in an efficient and automatic manner this projection for any pattern of data availability in Ψ_v. Another important feature of the nowcasting process is that a sequence of nowcasts is updated as new data arrive. In other words we, in general, perform a sequence of projections $\mathbb{E}[y_t^M|\Psi_v], \mathbb{E}[y_t^M|\Psi_{v+1}], ...,$. Nowcast updates are generally influenced by model's forecast errors corresponding to each data release and the effects of parameters re-estimation. Suppose at time $v+1$ new data are released, $x_{j,v+1}, j \in \mathscr{J}_{v+1}$, where j is the variable for which data are released and $\mathscr{J}$ the set of data released, then $\Psi_v \subset \Psi_{v+1}$ and $\Psi_v \setminus \Psi_{v+1} = x_{j,v+1}, j \in \mathscr{J}_{v+1}$. The nowcast update is given by a revision effect plus a parameter re-estimation effect:

$$\underbrace{\mathbb{E}(y_t^M|\Psi_{v+1}, \theta_{v+1})}_{\text{Update Nowcast}} = \underbrace{\mathbb{E}(y_t^M|\Psi_v, \theta_v)}_{\text{Old Nowcast}} + \underbrace{\mathbb{E}(y_t^M|\Psi_v, \theta_{v+1})}_{\text{Parameters Re-estimation}} + \underbrace{\sum_{j \in \mathscr{J}_{v+1}} \delta_{j,t,v+1}[x_{j,v+1} - \mathbb{E}(x_{j,v+1}|\Psi_v)]}_{\text{News Impact}} \tag{2}$$

Hence, the nowcast revision is a weighted sum of the news associated with the data release for each variable, while the effect of re-estimation is the difference between

the nowcast obtained using the old information set,Ψ_v, and the old parameter estimates, θ_v, and the nowcast using the old information set, Ψ_v, and the new parameter estimates, θ_{v+1}.

3 Data and Empirical Findings

The lockdown measures imposed in the countries exerted a dramatic impact on different dimensions of human mobility. We observe the different geographic distribution of commuting patterns in the countries' administrative regions before the first lockdown measures were imposed and the variations registered immediately afterwards. We find an interesting substitution effect regarding the mobility between and within places. Indeed, we observe a striking decrease in the commuting flows between administrative regions, accompanied by a surge in the commuting flows within them. On the one hand, the difference in the mobility between geographic zones comes as a natural consequence of the fact that restrictive measures did not allow people to move outside their own administrative regions. On the other hand, evidence suggests that people who are banned to travel outside their administrative tiles tend to move more around their neighborhood, being that for (rushed) grocery shopping or for a simple walk, thus increasing the degree of mobility within their region. The impact of lockdowns is not only tied to the magnitude of commuting flows, but also to the mobility trends across different categories of locations. Indeed, the effects of policy interventions and virus spread have been determinant to the evolution of human mobility flows.

In Figure 1 we present our nowcasting and forecasting results for the countries' industrial production in October, November and December 2020. Within our framework, the econometric model is re-estimated each day with the input of new data, hence variations of a country's industrial production result from the combination of news and model re-estimate effects. Evidence shows pretty accurate results, as well as that shocks in commuting flows and mobility trends impact the dynamics of the industrial production in an heterogeneous way, depending on the country under consideration, and the time span analyzed. On the one hand, we observe that, overall, Spain seems to be the country mostly affected by mobility shocks, followed by Germany and France. On the other hand, the dynamics of the Italian industrial production is mostly determined by the model re-estimation procedure. In general, news deriving from commuting patterns and mobility flows were particularly relevant to the industrial production in December, following the entangled restrictive measures in view of Christmas. Empirical outcomes also highlight the large intra-month variability of the industrial production dynamics, of which, in general, only the monthly level is known by policymakers, and not without any delay. This highlights the importance of our nowcasting tool as a higher-frequency indicator of economic activity during pandemic and emergency times, enhancing government's decision-making processes based on real-time data evidence.

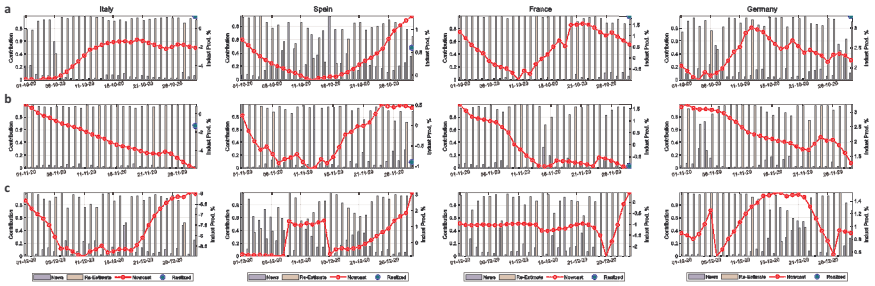


Fig. 1: **Nowcasting and forecasting results.** The figure shows the time series of nowcasts and forecasts of the Industrial Production Index for Italy, Spain, France and Germany. (**a**) and (**b**) show the nowcasting results for the months of October and November 2020, respectively, while (**c**) displays the forecasting results for December 2020. Black lines illustrate the dynamics of the daily nowcasts (and forecasts) for the change in Industrial Production. Colored bars represent the contribution of each component (news and model re-estimate components) to the daily change in Industrial Production in relative terms.

4 Conclusion

Mobility restrictions has been identified as key non-pharmaceutical interventions to limit the spread of the SARS-COV2 epidemics. On the other hand, these interventions present significant drawbacks to the social fabric and negative outcomes for the real economy. In this paper we propose a real time monitoring framework for tracking the economic consequences of various forms of mobility reductions across European countries. Our results first show the ability of mobility related policy to induce a contraction of the travelled distance and of mobility patterns across jurisdictions. Beside this contraction, we observe a substitution effect which increases mobility within jurisdictions. Secondly, we show how industrial production strictly follows the dynamics of population commuting patterns an of human mobility trends which thus provide information on day-by-day variation of countries' economic activities. Our work, besides shedding lights on how policy intervention targeted to induce a mobility contraction impact on industrial production, constitutes a practical toolbox for helping governs to design appropriate and balanced policy actions to timely respond SARS-COV2 while mitigating the detrimental effect on economy. Our study reveal how complex mobility patterns can have unequal consequences to economic losses across the countries and call for more tailored implementation of restrictions to balance the containment of contagion with the need to sustain economic activities.

References

[1] Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore y Piontti, A., Mu, K., Rossi, L., Sun, K., Viboud, C., Xiong, X., Yu, H., Halloran, M. E., Longini, I. M., Vespignani, A.: The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science, 368(6489), 395-400 (2020)

[2] Giannone, D., Reichlin, L., Small, D.: Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics, 55(4), 665-676 (2008)

[3] Banbura, M., Giannone, D., Reichlin, L.: Nowcasting with daily data. European Central Bank, Working Paper, 18 (2011)

Assessing the number of groups in consensus clustering by pivotal methods

Determinazione del numero di gruppi nel clustering di consenso mediante unità pivotali

Roberta Pappadà, Francesco Pauli, Nicola Torelli

Abstract We propose a tool for exploring the number of clusters based on pivotal methods and consensus clustering. K-means algorithm is used to learn the pairwise similarity via the co-occurrence of points in multiple partitions of the data. This similarity can be used to investigate the number of groups and detect arbitrary shaped clusters. Different criteria for identifying the pivots are discussed, as well as preliminary results concerning the selection of the optimal number of clusters.
Abstract *Viene proposto un metodo per la determinazione del numero di gruppi basato su unità 'pivotali' e clustering di consenso. L'algoritmo delle K-medie viene utilizzato per esplorare la similarità a coppie nei dati mediante la co-occorrenza in molteplici partizioni. Tale matrice può essere impiegata per scegliere il numero di gruppi e descrivere cluster di forma arbitraria. Vengono proposti diversi criteri per la selezione dei pivot, e presentati i risultati preliminari sulla scelta del numero ottimale di gruppi.*

Key words: consensus clustering, pivotal methods, K-means algorithm

1 Introduction

In order to cope with some critical issues typically faced when clustering data, new approaches based on the concept of *consensus* have been developed [7, 6]. It is well-recognized that different algorithms for grouping data have different qualities and shortcomings. Consensus clustering (or cluster ensemble problem) has been considered in a variety of different areas such as machine learning [9, 8], pattern recognition [4], data mining [5], to name a few. The general idea is that combining

Roberta Pappadà, Francesco Pauli, Nicola Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: rpappada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it

multiple clusterings could yield a final superior result. In the framework of ensemble methods for clustering, a typical way to summarize multiple groupings of the same data obtained by many runs of different methods is to evaluate a co-association matrix that contains, for each pair of observations, the proportion of groupings in which they are classified into the same cluster. Given a dataset of n observations $(y_1,\ldots,y_n)$, $y_i \in \mathbb{R}^d$, and a set of H data partitions, the entry (i,j) of the $(n \times n)$ co-association matrix $\mathscr{C}$ is $c_{ij} = n_{ij}/H$, where n_{ij} is the number of times the pair (y_i, y_j) is assigned to the same cluster among the H alternative partitions of the ensemble. A final clustering can then be obtained by using the co-association matrix itself as a similarity matrix for the data points.

From another perspective, the structure of the consensus matrix provides a natural way of identifying the appropriate number of clusters k, when no prior information is available. A key point is then how the ensemble is produced and how the information contained in the consensus matrix is summarized. In [1] the co-association matrix has been employed to find some specific data points (hereafter, pivots) which are representative of the group they belong to (because they never or very rarely co-occur with members of other groups). Such pivots can be used as cluster centers in K-means algorithm, in order to reduce the effect of random seeding, thus improving the quality of the clustering results [3].

In this paper, we propose an approach for choosing the more appropriate number of clusters based on a suitable index of separation between the pivots, as it might reflect the underlying structure of data. The rest of the paper is organized as follows: Sect. 2 presents pivotal methods and our approach for assessing the number of clusters, as well as an application to synthetic data. A small simulation study is discussed in Sect. 3, and some final remarks in Sect. 4 conclude the paper.

2 Determining the number of clusters via pivotal methods

A simple way to obtain a co-association matrix $\mathscr{C}$ is to combine multiple runs of K-means algorithm with random initialization of the k cluster centers. Based on $\mathscr{C}$ and an initial partition into k groups, $P_0 = \{G_1, G_2, \ldots, G_k\}$, we can identify k pivots–each pivot representing a different cluster–using one of the following criteria:

$$(a)\ i_g^* = \arg\max_{i \in G_g} s_j \quad (b)\ i_g^* = \arg\min_{i \in G_g} \tilde{s}_j \quad (c)\ i_g^* = \arg\max_{i \in G_g} (s_j - \tilde{s}_j), \tag{1}$$

for $g = 1,\ldots,k$, where $s_j = \sum_{j \in G_g} c_{ij}$ and $\tilde{s}_j = \sum_{j \notin G_g} c_{ij}$. We refer to these criteria as *pivotal methods*, as they return those units that are as far as possible from units that belong to the other groups and/or as close as possible to units that belong to the same group. To estimate the number of clusters we take advantage of the information the pivotal units give about the degree of separation between the clusters: if the majority of the input clusterings places the pivots in k different clusters, then the off-diagonal elements of the corresponding $k \times k$ submatrix of $\mathscr{C}$ will be 0's or very close to 0. It should be pointed out that the pivot i_g^* is identified by restricting

the search to the points in cluster G_g of P_0, $g = 1, \ldots, k$, to reduce the computational complexity. A couple of alternative algorithms have been considered to find P_0 (Partitioning Around Medoids (PAM) and Ward's hierarchical method): experimental evaluation shows that our method is not particularly sensitive to the choice of the initial partition.

As mentioned before, the extent to which the pivots are perfectly separated is taken as an indication of the quality of clustering. Let $\mathbf{Y}$ be a dataset of interest and $R^K = \{2, \ldots, k_{max}\}$, $k_{max} \leq n$, the range for the number of clusters. The proposed method works as follows. For all $k \in R^K$,

1. Build the cluster ensemble of dimension H using the K-means algorithm with random seeds and k as input, then construct the $n \times n$ consensus matrix $\mathscr{C}^k$.
2. Find a partition P_0 of $\mathbf{Y}$ into k groups performing single run algorithm (e.g. PAM or Ward's method) and extract the pivots using (a), (b) or (c). Let $I^k = \{i_1^*, \ldots, i_k^*\}$ denote the set of indices of the pivotal units for the solution with k groups.

The algorithm has been implemented using the `pivmet` R package that is freely available from the CRAN contributed packages repository [2]. Given that perfect separation between the pivots translates into an identity matrix, we propose to compare the resulting consensus matrices of pivots based on the quantity

$$m_k = \max\{c_{ij}, i, j \in I^k, i \neq j\} \tag{2}$$

expressing the highest similarity associated with the extracted pivots obtained by forcing k-cluster solutions. The values m_k as k varies in $[2, k_{max}]$ can be used as a tool for choosing k^* taking the fact that the value m_{k^*} is relatively low and a steep increase from k^* to $k^* + 1$ provides an indication that k^* groups are present. Note that, in situations that are poorly handled by the K-means method or when the clusters are not well-separated, the plot of m_k may suggest more than one value of k^*, due to the difficulty in finding perfectly separated pivots. As an illustration, consider the synthetic dataset [5] in Fig. 1, which consists of seven groups. As expected, K-means is not able to identify such a complex structure. We perform the steps described above using an ensemble of dimension $H = 500$, for each $k \in \{2, \ldots, 12\}$. By plotting the values m_k we observe that adding another cluster to $k = 3$ or $k = 7$ would lead to a large increase in the index (see Fig. 1-right), a further inspection shows that the largest jump in m_k occurs after $k = 7$ so a naive, yet effective, criterion is to select 7 groups (which corresponds to the 'actual' number of groups).

3 Experimental evaluation

The goal of this section is to show how pivot separation can be used to identify the appropriate number of clusters by performing the procedure described in Sect. 2 on a collection of simulated datasets. Here, we restrict our discussion to the following three scenarios in two dimensions (see Fig. 2):

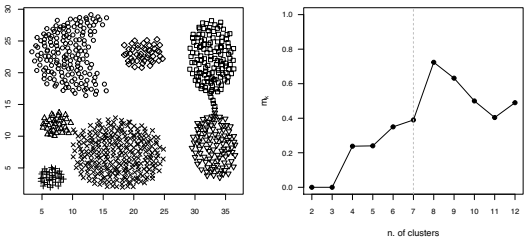


Fig. 1 The `aggregation` dataset ($n = 788$) on the left and plot of the m_k index versus the number of clusters on the right. The vertical line is the true value $k = 7$.

S1 Dataset It consists of $n = 600$ points that form two oblong clusters and one spherical cloud generated from a bivariate Gaussian distribution with diagonal covariance matrix $\Sigma = 0.01\mathbf{I}$, where $\mathbf{I}$ is the identity matrix; each group is composed by 200 observations.

S2 Dataset It is formed by four unequal-size clusters of $n = 400$ points generated by the union of observations from two mixtures of bivariate Gaussian distributions. For the first mixture we use weights $\pi = \{0.3, 0.7\}$, for the second mixture the weights are $\pi' = \{0.4, 0.6\}$, in both cases the covariance matrices of the mixtures' components are $\Sigma_1 = 25\mathbf{I}$ and $\Sigma_2 = 75\mathbf{I}$.

S3 Dataset It consists of a total of $n = 250$ points, 50 per group, where three out of five clusters are generated by spherical Gaussians with $\Sigma = 0.5\mathbf{I}$, while two clusters takes on different shapes having unit variance and correlation coefficient -0.7 and 0.9, respectively.

Note that some of these scenarios (notably $S1$ and $S3$) have been proved extremely challenging for classical non hierarchical and hierarchical clustering algorithms. We perform $B = 100$ simulations for each dataset, using $H = 500$ and $k_{max} = \sqrt{n}$. Fig. 3 shows the boxplots of m_k in Eq.(2) computed for the simulated datasets using pivotal methods (a)–(c), as k increases. Table 1 reports the proportion of simulations in which the pivot-based methods identify the correct number of clusters by choosing the value k^* before the biggest "jump" in the plot (this being a rough automatization of the use of the proposed diagnostic). The same proportion is computed for three well-known criteria, i.e. Average Silhouette Width, Dunn index, and CH, using K-means and hierarchical method with complete linkage (CL). Pivotal methods are consistent across a high proportion of simulated datasets, similarly to the silhouette index with CL in scenarios $S2$, $S3$. Moreover, our approach outperforms the other criteria tested in the majority of cases, especially in scenarios $S1$ and $S3$.

4 Conclusion and future work

The use of pivotal methods in consensus clustering is a promising and effective strategy to summarise the co-association matrix and to select the number of clusters. The proposed approach proved to be useful to assess the appropriate number of clusters,

(a) $S1$ Dataset (b) $S2$ Dataset (c) $S3$ Dataset

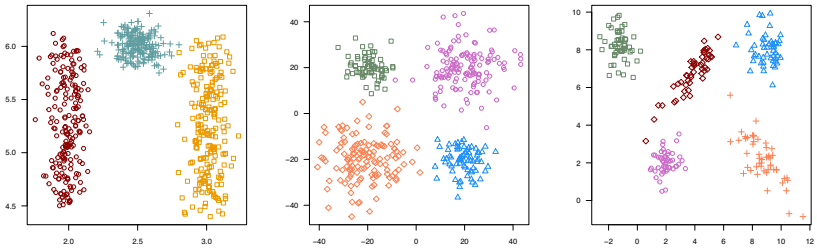


Fig. 2 Synthetic datasets considered in the simulation evaluation. The true clusters (denoted by different colors) differ in shape, size, and density.

Table 1 Proportion of simulations in which the correct number of clusters is identified by pivotal methods (with initial partition P_0 given by PAM or Ward's method) and Silhouette Width (Sil), Dunn index, CH index applied with K-means and Complete Linkage, respectively.

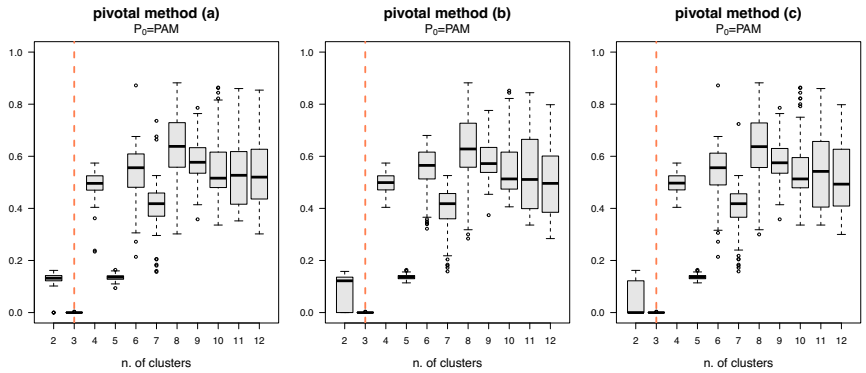
	P_0 : *PAM*			P_0 : *WARD*			K-means			AHC (CL)		
Data\ Method	(a)	(b)	(c)	(a)	(b)	(c)	Sil	Dunn	CH	Sil	Dunn	CH
$S1$ ($K_{true} = 3$)	0.86	0.86	0.89	0.75	0.76	0.80	0.03	0.03	0.00	0.18	0.00	0.00
$S2$ ($K_{true} = 4$)	0.96	0.85	0.96	0.95	0.83	0.93	0.98	0.45	0.98	0.89	0.44	0.78
$S3$ ($K_{true} = 5$)	0.99	0.85	0.96	0.97	0.79	0.87	0.01	0.01	0.00	0.85	0.47	0.57

even when the clustering method adopted for the ensemble generation consistently fails to recover the true data partition. The selection rule can be furtherly improved by jointly considering the level of the index and the magnitude of the jumps, an issue that deserves further investigation.

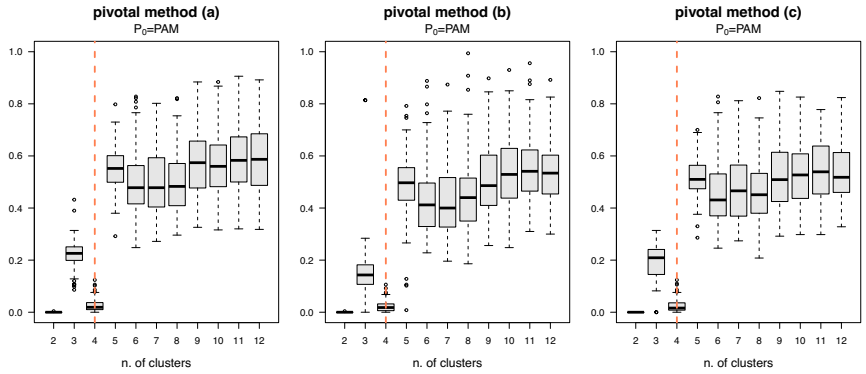
References

1. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Relabelling in Bayesian mixture models by pivotal units. Stat. Comput. **28**, 957–969 (2018)
2. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: pivmet: Pivotal methods for Bayesian relabelling and k-means clustering. R package version 0.3.0.
3. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Consensus clustering via pivotal methods. In G. C. Porzio, F. Greselin, S. Balzano (Eds.), Cladag 2019 - Book of Short Papers, pp 1–4 (2019)
4. Fred, A. L., Jain, A. K.: Combining multiple clusterings using evidence accumulation. IEEE Trans Pattern Anal Mach Intell **27**(6), 835–850 (2005)
5. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Transactions on Knowledge Discovery from Data. **1**, 1–30 (2007)
6. Jain, A.: Data clustering: 50 years beyond K-means. Patt. Recog. Lett. **31**(8), 651–666 (2010)
7. Jain, A. K., Dubes, R. C.: Algorithms for Clustering Data. Prentice Hall (1988)
8. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning, **52**, 91–118 (2003)
9. Strehl, A., Ghosh, J.: Cluster ensembles–a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2002)

(a) $S1$ Dataset ($n = 600, K_{true} = 3, d = 2$)



(b) $S2$ Dataset ($n = 400, K_{true} = 4, d = 2$)



(c) $S3$ Dataset ($n = 250, K_{true} = 5, d = 2$)

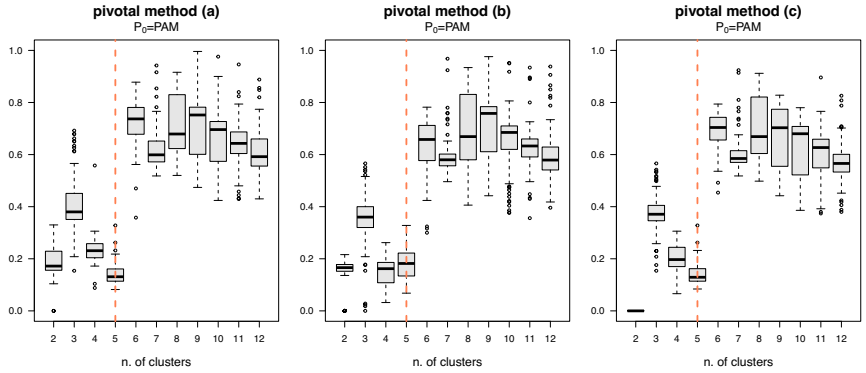


Fig. 3 Boxplots of the index m_k for k in the interval $[2, 12]$, for 100 simulations. The vertical line represents the 'actual' number of clusters. Overall the similarity between the pivots quantified by the index m_k takes on small values when k is equal to K_{true}, and it can be seen that a sharp increase occurs for $K_{true} + 1$. The plots refer to the three scenarios $S1$, $S2$ and $S3$ and pivotal methods (a), (b), (c). Similar results are obtained when the initial partition is derived by Ward's method, and are not reported here.

Clustering of data recorded by Distributed Acoustic Sensors to identify vehicle passage and typology

Antonio Balzanella[1] and Stefania Nacchia[1]

Abstract There is an increasing interest in researchers on the use of modern sensor networks deployed in smart cities to collect data that can be used for a better management of the transportation systems, and, more generally, to improve its impact on the environment. A fairly recent technology, Distributed Acoustic Sensors is being used more and more in the field of transportation system, especially in the field of traffic management. In this paper we propose a methodology for exploiting the data of moving vehicles, captured through these sensors to detect and classify the types of passing vehicles. The methodology is based on a data processing pipeline, whose purpose is to exploit signal processing algorithms to clean the data and to use a clustering algorithm to detect the types of vehicle.
The data used for testing the methodology is collected through a series of experiments with the DAS technology, in a real city environment.

Key words: Distributed Acoustic Sensors, Clustering, spatio-temporal sequences

1 Introduction

Intelligent Transportation Systems (ITS) have been developed over the past decades to improve transportation safety and mobility, to reduce the impact on the environment, to promote sustainable development of transportation and increase productivity.[5]. As the requirements for transportation capability rise annually, roads are becoming saturated, and, as the situation worsen, more and more problems are exposed. Some problems are ancient, like congestion, while others are new like environmental impacts. Among the most notable transport problems are, [2]: *traffic congestion*,*energy*

Università degli Studi della Campania "L. Vanvitelli"
Dipartimento di Matematica e Fisica
viale Lincoln 5, 81100 Caserta, Italy
e-mail: {name.surname}@unicampania.it

consumption, *high maintenance costs*. All these issues are often being addressed by complex algorithms that try to exploit data to find the best solution. In fact ITS is lately strongly based on the ongoing paradigm of smart cities, where its pervasive sensors networks enable the collection of a great amount of static and dynamic data that can be used to improve the ITS decision making and problem solution mechanisms.
ITS now contains a wide spectrum of elements such as smart traffic infrastructure, vehicle connectivity, real-time information service and big data analytics [1]. Nowadays an emerging technology is being largely used as a sensor to capture traffic real-time data: the distributed acoustic sensing (DAS). Recently, DAS are used to explore traffic flow and counting of vehicles [9]. DAS systems use fiber optic cables to provide distributed strain sensing, so that the optical fiber cable becomes the sensing element and measurements are made, and in part processed, using an attached optoelectronic device. Such a system allows acoustic frequency strain signals to be detected over large distances and in harsh environments. Traffic flow detection based on distributed acoustic sensing technology is more sensitive and discreet compared to conventional traffic flow detection, and provides lower costs and higher resistance to temperature, corrosion, and electromagnetic interference.
The main purpose of this work is to propose a pipeline for mining the data collected through these distributed acoustic sensors in order to *count and identify the type of vehicles(e.g. cars, trucks, SUVs etc)* flowing on a road.
In the data collection phase we have setup a real-life experiment where we use a central unit connected to a fiber placed along the road, as shown in the figure 1. It is worth noticing that the environment for the experiment is not controlled but it is an actual road in the city, where traffic flows with no interruption and the collected traffic data reflects the unconstrained experimental setup.
The raw data collected by the central unit is a space-time matrix in which each row represents the observations along a specific section of the fiber and columns record measurements over the time. The dataset used is just a sample of the whole experiment and it comprehends *1081 observations* collected over a *6 minutes and 40 seconds* period of time with a sampling frequency of $78Hz$. The most challenging issue of the used data set is that being the data collected in an unconstrained environment, records include a lot of environmental noise. Moreover, the high dimensionality of the raw data that is acquired, requires the use of appropriate dimensionality reduction techniques to support data transmission on reduced bandwidth networks and almost real-time processing.

Distributed acoustic sensing (DAS) was born mostly as a geophysical method that turns optical fibers into dense seismic recording arrays with virtual receiver points spaced every *1–10m* along the fiber. In fact most research activities focus on the use of DAS as a sensing and monitoring system for seismic activities, [3, 4, 8]. On the other hand, lately some research activities have tried to use the DAS data to monitor traffic flow. Some interesting results have been presented in [6, 9, 7], however all these papers focus mostly on counting the vehicles on the road and compute their speed. As opposed to these papers, the presented strategy aims at not just counting, but evaluating the type of the passing vehicles.

Fig. 1 Fiber positioning. The red spot indicates where the fiber starts.

2 Methodology & Application

In this section we present our strategy for discovering the vehicle type analysing the raw data recorded by the DAS. The strategy is made up of two steps. The first one - the *pre-processing step* - reduces the dimensionality of data and filters the noise coming from the city environment. The second one - the *vehicle type detection step* - is based on a clustering algorithm for providing a partition of the vehicles according to their typology.

To analyse the data recorded by the distributed acoustic sensor, we monitor the optic fiber at n observation points $i = 1,\dots,n$. For each i, we have a series $Y_i = \{y_i^1,\dots,y_i^t,\dots,y_i^T\}$ which records the optic fiber strain due to vehicle passage in the time frame $t = 1,\dots T$.

We organise the DAS data in a matrix $Y_{n\times T} = \{Y_1,\dots,Y_i,\dots,Y_n\}$.

We assume the optical fiber to be placed parallel to the road so that the noise emitted by a moving vehicle is sensed over the time by consecutive sections of the fiber. That is, if the series Y_i records some information about the passage of a vehicle, the series Y_{i+1} also records it with some phase shift depending on the vehicle speed. Of course, as soon as the vehicle leaves the monitored road, no series will record its movement.

The pre-processing step consists in performing a 2-dimensional wavelet decomposition of the matrix Y. Wavelet analysis allows to analyse signals and images at different resolutions to detect change points, discontinuities, and other events not readily visible in raw data. A key advantage it has over other signal processing techniques, e.g. Fourier transform, is temporal resolution: it captures both frequency and location information (location in time).

The 2-dimensional wavelet transform allows to get a multi-level decomposition of the matrix Y such that for each level we have an approximation of the matrix and three sets of coefficients: horizontal, vertical, and diagonal coefficients. It is interesting to note that each set of coefficients highlights specific features of the data matrix.

In our specific application, since the trace of vehicle passage is recorded by phase shifted signals, we are interested in using the information captured by the diagonal coefficients.

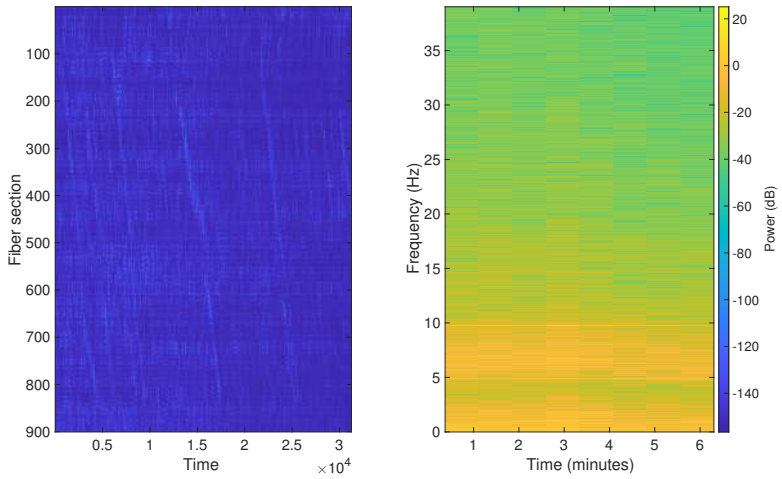


Fig. 2 Diagonal wavelet coefficients and time-frequency spectrum for $i = 400$

Specifically, we choose the Coiflets wavelets to perform a decomposition of the matrix Y into 8 levels. We select only the diagonal coefficients of the levels $5,\ldots,8$ in order to get a matrix $Z_{n\times T}$ which filters the background noise.

As we show in the left side of Fig.2, the passage of vehicles emerge as oblique straight lines in the plot of the matrix Z. Another interesting aspect is the time-frequency spectrum of the signals that allows us to see which is the informative frequency range for our data. In the right side of Fig. 2, we see that the most informative content is related to frequencies lower than $10Hz$.

While the *pre-processing step* allows to highlight the passage of vehicles, the *vehicle type detection step* can provide the vehicle type through an appropriate clustering strategy on the matrix Z.

The approach we propose to address this challenge is based on detecting the peaks (local maxima) in each signal Z_i (for $i = 1,\ldots,n$), assuming that every peak which is higher than a threshold value, corresponds to the passage of a vehicle. Note that the threshold value is required for removing previously unfiltered background noise.

Our idea is to use the time stamp of each peak as the center of a time window so that each signal Z_i is represented by a set of sub-sequences detected through the selection of data around the peak.

Formally, for each Z_i, we detect the peaks $PK_i = \left\{pk^j\right\}_{j=1,\ldots,l_i}$, where pk^j is the time stamp of a peak. By considering a time window having size w, we can recover for each pk^j a time interval $\left[a^j = pk^j - w/2; b^j = pk^j + w/2\right]$ and the corresponding subsequence $s_i^j = \left\{z_i^{a^j},\ldots,z_i^{b^j}\right\}$.

The clustering of sub-sequences allows to allocate vehicles to typologies. We use an algorithm based on BIRCH [10] since it is very effective in analysing huge

amounts of data. A first phase of the algorithm obtains a partition of data into a high number of low variability clusters, performing a single scan. A second phase consists in running a k-means algorithm on the centroids of the clusters discovered by the first phase. The final reduced set of clusters corresponds to vehicle typologies.

The pseudo-code of the first phase of the algorithm is the following:

Initialization:
$i = 1$
Detect the peaks PK_i
Detect the sub-sequences s_i^j for each peak pk^j
Run a k-means algorithm on the s_i^j $(j = 1, \ldots, l_i)$ to get a partition in K clusters C_k and the centroids G_k
Main:
for all Z_i such that $i > 1$ **do**
  Detect the peaks PK_i
  Detect the sub-sequences s_i^j for each peak pk^j
  for all s_i^j **do**
    Allocate s_i^j to the cluster C_k such that:
    $d^2(s_i^j; G_k) < d^2(s_i^j; G'_k)$ if $d^2(s_i^j; G_k) < u$
  end for
  Update the cluster centroids G_k $(k = 1, \ldots, K)$
end for

We have tested this algorithm on our dataset. In Fig. 3 we show the strain of the optic fiber at the observation point $i = 400$. The peaks represent the passage of vehicles. The coloured dots show the membership of each vehicle to a typology as result of a clustering of data into 4 clusters.

3 Conclusions

In this short paper we have shown a strategy for detecting the passage of vehicles and for clustering such vehicles into typologies starting from DAS data. We have run our algorithms on real data to evaluate the effectiveness of the strategy. By means of a camera, we have recorded the vehicle passage on the road to compare our results with the ground truth. Despite the data being affected by noise, we were able to obtain encouraging results. Further validations and tests to validate the input parameters of the procedure will be the subject of future works.

References

1. Barth, M., Todd, M., Shaheen, S.: Intelligent transportation technology elements and operational methodologies for shared-use vehicle systems. Transportation research record **1841**(1), 99–108 (2003)

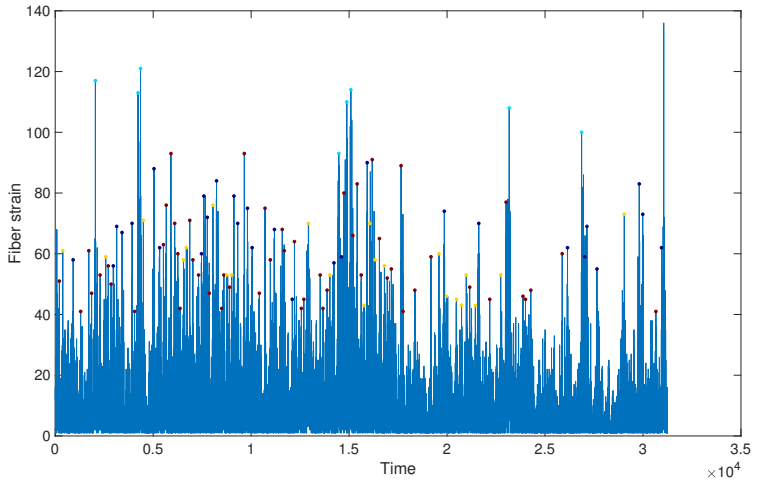


Fig. 3 Plot of the optic fiber strain for $i = 400$. The coloured dots represent the vehicle type provided as output by the clustering procedure

2. Comtois, C., Slack, B.: The geography of transport systems. Routledge (2009)
3. Daley, T.M., Freifeld, B.M., Ajo-Franklin, J., Dou, S., Pevzner, R., Shulakova, V., Kashikar, S., Miller, D.E., Goetz, J., Henninges, J., et al.: Field testing of fiber-optic distributed acoustic sensing (das) for subsurface seismic monitoring. The Leading Edge **32**(6), 699–706 (2013)
4. Dou, S., Lindsey, N., Wagner, A.M., Daley, T.M., Freifeld, B., Robertson, M., Peterson, J., Ulrich, C., Martin, E.R., Ajo-Franklin, J.B.: Distributed acoustic sensing for seismic monitoring of the near surface: A traffic-noise interferometry case study. Scientific reports **7**(1), 1–12 (2017)
5. Lin, Y., Wang, P., Ma, M.: Intelligent transportation system (its): Concept, challenge and opportunity. In: 2017 ieee 3rd intl conference on big data security on cloud (bigdatasecurity), ieee intl conference on hpsc, and ieee intl conference on ids, pp. 167–172. IEEE (2017)
6. Lindsey, N.J., Yuan, S., Lellouch, A., Gualtieri, L., Lecocq, T., Biondi, B.: City-scale dark fiber das measurements of infrastructure use during the covid-19 pandemic. Geophysical research letters **47**(16), e2020GL089,931 (2020)
7. Liu, H., Ma, J., Yan, W., Liu, W., Zhang, X., Li, C.: Traffic flow detection using distributed fiber optic acoustic sensing. IEEE Access **6**, 68,968–68,980 (2018)
8. Martins, H.F., Fernández-Ruiz, M.R., Costa, L., Williams, E., Zhan, Z., Martin-Lopez, S., Gonzalez-Herraez, M.: Monitoring of remote seismic events in metropolitan area fibers using distributed acoustic sensing (das) and spatiotemporal signal processing. In: Optical Fiber Communication Conference, pp. M2J–1. Optical Society of America (2019)
9. Parker, T., Shatalin, S., Farhadiroushan, M.: Distributed acoustic sensing–a new tool for seismic applications. first break **32**(2) (2014)
10. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: An efficient data clustering method for very large databases. SIGMOD Rec. **25**(2), 103–114 (1996). DOI 10.1145/235968.233324. URL https://doi.org/10.1145/235968.233324

2.11 New developments in latent variable models

A Hidden Markov Model for Variable Selection with Missing Values

Un Modello Hidden Markov per la Selezione delle Variabili con Valori Mancanti

Fulvia Pennoni, Francesco Bartolucci, and Silvia Pandolfi

Abstract We propose a hidden Markov model for longitudinal multivariate continuous responses, accounting for missing data under the missing at random assumption. Maximum likelihood estimation of this model is carried out through the Expectation-Maximization algorithm. To address the problem of dimensionality reduction, we develop a greedy search algorithm based on the Bayesian Information Criterion. We illustrate the proposal through a dataset collected by the World Bank and UNESCO Institute for Statistics on the basis of which we dynamically cluster countries according to the selected variables observed during the period 2000-2017.
Abstract *Viene proposto un modello hidden Markov per risposte continue multivariate longitudinali e possibili dati mancanti sotto l'assunzione* missing at random*. Il metodo della massima verosimiglianza è utilizzato per la stima dei parametri attraverso l'algoritmo Expectation-Maximization. Si implementa anche un algoritmo per la selezione delle variabili e del modello basato sul Bayesian Information Criterion. La proposta è illustrata tramite dati raccolti dalla Banca Mondiale e dall'Istituto di Statistica dell'UNESCO nel periodo 2000-2017, sulla base dei quali i paesi vengono classificati in modo dinamico considerando le variabili selezionate.*

Key words: development changes, Gaussian distribution, longitudinal data, missing at random assumption

Fulvia Pennoni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca
e-mail: fulvia.pennoni@unimib.it

Francesco Bartolucci, Silvia Pandolfi
Department of Economics, University of Perugia
e-mail: francesco.bartolucci@unipg.it, e-mail: silvia.pandolfi@unipg.it

1 Introduction

We consider hidden (or latent) Markov models (HMMs) for the analysis of time-series and panel data [1, 2]. The main assumption is that the observed data depend on a latent process that follows a first-order Markov chain that may be time homogeneous or heterogeneous. In this way, we can account for the unobserved heterogeneity in a time-varying fashion, and we are able to cluster the units in the panel into homogeneous groups corresponding to comparable unobservable characteristics. Given each latent state, the continuous responses at the same time occasion are assumed to follow a multivariate Gaussian distribution with specific mean vector and variance-covariance matrix. We focus on the problem of non-monotone missing data patterns considering partially or totally missing responses at one or more time occasions, under the *missing at random* (MAR) assumption [3]. Following the idea proposed in [4], we implement a greedy forward-backward procedure based on an approximation of the Bayes factor so as to select the subset of the most useful responses for clustering and simultaneously choose the optimal number of latent states. A modified Expectation-Maximization (EM) algorithm [5] is employed to obtain maximum likelihood estimates of the model parameters.

To illustrate the proposal we consider data derived from the World Bank and UNESCO Institute for Statistics to study countries' economic conditions over the period 2000-2017. We use several variables, including GDP per capita, educational levels, life expectancy at birth, and others related to the human development index proposed by the United Nations Development Programme[1] for measuring the well-being at the country level. The proposed approach allows us to characterize disparities among countries in a dynamic fashion and to evaluate development changes.

In the following section we show the proposed HMM accounting for missing data. In Section 3, we outline the main features of the greedy search algorithm for variable and model selection, and in Section 4, we describe the application.

2 Model Formulation and Estimation

Let $\boldsymbol{Y}_{it} = (Y_{i1t}, \ldots, Y_{irt})'$ denote the vector of r continuous response variables measured at time t, $t = 1, \ldots, T_i$, where T_i denotes the number of occasions of observation for unit i, $i = 1, \ldots, n$. Also, let $\boldsymbol{Y}_i$ be the vector obtained by stacking $\boldsymbol{Y}_{it}$ for $t = 1, \ldots, T_i$. The latent process denoted as $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iT_i})'$ is assumed to follow a first-order Markov chain with state-space ranging from 1 to k. This process is characterized by the initial probabilities

$$\pi_u = p(U_{i1} = u), \quad u = 1, \ldots, k,$$

and the transition probabilities

[1] Data are available at https://datacatalog.worldbank.org/dataset/world-development-indicators

$$\pi_{u|\bar{u}} = p(U_{it} = u | U_{i,t-1} = \bar{u}), \quad t = 2, \ldots, T_i, \quad u, \bar{u} = 1, \ldots, k,$$

where u denotes a realization of U_{it} and $\bar{u}$ a realization of $U_{i,t-1}$. Under the *local independence assumption*, the response vectors $\boldsymbol{Y}_{it}$ collected in $\boldsymbol{Y}_i$ are conditionally independent given the latent process $\boldsymbol{U}_i$. A conditional multivariate Gaussian distribution is assumed for the responses:

$$\boldsymbol{Y}_{it} | U_{it} = u \sim N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u),$$

where $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}_u$ are latent state specific mean vectors and variance-covariance matrices. These matrices are constrained to be equal each other when homoscedasticity is assumed, as is usually done in the HMM and finite mixture context [6].

In presence of partially incomplete data, the response variables may be partitioned as $(\boldsymbol{Y}_{it}^o, \boldsymbol{Y}_{it}^m)'$, where $\boldsymbol{Y}_{it}^o$ corresponds to the observed variables and $\boldsymbol{Y}_{it}^m$ corresponds to the missing ones. Accordingly, the conditional mean vectors and variance-covariance matrices may be decomposed as follows

$$\boldsymbol{\mu}_u = \begin{pmatrix} \boldsymbol{\mu}_u^o \\ \boldsymbol{\mu}_u^m \end{pmatrix}, \quad \boldsymbol{\Sigma}_u = \begin{pmatrix} \boldsymbol{\Sigma}_u^{oo} & \boldsymbol{\Sigma}_u^{om} \\ \boldsymbol{\Sigma}_u^{mo} & \boldsymbol{\Sigma}_u^{mm} \end{pmatrix},$$

where the single blocks are identified by letters o and m when referred to observed and missing components, respectively.

Likelihood based inference with missing data is performed under the MAR assumption and independence between sample units. The log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_{it}^o) = \sum_{i=1}^{n} \log \sum_{\boldsymbol{u}_i} \left(\prod_{t=1}^{T_i} f(\boldsymbol{y}_{it}^o | u_{it}) \right) \left(\pi_{u_{i1}} \prod_{t=2}^{T_i} \pi_{u_{it}|u_{i,t-1}} \right),$$

where $\boldsymbol{\theta}$ is the vector of all the model parameters, $f(\boldsymbol{y}_{it}^o)$ is the manifest distribution of the observed responses $\boldsymbol{y}_{it}^o$, and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iT_i})'$.

The EM algorithm maximizes the above likelihood by alternating two steps until convergence. In particular, at the *E-step* we compute the posterior expected value of the *complete data log-likelihood*, $\ell^*(\boldsymbol{\theta})$, given the observed data and the current value of the parameters. With missing data, this step includes the computation of $\mathrm{E}(\boldsymbol{Y}_{it} \mid \boldsymbol{y}_{it}^o, u)$ and $\mathrm{E}[(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_u)(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_u)' \mid \boldsymbol{y}_{it}^o, u]$. At the *M-step* we update the estimate of $\boldsymbol{\theta}$ by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$ obtained at the *E-step*. We combine deterministic and random initializations of the EM algorithm to limit the problem of multimodality of the log-likelihood function.

3 Variable and Model Selection

We implement an algorithm to perform variable and model selection in line with the proposal in [4], which is based on assessing the importance of each variable, among those available, by comparing two suitably chosen models. In the first of these models, the candidate variable is assumed to provide additional information

about clustering allocation beyond that contained in the already selected variables; in the second model, this variable is not used for clustering. The two models are compared through the Bayesian Information Criterion (BIC) [7], which is related to the Bayes factor and is based on the following index

$$BIC_k = -2\hat{\ell}_k + \log(n)\#par,$$

where $\hat{\ell}_k$ denotes the maximum of the log-likelihood of the HMM with k states and $\#par$ denotes the number of free parameters.

We propose a greedy forward-backward procedure that starts with an initial set of clustering variables, denoted by $\mathscr{Y}^{(0)}$, and a number of latent states, denoted by $k^{(0)}$. At the h-th iteration, the algorithm performs the following three steps:

- *Inclusion step*: each variable j in the remaining set of variables, is singly proposed for inclusion in $\mathscr{Y}^{(h)}$. The variable to be included is selected on the basis of the following difference between *BIC* indexes:

$$BIC_{diff} = BIC_{k^{(h-1)}}(\mathscr{Y}^{(h-1)} \cup j) - \left[BIC_{k^{(h-1)}}(\mathscr{Y}^{(h-1)}) + BIC_{reg}(j \sim \mathscr{Y}^{(h-1)})\right],$$

  where BIC_k is the index computed under the proposed HMM with k states, and BIC_{reg} is the index related to the multivariate linear regression of the candidate variable on the currently selected set of variables. The variable with the smallest negative BIC_{diff} is included in $\mathscr{Y}^{(h-1)}$, and this set is updated.
- *Exclusion step*: each variable j in $\mathscr{Y}^{(h)}$ is singly proposed for the exclusion on the basis of the following index:

$$BIC_{diff} = BIC_{k^{(h)}}(\mathscr{Y}^{(h)}) - \left[BIC_{k^{(h)}}(\mathscr{Y}^{(h)} \setminus j) + BIC_{reg}(j \sim \mathscr{Y}^{(h)} \setminus j)\right].$$

  The variable with the highest positive value of the BIC_{diff} is removed from $\mathscr{Y}^{(h)}$.
- *Model selection*: the current value of $k^{(h-1)}$ is updated by minimizing the BIC_k index of the HMM for the current set of clustering variables $\mathscr{Y}^{(h)}$ over k, from $(k^{(h-1)} - 1)$ to $(k^{(h-1)} + 1)$, so as to obtain the new value of $k^{(h)}$.

The algorithm ends when no variable is added to or is removed from $\mathscr{Y}^{(h)}$. It is worth mentioning that the proposed approach may be influenced by the choice of the initial set of responses, therefore some preliminary or sensitivity analyses at this aim are needed.

Once the variables and the number of states have been selected, the EM algorithm directly provides the estimated posterior probabilities of U_{it} used to obtain a prediction of the latent states of each unit i at every time occasion t. The code implemented to perform the estimation and the selection of the proposed HMM is developed by extending the functions included in the `R` package `LMest` [8].

4 Application

Data referred to $n = 217$ countries followed for $T = 18$ years over a set of $r = 25$ responses with missing values are used to illustrate the proposal, which is based on a model assuming a constant variance-covariance matrix across latent states. The greedy search algorithm is applied starting from a model with only one response variable and $k = 6$ latent states chosen on the basis of a preliminary analysis. In the end, this algorithm leads us to choose a model including $r = 15$ responses with $k = 9$ latent states and heterogeneous transition probabilities.

The selected responses are reported in Table 1 along with the estimated cluster conditional means. The latent states are ordered according to increasing values of the estimated means of the variables highlighted in bold and are able to discriminate between countries with different income levels. The estimated parameters of the latent model are reported in Table 2. We notice that the first group of countries (about 11% in 2000) is characterized mainly by low values of GDP, current health expenditure, and school enrollment in tertiary education. However, we estimate that in 2017 the 43% of countries moves to the 3rd cluster referred to countries having especially a higher coverage of social safety net programs in the poorest quintile. The 5th group of countries (about 13% in 2000) shows intermediate levels of development with a remarkable high rate of primary school enrollment. For these countries, we observe a probability of around 0.03 of moving towards the 6th state in 2017.

Table 1 Estimated conditional means of the HMM with $k = 9$ latent states (in bold variables with increasing means across states).

	1	2	3	4	5	6	7	8	9
Ele	13.25	14.45	34.64	54.54	77.15	96.84	99.64	100.00	99.99
GDP	1457.75	1717.51	3228.88	5689.88	6456.77	9816.22	25361.15	41879.28	79947.84
Hea	68.78	106.66	152.39	242.18	313.16	545.37	1493.95	3801.95	2673.04
Lex	52.72	57.01	59.59	60.55	67.64	72.12	76.00	80.58	78.92
Sav	10.72	8.01	19.76	21.84	21.44	21.81	20.51	24.74	42.97
Imp	34.29	51.55	44.93	39.58	51.10	46.64	56.47	38.80	96.42
Sch3	2.89	4.09	8.03	9.45	20.59	33.01	56.35	70.06	32.32
Rese	0.14	0.13	0.36	0.35	0.30	0.39	0.70	2.43	0.61
Trade	58.95	78.29	79.55	72.77	87.08	83.32	110.38	81.14	217.73
Edu	2.91	4.81	3.84	4.77	4.57	4.52	4.88	5.72	3.00
Sch1	71.69	117.63	98.24	95.58	107.49	105.20	101.75	102.28	102.14
Int	1.11	3.49	6.18	9.39	11.69	22.39	52.43	73.82	61.19
Sch2	19.16	31.39	44.12	44.93	73.25	83.69	97.30	112.36	94.59
Safe	7.72	19.99	22.50	20.25	58.25	51.51	68.94	24.08	29.87
Lit	36.21	65.86	59.02	63.89	82.11	91.60	95.19	83.60	96.64

Note: Ele: access to electricity; GDP: gross domestic product per capita; Hea: current health expenditure; Lex: life expectancy at birth; Sav: gross savings; Imp: import of goods, and services; Sch3: school enrollment, tertiary; Rese: research and development expenditure; Trade: exports and imports of goods, and services; Edu: government expenditure on education; Sch1: school enrollment, primary; Int: individuals using the Internet; Sch2: school enrollment, secondary; Safe: coverage of social safety net programs in poorest quintile; Lit: literacy rate.

Table 2 Estimated averaged initial and transition probabilities for the HMM with $k = 9$ states referred to the period 2016-2017.

	1	2	3	4	5	6	7	8	9
$\hat{\pi}_u$	0.11	0.06	0.06	0.05	0.13	0.34	0.11	0.08	0.06
$\hat{\pi}_{u\|1}$	0.57	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\pi}_{u\|2}$	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\pi}_{u\|3}$	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\pi}_{u\|4}$	0.00	0.00	0.00	0.94	0.06	0.00	0.00	0.00	0.00
$\hat{\pi}_{u\|5}$	0.00	0.00	0.00	0.00	0.97	0.03	0.00	0.00	0.00
$\hat{\pi}_{u\|6}$	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
$\hat{\pi}_{u\|7}$	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.02	0.00
$\hat{\pi}_{u\|8}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.03
$\hat{\pi}_{u\|9}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

The 6th group differs from the 5th mainly for higher values of GDP, electricity access, and health expenditure. Most countries (about 34%) are allocated to this cluster in 2000 with a persistence probability of around 1.00. The 8th cluster is that of high-income countries (about 8% in 2000), and we estimate a probability of 0.03 of moving towards the 9th cluster in 2017 that is characterized by the highest average values of GDP, trade, import of goods and services, and literacy rate.

Using *local decoding*, we identify development changes of each country over time. For example, the following countries are allocated in the 1st cluster in 2000: Afghanistan, Angola, Benin, Burkina Faso, Burundi, Central African Republic, Chad, Congo, Democratic Republic of Congo, Ethiopia, Guinea, Guinea-Bissau, Mali, Mauritania, Mozambique, Niger, Papua New Guinea, Sierra Leone, Solomon Islands, Somalia, South Sudan, Tanzania, Zambia. We estimate that only Chad and Niger remain in this cluster at the end of 2017, revealing that their economic and social conditions have not changed over time.

Acknowledgements We thank A. Serafini for the support in the preliminary data analysis.

References

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman & Hall/CRC press (2013)
2. Zucchini, W., MacDonald, I.L., Langrock, R.: *Hidden Markov Models for Time Series: An Introduction using* `R`. Boca Raton, FL: CRC press (2017)
3. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, 3rd Ed. Wiley Series in Probability and Statistics, Hoboken, NJ: John Wiley Sons (2019)
4. Raftery, A.E., Dean, N.: Variable selection for model-based clustering. *J. Am. Stat. Assoc.*, **101**, 168–178 (2006)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–38 (1977)
6. McLachlan G, Peel D. *Finite Mixture Models*. New York: Wiley (2000)
7. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464 (1978)
8. Bartolucci, F., Pandolfi, S., Pennoni, F.: `LMest`: An `R` package for latent Markov models for longitudinal categorical data. *J. Stat. Softw.*, **81**, 1–38 (2017)

Comparison between Different Likelihood Based Estimation Methods in Latent Variable Models for Categorical Data

Un Confronto tra Metodi di Stima Basati sulla Verosimiglianza nei Modelli a Variabili Latenti per Dati Categorici

Silvia Bianconcini and Silvia Cagnone

Abstract Latent variable models represent a useful tool in different fields of research in which the constructs of interest are not directly observable. In presence of many latent variables and/or random effects, problems related to the integration of the likelihood function can arise since analytical solutions do not exist. In literature, different remedies have been proposed to overcome these problems. Among these, the pairwise likelihood method and, more recently, the dimension-wise quadrature have been shown to produce estimators with desirable properties. We compare the performance of the two methods for a class of dynamic latent variable models for count data.

Abstract *I modelli a variabili latenti rappresentano uno strumento di analisi molto utile in ambiti di applicazione nei quali i costrutti di interesse non sono direttamente osservabili. In particolare, in presenza di molte variabili latenti e/o effetti casuali e di dati categorici, possono sorgere problemi relativi al calcolo di integrali presenti nella funzione di verosimiglianza poich non esistono soluzioni analitiche. Per superare questi problemi, in letteratura sono state proposte diverse soluzioni. Tra queste, i metodi basati su verosimiglianze composite e, pi recentemente, i metodi basati sulla riduzione dimensionale degli integrali producono stimatori accurati. In questo lavoro confrontiamo la performance dei due metodi per una classe di modelli a variabili latenti dinamici per dati di conteggio.*

Key words: latent autoregressive models, count data, pairwise likelihood, dimension-wise quadrature.

Silvia Bianconcini
Department of Statistical Sciences, University of Bologna, e-mail: silvia.bianconcini@unibo.it

Silvia Cagnone
Department of Statistical Sciences, University of Bologna e-mail: silvia.cagnone@unibo.it

1 Introduction

Latent variable models represent a useful tool in different fields of research in which the constructs of interest are not directly observable. However, in presence of many latent variables and/or random effects, problems related to the integration of the likelihood function can arise since analytical solutions do not exist. Common examples are represented by latent variable models for panel data or time series that involve many continuous time-varying latent variables whose dynamics is typically modeled by an autoregressive process of order 1 ([5],[6]). In these models, inference is cumbersome because the likelihood function depends on an intractable high-dimensional integral.

Alternative methods that produce estimators with desired statistical properties and that, in addition, simplify the estimation process, are greatly needed. The most popular method that offers reduction in estimation complexity is the composite likelihood approach, introduced by [7] and further discussed, among the others, by [9]. The composite likelihood estimator is obtained by maximizing the univariate and/or bivariate likelihood products that contain the greatest quantity of model parameter information. The immediate effect of the composite likelihood estimation is the reduction of the number of integrations required in the likelihood computation. Another approach that has been recently proposed in the literature is the dimension-wise quadrature (DWM), developed by [2]. It consists in reducing the dimension of the multidimensional integrals by truncating the Taylor series expansion of the integrand. This makes the computation feasible also when the number of latent variables is large. The proposed approach provides a higher order approximation than the Laplace one but does not require any derivative computation, hence it is very simple to implement. Furthermore, the corresponding estimators are asymptotically as accurate as the adaptive Gauss Hermite estimators.

A first comparison between the pairwise likelihood approach and DWM has been recently proposed by [1] for latent variable models for multivariate longitudinal ordinal data. The results highlighted that DWM outperforms the pairwise likelihood approach when the dimension reduction involves up to two integrals. Moreover, DWM appears to be more advantageous since, unlike the pairwise likelihood approach, it is always feasible, even in presence of very complex models. However, the pairwise estimator considered in this study involved all the possible log pairwise components and it can result less efficient than weighted pairwise likelihood estimators that typically involve only the most informative pairs of observations.

In this paper, we compare the two methods for a class of dynamic latent variable models for count data considered by [8] to analyze infectious disease data. The authors proposed a pairwise likelihood estimator that involves only a selected number of pairs. Moreover, the pairwise likelihood components are properly weighted. A preliminary simulation study is done to compare the performance of the estimators under the two methods.

2 Latent Autoregressive Models for Count Data

Let $y_1, \ldots, y_T$ denote an observed time series of count data of length T and $\alpha_1, \ldots, \alpha_T$ represent a vector of time-dependent latent variables. The relation between observed and latent variables can be expressed through a latent autoregressive model as follows

$$\mu_t = E(y_t|\alpha_t) = \exp(\mathbf{x}_t'\beta + \alpha_t) \quad (1)$$

$$\alpha_t = \gamma + \phi\alpha_{t-1} + \varepsilon_t \quad (2)$$

Equation (1) specifies the conditional distribution of observed variables given the latent ones. For count data, it is a Poisson distribution of parameter μ_t depending on a set of time dependent covariates $\mathbf{x}_t$ and on the latent variable α_t that induces serial correlation and overdispersion. The dynamics of α_t is modelled through a stationary autoregressive process of order 1 specified in equation (2), with $|\phi| < 1$ and $\varepsilon_t \sim N(0, \sigma^2)$.

3 Model Estimation

Model estimation is usually performed by using a full maximum likelihood method. The likelihood is given by:

$$L(\theta) = \int_{R^T} \prod_{t=1}^{T} f_y(y_t|\alpha_t) f_\alpha(\alpha_t|\alpha_{t-1}) d\alpha_T \ldots d\alpha_1, \quad (3)$$

where $\theta = (\gamma, \beta, \phi, \sigma^2)$ is the vector of parameters to be estimated and $f_\alpha(\alpha_{i1}|\alpha_{i0}) = f_\alpha(\alpha_{i1})$. A problem related to the maximization of the likelihood is that, in general, the multidimensional integral in (3) is not solvable analytically.
Among the remedies proposed in the literature, numerical quadrature-based methods represent a widespread solution to this problem and, among them, the adaptive Gauss Hermite quadrature has been found to be very accurate for latent autoregressive models for ordinal data when a non-linear filter technique is applied ([3]). However, the adaptive quadrature is computationally demanding if it is based on the computation of the mode of the integrand in equation (3). Alternative solutions can be the pairwise likelihood approach recently proposed by [8] for the latent autoregressive model considered in this work and the DWM proposed by [2].

3.1 Pairwise Likelihood Approach

The pairwise likelihood estimator is obtained by maximizing bivariate likelihood products that contain the greatest quantity of model parameter information ([7],

[4]). The immediate effect of the pairwise likelihood estimation is the reduction of the number of integrations in the expression of the likelihood (3).

[8] proposed a pairwise log-likelihood of order d, defined as the sum the log bivariate densities for all the pairs of observations that are separated at most by d units as follows

$$l(\theta) = \sum_{t=d+1}^{T} \sum_{i=1}^{d} wg_i \log f(y_{t-i}, y_t, \theta) \tag{4}$$

where

$$f(y_{t-i}, y_t, \theta) = \int_{R^2} f(y_t|\alpha_t) f(y_{t-i}|\alpha_{t-1}) f_\alpha(\alpha_{t-1}, \alpha_t) d\alpha_{t-1} d\alpha_t \tag{5}$$

and wg_i are rectangular or trapezoidal non-negative weights. The number of pairs involved in expression (4) are $(T-d)d$ implying a significant reduction of the computational effort. In the case of $d = 1$, only consecutive pairs are considered, whereas, as d increases, the pairwise likelihood involves an increasing number of pairs of independent observations.

3.2 Dimension-wise Quadrature Method

Consider the following representation of the marginal density function

$$\begin{aligned} f(y_1, \ldots, y_T; \theta) &= \int_{R^T} \frac{[\prod_{t=1}^{T} f(y_t|\alpha)] f_\alpha(\alpha)}{\phi(\alpha; \alpha_{mo}, \Sigma_{mo})} \phi(\alpha; \alpha_{mo}, \Sigma_{mo}) d\alpha = \\ &= |C_{mo}| \int_{R^T} \frac{\prod_{t=1}^{T} f(y_t|C_{mo}\alpha^* + \alpha_{mo}) f_\alpha(C_{mo}\alpha^* + \alpha_{mo})}{\phi(\alpha^*; 0, I)} \\ &\quad \phi(\alpha^*; 0, I) d\alpha^* = \\ &= |C_{mo}| \int_{R^T} m(\alpha^*) \phi(\alpha^*; 0, I) d\alpha^* = \\ &= |C_{mo}| E_\phi[m(\alpha^*)] \end{aligned} \tag{6}$$

where $f_\alpha(\alpha) = f_\alpha(\alpha_1, \ldots, \alpha_{t-1}, \alpha_t)$, α_{mo} is the maximum of the logarithm of the integrand $\prod_{t=1}^{T} f(y_t|\alpha) f_\alpha(\alpha)$, and $\Sigma_{mo} = C_{mo} C'_{mo}$ is minus the inverse of the corresponding Hessian matrix evaluated in the mode α_{mo}. $\phi(\cdot)$ is the normal density function.

The dimension-wise method is applied to the expected value $E_\phi[m(\alpha^*)]$. It is based on the Taylor expansion of $m(\alpha^*)$ around 0 up to the s term as follows

$$\hat{m}(\alpha^*) = \sum_{w=1}^{s} t_w \tag{7}$$

where each component t_w considers all the derivatives of $m(\alpha^*)$ taken with respect to w latent factors, that is

$$t_w = \sum_{j_1,j_2,\ldots,j_w} \sum_{k_1<k_2<\ldots<k_w} \frac{1}{j_1!j_2!\ldots j_w!} \frac{\partial^{j_1+j_2,\ldots,+j_w} m(0)}{\partial \alpha_{k_1}^{*j_1} \partial \alpha_{k_2}^{*j_2} \ldots \partial \alpha_{k_w}^{*j_w}} \alpha_{k_1}^{*j_1} \alpha_{k_2}^{*j_2} \ldots \alpha_{k_w}^{*j_w} \quad (8)$$

The approximated function $\hat{m}(\alpha^*)$ admits the following equivalent representation ([2])

$$\hat{m}(\alpha^*) = \sum_{l=0}^{s} (-1)^l \binom{T-s+l-1}{l} m_{s-l}(\alpha^*) \quad (9)$$

where $m_{s-l}(\alpha^*) = m(0,\cdots,\alpha_{k_1}^*,0\cdots,0,\alpha_{k_{s-l}}^*,0,\cdots,0)$. Thus, m_{s-l} is a function of just $s-l$ variables being all the remaining fixed to 0. Replacing eq. (9) in eq. (6), we obtain the approximate density function

$$f_a(\mathbf{y};\theta) = |C_{mo}| \left[\sum_{l=0}^{s-1} (-1)^l \binom{T-s+l-1}{l} \int_{R^{s-l}} \sum_{k_1<\ldots<k_{s-l}} m_{s-l}(\alpha^*) \times \right. \quad (10)$$
$$\left. \phi(\alpha_{k_1}^*) \cdots \phi(\alpha_{k_{s-i}}^*) d\alpha_{k_1}^* .. d\alpha_{k_{s-l}}^* \right].$$

The dimension of the integrals in expression (10) depends on the choice of s. If $s=1$, we obtain a linear combination of unidimensional integrals, if $s=2$, we obtain a linear combination of uni- and bi-dimensional integrals and so on. For small values of s, the integrals can be easily approximated using the Gauss Hermite quadrature method. In the extreme cases of $s=0$ and $s=q$ the solution is equivalent to the classical Laplace approximation and to the adaptive Gauss-Hermite quadrature, respectively. The dimension-wise quadrature estimators share the same accuracy as the adaptive Gauss-Hermite method, but avoiding the main computational limitations of the latter [2].

4 Simulation Study: Preliminary Results

The performance of the two approximation methods are compared through a simulation study.

We generated 500 time series of length $T=90$, with true values of the parameters $\phi=0.5$, $\sigma^2=1.528$ and no covariates. For the pairwise likelihood method we consider rectangular weights, $d=1$ and $d=10$. The pairwise of order 1 was shown to have a better performance than higher orders by [8], whereas the pairwise of order 10 is expected to be less efficient. For DWM, we consider $s=1$ and $s=2$ since [1] showed that DWM with $s=1$ produces similar results to the unweighted pairwise likelihood estimation method, whereas DWM with $s=2$ outperforms it.

Table 1 reports the bias and the root mean square error (rmse) of the parameter estimates. We can observe that, in all the conditions and for both the parameters, DWM produces less biased estimates than the pairwise method. The differences are more relevant for σ^2. On the other hand, the pairwise method appears more efficient than DWM. Moreover, for this particular design, there are no relevant differences

between $s = 1$ and $s = 2$ and between $d = 1$ and $d = 10$ apart from the bias of ϕ that for $d = 10$ is more than doubles than for $d = 1$. This is consistent with the fact that, as the order d diverges, the pairwise likelihood involves an increasing number of pairs that do not contain any information about the parameter ϕ that cannot be consistently estimated [8]. We have also applied the pairwise method using trapezoidal weights, but there were no difference in the performance respect to the use of rectangular weights. In Table 1 the average computational time in seconds taken by each method is also reported. DWM appears slower than pairwise since it requires the computation of the mode of the integrand.

This is a preliminary simulation study that gives a first insight on the performance of the two methods. Further work needs to be done to corroborate the findings of this study. More scenarios have to be considered and the properties of the dimension-wise based and pairwise estimators should be investigated and compared theoretically.

Table 1 Estimated mean, bias and rmse for $T = 90$, 500 replications

	Pairwise		Pairwise		DWM		DWM	
	$d = 1$		$d = 10$		$s = 1$		$s = 2$	
True	bias	rmse	bias	rmse	bias	rmse	bias	rmse
$\phi = 0.500$	-0.047	0.184	-0.106	0.202	-0.001	0.380	0.003	0.383
$\sigma^2 = 1.528$	-0.499	0.585	-0.379	0.528	0.026	0.424	-0.030	0.455
Time (sec)	1.04		3.00		6.42		715.40	

References

1. Bianconcini S., Cagnone S.: Comparison between different estimation methods of factor models for longitudinal ordinal data. In: Quantitative Psychology: The 85th Annual (Virtual) Meeting of the Psychometric Society. Springer. Forthcoming
2. Bianconcini S., Cagnone S., Rizopoulos D.: Approximate likelihood inference in generalized linear latent variable models based on the dimension-wise quadrature. Electron. J. Stat. 11, 4404–4423 (2017)
3. Cagnone S, Bartolucci F.: Adaptive quadrature for maximum likelihood estimation of a class of dynamic latent variable models. Computat Econ **49**, 599-622 (2017)
4. Cox D.R., Reid N.: A note on pseudolikelihood constructed from marginal densities. Biometrika **91**, 729–737 (2004)
5. Cox D.R.: Statistical analysis of time series: some recent developments. Scand J Stat, **8**, 93-115 (1981)
6. Frees, E.: Longitudinal and panel data: Analysis and application in social sciences. Cambridge: Cambridge University Press (2004).
7. Lindsay B.: Composite likelihood methods. In: N.U. Prabhu (eds.) Statistical inference from stochastic processes. Providence: Am. Math. Soc., 221–239 (1988)
8. Pedeli X., Varin C.: Pairwise likelihood estimation of latent autoregressive count models. Stat Methods Med Res, 3278-3293 (2020)
9. Varin C., Vidoni P.: A note on composite likelihood inference and model selection. Biometrika.**92**, 519–528 (2005)

A Comparison of Estimation Methods for the Rasch Model

Alexander Robitzsch[1,2]

Abstract The Rasch model is one of the most prominent item response models. In this article, different item parameter estimation methods for the Rasch model are compared through a simulation study. The type of ability distribution, the number of items, and sample sizes were varied. It is shown that variants of joint maximum likelihood estimation and conditional likelihood estimation are competitive to marginal maximum likelihood estimation. However, efficiency losses of limited-information estimation methods are only modest. It can be concluded that in empirical studies using the Rasch model, the impact of the choice of an estimation method with respect to item parameters is almost negligible for most estimation methods. Interestingly, this sheds a somewhat more positive light on old-fashioned joint maximum likelihood and limited information estimation methods.

Key words: Rasch model, item parameter estimation, maximum likelihood estimation, item response model

1 Rasch Model

The Rasch model [9, 18] is likely the most important item response model. It is of interest to select appropriate estimation methods in diverse applications. A variety of estimation methods has been proposed. In this article, a comprehensive comparison of different estimation methods for the Rasch model is conducted. We manipulate the factors test length (i.e., number of items), sample size, and type of ability of distribution.

Alexander Robitzsch
[1] IPN – Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, D-24118 Kiel, Germany, e-mail: robitzsch@leibniz-ipn.de
[2] Centre for International Student Assessment (ZIB), Kiel, Germany

For a number of items X_i ($i = 1, \ldots, I$) and a random variable θ (ability), the item response function for the Rasch model is given as

$$\mathrm{P}(X_i = 1|\theta; b_i) = \Psi(\theta - b_i) \quad , \quad \theta \sim F \tag{1}$$

where Ψ is the logistic link function, b_i is the item difficulty, and F is some distribution for ability θ. In addition, items X_i are assumed to be locally independent, that is $\mathrm{P}(X_1, \ldots, X_I|\theta) = \prod_{i=1}^{I} \mathrm{P}(X_i|\theta)$. Importantly, the sum score $S = \sum_{i=1}^{I} X_i$ is a sufficient statistic for θ if maximum likelihood (ML) estimation is employed. Hence, all items are equally weighted in θ, which eases the interpretation of Rasch model parameters. Moreover, because in the Rasch model, only a single parameter is estimated per item, low sample sizes are required for reliable estimation.

2 Estimation Methods for the Rasch Model

A variety of estimation methods has been proposed for the Rasch model [16]. In the Rasch model, item parameters $\mathbf{b} = (b_1, \ldots, b_I)$ and distribution parameters of F are estimated. Assume that item responses x_{pi} are available for persons $p = 1, \ldots, P$ and items $i = 1, \ldots, I$. Denote by $\mathbf{x}_p$ the vector of item responses and by s_p the sum score of person p.

In marginal maximum likelihood estimation (MML; [4]), latent variables θ are integrated out by posing some distributional assumption $G_{\boldsymbol{\gamma}}$ for θ, where distribution parameters $\boldsymbol{\gamma}$ are simultaneously estimated with $\mathbf{b}$. The log-likelihood function $l(\mathbf{b}, \boldsymbol{\gamma})$ is maximized. The likelihood contribution for person p is given by $l_p(\mathbf{b}, \boldsymbol{\gamma}) = \log \left[\int \prod_{i=1}^{I} \mathrm{P}(X_i = x_{pi}|\theta; b_i) \mathrm{d}G_{\boldsymbol{\gamma}}(\theta) \right]$. If $G_{\boldsymbol{\gamma}}$ differs from the data-generating distribution F, biased item parameters can occur. Frequently, a normal distribution for θ is posed (MML-N), and a standard deviation σ is estimated. The integral in the likelihood function is evaluated by numerical integration. Alternatively, a multinomial distribution for θ can be estimated. This approach starts with a fixed grid of θ points $\theta_1, \ldots, \theta_C$ and estimates probabilities $\gamma_c = \mathrm{P}(\theta = \theta_c)$. A log-linear smoothing of these probabilities has been proposed in the so-called general diagnostic model (MML-LM; [19, 22]). Typically, smoothing is performed for up to three or four moments. In a located latent class model with C classes, the values of the grid points θ_c are estimated in addition to probabilities γ_c (MML-LC; [7, 10]). It has been shown that in the Rasch model with I items, at most $C = I/2$ latent classes can be identified. The MML-LC approach imposes the weakest assumptions about F.

In conditional maximum likelihood estimation (CML; [1]), a conditioning step on the sum score S is performed that eliminates θ from estimation equations. In more detail, $l_p(\mathbf{b}) = \log \mathrm{P}(\mathbf{X} = \mathbf{x}_p|S = s_p)$ is evaluated that is independent of θ. In joint maximum likelihood estimation (JML; see [16] for an overview), persons are regarded as fixed effects, and person parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_P)$ are simultaneously estimated with item parameters $\mathbf{b}$. The estimation JML algorithm alternates between $\boldsymbol{\theta}$ and $\mathbf{b}$ parameter estimation in one iteration. Because the number of estimated

parameters grows with sample size, a bias correction for item parameters is required [14, 21]. With obtained item parameters $\hat{b}_i$, the bias-corrected item parameter is computed as $(I-1)/I \cdot \hat{b}_i$. In order to include all persons in the estimation (because an ML estimate for θ is not defined for persons with extreme scores $s_p = 0$ or $s_p = I$), weighted likelihood estimation (WLE; [20]) can be used for obtaining person parameter estimates. As an alternative to WLE, the ε-algorithm of Bertoli-Bersotti (JMLε; [3]) employs a modified likelihood by replacing the sufficient statistic s_p with $\varepsilon + (s_p - 2\varepsilon)/I$ using an appropriate $\varepsilon > 0$. In penalized JML (PJML; see [5] for a related approach), a ridge penalty term is added to the log-likelihood function. This approach corresponds to assuming a normal prior distribution $\theta \sim \mathrm{N}(0, \sigma^2_{\text{prior}})$ with an appropriate choice of the regularization parameter $\sigma_{\text{prior}} > 0$. This approach also circumvents the exclusion of persons with extreme scores from CML. It has been demonstrated that JML and CML can be considered variants of MML estimation [13].

Several simpler estimation alternatives (so-called limited information methods) do not rely on the full item response pattern $\mathbf{x}_p$. In pairwise MML (PMML; [15]) person contributions $\mathrm{P}(X_i = x_{pi}, X_j = x_{pj})$ are considered by integrating out the latent variable θ as in MML. Typically, a normal distribution is employed. In pairwise CML (PCML; [23]), the conditioning $\mathrm{P}(X_i = x_{pi}, X_j = x_{pj}) / \mathrm{P}(X_i + X_j = x_{pi} + x_{pj})$ is used for optimization that also removes θ from estimation equations as in CML. The row averaging approach (RA; [6]), the eigenvector method (EVM; [12]; see also [2]) as well as the MINCHI method [8] only rely on the evaluation of bivariate frequencies $\mathrm{P}(X_i = x, X_j = y)$ $(x, y = 0, 1)$ and do not require assumptions about the distribution F of θ.

3 Simulation Study

3.1 Method

In the simulation study, item response data has been generated for the Rasch model. We varied the number of items ($I = 10$, and 30) and sample sizes ($N = 100$, 250, 500, and 1,000). We chose I equidistant item parameters in the interval -1.5 and 1.5. Three types of ability distributions were simulated. First, we assumed a normal distribution $\mathrm{N}(0, 1)$ (Normal) for θ. Second, we simulated a standardized chi-square (Chi^2) distribution with one degree of freedom. Third, we simulated a located latent class Rasch model with three classes (LC3) and θ points -0.790, 1.033, 2.248 with corresponding probabilities .60, .35, and .05.

As analysis models, we implemented the estimation methods described in Section 2. For MML-LM estimation, we used a log-linear smoothing up to three and four moments. We specified MML-LC with 3, 4, and 5 located latent classes. For JMLε estimation, we tried values $\varepsilon = 0.1$, 0.3, and 0.5. In PJML estimation, we

chose normal priors $N(0, \sigma^2_{\text{prior}})$ with $\sigma_{\text{prior}} = 1$, 1.5, and 2. Notably, an optimal value of σ_{prior} could also be estimated by cross-validation or empirical Bayes methods.

The whole simulation was carried out in R [17] utilizing the R packages immer, pairwise and sirt. To enable comparisons of estimated item parameters across estimation methods, the set of item parameters were centered after estimation (i.e., they have a mean of 0). In total, 5,000 replications were conducted in each cell of the simulation design. Bias, standard deviation (SD), and root mean square error (RMSE) were estimated for all item parameters. We consider two summary measures of item parameter recovery. First, the mean absolute bias $\text{MAB}(\hat{\mathbf{b}}) = I^{-1} \sum_{i=1}^{I} |\text{Bias}(\hat{b}_i)|$ quantifies the average bias of item parameters. Second, bias and variability is summarized in the average relative RMSE (RRMSE) that is defined as $\text{RRMSE}(\hat{\mathbf{b}}) = \left[\sum_{i=1}^{I} \text{RMSE}(\hat{b}_i)\right] / \left[\sum_{i=1}^{I} \text{SD}_{\text{MML-N}}(\hat{b}_i)\right]$, where $\text{SD}_{\text{MML-N}}$ is the SD of item parameters using MML-N estimation. Hence, MML estimation using the normal distribution serves as the reference method.

3.2 Results

We only report JMLε with $\varepsilon = .3$ and PJML with $\sigma_{\text{prior}} = 1.5$ that performed best on average across conditions for lack of space. We also only state results for MML-LM with smoothing 4 moments (MML-LM4) which was superior to only using three moments). MML-LC is reported for 3 located latent classes (MML-LC3), but there were only low efficiency losses when using 4 or 5 classes.

The bias (i.e., the MAB) of item parameters was highest for JML using WLE (JMLW) for short test length ($I = 10$) but vanished in a long test ($I = 30$). However, MML using an incorrect normal distribution (MML-N) produced slightly biased item parameters in the case of non-normal distributions (Chi^2 and LC3). Surprisingly, the normal distributional misspecification in pairwise MML (PMML) had even worse consequences than in MML-N. Bias and RRMSE values were averaged across conditions for each methods and ranked. These ranks are shown in Table 1. Overall, CML, the limited information methods EVM, RA, and CCML as well as MML-LC3 and MML-LM4 performed best in terms of bias. It may also be surprising that MML with located latent classes (MML-LC3) also performs well for continuous ability distributions.

In Table 1, the ranks of estimation methods across all conditions and results for 10 items are shown for the RRMSE. The findings for 30 items were similar but less pronounced. Overall, JML estimation methods performed well, in particular the ε-algorithm JMLε. Notably, MML with more flexible distributions and CML produced low RRMSE values. Interestingly, misspecified MML using a normal distribution (MML-N) outperformed limited information estimators with respect to variability (PMML, CMML, EVM, RA, MINCHI). Hence, the potential bias introduced by MML-N compared to the latter estimation methods can be compensated by smaller variability. It is likely that these findings also transfer to test designs with missing data.

Table 1 Performance ranks and relative root mean square error of item parameters in the Rasch model for different ability distributions and estimation methods as a function of the number of items (I) and sample size (N)

	Rank		Relative RMSE								
			Normal, $I = 10$			Chi^2, $I = 10$			LC3, $I = 10$		
			N			N			N		
Method	Bias	RRMSE	100	250	500	100	250	500	100	250	500
MML-N	9	6	100.1	100.1	100.0	100.4	100.8	101.6	100.1	100.3	100.6
MML-LM4	6	3	100.2	100.1	100.1	101.1	100.7	100.3	100.8	100.1	99.6
MML-LC3	4	2	100.2	99.7	99.5	100.8	100.2	99.7	100.3	99.6	99.1
CML	1	4	100.2	100.2	100.1	100.8	100.5	100.1	100.0	99.9	99.8
JMLW	12	5	97.0	98.8	102.0	97.6	98.6	100.7	97.0	98.7	102.1
JMLε	7	1	98.2	98.4	98.6	99.0	99.3	99.5	98.0	98.2	98.4
PJML	10	7	99.3	99.5	99.5	102.6	103.6	104.8	98.7	99.0	99.2
PMML	11	8	100.1	100.1	100.0	100.5	101.0	102.0	100.4	100.7	101.2
CCML	5	9	103.2	102.7	102.4	103.5	102.7	102.4	103.1	102.8	102.2
EVM	2	10	104.0	103.5	103.2	104.3	103.5	103.2	104.1	103.7	103.1
RA	3	11	104.1	103.6	103.3	104.4	103.6	103.3	104.2	103.8	103.2
MINCHI	8	12	106.1	104.4	103.9	106.2	104.3	103.7	106.1	104.8	103.8

Note. Bias = mean absolute bias (MAB); RRMSE = relative root mean square error; N = sample size; I = number of items; Normal = $\theta \sim \text{N}(0,1)$; Chi^2 = $\theta \sim \chi^2(df = 1)$ with subsequent transformation such that $\text{E}(\theta) = 0$ and $\text{Var}(\theta) = 1$; Discrete = discrete ability distribution with 3 support points (see "Method"); MML-N = marginal maximum likelihood estimation (MML) with normal distribution; MML-LM4 = MML with log-linear smoothing up to 4 moments; MML-LC3 = MML with 3 located latent classes; CML = conditional maximum likelihood; JMLW = joint maximum likelihood estimation (JML) with WLE person parameter estimation and bias correction; JMLε = JML with ε-algorithm using $\varepsilon = 0.3$; PJML = penalized maximum likelihood estimation with prior $\text{N}(0, 1.5^2)$; PMML = pairwise MML; PCML = pairwise CML; EVM = eigenvector estimation method; RA = row averaging method; MINCHI = Fischer's Minchi estimation method. Ranks smaller than 7 and RRMSE values smaller than 100.5 are colored in gray.

4 Discussion

In this article, we compared several estimation methods for the Rasch model. It has been shown that the choice of the ability distribution impacts estimated item parameters. However, differences between estimation methods are only modest, in particular for longer test lengths. Interestingly, joint maximum likelihood estimation methods outperformed conditional and marginal maximum likelihood as well as limited information estimation methods. Prior distributions for item parameters can further improve estimation in small samples [11].

References

1. Andersen, E.B.: The numerical solution of a set of conditional estimation equations. J. R. Stat. Soc. Series B Stat. Methodol. **34**(1), 42–54 (1972). DOI 10.1111/j.2517-6161.1972.tb00887.x
2. Andrich, D., Luo, G.: Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. J. Appl. Meas. **4**(3), 205–221 (2003)
3. Bertoli-Barsotti, L., Lando, T., Punzo, A.: Estimating a Rasch Model via fuzzy empirical probability functions. In: D. Vicari, A. Okada, G. Ragozini, C. Weihs (eds.) Analysis and modeling of complex data in behavioral and social sciences, pp. 29–36. Springer, Cham (2014). DOI 10.1007/978-3-319-06692-9_4
4. Bock, R.D., Aitkin, M.: Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika **46**(4), 443–459 (1981). DOI 10.1007/BF02293801
5. Chen, Y., Li, X., Zhang, S.: Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. Psychometrika **84**(1), 124–146 (2019). DOI 10.1007/s11336-018-9646-5
6. Choppin, B.: A fully conditional estimation procedure for Rasch model parameters. Eval. Educ. **9**(1), 29–42 (1982)
7. De Leeuw, J., Verhelst, N.: Maximum likelihood estimation in generalized Rasch models. J. Educ. Behav. Stat. **11**(3), 183–196 (1986). DOI 10.3102/10769986011003183
8. Fischer, G.H.: Rasch models. In: C.R. Rao, S. Sinharay (eds.) Handbook of statistics, Vol. 26: Psychometrics, pp. 515–585 (2007). DOI 10.1016/S0169-7161(06)26016-4
9. Fischer, G.H., Molenaar, I.W.: Rasch models. Foundations, recent developments, and applications. Springer, New York (1995). DOI 10.1007/978-1-4612-4230-7
10. Formann, A.K.: Constrained latent class models: Theory and applications. Brit. J. Math. Stat. Psychol. **38**(1), 87–111 (1985). DOI 10.1111/j.2044-8317.1985.tb00818.x
11. Fox, J.P.: Bayesian item response modeling. Springer, New York (2010). DOI 10.1007/978-1-4419-0742-4
12. Garner, M.: An eigenvector method for estimating item parameters of the dichotomous and polytomous Rasch models. J. Appl. Meas. **3**(2), 107–128 (2002)
13. Holland, P.W.: On the sampling theory foundations of item response theory models. Psychometrika **55**(4), 577–601 (1990). DOI 10.1007/BF02294609
14. Jansen, P.G.W., van den Wollenberg, A.L., Wierda, F.W.: Correcting unconditional parameter estimates in the Rasch model for inconsistency. Appl. Psychol. Meas. **12**(3), 297–306 (1988). DOI 10.1177/014662168801200307
15. Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., Jöreskog, K.G.: Pairwise likelihood estimation for factor analysis models with ordinal data. Comp. Stat. Data An. **56**(12), 4243–4258 (2012). DOI 10.1016/j.csda.2012.04.010
16. Molenaar, I.W.: Estimation of item parameters. In: G.H. Fischer, I.W. Molenaar (eds.) Rasch models. foundations, recent developments, and applications, pp. 39–52. Springer, New York (1995). DOI 10.1007/978-1-4612-4230-7_3
17. R Core Team: *R: A language and environment for statistical computing* (2020). Vienna, Austria. https://www.R-project.org/
18. Rasch, G.: Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen (1960)
19. von Davier, M.: A general diagnostic model applied to language testing data. Brit. J. Math. Stat. Psychol. **61**(2), 287–307 (2008). DOI 10.1348/000711007X193957
20. Warm, T.A.: Weighted likelihood estimation of ability in item response theory. Psychometrika **54**(3), 427–450 (1989). DOI 10.1007/BF02294627
21. Wright, B.D., Douglas, G.A.: Best procedures for sample-free item analysis. Appl. Psychol. Meas. **1**(2), 281–295 (1977). DOI 10.1177/014662167700100216
22. Xu, X., von Davier, M.: *Fitting the structured general diagnostic model to NAEP data.* (Research Report No. RR-08-28). Educational Testing Service (2008). DOI 10.1002/j.2333-8504.2008.tb02113.x
23. Zwinderman, A.H.: Pairwise parameter estimation in Rasch models. Appl. Psychol. Meas. **19**(4), 369–375 (1995). DOI 10.1177/014662169501900406

2.12 New issues on multivariate and univariate quantile regression

Directional M-quantile regression for multivariate dependent outcomes

Regressione M-quantile direzionale per dati multivariati dipendenti

Merlo Luca, Petrella Lea and Tzavidis Nikos

Abstract In the present work we generalize the univariate M-quantile regression to the analysis of multivariate dependent outcomes. Extending the notion of directional quantiles, we introduce directional M-quantiles which are obtained as projections of the original data on a specified unit norm direction. In order to take into consideration the correlation within grouped measurements and to increase efficiency, we develop a marginal M-Quantile regression model extending the well-known generalized estimating equations approach. We build M-quantile regions and contours which allow us to investigate the effect of the covariates on the location, spread and shape of the distribution of the responses. To identify potential outliers and provide a simple visual representation of the variability of the M-quantile contours estimator, we construct confidence envelope via nonparametric bootstrap. The validity of our method is analyzed through the study of the wages data from the National Longitudinal Survey of Youth.

Abstract In questo lavoro si estende la regressione M-quantilica univariata per l'analisi di dati multivariati dipendenti introducendo la definizione di M-quantile direzionale associato a variabili risposta vettoriali. Al fine di incorporare la struttura di correlazione dei dati nella procedura di stima e determinare stimatori più efficienti, si considera un modello marginale M-quantile estendendo l'approccio delle equazioni di stima generalizzate. Inoltre, proponiamo di utilizzare i contorni M-quantile per investigare l'effetto delle covariate sulla distribuzione delle variabili risposta e, per esaminare la loro variabilità, costruiamo degli insiemi di confidenza attraverso l'approccio bootstrap. L'analisi empirica si concentra sulle retribuzioni salariali di giovani americani ottenute dal National Longitudinal Survey of Youth.

Merlo Luca
Department of Statistical Sciences, Sapienza University of Rome, e-mail: luca.merlo@uniroma1.it

Petrella Lea
MEMOTEF Department, Sapienza University of Rome, e-mail: lea.petrella@uniroma1.it

Tzavidis Nikos
Department of Social Statistics and Demography, Southampton Statistical Sciences Research Institute, University of Southampton, e-mail: N.TZAVIDIS@soton.ac.uk

Key words: Correlated data, GMQEE, Marginal approach, M-quantile contours

1 Introduction

In the univariate setting, the quantile regression approach proposed by [7] has attracted considerable interest in many applications because it provides a way to model the conditional quantiles of a response as a function of explanatory variables in order to have a more complete picture of the entire conditional distribution compared to the classical mean regression. For a detailed review and list of references see [8]. Within the quantile regression framework, a possible alternative is represented by the M-quantile regression approach proposed by [2]. This method provides a "quantile-like" generalization of mean regression based on influence functions combining in a common framework the robustness and efficiency properties of quantiles and expectiles [12], respectively. Although M-quantiles have a less intuitive interpretation than standard quantiles, with respect to the latter, they are very versatile. Specifically, they allow for robust estimation in the presence of influential observations, they can trade robustness for efficiency, ensure uniqueness of the Maximum Likelihood solutions and offer greater stability as a wide range of continuous influence functions can be employed. Unfortunately, M-quantiles have remained relegated to univariate problems due to the lack of a natural ordering in a p-dimensional space, $p > 1$, which preclude the laying down of pertinent concepts of multivariate M-quantiles, ranks and signs. Yet, an extension to higher dimensions could prove to be very useful in many fields of applied statistics when the problem being studied involves the characterization of the distribution of a multivariate response. In the literature some proposals for defining the multivariate M-quantile have been put forward by [2], [9] and [1], for example.

In the present paper we generalize the univariate M-quantile regression to the multivariate setting for the analysis of dependent data by extending the notion of directional quantiles in [10]. More in detail, we introduce directional M-quantiles which are obtained as projections of the original data on a specified unit norm direction. In real world scenarios, observations are often correlated with each other across time, space, or other dimensions, like groups, and their analysis deserves specific instruments which have received enormous attention over the years [3, 4]. In order to take into consideration the correlation within grouped measurements and to increase efficiency, we develop a Marginal M-Quantile (MMQ) regression model. The marginal approach refers to a general class of statistical methods that are used to model dependent data where observations within a cluster are correlated with each other ([11, 5, 3, 4]). A popular estimation procedure for estimating the marginal model parameters is the Generalized Estimating Equations (GEE) approach of [11]. Because the true correlation structure of the data is unknown, the GEE formulates a "working covariance matrix" to capture dependence between observations and incorporate that structure into the model. To estimate the model parameters, we extend the well-known GEE approach of [11] and present the Generalized M-Quantile Estimating Equations (GMQEE). For a fixed direction, we derive the asymptotic properties for the proposed estimator and establish consistency and asymptotic nor-

mality. We also investigate M-quantile regions and contours for a given quantile level and we propose to use M-quantile contour lines to investigate the effect of the covariates on the response variables. In order to visualize the sample variability of the M-quantile contours estimator, we construct confidence envelopes via nonparametric bootstrap. From an empirical point of view, we exploit the proposed MMQ regression model to track the labor-market experiences of male high school dropouts collected by the National Longitudinal Survey of Youth (NLSY).

2 Methodology

Let $\mathbf{Y}_{ij} = (Y_{ij}^{(1)}, \ldots, Y_{ij}^{(p)})'$ and $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(k)})$ denote a continuous p-variate response variable and a k-dimensional vector of explanatory variables for the i-th statistical unit in the j-th cluster of size n_j, for $j = 1, ..., d$ and $i = 1, ..., n_j$ with $n = \sum_{j=1}^{d} n_j$, respectively. We define $\mathbf{u}$ a unit norm direction vector ranging over $\mathscr{S}^{p-1} = \{\mathbf{z} \in \mathbb{R}^p : ||\mathbf{z}|| = 1\}$. To simplify the notation, we stack up the projected responses on $\mathbf{u}$ to the n_j dimensional vector $\tilde{\mathbf{Y}}_j = (\mathbf{u}'\mathbf{Y}_{1j}, \ldots, \mathbf{u}'\mathbf{Y}_{n_j j})'$, while $\mathbf{X}_j = (\mathbf{X}_{ij}, \ldots, \mathbf{X}_{n_j j})$ is a $n_j \times k$ matrix collecting the covariates for group j. Extending [10], we define the directional M-quantile for multivariate distributions as follows.

Definition 1. Let $\mathbf{Y}$ be a p-dimensional random vector with absolutely continuous distribution function. For any $\tau \in (0,1)$ and direction $\mathbf{u} \in \mathscr{S}^{p-1}$, let $\psi_\tau(u) = |\ \tau - \mathbf{1}_{(u<0)}\ |\ \psi(u)$ denote the asymmetric Huber influence function with $\psi(u) = u\mathbf{1}_{(|u|\leq c)} + c\,\mathrm{sign}(u)\mathbf{1}_{(|u|>c)}$, where c denotes a tuning constant bounded away from zero. Then, the directional M-quantile of order τ in the direction $\mathbf{u}$, $\theta_{\mathbf{u}}(\tau)$, is the τ-th M-quantile of the corresponding projection of the distribution of $\mathbf{Y}$, namely:

$$\int \psi_\tau(\mathbf{u}'\mathbf{y} - \theta_{\mathbf{u}}(\tau))\mathrm{d}F_{\mathbf{u}'\mathbf{Y}} = 0. \tag{1}$$

The proposed directional M-quantile is real-valued and it corresponds to the univariate τ-th M-quantile of the distribution of $\mathbf{u}'\mathbf{Y}$ where the direction $\mathbf{u}$ can be interpreted as a weight vector for each marginal distribution of $\mathbf{Y}$ involved in the regression problem. In addition, directional M-quantiles inherit the computational advantages, robustness and efficiency properties of standard univariate M-quantiles. Specifically, by varying the tuning constant c in $\psi_\tau(\cdot)$, directional M-quantiles reduce to directional quantiles of [10] when $c \to 0$ and reduce to directional expectiles for c large. Clearly, Definition 1 includes the traditional univariate one when $p = 1$. In the regression context, the proposed definition can be easily extended to conditional distributions when covariates are available. For a given $\tau \in (0,1)$ and $\mathbf{u} \in \mathscr{S}^{p-1}$, the directional M-quantile model is defined as:

$$\theta_{\mathbf{u},\mathbf{X}}(\tau) = \mathbf{X}_{ij}'\beta(\tau), \qquad i = 1, ..., n_j \text{ and } j = 1, ..., d, \tag{2}$$

where $\beta(\tau)$ is the k-dimensional vector of regression coefficients.

To account for the dependence structure of the data we consider the so called marginal modeling approach and estimate the parameters using the GEE. By introducing a suitable correlation matrix $\mathbf{C}_j(\mathbf{r}_j)$ of size n_j indexed by the s_j-dimensional vector $\mathbf{r}_j$ which characterizes the correlation within groups, $j = 1, ..., d$, we are able to capture within group dependence and enhance the efficiency of the regression coefficients estimator [11]. Following [13] and [11], for a given τ and direction $\mathbf{u}$, we define the estimator $\widehat{\beta}_{MMQ}(\tau)$ as the solution of the following Generalized M-quantile Estimating Equations (GMQEE):

$$\mathbf{U}(\beta(\tau)) = \sum_{j=1}^{d} \mathbf{U}_j(\beta(\tau)) = \sum_{j=1}^{d} \mathbf{X}_j' \boldsymbol{\Sigma}_j^{-1}(\mathbf{r}_j) \mathbf{V}_j^{\frac{1}{2}} \psi_\tau(\mathbf{z}_j) = \mathbf{0}, \tag{3}$$

where $\mathbf{z}_j = \mathbf{V}_j^{-\frac{1}{2}}(\tilde{\mathbf{Y}}_j - \mathbf{X}_j\beta(\tau))$ denotes the n_j-dimensional vector of standardized residuals, $\mathbf{V}_j$ is the diagonal matrix of size n_j which contains the scale parameter σ_τ^2 for the residuals' distribution $\tilde{\mathbf{Y}}_j - \mathbf{X}_j\beta(\tau)$ and $\boldsymbol{\Sigma}_j(\mathbf{r}_j) = \phi \mathbf{V}_j^{\frac{1}{2}} \mathbf{C}_j(\mathbf{r}_j) \mathbf{V}_j^{\frac{1}{2}}$ is the "working" covariance matrix with ϕ being a positive nuisance parameter. It is worth noticing that, when $\mathbf{C}_j(\mathbf{r}_j) = \mathbf{I}_{n_j}$, with $\mathbf{I}_{n_j}$ being the identity matrix of size n_j, independence between clustered observations is assumed. Several choices for $\mathbf{C}_j(\mathbf{r}_j)$ have been proposed in the related literature, such as the exchangeable correlation structure, the AR(1) structure, or the totally unspecified structure. Their specification and parameters interpretation generally depend on the application under investigation. For fixed τ and $\mathbf{u}$, under mild regularity conditions the estimator $\widehat{\beta}_{MMQ}(\tau)$ is consistent and asymptotically normally distributed. In addition, an estimate of the model parameters $(\beta_{MMQ}(\tau), \sigma_\tau, \phi, \mathbf{r}_j)$ can be obtained using a Newton-Raphson algorithm to solve the GMQEE in (3).

To provide a full description of the dependence of the responses $\mathbf{Y}$ on the regressors $\mathbf{X}$, we investigate how directional M-quantiles can provide a summary when, theoretically, all directions over $\mathscr{S}^{p-1}$ are investigated simultaneously, for fixed τ. Let $\mathbf{y}$ denote the realization of the random vector $\mathbf{Y}$. For a given $\tau \in (0,1)$ and $\mathbf{u} \in \mathscr{S}^{p-1}$, we first define the τ-th directional M-quantile regression hyperplane $\pi_{\mathbf{u},\mathbf{x}}(\tau) = \{\mathbf{y} \in \mathbb{R}^p : \mathbf{u}'\mathbf{y} = \theta_{\mathbf{u},\mathbf{x}}(\tau)\}$. Each hyperplane $\pi_{\mathbf{u},\mathbf{x}}(\tau)$ characterizes a lower (open) and an upper (closed) M-quantile regression halfspace $H^-_{\mathbf{u},\mathbf{x}}(\tau) = \{\mathbf{y} \in \mathbb{R}^p : \mathbf{u}'\mathbf{y} < \theta_{\mathbf{u},\mathbf{x}}(\tau)\}$ and $H^+_{\mathbf{u},\mathbf{x}}(\tau) = \{\mathbf{y} \in \mathbb{R}^p : \mathbf{u}'\mathbf{y} \geq \theta_{\mathbf{u},\mathbf{x}}(\tau)\}$, respectively. For $\tau \in (0, \frac{1}{2}]$, the τ-th M-quantile region conditional on $\mathbf{X} = \mathbf{x}$, $R_{\mathbf{x}}(\tau) \subset \mathbb{R}^p$, is defined as:

$$R_{\mathbf{x}}(\tau) = \bigcap_{\mathbf{u} \in \mathscr{S}^{p-1}} H^+_{\mathbf{u},\mathbf{x}}(\tau). \tag{4}$$

The region defined in (4) is convex, compact and bounded, and the corresponding conditional M-quantile contour of order τ is defined as the boundary $\partial R_{\mathbf{x}}(\tau)$ of $R_{\mathbf{x}}(\tau)$. Such quantities are of crucial interest as they are able to detect covariate-dependent features of the distribution of the responses given $\mathbf{X}$, while ensuring robustness to outlying data. Specifically, for fixed τ, when $c \to 0$, M-quantile contours reduce to directional quantile envelopes illustrated in [10]; on the other hand, when

$c \to \infty$ they generate expectile contours. Meanwhile, for a given c, as $\tau \to 0$ the M-quantile contour of order τ approaches the convex hull of the sample data providing valuable information about the extent of extremality of points.

3 Application

The proposed methodology has been applied to the NLSY data (NLS79.txt). The NLSY is a longitudinal study that follows the lives of a sample of American youth born between 1980-84. The considered data contains measurements on hourly log-wages (Lnw), years of experience (Exper) in the workforce, unemployment rates in the local geographic region (Uerate) and race (White (baseline), Black) of male high-school dropouts, aged between 14 and 17 years when first measured. The considered sample consists of $d = 500$ men for a total of $n = 3749$ observations. The aim of this analysis is to investigate how the local area unemployment rate and men race affect differently hourly earning and the workers' experience of disadvantages young Americans (low quantiles) and high earners (high quantiles). To handle dependence between repeated measurements and account for stronger dependence between adjacent measurements than for distant ones, we assume an AR(1) structure, $[\mathbf{C}_j(\mathbf{r}_j)]_{ik} = r^{|i-k|}$, working correlation structure and, the tuning constant c in Definition 1 has been set to 1.345 which gives reasonably efficiency under normality and protects against outliers (see [6]).

Table 1 shows point estimates of the regression coefficients and of the correlation parameter for the MMQ model at $\tau = (0.1, 0.25, 0.5, 0.75, 0.9)$ and for two directions, $\mathbf{u}_1 = (1,0)$ and $\mathbf{u}_2 = (0,1)$, which reduces the multidimensional problem to two MMQ regressions on each component of the bivariate response. In addition, the results of a Marginal Mean (MM) model fitted with the standard GEE approach are reported. We observe that the MM and MMQ models produce comparable estimates at the center of the distribution (the MM model cannot be applied to estimate the covariates' effects in the tails of the distribution). The results show that there is evidence of a negative association between the considered covariates and either log wage and working experience. In particular, the effect is statistically significant at the investigated quantile levels and it is more pronounced at the high quantiles. By looking at the estimated correlation parameters, as expected, there is a high within-subject correlation which is consistent with the repeated measures design.

To provide a graphical representation of the effects of the included covariates in the tails of the responses distribution, we fit the MMQ model at $\tau = (0.01, 0.25)$ for 200 equispaced directions in $\mathscr{S}^{p-1}$ and construct M-quantile regression contours using (4). Figure 1 illustrates the estimated $\partial R_{\mathbf{x}}(\tau)$ conditional on race, white (red curves) and black (blue curves), at the 0.50-th (left), 0.75-th (center) and 0.99-th (right) empirical quantiles of the unemployment rate which correspond to an unemployment rate of 6.9%, 12.2% and 22.9%. The shaded areas represent 95% confidence envelopes for M-quantile contours obtained using nonparametric block bootstrap. The contours for smaller τ capture the effects for more extreme workers e.g., men with low levels of income and experience and those with exceptionally high values of income and experience. The elongated curves indicate that there is

u	Variable	MM	MMQ				
			0.1	0.25	0.5	0.75	0.9
$\mathbf{u}_1$	Intercept	5.204^{***}	2.035^{***}	2.759^{***}	4.352^{***}	6.382^{***}	7.808^{***}
	Black	-0.681^{***}	-0.270^{***}	-0.358^{***}	-0.569^{***}	-0.826^{***}	-0.876^{***}
	Uerate	-0.156^{***}	-0.070^{***}	-0.078^{***}	-0.105^{***}	-0.117^{***}	-0.103^{***}
	r	0.817^{***}	0.637^{***}	0.778^{***}	0.853^{***}	0.824^{***}	0.741^{***}
$\mathbf{u}_2$	Intercept	2.031^{***}	1.702^{***}	1.829^{***}	2.017^{***}	2.245^{***}	2.473^{***}
	Black	-0.089^{***}	-0.071^{***}	-0.077^{***}	-0.094^{***}	-0.094^{***}	-0.074
	Uerate	-0.017^{***}	-0.017^{***}	-0.016^{***}	-0.018^{***}	-0.022^{***}	-0.024^{***}
	r	0.656^{***}	0.480^{***}	0.528^{***}	0.620^{***}	0.539^{***}	0.433^{***}

Table 1 MM and MMQ model parameters estimates at the investigated quantile levels. ***, ** and * denote statistical significance at the $0.01, 0.05$ and 0.1 levels, respectively.

more variability in years of working experience and it can also be easily seen that blue curves always lie below and to the left of the red ones, suggesting the existence of a significant racial wage gap that disadvantages young African-American males, especially at $\tau = 0.25$. Finally, as the unemployment rate increases the M-contours rapidly descend downward from right to left and become cone-shaped which exert downward pressure on both wages and years of working experience.

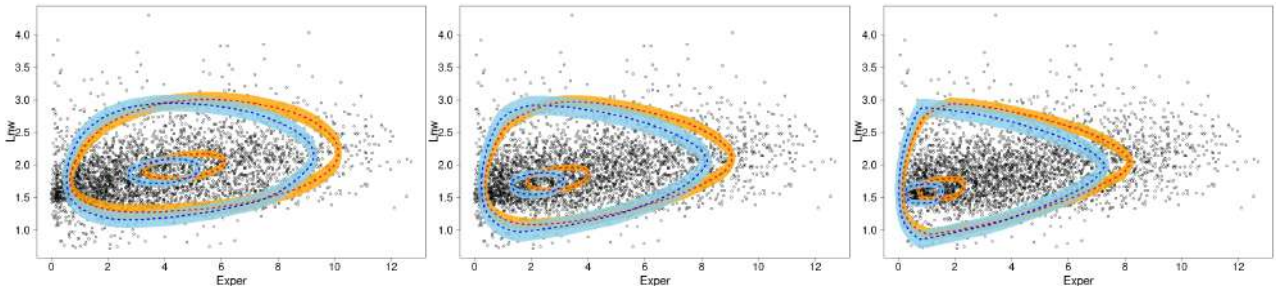


Fig. 1 Estimated M-quantile contours at $\tau = (0.01, 0.25)$ conditional on race, white (red curves) and black (blue curves), at the 0.50-th (left), 0.75-th (center) and 0.99-th (right) empirical quantiles of the unemployment rate. The shaded surfaces represent 95% confidence envelopes for M-quantile contours obtained using nonparametric block bootstrap.

References

[1] Alfò, M., Marino, M. F., Ranalli, M. G., Salvati, N. and Tzavidis, N. [2021], 'M-quantile regression for multivariate longitudinal data with an application to the Millennium Cohort Study', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **70**(1), 122–146.
[2] Breckling, J. and Chambers, R. [1988], 'M-quantiles', *Biometrika* **75**(4), 761–771.
[3] Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Heagerty, P. J., Zeger, S. et al. [2002], *Analysis of longitudinal data*, Oxford University Press.
[4] Goldstein, H. [2011], *Multilevel statistical models*, Vol. 922, John Wiley & Sons.
[5] Heagerty, P. J., Zeger, S. L. et al. [2000], 'Marginalized multilevel models and likelihood inference', *Statistical Science* **15**(1), 1–26.
[6] Huber, P. and Ronchetti, E. [2009], *Robust Statistics*, Wiley.
[7] Koenker, R. and Bassett Jr, G. [1978], 'Regression quantiles', *Econometrica: Journal of the Econometric Society* pp. 33–50.
[8] Koenker, R., Chernozhukov, V., He, X. and Peng, L. [2017], *Handbook of Quantile Regression*, CRC press.
[9] Kokic, P., Breckling, J. and Lübke, O. [2002], A new definition of multivariate M-quantiles, *in* 'Statistical data analysis based on the L_1-norm and related methods', Springer, pp. 15–24.
[10] Kong, L. and Mizera, I. [2012], 'Quantile tomography: using quantiles with multivariate data', *Statistica Sinica* pp. 1589–1610.
[11] Liang, K.-Y. and Zeger, S. L. [1986], 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.
[12] Newey, W. K. and Powell, J. L. [1987], 'Asymmetric least squares estimation and testing', *Econometrica* pp. 819–847.
[13] Sinha, S. K. and Rao, J. [2009], 'Robust small area estimation', *Canadian Journal of Statistics* **37**(3), 381–399.

2.13 Semi-parametric and non-parametric latent class analysis

Stepwise Estimation of Multilevel Latent Class Models

Stima stewise di modelli multilivello a classi latenti

Zsuzsa Bakk and Roberto di Mari and Jennifer Oser and Jouni Kuha

Key words: multilevel latent class analysis, covariates, stepwise estimation
Abstract We propose a two-step estimator for multilevel latent class analaysis with co-variates that separates the estimation of the measurement and structural model. Keeping the measurement model fixed in step 2 when covariates are added it is possible to obtain an unbiased and efficient stepwise estimator. We investigate the bias and the efficiency of the proposed estimator via a simulation study.
Abstract In questo lavoro viene proposto uno stimatore a due steps per modelli multilivello a classe latente, che separa la stima del modello di misurazione da quello strutturale. Mantenendo i parametri del modello di misurazione fissi nel secondo step, quando le covariate sono usate come predittori della variabile di classe latente, è possibile ottenere uno stimatore corretto ed efficiente. La distorsione e la variabilità in campioni finiti dello stimatore proposto vengono analizzate tramite uno studio di simulazione.

Key words: multilevel latent class analysis; covariates; two-step estimation

Zsuzsa Bakk
Leiden University, Leiden, The Netherlands, e-mail: z.bakk@fsw.leidenuniv.nl

Roberto Di Mari
Catania University, Catania, Italy e-mail: roberto.dimari@unict.it

Jennifer Oser
Ben Gurion University of the Negev e-mail: oser@post.bgu.ac.il

Jouni Kuha
London School of Economics, London, UK e-mail: J.Kuha@lse.ac.uk

1 Introduction

Latent class (LC) analysis is an approach used to create a clustering of a set of observed variables, based on an underlying unknown classification. In multilevel LC analysis the respondents are assumed to belong to higher level groups, such as students nested in schools, or entrepreneurs in countries. Multilevel LCA is increasingly popular in fields such as educational research (modeling students learning profiles in different school types [4] in economics ([10], epidemiology (substance abuse profiles nested in communities [11]) Usually the interest lies at the lower level clustering, and the difference in the distribution of the lower level classes at the higher level unit.

In LCA creating a clustering is usually only the first step. The research want to explain the clustering by co-variates. For example relating substance abuse profiles to community and person characteristics ([11]). Classically a one-step or a three step approach is used. Using the first approach every time a new co-variate is added to the model, the full model needs to be re-estimated making modeling cumbersome. Alternatively using the three step approach first the measurement model is established based on the indicators of the LC variable, in the second step respondents are assigned to posterior classes. In the last step the assigned class membership is used in a multilevel regression analysis. The problem with this approach is that in the classification step a classification error is introduced, that leads to biased estimates in the third step. For single level LC models several bias-adjusted stepwise estimators were proposed [2, 13],among which the most straightforward to extend to multilevel setting is the two-step approach [1].

In the current paper we extend the two-step approach to the multilevel LC model [1], as such proposing a bias-adjusted stepwise estimator for these family of models. We focus on the definition of the non parametric multilevel LC model introduced by [12], that contains 1 latent variable per random coefficient and one latent variable per level 1 unit within a level 2 unit. While multiple definitions are available in literature, the non parametric LC model is commonly used in applied research due to it's simplicity.

2 The multilevel latent class model

Consider the vector of responses $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijK})$, where Y_{ijk} denotes the response of individual i in group j on the k-th categorical indicator variable, with $1 \leq k \leq K$ and $1 \leq j \leq J$, where K denotes the number of categorical indicators and J the number of level 2 units. In addition, we let n_j denote the number of level 1 units within the j-th level 2 unit, with $1 \leq j \leq J$. For simplicity of exposition, we focus on dichotomous indicators.

LC analysis assumes that respondents belong to one of the T categories ("latent classes") of an underlying categorical latent variable X which affects the responses ([9, 6]). The model for $\mathbf{Y}_{ijk}$ can then be specified as

$$P(\mathbf{Y}_{ijk}) = \sum_{t=1}^{T} P(X_{ij} = t) \prod_{k=1}^{K} P(Y_{ijk}|X_{ij} = t). \tag{1}$$

where the weight $P(X_{ij} = t)$ is the probability of person i in group j to belong to latent class t. The term $P(\mathbf{Y}_{ijk}|X = t)$ is the class-specific response probability on indicator K given that a person belongs to class t. We make the "local independence" assumption that the K indicator variables are independent within the latent classes.

The general definition in Equation 1 applies to both the standard and multilevel LC model. Re-expressing it in terms logit equations we define the simple LC model as:

$$P(X_{ij} = t) = \frac{\exp(\gamma_t)}{1 + \sum_{t=2}^{T} \exp(\gamma_t)}, \tag{2}$$

for $1 < t \leq T$ - where we have taken the first class as reference - and

$$P(Y_{ijk}|X_{ij} = t) = \frac{\exp(\beta_t^k)}{1 + \exp(\beta_t^k)}. \tag{3}$$

Extending the simple LC model to account for the multilevel data structure is possible by allowing the parametrizations in equations 2 and 3 to take the grouping into account. Let W to be a multinomial group latent variable with M mass points each with probability $P(W = m) = \pi_m$. By letting W_j be the value of W for group j, in the nonparametric approach the model for the (individual) latent class probabilities is specified as follows

$$P(X_{ij} = t|W_j = m) = \frac{\exp(\gamma_{tm})}{1 + \sum_{s=2}^{T} \exp(\gamma_{sm})}. \tag{4}$$

Also the mixing probabilities $P(W = m)$ can be parametrized by means of logistic regressions as follows

$$P(W = m) = \frac{\exp(\delta_{0m})}{1 + \sum_{l=2}^{M} \delta_{0l}}, \tag{5}$$

where parameters for $m = 1$ are set to zero for identification and the related class is set as reference.

While the conditional response probabilities can also directly depend on the higher level LC variable, a restricted version is common ([12, 7]) that assumes item-conditional probabilities do not depend on the level 2 unit leading to the following specification for $\mathbf{Y}_j$

$$P(\mathbf{Y}_j) = \sum_{m=1}^{M} P(W_j = m) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m) \prod_{k=1}^{K} P(Y_{ijk}|X_{ij} = t), \tag{6}$$

where $\mathbf{Y}_j = \{\mathbf{Y}_{j1}, \ldots, \mathbf{Y}_{jn_j}\}$.

We do not discuss model selection in the current paper, interested readers can refer to [7]. In the stage of adding covariates the number of classes should be fixed, also to be in line with general recommendations for LCA with covariates [8].

2.1 Extending the model with covariates. Classical approaches

Level 1 and level 2 covariates can be included to predict class membership. Denoting one level 2 covariate by Z_{1j} and a level 1 covariate by Z_{2ij} the multinomial logistic regression for X_{ij} with a random intercept can be written as:

$$P(X_{ij} = t|W_j, Z_{1j}, Z_{2ij}) = \frac{\exp(\gamma_{0tm} + \gamma_{1t}Z_{1j} + \gamma_{2t}Z_{2ij})}{\sum_{s=1}^{T}\exp(\gamma_{0sm} + \gamma_{1s}Z_{1j} + \gamma_{2s}Z_{2ij})}. \tag{7}$$

A random slope for the level 1 covariate can be obtained by replacing γ_{2t} by γ_{2jt}. Level 2 covariates can be used also to predict group class membership extending Equation 5 similarly to the extension in Equation 7. The extended model for $P(\mathbf{Y}_j|\mathbf{Z}_j)$, where $\mathbf{Z}_j = (Z_{1j}, Z_{2ij})'$, can be specified as

$$P(\mathbf{Y}_j|\mathbf{Z}_j) = \sum_{m=1}^{M} P(W_j = m|Z_{1j}) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, Z_{1j}, Z_{2ij}) \prod_{k=1}^{K} P(Y_{ijk}|X_{ij} = t). \tag{8}$$

Using the one-step approach the full model needs to be re-estimated every time a new co-variate is added keeping the number of lower and higher level classes fixed. The one-step multilevel model has two main drawbacks: 1) estimating the full model multiple times can be time consuming, and 2) misspecifications in a part of the model may destabilize also parameters in other parts of the model.

Due to this complexities in practice often the classical three step approach is used. Using this approach in step 1 the measurement model is estimated. In step 2 the respondents are assigned to posterior classes ($P(X|\mathbf{Y})$ based on their response probabilities on the indicators. The assignment method for the multilevel LCA approach is described in detail in [12]. In step three the posterior class assignment is related to external variables in a logistic regression ignoring the misclassification probabilities. Due to the existence of bias-adjusted approaches the classical three step approach in single level LCA is not recommended anymore, and will not be described in more detail here.

2.1.1 Two-step estimation of models with co-variates

An alternative option that would fit the logic of the stepwise modeling procedure is to apply the two-step LC approach proposed for simple LC models by [1] and applied to latent Markov models by [3].

Using the two step approach the measurement model as described in Equation 6 is fitted, and the number of lower and higher level classes are selected using fit measures, such as AIC or BIC.

As a next step the covariates can be added to the model. At this stage also a decision needs to be taken whether a stepwise approach is preferred (adding first lower level covariates, and after fixing those adding at the higher level) or all covariates can be added in a single step. The benefit of the first option can be robustness, however no simulation or theoretical results are available - this still needs further research. For sake of conciseness, we will present the simultaneous step 2 - its split counterpart can be derived analogously.

Let us define $\theta_2 = (\gamma_{12}, \ldots, \gamma_{1T})$. With the parametrizations specified in Equations 6, 7 the model log-likelihood can be written as follows

$$\log L(\theta_2|\theta_1 = \widehat{\theta}_1) = \sum_{j=1}^{J} \log P(\mathbf{Y}_j|\mathbf{Z}_j), \tag{9}$$

An issue to take into account with two-step estimation is how to account for the uncertainty about the fixed parameters in the calculation of the step two standard errors. Pseudo ML estimates have two sources of variability: the variability due to sampling in step two, but also that of the sampling variability of step one [5]. We propose to apply the approach proposed by [1] for single level LC models to the multilevel setting.

Let the Fisher information matrix of the joint (one-step) model for θ be denoted by

$$\mathscr{I}(\theta^*) = \mathscr{I}_{11}\mathscr{I}'_{12}\mathscr{I}_{22}$$

where θ^* denotes the true value of θ and the partitioning corresponds to θ_1 and θ_2. The asymptotic variance matrix of the one-step estimator $\hat{\theta}$ is thus $\mathbf{V}_{ML} = \mathscr{I}^{-1}(\theta^*)$, which is estimated by $\hat{\mathbf{V}}_{ML} = \mathscr{I}^{-1}(\hat{\theta})$. Let Σ_{11} denote the asymptotic variance matrix of the step-1 estimates $\tilde{\theta}_1$ of the two-step method, obtained similarly from the Fisher information matrix of model described in Equation (6) and estimated by substituting $\tilde{\theta}_1$. The asymptotic variance matrix of the step-2 estimator $\tilde{\theta}_2$ is then given by

$$\mathbf{V} = \mathscr{I}_{22}^{-1} + \mathscr{I}_{22}^{-1}\mathscr{I}_{12}\Sigma_{11}\mathscr{I}'_{12}\mathscr{I}_{22}^{-1} \equiv \mathbf{V}_2 + \mathbf{V}_1. \tag{10}$$

Here V_2 describes the variability in $\tilde{\theta}_2$ if the step-1 parameters θ_1 were actually known, and $\mathbf{V}_1$ the additional variability arising from the fact that θ_1 are not known but estimated by $\tilde{\theta}_1$. The variance matrix $\mathbf{V}$ is estimated by subtituting $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$ for θ^*. It can then also be used to calculate confidence intervals for the parameters in θ_2, and Wald test statistics for them.

If we use the estimated standard errors that are routinely produced by software which fits the step-2 model, this amounts to using $\mathbf{V}_2$ only. Because they omit the contribution from $\mathbf{V}_1$, these standard errors will underestimate the full uncertainty of $\tilde{\theta}_2$.

3 Simulation study set up and statement of main results

We will setup a simulation study to compare the proposed two-step approach to the one step and three step approaches with regard to parameter bias and efficiency. We will generate data from models with varying sample sizes at both the lower and higher level, and varying entropy manipulated by changing the strength of the $Y|X$ relationship. The strength of the association between $Z-X$ will also be varied from no, weak to strong association. We expect that the proposed two-step estimator will be unbiased (similarly to the one-step approach, and showing much lower levels of bias as the three-step approach), and will be slightly less efficient than the one-step estimator.

References

1. BAKK, Z., AND KUHA, J. Two-step estimation of models between latent classes and external variables. *Psychometrika 83* (2018), 871–892.
2. BOLCK, A., CROON, M., AND HAGENAARS, J. Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis 12* (2004), 3–27.
3. DI MARI, R., AND BAKK, Z. Mostly harmless direct effects: A comparison of different latent markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal 25*, 3 (2018), 467–483.
4. FAGGINGER AUER, M. F., HICKENDORFF, M., VAN PUTTEN, C. M., BÈGUIN, A. A., AND HEISER, W. J. Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students' mathematical strategies. *Applied Measurement in Education*, 29 (2016), 144–159.
5. GONG, G., AND SAMANIEGO, F. J. Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics* (1981), 861–869.
6. GOODMAN, L. A. The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology* (1974), 79–259.
7. LUKOCIENE, O., VARRIALE, R., AND VERMUNT, J. The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology 40*, 1 (2010), 247–283.
8. MASYN, K. E. Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal 24*, 2 (2017), 180–197.
9. MCCUTCHEON, A. L. *Latent Class Analysis*. Newbury Park, CA:Sage, 1987.
10. PACCAGNELLA, O., AND VARRIALE, R. *Asset Ownership of the Elderly Across Europe: A Multilevel Latent Class Analysis to Segment Countries and Households*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 383–393.
11. TOMCZYK, S., HANEWINKEL, R., AND ISENSEE, B. Multiple substance use patterns in adolescents—a multilevel latent class analysis. *Drug and alcohol dependence 155* (2015), 208–214.
12. VERMUNT, J. K. Multilevel latent class models. *Sociological Methodology 33*, 1 (2003), 213–239.
13. VERMUNT, J. K. Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis 18* (2010), 450–469.

Distance learning, stress and career-related anxiety during the Covid-19 pandemic: a students perspective analysis

Iodice D'Enza A., Iannario M., Romano R.

Abstract The Covid-19 pandemic made distance learning (DL) the only way to consistently provide an education to students of any age and level. The sudden switch from classroom learning to DL surely had an impact on the students learning experience as well as on their social and psychological spheres. In fact, students adaptation to DL process also depends on Covid-19 induced stress and on the anxiety for future career. The aim of the paper is to analyse an Italian university students survey on DL perception and Covid-19 related stress and anxiety. The proposed approach implements a stacked ensemble method combining data reduction and the Samejima's Graded Response Model.

Abstract *La pandemia da Covid-19 ha reso la didattica a distanza (DAD) uno strumento fondamentale per consentire a studenti di ogni età e grado di continuare a ricevere un'istruzione. Il drastico cambio di erogazione della didattica ha avuto un impatto sia sul processo di apprendimento degli studenti, sia sulla sfera sociale e psicologica di questi ultimi. Il processo di adattamento alla DAD da parte degli studenti è legato anche al loro livello di stress e di incertezza per il futuro. In questo lavoro vengono analizzati i risultati di una indagine tra gli studenti di università italiane su percezione, stress ed ansia legati alla DAD. L'approccio proposto è a due stadi e combina verticalmente metodi di data reduction e il Graded Response Model di Samejima.*

Key words: distance learning; joint data reduction; latent variables

Iodice D'Enza A., Iannario M.
Dipartimento di Scienze Politiche, Università degli Studi di Napoli Federico II
e-mail: iodicede@unina.it; maria.iannario@unina.it

Romano R.
Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Napoli Federico II
e-mail: rosaroma@unina.it

1 Introduction

The Covid-19 pandemic had a major impact on all human activities, and education makes no exception. Distance learning (DL) became the only way to consistently provide an education to students of any age and level. The sudden switch from classroom learning to DL surely had an impact on the students learning experience; the technical setbacks, such as poor internet connection or lack of tools (computers, tablets), are relatively easy to identify, and their effects on the learning process are rather obvious; instead, it is more difficult to study the effects of DL transition on students from a social and psychological perspective. In fact, it is fair to consider the level of adaptation of the students to the DL process as related to the stress for the fear of contagion and the social limitations, and to the anxiety for the future career. In order to investigate the faceted DL impact on students, we considered three different scales proposed and validated in the literature. In particular, we considered the scale proposed by [1] to study the perspective of DL high education students; two further scales were considered, the 'student stress scale', proposed and validated by [11], and the 'fear of Covid-19' scale, proposed by [4], that investigates the future career anxiety. Therefore, a survey is considered that consists of items from the three aforementioned scales, and it is structured in four item-blocks: the first block contains 19 items on students demographics and their proximity to Covid-19 cases; the second block is of 23 items that measure the DL perception of the students; the third and fourth blocks, respectively with 7 and 5 items, aim at measuring students stress and anxiety induced by Covid-19. The survey refers to 1592 students from 60 Italian Universities, with University of Naples and University of Bologna being the most represented, with a 25.9% and 18.5% share, respectively. The response option for the majority of items is a 4 levels Likert-type scale, ranging from *strongly disagree* to *strongly agree*.

The aim of the paper is to analyse the survey results via a stacked ensemble model. In particular, the results from the first scale, concerning the students perspective on different aspects of DL, are synthesized into three ordinal responses, defined via a joint data reduction approach; an IRT model for ordered polytomous variables is considered in order to investigate the item/factor properties and to evaluate the student achievement. Particularly, the Graded Response Model (GRM) is taken into account in the analysis of the three different synthesis to assess the latent continuous variable (the DL perception). Finally, the impact of stress and anxiety on the DL perception is assessed by means of a latent regression GRM. The paper is structured as follows: Section 2 briefly describes the generation process of the ordinal responses; Section 3 introduces the graded response model while Section 4 illustrates the main results and concludes the paper.

2 JDR-based ordinal response

In order to synthesize the students perspective on DL, we apply on the the DL-related items a joint data reduction approach. Under the umbrella of data reduction (DR) fall both dimension reduction (column-wise DR) and clustering (row-wise DR) methods. In the context of unsupervised learning, it is common practice to apply column and row-wise DR one after the other: such practice is referred to as tandem analysis. The first step (data reduction) is independent from the second, and this may lead the tandem analysis to fail retrieving the underlying structure of the data. Joint DR methods seek for a solution that is optimal for both steps: to this end, JDR methods consist of an iterative procedure that alternatively optimise one step given the other. Different JDR methods have been proposed, for continuous ([2],[9]), categorical [3] and mixed-type variables (see [7] for a review). In this paper we refer to cluster correspondence analysis (cluster CA, [8]), a JDR method suitable for survey data. Let $\mathbf{Z}_j$ denote an $n \times p_j$ indicator matrix. That is, each row corresponds to a respondent, and the columns represent the p_j levels of agreement for the j^{th} item. Observed responses are coded by ones and all other elements are zero. Data from multiple items are collected in the block matrix $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$. The application of cluster CA on the DL-related item leads to the definition of a cluster membership variable, and the cluster CA objective is

$$\min \phi_{CCA}(\mathbf{B}^*, \mathbf{Z}_K) = \left\| \mathbf{D}_z^{-1/2} \mathbf{M}\mathbf{Z} - \mathbf{Z}_K \mathbf{G} \mathbf{B}^{*\prime} \right\|^2 \quad s.t. \;\; \mathbf{B}^{*\prime}\mathbf{B}^* = \mathbf{I}_d \tag{1}$$

where $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n'/n$ is a centering operator, $\mathbf{B}^* = \frac{1}{\sqrt{np}} \mathbf{D}_z^{1/2} \mathbf{B}$, $\mathbf{D}_z = diag(\mathbf{Z}'\mathbf{Z})$, $\mathbf{B}$ is the item weights matrix, and $\mathbf{Z}_K$ is the indicator coding of the cluster membership categorical variable.
For the cluster CA application on DL item-blocks, the input parameter K is set to four, just like the levels of the Likert scales. For each block, the cluster membership levels are sorted out according to the group characterizations, in order to obtain a different ordinal response that synthesizes the observed perception of DL in each of the considered blocks. Figure 1 shows the items from the DL-related block that characterize each group: in particular for each item $\mathbf{Z}_j$, $j = 1, \ldots, p$, the standardized residuals matrix of the table $\mathbf{Z}_K'\mathbf{Z}_j$ is computed and the values for each of the p_j levels in each of the K groups are reported in the plot. Large residuals (in absolute value) indicate high group characterization (positive or negative) of the corresponding item levels.

3 The graded response model of DL perception

The three ordinal responses resulting from the JDR procedure are coded so that Y_{is}, with $i = 1, \ldots, n$ and $s = 1, \ldots, 3$, represents the grade for the respondent i to the s^{th} considered DL synthesis. The main aim is now to model the responses on the

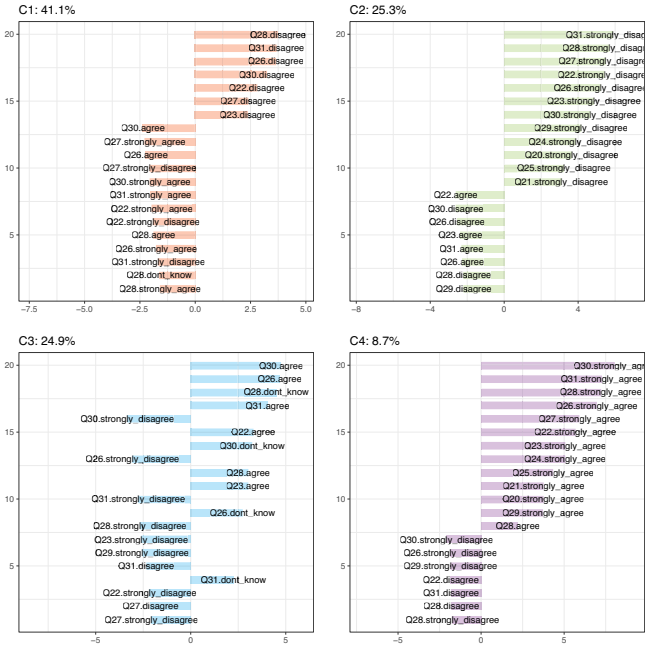


Fig. 1 Item scores for groups characterization: deviations from independence condition

obtained synthetic items without using explanatory variables: one of the candidate models is the Graded Response Model (GRM), introduced by [6] in the context of Item Response Theory (IRT). Such approach yields the cumulative model with proportional assumption where logistic link function is considered [5].
Assuming a constant number of ratings K for each item s, in our analysis $K = 4$, the probability that a respondent i with latent trait θ_i responds by category r or higher, $r = 2,\ldots,K$, to item s ($s = 1,\ldots 3$) is

$$P(Y_{is} \geq r|\theta_i) = F[\gamma_s(\theta_i - \delta_{sr})] \qquad r = 2,\ldots,K, \tag{2}$$

where γ_s are item-specific *discrimination indices* and δ_{sr} are item parameters denoting the *difficulty* of choosing the category r of item s. Parameters δ_{sr} may be also expressed in an additive way as $\delta_{sr} = \delta_s + \alpha_{sr}$, with δ_s representing the difficulty of item s and α_{sr} cut-off point between categories that denote the difficulty of passing from category $r-1$ or smaller to category r or higher. $F(\cdot)$ is a cumulative distribution function, that in the GRM is the logistic cumulative distribution, that is $F(\eta) = \frac{exp(\eta)}{1+\exp(\eta)}$, yielding global (or cumulative) logits.
Thus, the formulation of a *GRM* in terms of logit model follows as

$$\log \frac{P(Y_{is} \geq r|\theta_i)}{P(Y_{is} < r|\theta_i)} = \gamma_s(\theta_i - \delta_{sr}). \tag{3}$$

Notably, a measurement sub-model, such as the GRM in Equation 2, and an explanatory (structural) sub-model achieved by constructing a (multiple) regression model for the latent variable θ_i are identified. To investigate the effect of covariates on the latent variable (Distance Learning perception), it is possible to identify a measurement sub-model, such as the GRM in Equation 2, and an explanatory (structural) sub-model achieved by constructing a (multiple) regression model for the latent variable θ_i (see [10], among others).

4 Results

The main results from explanatory sub-model mentioned in Section 3 are reported in Table 1. In particular, the reported values refer to the estimates of the significant regression coefficients, together with standard errors and z-statistics of the covariates whose estimated regression (significance level at 5%). The Distance Learning

Table 1 Latent regression analysis: regression coefficients (coeff.), standard errors (s.e.), z-statistics.

Covariate	coeff.	s.e.	z-stat
Age	0.0007	0.0004	1.923
Mild stress Isolation	-0.1147	0.1552	-0.739
Stress Isolation	-0.3100	0.1458	-2.126
Intense stress Isolation	-0.5016	0.1419	-3.535
Heavy stress Isolation	-0.6572	0.1411	-4.658
Mild stress Academic	-0.1969	0.0700	-2.814
Stress Academic	-0.4093	0.0639	-6.400
Intense stress Academic	-0.5734	0.0697	-8.226
Heavy stress Academic	-0.7500	0.0761	-9.851
No Covid	-0.0868	0.0362	-2.394
Campania	0.1707	0.0445	3.835
Federico II	0.1426	0.0506	2.820

perception resulted to be significantly less accessible for respondents who did not result positive to Covid-19. Students with perceived heavy stress in relation to academic and isolation consider Distance Learning not accessible. People living in the Campania region, especially students of Federico II consider Distance Learning a good experience. This feeling increases with age of respondents. Finally, it is worth to outline that no significant association resulted for Anxiety scale and other aspects concerning stress scale.

References

1. Amir, L.R., Tanti, I., Maharani, D.A., Wimardhani, Y.S., Julia, V., Sulijaya, B., Puspitawati, R.: Student perspective of classroom and distance learning during covid-19 pandemic in the undergraduate dental study program universitas indonesia. BMC medical education **20**(1), 1–8 (2020)
2. De Soete, G., Carroll, J.D.: K-means clustering in a low-dimensional euclidean space. In: New approaches in classification and data analysis, pp. 212–219. Springer (1994)
3. Hwang, H., Dillon, W.R., Takane, Y.: An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. Psychometrika **71**(1), 161–171 (2006)
4. Mahmud, M.S., Talukder, M.U., Rahman, S.M.: Does 'fear of covid-19'trigger future career anxiety? an empirical investigation considering depression from covid-19 as a mediator. The International journal of social psychiatry (2020)
5. McCullagh, P.: Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological) **42**(2), 109–127 (1980)
6. Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement (1969)
7. van de Velden, M., Iodice D'Enza, A., Markos, A.: Distance-based clustering of mixed data. Wiley Interdisciplinary Reviews: Computational Statistics **11**(3), e1456 (2019)
8. van de Velden, M., Iodice D'Enza, A., Palumbo, F.: Cluster correspondence analysis. Psychometrika **82**(1), 158–185 (2017)
9. Vichi, M., Kiers, H.A.: Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis **37**(1), 49–64 (2001)
10. Wilson, M., De Boeck, P., Carstensen, C.H.: Explanatory item response models: A brief introduction. Assessment of competencies in educational contexts pp. 91–120 (2008)
11. Zurlo, M.C., Cattaneo Della Volta, M.F., Vallone, F.: Covid-19 student stress questionnaire: Development and validation of a questionnaire to evaluate students' stressors related to the coronavirus pandemic lockdown. Frontiers in Psychology **11**, 2892 (2020)

A Tempered Expectation-Maximization Algorithm for Latent Class Model Estimation

Un Algoritmo Tempered Expectation-Maximization per la Stima del Modello a Classi Latenti

Luca Brusa, Francesco Bartolucci and Fulvia Pennoni

Abstract We consider maximum likelihood estimation of the Latent Class model, which is formulated through individual discrete latent variables. We explore tempering techniques to overcome the problem of multimodality of the log-likelihood function. A Tempered Expectation-Maximization algorithm is proposed, which can adequately explore the parameter space and reach the global maximum more frequently than the standard EM algorithm. We assess the performance of the proposed approach by a Monte Carlo simulation study and an application based on data about anxiety and depression in oncological patients.

Abstract *Consideriamo la stima di massima verosimiglianza del modello a classi latenti che è formulato attraverso variabili latenti discrete a livello individuale. Esploriamo le tecniche di tempering per fronteggiare il problema della multimodalità della funzione di log-verosimiglianza. Proponiamo un algoritmo denominato Tempered Expectation-Maximization che permette di esplorare adeguatamente lo spazio dei parametri e di raggiungere il massimo globale più frequentemente rispetto all'usuale algoritmo EM. Per valutare l'efficacia della proposta utilizziamo uno studio di simulazione Monte Carlo e un'applicazione basata su dati reali riguardanti misure di ansia e depressione in pazienti oncologici.*

Key words: annealing, finite mixture models, latent variables, local maxima.

Luca Brusa,
Department of Economics, Management and Statistics, University of Milano-Bicocca,
e-mail: luca.brusa@unimib.it

Francesco Bartolucci,
Department of Economics, University of Perugia,
e-mail: francesco.bartolucci@unipg.it

Fulvia Pennoni,
Department of Statistics and Quantitative Methods, University of Milano-Bicocca,
e-mail: fulvia.pennoni@unimib.it

1 Introduction

The Latent Class (LC) model [1] is very popular for the analysis of categorical, and in particular binary, response variables. It is formulated by assuming the existence of individual-specific latent variables having a discrete distribution. This model may be seen as semi-parametric since, differently from other models based on continuous latent variables, no parametric assumptions are formulated on the distribution of such variables. The LC model may be seen as a finite mixture model, and it is employed to cluster subjects on the basis of a set of categorical, typically binary, responses.

Despite maximum likelihood estimation of the LC model may be simply performed using the Expectation-Maximization (EM) algorithm [2, 3], a well-known drawback of this estimation method is related to the multimodality of the likelihood function that is due to the inclusion of discrete latent variables. The consequence is that the *global* maximum of the likelihood is not ensured to be reached, and a proper initialization of the estimation algorithm is crucial. A multi-start strategy is typically adopted based on deterministic and random rules to explore the parameter space adequately. However, this approach may be computationally intensive, and it does not guarantee convergence to the global maximum.

In order to face the multimodality of the likelihood function, we propose a Tempered EM (T-EM) algorithm able to explore the parameter space adequately. In an optimization context, tempering [4] consists of re-scaling the objective function depending on a variable, known as temperature, which controls the prominence of global and local maxima. High temperatures allow us to explore wide regions of the parameter space, avoiding the maximization algorithm being trapped in non-global maxima; low temperatures, instead, guarantee a sharp optimization in a local region of the parameter space. By properly tuning the sequence of temperature values, the procedure is gradually attracted toward the global maximum, escaping in this way local sub-optimal solutions. As a future development, this procedure will also be applied to estimate the parameters of the hidden Markov (HM) models for the analysis of longitudinal data [5].

The rest of the paper is organized as follows. Section 2 outlines the LC model formulation and maximum likelihood estimation through the EM algorithm. Section 3 provides details on the proposed T-EM algorithm. Section 4 summarizes the main findings of the simulation study and the results of an application concerning patients' responses to ordinal items measuring anxiety and depression.

2 Latent Class Model and Expectation-Maximization Algorithm

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ir})'$ denote the vector of r categorical response variables for individual $i = 1, \ldots, n$; each variable has the same number c of categories, labeled from 0 to $c-1$. The LC model relies on individual-specific discrete latent variables U_i with k support points that identify the latent classes in the population. The model

parameters are the conditional probabilities of each response variable given the latent variable, denoted by $\phi_{jy|u} = p(Y_{ij} = y|U_i = u)$, and the weight of each latent class, denoted by $\pi_u = p(U_i = u)$. The resulting *manifest distribution* is then

$$p(\boldsymbol{y}_i) = \sum_{u=1}^{k} \pi_u \prod_{j=1}^{r} \phi_{jy_{ij}|u},$$

where $\boldsymbol{y}_i$ denotes a realization of $\boldsymbol{Y}_i$.

In order to estimate the model parameters, collected in the vector $\boldsymbol{\theta}$, on the basis of a sample of n independent observations $\boldsymbol{y}_i$, we rely on the log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i).$$

This function is maximized through the EM algorithm on the basis of the *complete data log-likelihood*, which may be written as

$$\ell^*(\boldsymbol{\theta}) = \sum_{j=1}^{r} \sum_{u=1}^{k} \sum_{y=0}^{c-1} a_{juy} \log \phi_{jy|u} + \sum_{u=1}^{k} b_u \log \pi_u,$$

where $a_{juy} = \sum_{i=1}^{n} I(u_i = u, y_{ij} = y)$ is the frequency of subjects that are in latent class u and responded by y at the j-th response variable and $b_u = \sum_{i=1}^{n} I(u_i = u)$ is the number of sample units in latent class u, with $I(\cdot)$ denoting the indicator function. The EM algorithm alternates the following two steps until a suitable convergence criterion is satisfied:

- **E-Step**: compute the conditional expected value of $\ell^*(\boldsymbol{\theta})$, given the observed data and the value of the parameters at the previous step;
- **M-Step**: maximize the expected value of the log-likelihood function $\ell^*(\boldsymbol{\theta})$ and so update the model parameters.

In particular, the E-step is based on the posterior probabilities

$$q(u|\boldsymbol{y}_i) = p(U_i = u|\boldsymbol{Y}_i = \boldsymbol{y}_i) = \frac{\pi_u \prod_{j=1}^{r} \phi_{jy_{ij}|u}}{p(\boldsymbol{y}_i)},$$

on the basis of which the expected values of the frequencies a_{juy} and b_u are simply obtained.

The EM algorithm is straightforward to implement, it is able to converge in a stable way to a local maximum of the log-likelihood function, and it is used for parameter estimation in many available packages [3]. However, this log-likelihood function may be multimodal, especially when the model has many latent classes. For this reason, several starting values of the parameters in $\boldsymbol{\theta}$ are typically used, and the solution corresponding to the highest log-likelihood is then selected as the maximum likelihood estimate, denoted by $\hat{\boldsymbol{\theta}}$. In the next section, we show an alternative solution based on the proposed T-EM algorithm.

3 Tempered Expectation-Maximization Algorithm

We introduce a T-EM algorithm [6, 7] by defining the following modified posterior probabilities:

$$\tilde{q}^{(T_h)}(u|\boldsymbol{y}_i) \doteq \frac{q(u|\boldsymbol{y}_i)^{1/T_h}}{\sum_{u=1}^{k} q(u|\boldsymbol{y}_i)^{1/T_h}},$$

where $(T_h)_{h\geq 1}$ is a suitable sequence of temperature values, under the constraint that $T_h \to 1$ as $h \to \infty$, where h is the algorithm iteration number.

The E-step and M-step of the T-EM algorithm are implemented as follows by modifying those of the original EM algorithm:

- **E-Step**: compute

$$\tilde{b}_u^{(T_h)} = \sum_{i=1}^{n} \tilde{q}^{(T_h)}(u|\boldsymbol{y}_i) \qquad \text{and} \qquad \tilde{a}_{juy}^{(T_h)} = \sum_{i=1}^{n} I(y_{ij} = y)\, \tilde{q}^{(T_h)}(u|\boldsymbol{y}_i);$$

- **M-Step**: update the parameters as

$$\pi_u^{(T_h)} = \frac{\tilde{b}_u^{(T_h)}}{n} \qquad \text{and} \qquad \phi_{jy|u}^{(T_h)} = \frac{\tilde{a}_{juy}^{(T_h)}}{\tilde{b}_u^{(T_h)}}.$$

Given the above setting, it is clear that the tempering profile (i.e., the sequence $(T_h)_{h\geq 1}$) may have a deep impact on the performance of the proposed algorithm. In fact, increasing the temperature value has the effect of flattening the profile of the log-likelihood, thereby reducing the chance that the algorithm will get trapped into local maxima. In particular, $T_h \to +\infty$ yields $\tilde{q}^{(T_h)}(u|\boldsymbol{y}_i)$ to a uniform distribution, while $T_h = 1$ makes $\tilde{q}^{(T_h)}(u|\boldsymbol{y}_i)$ equal to the standard posterior probability $q(u|\boldsymbol{y}_i)$. Therefore, the only necessary condition for proper convergence is that the temperature value T_h tends towards 1 as the iteration counter increases.

We consider the following two tempering profiles: (*i*) a decreasing exponential profile:

$$T_h = \frac{1 + e^{h/\alpha - \beta}}{e^{h/\alpha - \beta}}, \tag{1}$$

with constants $\alpha \geq 1$ and $\beta \geq 0$, which has the advantage to be easy to tune; (*ii*) a non-monotonic profile [6] with oscillations of gradually smaller amplitude:

$$T_h = \tanh\left(\frac{h}{2r}\right) + \left(T_0 - \beta \cdot \frac{2\sqrt{2}}{3\pi}\right) \cdot \alpha^{h/r} + \beta \cdot \operatorname{sinc}\left(\frac{3\pi}{4} + \frac{h}{r}\right), \tag{2}$$

with constants $r,\ T_0,\ \beta > 0$, and $0 < \alpha < 1$. The latter choice has more parameters to tune, but it guarantees a very high level of flexibility. The proposed procedure requires selecting the set of tempering parameters by a grid-search.

4 Simulation Results and Applicative Example

Within the simulation study, we randomly drew several samples of size $n = 500$ from an LC model with $r = 6$ responses, having $c = 3$ categories, assuming $k = 3$ latent classes, and for each of these samples we estimated a misspecified LC model with $k = 4$ latent classes. In particular, we fitted the LC model 100 times for each sample, always using different sets of random starting values. We used the standard EM algorithm and the proposed T-EM algorithms denoted by M. T-EM, when the monotonic tempering profile (1) is used, and by O. T-EM, when the oscillating tempering profile (2) is employed. The convergence of the algorithms is checked on the basis of the relative log-likelihood difference; regarding the algorithm initialization, we adopted a random starting rule based on normalized random numbers drawn from a uniform distribution from 0 to 1.

We carried out a grid-search for the tempering parameters for each sample, and we evaluated the setting that ensures the best performance. We noticed that the method is not excessively sensitive to the tempering parameters: once the grid-search sets such parameters, they remain valid over datasets sharing the same features (e.g., the same number of response variables and categories). Therefore, this preliminary procedure may be less time consuming than the current practice of estimating the model many times with random initial parameters.

In Table 1 we show some results obtained as described above about the EM and T-EM algorithms: for each of six considered samples, we report the mean and the median of the 100 log-likelihood values at convergence. From this table, it is clear the advantage of the use of the tempering modification. In particular, the oscillating version of the T-EM algorithm exhibits the best performance, slightly outperforming also the monotonic version in most cases. We also considered the following criteria: (*i*) dispersion of the resulting maxima measured by the standard deviation; (*ii*) proportion of times the obtained maximum is close enough to the global one (based on the 100 repetitions); (*iii*) dispersion of the estimated probability vectors (π, sorted into descending order). From the results reported in Table 2 we notice a clear superiority of T-EM algorithm over the standard EM algorithm: in each scenario the best results are obtained with the modified algorithm, and only for the fourth sample the improvement is mild.

Table 1 Mean and median of log-likelihood values at the maximum, with EM and T-EM algorithm using monotonic (M. T-EM) and oscillating (O. T-EM) tempering profiles on simulated data; each row refers to a specific sample, and values in bold highlight the best results.

Mean			Median		
EM	M. T-EM	O. T-EM	EM	M. T-EM	O. T-EM
-2,847.2879	-2,846.5392	**-2,845.4207**	-2,846.7726	-2,844.9000	**-2,844.8369**
-2,864.7102	-2,864.8438	**-2,864.6754**	-2,864.8575	-2,864.7875	**-2,864.7336**
-2,848.3929	-2,846.4819	**-2,846.4817**	-2,849.0982	-2,846.4819	**-2,846.4817**
-2,798.8988	-2,798.7510	**-2,798.3810**	-2,799.5792	-2,797.9326	**-2,797.5355**
-2,846.4159	**-2,843.4433**	-2,843.4672	-2,847.5666	**-2,843.4433**	-2,843.4449
-2,832.5526	**-2,831.5140**	-2,831.5808	-2,831.9158	**-2,831.2970**	**-2,831.2970**

Table 2 Dispersion (SD) and proportion (Freq) of global maxima and dispersion of the estimated probabilities ($\boldsymbol{\pi}$); each row refers to a specific sample, and values in bold highlight the best results.

SD (Max.)			Freq. (Glob. Max.)			Dispersion of $\boldsymbol{\pi}$		
EM	M. T-EM	O. T-EM	EM	M. T-EM	O. T-EM	EM	M. T-EM	O. T-EM
2.2106	2.0383	**1.3565**	0.63	0.67	**0.91**	0.0057	0.0042	**0.0009**
1.1132	**0.7430**	1.0831	0.89	**0.93**	0.85	0.0029	0.0024	**0.0023**
1.9395	**0.0000**	0.0000	0.64	**1.00**	**1.00**	0.0067	**0.0000**	0.0000
1.7738	1.6139	**1.5742**	0.49	0.51	**0.62**	0.0038	0.0033	**0.0029**
2.8364	**0.0000**	0.0298	0.47	**1.00**	**1.00**	0.0024	**0.0000**	0.0002
1.5343	0.3525	**0.3404**	0.88	**1.00**	**1.00**	0.0043	0.0005	**0.0004**

Comparing the two types of tempering profile, we note that the oscillating profile often outperforms the monotonic one; only when the results exhibit an almost absolute perfection (dispersion approximately equal to 0 and proportion of global maxima close to 1, as in samples 3 and 5), the monotonic profile reaches a slightly better performance. This observation suggests that if the model is not too complex, this choice is generally preferable, while in other cases, the oscillating profile guarantees better results. Similar T-EM algorithms are implemented for estimating the HM model and preliminary results of the simulation study show the same improvements with respect to the standard EM algorithm.

We also considered data deriving from the administration of 14 ordinal items with three categories measuring anxiety and depression in 201 oncological patients [3]. A misspecified LC model is estimated with $k = 4$ latent classes with the following tempering parameters: $\alpha = 42$ and $\beta = 1.5$ for the monotonic profile; $r = 90$, $T_0 = 10$, $\beta = 20$, and $\alpha = 0.8$ for the oscillating profile. We notice that the T-EM algorithms outperform the classic version of the EM algorithm because they always converge to the same value that is presumably the global maximum, while the classical EM algorithm spreads out over a wide range of estimates.

References

1. Goodman, L.A.: Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biom.*, **61**, 215-231 (1974).
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from in-complete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B*, **39**, 1–38 (1977).
3. Bartolucci, F., Bacci, S., Gnaldi, M.: `MultiLCIRT`: An `R` package for multidimensional latent class item response models. *Comput. Stat. Data Anal.*, **71**, 971–985 (2014).
4. Sambridge, M.: A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophys. J. Int.*, **196**, 357–374 (2013).
5. Bartolucci, F., Farcomeni, A., Pennoni, F.: *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, Boca Raton (2013).
6. Lartigue, T., Durrleman, S., Allassonnière, S.: Deterministic approximate EM algorithm; Application to the Riemann approximation EM and the tempered EM. *arXiv:2003.10126* (2021).
7. Ueda, N., Nakano, R.: Deterministic annealing EM algorithm. *Neural Netw.*, **11**, 271–282 (1998).

2.14 Statistics for finance high frequency data, large dimension and networks

The Italian debt not-so-flash crash
Il flash crash del debito italiano

Maria Flora and Roberto Renò

Abstract We document a "flash" crash in the Italian debt market on May 29, 2018 which recovered in the subsequent two days. Selling pressure in the secondary market due to a change of the Italian political scenario was not absorbed properly and caused overreaction. Using a regime-switching model, we estimate a direct cost for Italian taxpayers around 450 million euros just in that week, due to auctions that were taking place on May 30, plus several long-term indirect costs in terms of increased volatility and harsher liquidity in the following months. Flash crashes represent thus a serious threat to financial stability even in systemic, economically central markets like sovereign debt.

Abstract *Studiamo il flash crash nel mercato secondario dei titoli di stato italiani del 29 maggio 2018, riassorbito nei due giorni successivi. Il crash è stato dovuto a forti richieste di vendite nel mercato secondario causate da un cambio dello scenario politico. Viene stimato, utilizzando un modello a cambio di regimi, un costo diretto per i contribuenti italiani di 450 milioni di euro, dovuto principalmente alle aste di titoli di stato del 30 maggio, e costi indiretti in termini di volatilità in eccesso e deterioramento della liquidità nei mesi successivi. I flash crash rappresentano quindi una seria minaccia alla stabilità del sistema finanziario anche in mercati liquidi come nel caso del debito sovrano.*

Key words: Financial fragility, Italian bonds, flash crash, drift explosion

Maria Flora
CREST, ENSAE, Institut Polytechnique de Paris, e-mail: maria.flora@ensae.fr

Roberto Renò
Università di Verona, e-mail: roberto.reno@univr.it

1 Introduction

The flash crash of May 6, 2010 in the US stock market, triggered by a huge selling trade in the E-mini futures market (CFTC and SEC, 2010), has attracted the attention of traders, institutions and academics (see, e.g., Madhavan, 2012; Kirilenko, Kyle, Samadi, and Tuzun, 2017; Menkveld and Yueshen, 2019). The event shed light on a market vulnerability which appears to affect financial markets quite often, and increasingly over time (Christensen, Oomen, and Renò, 2020; Golub, Keane, and Poon, 2017). The natural question is: what is the impact of flash crashes on market activity and social welfare? The transient impact of these events may lead to think they do not matter much. Bank of England (2019) writes: "Flash episodes have not, as yet, had financial stability consequences." On the other end, financial stability is defined as the "ability to facilitate and enhance economic processes, manage risks, and absorb shocks" (Schinasi, 2004).

Our research contributes to this literature by focusing on a deep crash that happened in the Italian sovereign bond markets on May 29, 2018, when a shock due to macro news (change of government after political elections) was not absorbed properly. We show that this event had large consequences on the market. Direct costs came from auctions which took place exactly at the bottom of the crash. The crash was indeed particularly unfortunate for Italian taxpayers, since the Treasury was auctioning at 11:00 of May 30. We estimate the money lost by the Treasury because of the crash (which involved one CCT and two BTPs, for a total of more than 6 billion euros offered) was roughly 0.45 billion euros (see Flora and Renò, 2020 for details).

Using a formal, recently developed statistical test, we show that the crash was due to a "drift burst" (Christensen, Oomen, and Renò, 2020), that is a large (downward, in this case) trend in prices localized in a short time interval. The distinction between a volatility move and a drift move is not immaterial. Indeed, large volatility is possible even in an efficient and perfectly liquid market. Large drift is instead typically associated with flash crashes, that is with inefficiency and market frictions.

2 Empirical analysis

We start our analysis by showing that the crash of May 29 was associated to a large drift, not to large volatility, using nonparametric statistics to show that this evidence is robust to model specification. To identify large drifts we use the recent test statistics of Christensen, Oomen, and Renò (2020), henceforth COR. This is a non-parametric test which uses $n+1$ log-price observations $X_0, \ldots, X_n$ observed at times $t_0, \ldots, t_n$. The test can be formally expressed, at time-point t, as:

$$T_t^n = \sqrt{\frac{h_n}{K_2}} \frac{\hat{\mu}_t^n}{\hat{\sigma}_t^n}, \tag{1}$$

where, for $t \in [0,T]$,

$$\hat{\mu}_t^n = \frac{1}{h_n} \sum_{i=1}^{n} K\left(\frac{t_{i-1}-t}{h_n}\right) \Delta_i^n X \tag{2}$$

is a localized estimator of the drift, in which h_n is a bandwidth parameter measuring the extent of the localization, and K is a suitable kernel, while $\hat{\sigma}_t^n$ is a localized, pre-averaged and HAC-corrected (Andrews, 1991) estimator of the spot volatility.

In a large sample of liquid futures data (including US Treasury bonds), COR show that large values of the test statistics are almost always associated with large trading volume, and short-term price reversals, which are typical of flash crashes. In a related paper, using data on the French market, Bellia, Christensen, Kolokolov, Pelizzon, and Renò (2019) show that large values of the test statistics are unambiguously associated with reversals and evaporating liquidity. Flora and Renò (2020) generalize the t-statistic in a V-statistic, which is proposed to test for market inefficiency.

We implement the test statistic (1) on selected Italian sovereing bonds in 2018 and 2019. We choose a bandwidth h_n for the drift equal to 2 days, while we base the volatility on a 10-day bandwidth. We adopt a left-sided exponential kernel $K(x) = \exp(-|x|)$, for $x \leq 0$. Finally, for each trading day, we compute the minimum of the calculated t-statistics.[1]

Figure 1 reports the results of the drift burst test. The daily minimum of the test statistics T_t^n for the five instruments first crosses the -3 value approximately two weeks before the crash event. The five T_t^n all peak on May 29, with BTP-1nv23 9% being the most affected: the value of the t-stat is slightly below -6, and thus provides strong evidence of a dominating trend in prices. Few subsequent negative peaks in the test-statistics are observed in 2019.

What may cause a large value of the test-statistics? The neatest economic interpretation comes from the intermediation theory of Grossman and Miller (1988). In their model, a trader looking for immediacy is willing to sell to M market makers a volume s of a security. Market makers accept to trade immediately but with a price concession. A key prediction of the model is that the transient mispricing should be more severe in a market with poor liquidity. We compute two realized liquidity measures that can be inferred directly from transaction prices. The first is a measure of price impact which is close to the Amihud (2002) measure (we modify it to take into account the irregular sampling of trades). The measure is implemented as follows. For each trade t and bond i in the sample, we compute

[1] To deal with overnight gaps, we construct a new time vector for each time series, associated to the original one, where the time (in milliseconds) elapsed from the closing of day $t-1$ to the next available open price is equal to

$$\tilde{t} = \frac{\sigma_{overnight}}{\sigma_{intraday}} \hat{t}\,.$$

Here, $\sigma_{overnight}$ is the standard deviation of overnight returns, defined as the price appreciation or depreciation between market close of day $t-1$ and market open of day t, while $\sigma_{intraday}$ is the standard deviation of intraday returns, defined as the price appreciation or depreciation between market open and close of the same day. Finally, $\hat{t}$ is the time, in milliseconds, elapsed from market open (9 a.m.) and close (5:30 p.m.).

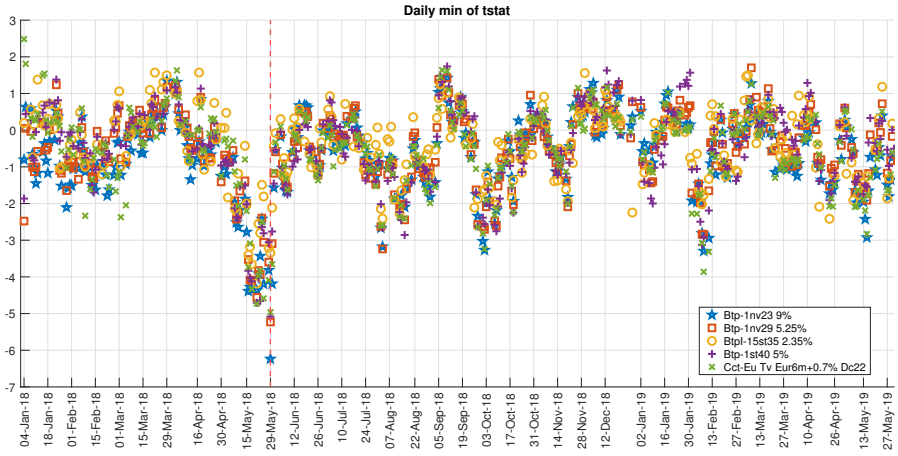


Fig. 1 Daily minimum of the drift burst test statistics proposed by Christensen, Oomen, and Renò (2020). Large values of the test statistics signal market distress. The dashed-red line is May 29, 2018.

$$\text{Amihud}^*_{i,t} = \frac{|\Delta \log(p_{i,t})|}{V_{i,t}\sqrt{T_{i,t}}}, \tag{3}$$

where $\Delta \log(p_{i,t}) = \log(p_{i,t}) - \log(p_{i,t-1})$ is the log-return between trade $t-1$ and trade t, $V_{i,t}$ is the volume of the $t-th$ trade, and $T_{i,t}$ is the time, in milliseconds, between trade $t-1$ and trade t. Figure 2 shows the daily median of the measure in (3) for four bonds. As expected, the measure peaks in the crash week. Most importantly, the impact of the crash is to increase persistently the illiquidity measure in the market, with a transient effect that lasts several weeks after the crash.

The second statistics we employ is the one proposed in Roll (1984), which we compute day-by-day:

$$\text{Roll} = 2\sqrt{-\text{Cov}\left(\Delta p_{i,t}\Delta p_{i,t-1}\right)}, \tag{4}$$

where $\Delta p_{i,t}$ is the price difference between two consecutive transactions. This measure can be regarded as a proxy for the effective bid-ask spread: the higher its value, the higher the costs in terms of immediacy for the investors. Figure 3 shows the daily value of the Roll measure, which has a very similar dynamics to that of the Amihud* measure. Again, there is a marked spike during the crash, and a persistent impact on the market illiquidity, with a transient impact that lasts several weeks.

Importantly, the test could have been used to identify market distress from market transactions themselves. Using the precautionary threshold of -4.5, for the BTP 1Nov23 this line would have been crossed, using the same procedure described here in real time, at 9:53 of May 29, more than one day ahead of the auction of May 30, and never more in our sample. For the BTP 1Nv29, the line was crossed twice in the sample: at 17:23 of May 21, 2018 and at 10:26 of May 29. For the BPT 15St40,

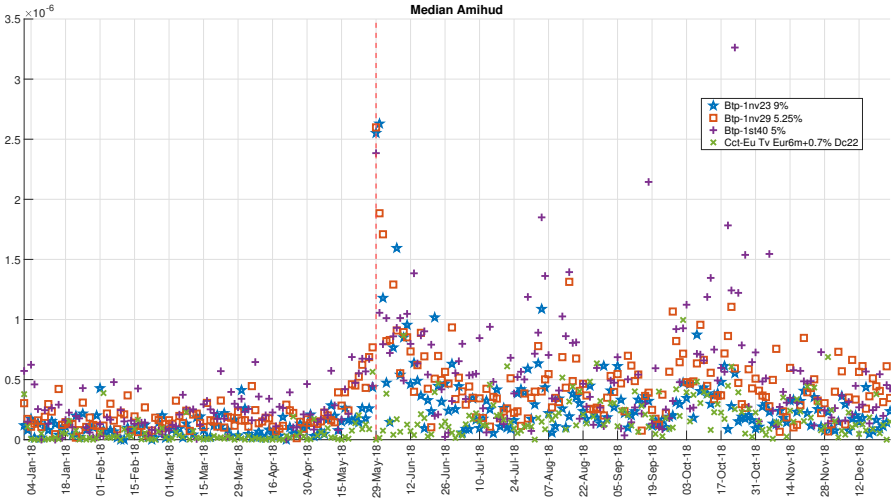


Fig. 2 Daily median of the Amihud* measure, as defined in Eq. (3). The dashed-red line is May 29, 2018.

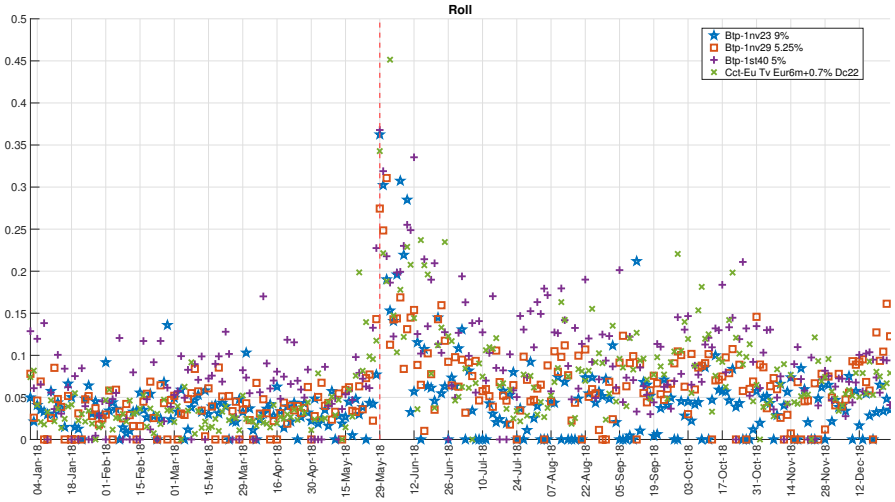


Fig. 3 Daily measures of the Roll illiquidity estimator. The dashed-red line is May 29, 2018.

the line was crossed three times: 17:15 of May 21, 11:39 of May 23 and 10:36 of May 29. The CCT also crossed the line three times: 16:44 of May 21, 15:57 of May 25 and 10:31 of May 29. The BTPI never crossed the line, being the instrument with by far less transactions in our sample. Thus, a simple monitoring of the market would have at least informed market regulators that the market was distressed in the morning of May 29, and even with some "tremors" in the previous days.

3 Conclusions

We document a severe crash occurred in the secondary Italian debt market, with huge consequences for the primary market and the Italian taxpayers, which we quantify in a loss for the Treasury around half a billon euros (Flora and Renò, 2020). Similar losses have been experienced during the COVID-19 pandemic (Ferrara, Flora, and Renò, 2021). This finding is particularly important for financial stability, since it illustrates that the occurrence of phenomena similar to flash crashes is likely even in a systemic, allegedly liquid market like that for the Italian bonds, and that their impact can be destructive. Our research thus contributes to the debate of whether regulators should worry about the occurrence of flash crashes, and the conclusion of this paper is that they definitively should.

References

Amihud, Y., 2002, "Illiquidity and stock returns: cross-section and time-series effects," *Journal of Financial Markets*, 5(1), 31–56.

Andrews, D. W. K., 1991, "Heteroscedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 59(3), 817–858.

Bank of England, 2019, *Financial Stability Report*.

Bellia, M., K. Christensen, A. Kolokolov, L. Pelizzon, and R. Renò, 2019, "High-Frequency Trading During Flash Crashes: Walk of Fame or Hall of Shame?," Working paper.

CFTC and SEC, 2010, *Findings regarding the market events of May 6, 2010*.

Christensen, K., R. C. A. Oomen, and R. Renò, 2020, "The Drift Burst Hypothesis," *Journal of Econometrics*, Forthcoming.

Ferrara, G., M. Flora, and R. Renò, 2021, "The COVID-19 Auction Premium," Working paper.

Flora, M., and R. Renò, 2020, "V-shapes," Working paper.

Golub, A., J. Keane, and S.-H. Poon, 2017, "High Frequency Trading and Mini Flash Crashes," Working paper.

Grossman, S., and M. Miller, 1988, "Liquidity and market structure," *Journal of Finance*, 43(3), 617–633.

Kirilenko, A., A. S. Kyle, M. Samadi, and T. Tuzun, 2017, "The Flash Crash: High frequency trading in an electronic market," *Journal of Finance*, 3, 967–998.

Madhavan, A. N., 2012, "Exchange-traded funds, market structure and the Flash Crash," *Financial Analysts Journal*, 68(4), 20–35.

Menkveld, A. J., and B. Z. Yueshen, 2019, "The Flash Crash: A cautionary tale about highly fragmented markets," *Management Science*, 10(10), 4470–4488.

Roll, R., 1984, "A simple measure of the implicit bid-ask spread in an efficient market," *Journal of Finance*, 39, 1127–1139.

Schinasi, G. J., 2004, "Defining Financial Stability," IMF Working Paper.

3 Solicited Sessions

3.1 Advances in social indicators research and latent variables modelling in social sciences

A composite indicator to measure frailty using administrative healthcare data

Un indicatore composito per misurare la fragilità utilizzando i dati sanitari amministrativi

Silan, M., Brocco, R., Boccuzzo, G.

Abstract The aim of this work is to propose an indicator to measure frailty in old subjects using administrative healthcare data. We propose a composite indicator that exploits a parsimonious set of variables, that are assembled through the use of partially ordered sets (poset) theory.

This proposal boasts several strengths that make it a valuable choice to identify frail old individuals: the use of administrative health data sources, the involvement of multiple health outcomes, the use of a parsimonious set of variables, and the use of poset theory that limits assumptions needed in the construction of the composite indicator.

Abstract *L'obbiettivo di questo lavoro consiste nel proporre un indicatore di fragilità che serva a misurare la fragilità nei soggetti anziani fragili utilizzando le banche dati amministrative sanitarie. La nostra proposta consiste in un indicatore composito formato da un insieme parsimonioso di variabili aggregate sfruttando l'approccio degli insiemi parzialmente ordinati (poset).*

Questa proposta vanta diversi punti di forza che la rendono una valida soluzione per identificare i soggetti anziani fragili: l'utilizzo di flussi amministrativi sanitari, il coinvolgimento di numerosi esiti di salute, l'utilizzo di un insieme

[1] Silan, M., Department of Statistical Sciences, University of Padua; email: margherita.silan@unipd.it

Brocco, R., Department of Statistical Sciences, University of Padua; email: rachele.brocco@unipd.it

Boccuzzo, G., Department of Statistical Sciences, University of Padua; email: giovanna.boccuzzo@unipd.it

parsimonioso di variabili e l'utilizzo della teoria dei poset che limita le assunzioni necessarie alla costruzione dell'indicatore composito.

Key words: Frailty indicator, Administrative healthcare data, Poset theory, Aging, Multiple outcomes

Introduction

Given the progressive ageing of Italian and European populations, chronic diseases attributable to ageing are rising steeply, calling for new strategies for health resources management and implementation of prevention policies. The COVID-19 pandemic has shown the importance of identifying frail subjects in order to safeguard their already compromised health. Moreover, the stratification of the population according to their health needs is a fundamental goal, claimed as the first step of the Italian National Program for Chronic Diseases. The main goal of our work is to provide a validated and reliable tool to identify frail old individuals, even those that are not already known by health or social services, to the Health Unit 6 (whose territory comprises 101 municipalities in the Padua province), possibly extendable to all Italian Local Health Units.

Despite the growing interest about the identification of frail individuals, frailty is defined as a syndrome in desperate need of description and analysis [2]. However, there are two fundamental aspects about it shared by majority of literature on this topic: frailty as a complex and multidimensional condition, involving multiple functional domains; and frailty as a state of susceptibility to adverse health outcomes, such as death or urgent hospitalization [1, 3].

Frail subjects are often older, they have special and wider care needs, but they are not always known and assisted by the health and social services. Indeed, their identification and the quantification of individual frailty level become really important in order to improve both the distribution of public resources and the quality of life of older individuals.

In literature, several works are focussed on building a frailty indicator to compute the prevalence of frailty in certain populations. Most of these studies use data collected on a sample of the reference population through self-administered questionnaires. However, from a policy implementation perspective, a measure of individual frailty level should be available for the whole population. This is possible only using healthcare administrative data. Thanks to an agreement with the Health Unit 6, we were able to use administrative data regarding their assisted population.

All in all, the characteristics that make a frailty indicator useful, efficient and well-focused may be summarized as three: multidimensionality, as frailty involves several functional domains; ability to predict negative outcomes, since frail individuals are more exposed to them than other people and since these events need to be prevented; and universality, as exploiting administrative data a frailty measure may be computed for the whole population.

Construction of the frailty indicator

The construction of our frailty indicator is articulated in six steps. Since, according to literature, frail individuals are more exposed to the risk of experiment negative outcomes related with frailty condition, the first step consists on a large literature recognition to pinpoint which are the negative events that should be considered and enlist all their risk factors that may be generated using only administrative data.
Even for the identification of negative outcomes related with frailty there is no agreement on literature. Thus, as a second step, we considered an extensive set of outcomes in order to not neglect important aspects of frailty, with the intention of evaluate a-posteriori the exclusion of some of them. Seven outcomes were selected as being related with the frailty condition and collected from administrative healthcare data-flows: death, urgent unplanned hospitalization, access to the emergency room (ER) with red code, avoidable hospitalization, hip fracture, dementia and disability.
Outcomes are not directly included in the computation of the frailty indicator. However, they are necessary for the selection of the variables that constitute the composite indicator. Thus, the third step is the codification of the variables that predict selected outcomes and their construction with administrative healthcare databases. These variables are the result of the union of the information coming from several data sources, such as the participation in the prescription charges, as well as the territorial drug prescriptions and diagnosis assigned in hospital discharge records and in accident and emergency databases. The final dataset contains: a) 67 variables collected and computed through a deterministic record linkage of several different data sources (2016–2017); b) seven outcomes observed on administrative databases in 2018, with reference to all the individuals assisted by Health Unit 6 in the whole period 2016-2018 aged 65 years and over in 2018. Data referred to 2019 were used to evaluate indicator's robustness through time. In summary, the years 2016 and 2017 were used for the calculation of the variables that compose the indicator, the year 2018 for the calculation of outcomes, and the year 2019 for the robustness analysis. These are the most recent available data.
Having such a large amount of variables, the main aim of the fourth step of the frailty indicator's construction is focused on variables selection, based on their ability to predict all the seven considered outcomes. In order to do so, we estimated for every outcome 100 logistic regression models on different balanced samples of the whole population, selecting variables with a stepwise criterion and saving two variables: a dummy that records the presence of every explanatory variable in the final model and the order of entrance of every variable in the final model.
Then, the selection of variables is guided by both the median order of entrance and the percentage of presence of every variable in the models. Indeed, a smaller set of variables were selected including all the variables with median order of entrance below or equal 20 and percentage of presence higher than 60% for at least three outcomes, considering all the seven outcomes. However, we performed also a

sensitivity analysis, to better consider the choice of those thresholds and evaluate how much the set of outcomes affects the variables selection.
Then, the fifth step involves the aggregation of variables using partially ordered set (poset) theory. This method exploits the ordinal information that comes from the dataset and it orders profiles attributing to all of them an approximation of their average rank (AR), for more information see [4]. The normalized AR becomes an indicator that describes the relative position of an individual in the distribution of a latent concept, which is frailty condition.
However, increasing the number of variables, profiles are more difficult to compare and the number of incomparable profiles in the poset grows. In practical terms, this means that the approximation of the average rank is less sharp and the computational time to get it greatly increases. Thus, an additional variables selection is needed. In the sixth step variables selection is conduced following a forward logic to maximize the sum of the Area Under the ROC Curves (AUC) of the chosen outcomes.
The final composite indicator for frailty comprises eight variables: age, disability, kidney failure, mental diseases, Parkinson disease, number of accesses to the emergency room (ER) with yellow code, number of different drug prescriptions and Charlson index, that is a co-morbidity index.

Performance and use of the frailty indicator

In order to evaluate the frailty indicator's ability to predict negative outcomes, we first observe the AUC produced by the indicator for each outcome: 0.84 for death, 0.67 for urgent unplanned hospitalization, 0.81 for access to the emergency room (ER) with red code, 0.79 for avoidable hospitalization, 0.77 for hip fracture, 0.83 for dementia and 0.64 for disability.
Then, we compared indicator's values obtained for subjects who have undergone at least one (AUC equal to 0.69), two (AUC 0.78), three (AUC 0.82) or four (AUC 0.83) outcomes with those obtained for the rest of the population. In all cases the distribution of values of the indicator is higher for subjects who have suffered one or more outcomes. In particular, considering the last percentile of individuals with the highest values of the frailty indicator (2045 people): in 2018 almost 32% of them died, 10.3% underwent access to the emergency room with a red code and 46% underwent urgent hospitalization; over 56% suffered at least one outcome linked to frailty, 40.6% suffered at least two outcomes and 23.6% at least three outcomes.
Moreover, even if the variables included in the computation of the frailty indicator only concern individual health and use of health services, the indicator is higher also for those individuals who requested the exemption of health expenses for low-income. Thus, it includes also an implicit dimension that is social frailty.
From a practical point of view, the frailty indicator is extremely useful to stratify individuals according to their healthcare needs. In this sense, the whole population may be seen as a sort of pyramid. At its base there is the majority of the population, with old individuals that are mainly healthy. In the central portion of the pyramid, on

the other hand, there are patients characterized by a single chronic pathology. At the top of the pyramid we find the so-called frail and complex subjects, who have more than one chronic condition that must be taken care of with personalized care programs and a more extensive use of resources, but who also represent a very small portion of the total assisted population.

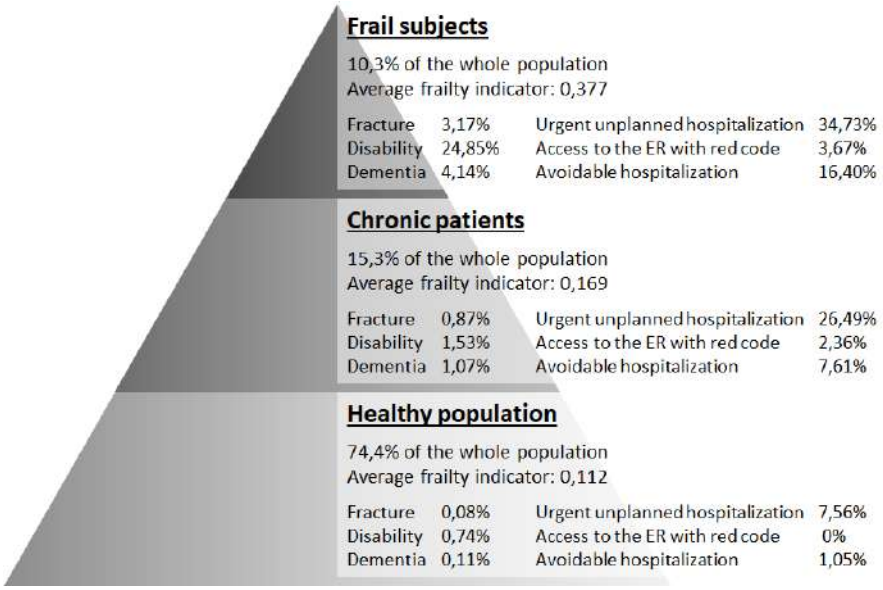


Figure 1: Stratification of elderly assisted by Health Unit 6, according to their use of services in 2018.

In order to reproduce this stratification also on old population assisted by the Health Unit 6, in Padua province, we divided the population into three groups through a regression tree based on the frailty indicator. The three groups are distinguished by individuals' use of services in 2018 (thus, we excluded individuals dead in 2018), in particular by their use of home assistance and their access to the emergency room. At the base of the pyramid we find about three quarters of the total old population: these are individuals who have not had access to the emergency room and have not benefited from home care in 2018. The average frailty indicator in this portion of the population is very low, equal to 0.112 and, as can be seen in Fig. 1, the proportion of individuals experiencing negative outcomes related to frailty is also limited. In this subgroup of individuals, we also find very low percentages of chronic diseases, a low number of hospitalizations that also last less than in other groups (less than 3 days on average). In the central portion of the pyramid there is the 15.3% of the whole population, characterized by intermediate health conditions and by an average frailty indicator of 0.169, which is already higher than the average of the total population (0.159). In this stratum we find higher percentages of both negative outcomes and chronic diseases than the previously described healthy population stratum.
At the top of the pyramid, on the other hand, we find frail subjects who represent one tenth of the total old population in Health Unit 6. These subjects all experienced at

least one access to the emergency room and used home care assistance in 2018. They have a higher average frailty indicator, equal to 0.377. These are individuals characterized by the coexistence of several chronic diseases, in fact the percentages of chronic diseases in this population stratum are much higher than in the other two subgroups. Even the occurrence of negative outcomes related to frailty is more common in this small portion of the population, as stated by the adopted definition. Moreover, these individuals are more commonly hospitalized than the rest of the population and their hospitalizations last almost 11 days on average.
The indicator was also computed for the old population assisted by the Health Unit 6 in 2019. It shows good performances in terms of outcome prediction even if applied to a different time lag than the one considered for its development.

Conclusions

The proposed frailty indicator boasts several strengths and seems a valuable choice to identify frail old individuals: the use of administrative health data, the involvement of multiple health outcomes, the use of a parsimonious set of variables, and the use of poset theory that limits assumptions needed in the construction of the indicator. However, the possibility of extending our proposal to other territories is conditioned on homogeneity of methods of collecting and coding clinical information across all regional health systems. For instance, criticalities emerged regarding the definition of urgent hospitalization, as the one used in Padua is different from the one used in Piedmont [4].
Moreover, an interesting aspect for further exploration is the use of this frailty indicator in a longitudinal perspective, to follow ageing process of the population in a more dynamic way. In a longitudinal scenario, it would be interesting to study also the role of negative health outcomes in worsening individual health status.

References

1. Fried, L. P., Tangen, C. M., Walston, J., Newman, A. B., Hirsch, C., Gottdiener, J., et al. Frailty in older adults: evidence for a phenotype. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 56, 46–56. (2001) doi: 10.1093/geron a/56.3.M146
2. Gillick, M. Guest editorial: Pinning down frailty. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 56(3), M134–M135. (2001) doi: 10.1093/geron a/56.3.M134
3. Gobbens, R. J. J., Luijkx, K. G., Wijnen-Sponselee, M. T., & Schols, J. M. G. A. In search of an integral conceptual definition of frailty: Opinions of experts. Journal of the American Medical Directors Association, 11, 338–343. (2010) doi:10.1016/j.jamda .2009.09.015
4. Silan, M., Signorin, G., Ferracin, E., Listorti, E., Spadea, T., Costa, G., & Boccuzzo, G. Construction of a Frailty Indicator with Partially Ordered Sets: A Multiple-Outcome Proposal Based on Administrative Healthcare Data. Social Indicators Research. (2020) doi: 10.1007/s11205-020-02512-7

Clusters of contracting authorities over time: an analysis of their behaviour based on procurement red flags

Cluster di stazioni appaltanti nel tempo: un'analisi basata su red flags negli appalti pubblici

Simone Del Sarto and Paolo Coppola and Matteo Troìa

Abstract In this paper we aim at clustering Italian contracting authorities in terms of their attitude over time (2015-2017) in managing public procurements, a field particularly prone to the occurrence of corrupt acts. For our purpose we rely on an approach based on red flag indicators, which considers public procurement data and points out possible anomalies in order to alert the system to the possible risk of corruption. As such, this approach allows us to perform the analysis at every moment of the procedure, from the call for tender until the final realisation of the work. By exploiting the richness of information contained in the Italian *Banca Data Nazionale dei Contratti Pubblici*, a number of red flag indicators proposed by the international literature are computed for the three-year period. By means of a latent Markov model for multivariate continuous responses, we aim at: *i*. identifying clusters of contracting bodies and *ii*. quantifying the probability for a contracting authority belonging to a certain cluster to move to a different cluster (or to persist in the same cluster) over time. First results show that several clusters of administrations may be highlighted. Among them, one profile draws attention, as it includes administrations with extreme values for all the red flags.

Abstract *L'obiettivo di questo lavoro consiste in raggruppare le stazioni appaltanti in base alla loro attitudine nella gestione degli appalti pubblici nel tempo (2015-2017). La nostra analisi è basata su indicatori red flag, che considerano dati sugli appalti pubblici per segnalare eventuali anomalie e allertare il sistema per un possibile rischio di corruzione. Questo approccio consente di analizzare ogni singola fase della procedura di appalto, dalla pubblicazione del bando fino alla realizzazione finale dell'opera. Sfruttando il contenuto della Banca Dati Nazionale dei Contratti Pubblici, sono stati calcolati alcuni degli indicatori "red flag" proposti*

Simone Del Sarto
Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence
e-mail: simone.delsarto@unifi.it

Paolo Coppola
Department of Mathematics, Computer Science and Physics - University of Udine
e-mail: paolo.coppola@uniud.it

Matteo Troìa
Capgemini Italia - Insights&Data
e-mail: matteo.troia@capgemini.com

in letteratura. Utilizzando un modello latent Markov per risposte continue multivariate, il nostro obiettivo è duplice: i. identificare gruppi di stazioni appaltanti e ii. quantificare la probabilità per una stazione collocata in un certo cluster di spostarsi in un altro cluster (o di rimanere nello stesso) nel corso del tempo. I primi risultati evidenziano diversi profili di amministrazioni. Tra di essi, è possibile individuare un profilo "estremo", poiché include stazioni appaltanti con valori critici per tutti gli indicatori considerati.

Key words: public procurement; red flags; latent Markov; longitudinal data

1 Introduction

The corruption phenomenon is particularly challenging to capture. Its nature is elusive since it is difficult to observe and unlike other crimes, both participants obtain a benefit and have no incentive to bring the illegal agreement to light. In addition, the injured party is often not identifiable as a physical or legal person, and the consequences of bribery can remain under the radar for a long time. This is probably one of the reasons why the best-known corruption measurement indices are perception-based. Transparency International's Corruption Perception Index (CPI), for example, analyses 180 territories and countries based on perceived levels of corruption in the public sector by interviewing experts and business leaders. Despite the undeniable advantage of raising public awareness on the issue of corruption, the CPI, like all subjective indicators, has, however, obvious limitations, because perceptions are affected by the cultural context that can lead to sudden changes due, for example, to scandals or respondents can be influenced by the formulation of questions [5].

There are also methods based on objective indicators, such as, say, the number of crimes reported or actually prosecuted, but on the one hand they underestimate the magnitude of the phenomenon, even if they can be useful in analysing trends, and on the other hand they photograph an event that in some cases can be very distant in time and their analysis has limited applications in the fight against corruption. Other objective methods involve comparing public money spent with the physical infrastructure actually implemented [6], but while it is not straightforward to correctly estimate the value of the goods or services implemented, even this type of analysis can only be done after the fact.

A more promising approach – which we are going to pursue in this paper – is the one based on red flag indicators [1, 7] that analyse public procurement data and highlight anomalies. Red flags do not always correspond to corrupt behaviour, but they can still be useful in pointing out errors and degenerations in the administration to which they refer, and they have the important advantage of timeliness compared to previous techniques, because the analysis can be done at every moment of the procedure, from the call for tender (e.g., by analysing the timing and types of tenders), to the submission of bids (e.g., by assessing the total number or the possible absence of participants), to the awarding (e.g., by measuring markdowns), until the implementation (by taking into account, for example, any variants during the work). Such an approach can be integrated into information systems that help in the fight

against corruption. The simple fact of knowing that certain indicators are constantly being monitored could lead to greater attention and improved procedures. The disadvantage of this approach lies in the difficulty of finding data and the quality of the data itself, which too often is still hand-tagged in different systems with delays and, invariably, introducing errors. A total computerisation of processes with a decisive push towards interoperability will make these approaches and, consequently, the fight against corruption more effective.

In this regard, we exploit the richness of the Italian database of public procurement (BDNCP in the following, which stands for *Banca Dati Nazionale dei Contratti Pubblici*), managed by the Italian anticorruption authority[1]. This huge database systematically collects data about every single phase about the complex process of the realisation of a tender by a contracting body, from the publication of the call, to the award notice, to the management of possible variants, which may occur on specific contract terms, until the final test of realised work.

Data extracted from BDNCP allow us to build a set of red flag indicators [4, 9], which are used to assess different profiles of Italian contracting bodies in terms of public procurement management, and, also, to verify whether an evolution over time occurs in terms of transitions across profiles. A very suitable statistical tool for this purpose is the latent Markov model [10, 2], which allows us to extract latent states for observed response variables (red flag indicators in our context) and to estimate transitions across states over time.

This paper is organised as follows: Section 2 describes the data and the statistical model used for the analysis, while first results are shown in Section 3. Finally, Section 4 draws some concluding remarks and illustrates some future developments.

2 Data and statistical model

For this work we use data from BDNCP, which, as stated above, contains information about the entire life cycle of every tender that have involved Italian contracting bodies. After some data quality checks, we consider data about 2,851 contracting authorities, related to tenders with a call published in 2015, 2016 and 2017. Specifically, data about call for tender and contract award are retained in order to construct the following five red flag indicators for each contracting body and each year: *i*. proportion of procedures awarded through non-price related evaluation criteria, hence with a certain degree of subjectivity (labelled as *non_price_eval*); *ii*. proportion of non-open procedures (*non_open*); *iii*. proportion of contracts for which a single bid is received (*single_bid*); *iv*. average number of days between publication of call for tender and submission deadline (*submission_period*); *v*. average number of days between award notice and publication of call for tender (*award_publication*).

From a statistical point of view, we consider the above set of five indicators as our continuous multivariate response variable. Let Y_{ijt} be the value of indicator $j = 1,\dots,J$ computed on contracting body $i = 1,\dots,n$ at time $t = 1,\dots,T$. In our case, $n = 2,851$, $J = 5$ and $T = 3$, with $t = 1$ represents year 2015 (first year of analysis). By arranging response variables by statistical unit, we can consider

[1] https://dati.anticorruzione.it/opendata

Table 1 Estimated conditional means $\hat{\boldsymbol{\mu}}_u$ for each latent state and estimated initial probabilities $\hat{\pi}_u$

	Latent state u						
Indicator	1	2	3	4	5	6	7
non_price_eval	0.224	0.135	0.481	0.247	0.841	0.114	0.789
non_open	0.517	0.079	0.269	0.515	0.108	0.918	0.852
single_bid	0.196	0.260	0.170	0.236	0.283	0.219	0.304
submission_period	150.5	23.8	35.5	22.4	27.1	17.0	17.2
award_publication	183.4	66.6	277.6	67.9	78.0	48.2	53.7
$\hat{\pi}_u$	0.005	0.177	0.021	0.269	0.048	0.424	0.056

the J-dimensional response vector $\boldsymbol{Y}_{it}$, which collects response variables of unit i at time t. Moreover, we assume that the response process of unit i is affected by a latent process $\boldsymbol{U}_i = [U_{i1}, \ldots, U_{iT}]^\top$, which is assumed to follow a first-order Markov chain with k latent states. Another hypothesis concerns the response vector distribution, given the latent process, as follows:

$$f(\boldsymbol{Y}_{it} = \boldsymbol{y} | U_{it} = u) \sim N(\boldsymbol{\mu}_u, \Sigma),$$

where $\boldsymbol{y}$ and u are realisations of $\boldsymbol{Y}_{it}$ and U_{it}, respectively.

The model parameters are: the conditionals means, $\boldsymbol{\mu}_u$, with $u = 1, \ldots, k$; the variance-covariance matrix, Σ (supposed to be constant across states); the initial probabilities, $\pi_u = P(U_{i1} = u)$, with $u = 1, \ldots, k$; the transition probabilities, $\pi^{(t)}_{u|\bar{u}} = P(U_{it} = u | U_{it-1} = \bar{u})$, with $t = 2, \ldots, T$ and $u, \bar{u} = 2, \ldots, k$.

In order to estimate the parameters, we rely on a maximum likelihood approach, by exploiting an Expectation-Maximisation algorithm. Moreover, in order to overcome issues related to multimodality of the log-likelihood function, a combination of deterministic and random initialisations of model parameters is performed. The model fitting is performed by means of R package "LMest" [3].

3 Results

A well-known issue of this modelling approach is related to the selection of the number of latent states, denoted by k. A combination of subjective and objective criteria (the latter based on the Bayesian Information Criterion [8]) allows us to select $k = 7$ as suitable number of latent states (data not shown). As a consequence, seven profiles of Italian contracting bodies can be highlighted, each one characterised by the estimates of the vector of conditional means (see Table 1 by column).

As can be noticed from the table, single bidding indicator results to be the least discriminant, as it does not show a consistent variability across states, although contracting bodies clustered in latent state 7 exhibits around 30% share of single bidder contracts, almost twice than that of latent state 3. Furthermore, latent state 1 and 4 are very similar with respect to the first three indicators – related to proportion of

procedures *i*. evaluated with non-price related criteria, *ii*. of type "non-open" and *iii*. with a single bid – while they consistently differ as regards the last two red flags (related to the time passed *i*. between the publication of call for tender and the submission deadline, and *ii*. between the award notice and the publication of the call). Latent state 7 clusters contracting bodies with, say, "critical" values for all the indicators. In fact, they show, on average, the greatest values related to the first three indicators (i.e., high proportions of "most at risk" procedure types) and the lowest average time periods between the main phases of a public procurement process. On the other hand, latent state 1 and 2 can be labeled as the most virtuous ones. The former shows wide time periods between call for tender publication/submission deadline/award notice and low proportions of procedures awarded with non-price related evaluation criteria, whereas the latter is characterised by the lowest means of the first two red flags. As far as the estimated initial probabilities are concerned, in 2015 (beginning of the study) clusters 6 and 4 are the most numerous, as they include almost 70% of the contracting bodies, while latent state 1 is the smallest one (only 0.5% of administrations).

Looking at the transition matrices, reported in Table 2, we can observe high probabilities in the main diagonal for almost all the states and in both time transitions, that is, from 2015 to 2016, reported in Table 2(a), and from 2016 to 2017, as shown in Table 2(b). As a consequence, a certain degree of persistence in the current cluster arises, especially for latent states 2, 4, 5, 6 and 7. Finally, latent state 1 (characterised by extremely high time extension between the main phases of a public procurement process) can be considered as a non-persistent condition, as high probability of moving towards state 2 or state 6 can be noticed, while a transition from any other states to state 1 is very unlikely.

4 Conclusions

In this work a latent Markov approach for multivariate continuous response variables is applied to a selection of red flag indicators in public procurement, in order to study the evolution over time of a set of Italian contracting authorities in terms of management of the public procurement process.

First results identify several groups of contracting bodies, characterised by varied behaviours, measured in terms of average red flag indicators. Among the ascertained profiles, one includes administrations with, on average, extreme values for all the red flags, values which usually may be considered as "most at-risk". On the other hand, a definite virtuous group according to the entire set of indicators is not immediately identifiable.

This work is meant to be a first attempt for studying the evolution over time of red flag indicators in public procurement. Obviously, further considerations need to be taken into account, in order to better characterise the ascertained clusters. This can be performed, for example, by including covariates in the model (e.g., the contract object in order to identify the market involved in the procurement activity), or by considering the monetary value of each tender (i.e., using weighted indicators). Finally, as the purpose is to identify extreme behaviours (high/low values of certain

Table 2 Matrices of estimated transition probabilities from 2015 to 2016 (a) and from 2016 to 2017 (b)

(a)

State	1	2	3	4	5	6	7
1	0.000	0.483	0.000	0.000	0.129	0.322	0.066
2	0.011	0.548	0.017	0.134	0.092	0.161	0.036
3	0.000	0.174	0.204	0.079	0.108	0.264	0.171
4	0.008	0.111	0.018	0.507	0.067	0.215	0.074
5	0.028	0.169	0.014	0.055	0.439	0.141	0.154
6	0.014	0.076	0.010	0.108	0.031	0.680	0.080
7	0.000	0.058	0.025	0.136	0.085	0.145	0.551

(b)

State	1	2	3	4	5	6	7
1	0.031	0.208	0.000	0.248	0.109	0.404	0.000
2	0.004	0.534	0.003	0.151	0.071	0.198	0.040
3	0.019	0.228	0.063	0.241	0.100	0.282	0.067
4	0.000	0.072	0.010	0.525	0.097	0.224	0.070
5	0.005	0.170	0.008	0.091	0.418	0.160	0.149
6	0.005	0.071	0.008	0.149	0.043	0.642	0.083
7	0.000	0.054	0.004	0.127	0.126	0.209	0.480

red flags), a quantile regression approach could be very suitable for this context, as it allows us to focus on the tails of indicator distributions.

References

1. ANAC: Corruzione sommersa e corruzione emersa in Italia: modalità di misurazione e prime evidenze empiriche. Available at http://www.anticorruzione.it/portal/public/classic/Attivitadocumentazione/Pubblicazioni/RapportiStudi (2013)
2. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. CRC Press, Boca Raton, FL (2013)
3. Bartolucci, F., Pandolfi, S., Pennoni, F.: LMest: an R package for latent Markov models for longitudinal categorical data. J. Stat. Softw., **81**(4), 1–38 (2017)
4. Fazekas, M., Tóth, I.J., King, L.P.: An objective corruption risk index using public procurement data. Eur. J. Crim. Policy Res., **22**, 369—397 (2016)
5. Fiorino, N., Galli, E.: La corruzione in Italia. Il Mulino, Bologna (2013)
6. Golden, M.A., Picci, L.: Proposal for a New Measure of Corruption, Illustrated with Data. Econ. Polit., **17**(1), 37–75 (2005)
7. Office Européen de Lutte Anti-Fraude (OLAF): Identifying and Reducing Corruption in Public Procurement in the EU. PwEU Service – Ecoyrs - Utrecht University (2013)
8. Schwarz, G.: Estimating the Dimension of a Model. Ann. Stat. **6**, 461–464 (1978)
9. Troìa, M.: Data analysis e costruzione di indicatori di rischio di corruzione per la Banca Dati Nazionale dei Contratti Pubblici. Autorità Nazionale Anticorruzione ANAC, working paper no. 5 (2020)
10. Wiggins, L.M.: Panel Analysis: Latent Probability Models for Attitude and Behaviour Processes. Elsevier, Amsterdam (1973)

An Application of Temporal Poset on Human Development Index Data

Un'applicazione del Temporal Poset ai dati dell'Indice di Sviluppo Umano

Leonardo Salvatore Alaimo, Filomena Maggino and Emiliano Seri

Abstract Within the international debate on finding measures beyond GDP, the Human Development Index released by the United Nations Development Programme, has become a reference over the years. In order to get a synthetic view of human development, different aggregative procedures has been applied over time. The aggregative road to synthesis is however problematic, because it raises a number of conceptual and methodological issues. As a valuable alternative, in this paper we adopt a non-aggregative approach to synthesis over time, based on Partially Ordered Set Theory.

Abstract *Nel dibattito internazionale sulla ricerca di misure che vadano oltre il PIL, l'Indice di Sviluppo Umano sviluppato da* United Nations Development Programme*, è diventato un riferimento nel corso degli anni. Al fine di ottenere una visione sintetica dello sviluppo umano, sono state applicate nel tempo diverse procedure aggregative. La strada aggregativa verso la sintesi è tuttavia problematica, poichè pone una serie di problematiche concettuali e metodologiche. Come valida alternativa, in questo articolo adottiamo un approccio non aggregativo alla sintesi, basato sulla teoria degli insiemi parzialmente ordinati.*

Key words: Human Development Index, Synthesis of statistical indicators, Compensability, Partially Ordered Set - Poset

Leonardo Salvatore Alaimo
Corresponding author. Italian National Institute of Statistics - Istat, e-mail: leonardo.alaimo@istat.it

Filomena Maggino
Sapienza University of Rome e-mail: filomena.maggino@uniroma1.it

Emiliano Seri
Sapienza University of Rome e-mail: emiliano.seri@uniroma1.it

1 Introduction

Although it has been considered a *niche field*, the topic of synthesis of statistical indicators has a rich and varied scientific literature. The growing attention on this issue is, in a way, linked to the international debate on identifying measures that go *Beyond GDP*. From this perspective, the synthetic approach becomes the only one possible for a correct understanding of the phenomenon (7). The traditional statistical approach to synthesis is the so-called aggregative-compensative, according to which the synthetic measure is the result of the mathematical aggregation of the basic indicators (composite indicators). This has become the dominant framework over time. However, it poses a series of conceptual and methodological criticalities that have been highlighted in recent literature (3; 2). To try to overcome these limitations, research has focused on methods belonging to the so-called non-aggregative approach, in which the synthetic measure is obtained by not combining the basic indicators. Among those methods, the theory of partially ordered sets (poset theory) has become a reference over the years. This method, particularly suitable for the treatment of ordinal data, is useful even if we deal with indicators of different scaling levels. In this paper, we apply a poset-based synthesis method suitable for time series data (the so-called *temporal poset*) developed by (1) to the data of the synthetic measure elaborated by the United Nations Development Programme (UNDP), Human Development Index (HDI). This is a synthetic measure elaborated by the United Nations Development Programme (UNDP) and conceptually based on Sen's capabilities approach (8). HDI identifies three main dimensions, the basic capabilities crucial to human development: a long and healthy life, knowledge and a decent standard of living. The need to identify alternative aggregations (also UNDP posed the problem, changing the aggregation procedure, which was previously an arithmetic mean, in 2010 adopting a geometric one) is closely linked to a central issue in the context of composite indicators, the level of compensability or substitutability allowed between basic indicators. Generally, the basic indicators of a composite index are called *substitutable* if a deficit in one may be compensated by a surplus in another; on the contrary, the basic indicators are called *non-substitutable* if a compensation among them is not allowed. Consequently, an aggregation approach can be 'compensatory' or 'non-compensatory' depending on the adoption or not of compensation. The issue is not only methodological but also, and above all, conceptual. Let us consider HDI. If we admit full compensability, we implicitly affirm, for instance, that a surplus in the education dimension can compensate for a deficit in the economic one. This is, at least, highly questionable. On the other hand, if we affirm the non-compensability of the basic indicators, we risk flattening the results of our synthesis downwards (2). Finally, if we adopt a partially compensative method, i.e. allowing it "up to a certain point", the question would arise as to what is the permissible and tolerable threshold of compensability. The problem is not so much one the compensability as the aggregation, which generates a *flattening effect* (3) regardless of the method and, consequently, the level of compensability allowed. The method adopted in this paper not only addresses these issues but also offers a possible solution, as demonstrated by different empirical studies (1; 2; 3).

2 Data

As previously mentioned, the data used in this paper are those relating to the Human Development Index (http://hdr.undp.org/en/content/download-data). In particular, we refer to the time series from 2016 to 2019 of the four indicators that compose the three dimensions of the HDI: *life expectancy at birth*; *expected years of schooling*; *mean years of schooling* and *gross national income (GNI) per capita*. The indicators described are therefore those necessary to give the most complete and exhaustive representation of human development, understood as the advancement of human freedom, dignity and equality, and these include the broad comprehensive range of freedoms covering economic, social, political and civil areas. For a better illustration of the proposed methodology we only used the European countries, which correspond to our units in the four years considered.

3 Methodological aspects of the application

Poset supplies concepts and tools that appropriately adapt to the needs of synthesis. It is focused on *profiles*, which are the combinations of scores of each statistical units in the basic indicators considered, describing the status of units. This approach has some advantages and overcomes some limitations of the aggregative-compensative approach. We refer to the extensive literature on the basic aspects and definition of poset (4) and on the methodology about its use for data synthesis over time (1; 2). in this paper, we focus on describing the different steps leading to the construction of the synthetic measure according to this method.

First of all, the indicators must all have positive polarity[1]; where some have negative polarity, this must be reversed using a transformation. This is necessary to ensure that nodes in the highest positions of the Hasse diagrams will indicate better situations than those in the lowest positions. The system considered in our work is composed by 43 units (the European countries), 4 indicators and 4 temporal occasions (2016-2019). This is a three-way data array $\mathbf{Y} \equiv \{y_{ijt} : i = 1,\ldots,43;\ j = 1,\ldots,4;\ t = 1,\ldots,4\}$ that can be seen as a set of 4 matrices of order (43×4), each of which represents a temporal slice of $\mathbf{Y}$. For each of the 4 matrices independently, we can calculate the incidence matrix[2] and construct the Hasse diagrams[3] , reported in Figure 1. Just the graphical representations give important information; it is evident, for instance, that the relationship structure of the system is different in the two times considered.

[1] Polarity is the sign of the relation between the indicator itself and the phenomenon

[2] The incidence matrix is a matrix $\mathbf{Z}_P = (z_{ij}) \in \mathbb{Z}^{k\times k}$, where $|X| = k$ is the cardinality of X and z_{ij} is equal to 1 if $x_i \trianglelefteq x_j$, 0 otherwise, with $x_i, x_j \in X$. It defines the structure of comparabilities in the poset

[3] The *Hasse diagram* is the graphical representation of the directed acyclic graph representing the cover relation $\prec$: two elements are comparable $\trianglelefteq$ if a path connects them in the Hasse diagram.

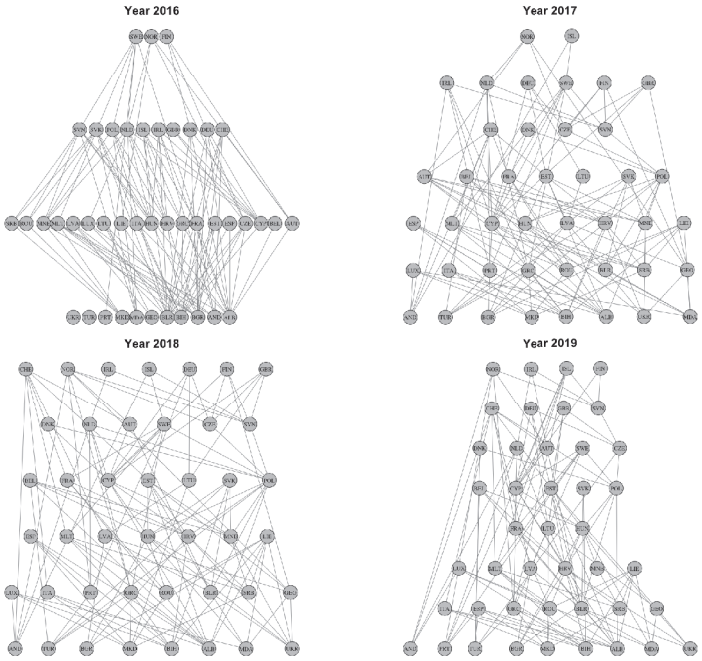


Fig. 1: Hasse diagrams of the European countries: years 2016 - 2019.

We want to obtain a synthetic measure. To do this, poset offer different possibility; in this paper, we use the so-called *average height*[4]. We can compute this measure for each of the 4 posets[5]. The results obtained allow an *intra-temporal* comparison of the units within the system. For example, we can affirm that Norway (NOR) is better than Italy (ITA) in 2017 (Figure 1). Anyway, it is impossible to make an *inter-temporal* comparison of units. For instance, Spain (ESP) worsens from 2018 to 2019, but we do not know why this happened. For instance, it may have happened that all the other countries had a very marked improvement in the indicators considered, while ESP could have increased slightly and been overtaken by the other units. Another possibility could be that Spain has been drastically reduced from 2018 to 2019 compared to the other units.

To make comparisons over time, we must *merge* the posets. Given two finite posets Λ and Π, we merge them by setting $x \leqslant_{\Lambda\Pi} y$ if and only if one of the following conditions is valid (1):

1. $x, y \subset \Lambda$ and $x \leqslant_{\Lambda} y$;
2. $x, y \subset \Pi$ and $x \leqslant_{\Pi} y$;
3. $x \subset \Lambda$; $y \subset \Pi$ and $x \leqslant_{\Lambda\Pi} y$;

[4] For a definition and the methodological step of its calculation, please see (1; 2).

[5] We do not include the results in the paper for question of space. They can be required to authors.

4. $x \subset \Pi$; $y \subset \Lambda$ and $x \leqslant_{\Lambda\Pi} y$.

In other words, by merging the two posets we maintain their initial structures of comparability, adding other comparabilities that are an expression of the temporal comparison among the elements. In this way, it will be possible to make intertemporal comparisons. Moreover, in order to make it possible to compare posets with different sets of nodes and to anchor the the average height computation to a *common reference system*, we introduce an *embedded scale* (6), i.e. some benchmark profiles that form a scale of increasing levels embedded in the original poset. They are points that help anchoring the comparisons between profiles in the Hasse diagram and the average heights to a reference scale. We can merge the 4 posets in Figure 1 in one temporal poset and add a 5 levels embedded scale[6]. By calculating the average height of the resulted *temporal poset*, we can make comparisons over time of different units, using benchmarks as a common reference system for the different years. For instance, we can observe that ESP passes from a value of average height of 147 in 2018 to a value of 126 (see Table 1) and compare this trend with those of other countries.

4 Conclusions

HDI aims to allow policy makers evaluation of national policies. Thus, it is important to have a measure of the human development not affected by compensation in order to observe how the phenomenon behaves in all its dimensions. The proposed method tries achieving this goal, giving an easy-to-read representation of the HDI for each country in all the considered years. In the synthetic index proposed, we do not focus on the values of basic indicators and on their aggregation, but on the profiles of each country. This allows the overcoming of the compensability issue. The results, shown in Table 1, are free from the *flattening effect*, typical in the mean–based aggregation methods: countries with different profiles present different average heights.

References

[1] Alaimo, L.S. Complexity of Social Phenomena: Measurements, Analysis, Representations and Synthesis. Unpublished doctoral dissertation, University

[6] The scale is defined as follows:

- *min*, with a profile given by the minimum value in all indicators;
- $B1$, with a profile given by the first quartile of all indicators;
- $B2$, with a profile given by the second quartile of all indicators;
- $B3$, with a profile given by the third quartile of all indicators;
- *max*, with a profile given by the maximum value in all indicators.

European countries	2016	2017	2018	2019	European countries	2016	2017	2018	2019
ALB	16.79	34.30	24.06	38.23	**LIE**	68.42	87.14	108.54	133.06
AND	38.46	21.45	45.66	73.93	**LTU**	89.73	113.87	125.06	136.52
AUT	103.32	115.61	127.81	112.37	**LUX**	81.81	132.79	124.02	147.55
BEL	91.96	112.81	135.16	153.46	**LVA**	69.08	77.92	89.58	101.79
BIH	11.89	11.30	22.34	34.39	**MDA**	7.37	15.00	26.59	19.37
BLR	40.58	48.48	56.35	49.04	**MKD**	4.83	9.37	14.58	20.21
BUL	20.67	21.09	15.19	32.80	**MLT**	67.49	82.11	95.86	111.37
CHE	148.50	152.79	162.71	169.67	**MNE**	41.31	48.21	55.83	62.58
CYP	58.87	72.00	85.14	92.05	**NLD**	107.88	120.87	133.65	150.29
CZE	105.42	100.15	109.45	120.12	**NOR**	141.52	152.60	161.00	168.86
DEU	148.20	155.77	162.63	169.65	**POL**	74.39	81.20	88.15	86.01
DNK	112.22	133.71	153.13	141.59	**PRT**	50.50	45.18	60.35	76.84
ESP	96.11	115.07	147.04	126.82	**ROU**	27.51	28.45	36.72	45.34
EST	110.41	105.13	131.72	117.48	**SRB**	13.07	23.44	32.55	33.13
FIN	139.34	149.19	159.02	168.48	**SVK**	89.20	113.96	141.33	158.64
FRA	83.44	96.38	109.34	123.21	**SVN**	85.42	99.49	113.16	127.56
GBR	132.78	145.19	157.91	168.23	**SWE**	121.37	133.05	141.32	167.13
GEO	28.65	41.81	54.89	78.56	**TUR**	9.79	18.93	29.05	39.74
GRC	63.24	76.59	90.50	104.13	**UKR**	7.66	17.81	18.64	29.92
HRV	52.09	58.33	64.10	70.01	**MIN**	1.00	1.00	1.00	1.00
HUN	54.33	55.62	66.78	74.20	**B1**	36.58	36.58	36.58	36.58
IRL	134.53	149.17	163.48	165.04	**B2**	73.65	73.65	73.65	73.65
ISL	143.10	153.78	157.02	160.45	**B3**	128.43	128.43	128.43	128.43
ITA	60.98	77.92	98.77	121.06	**MAX**	177.00	177.00	177.00	177.00

Table 1: Average height values: European countries; years 2016-2019.

of Rome "La Sapienza", Rome, Italy (2020).

[2] Alaimo, L.S., Arcagni, A., Fattore, M ., Maggino, F. Synthesis of Multi-indicator System. Soc. Ind. Res. (2020) doi: 10.1007/s11205-020-02398-5.

[3] Alaimo, L. S., Maggino, F. Sustainable development goals indicators at territorial level: Conceptual and methodological issues—The Italian perspective. Soc. Ind. Res., 147(2), 383–419. (2020) doi: 10.1007/s11205-019-02162-4.

[4] Bruggemann, R., Patil, G.P. Ranking and prioritization for multiindicator systems: Introduction to partial order applications. Springer, Dordrecht. (2011).

[5] D'Urso, P., Alaimo, L. S., De Giovanni, L., Massari, R. Well-being in the Italian regions over time. Soc. Ind. Res. (2020) doi: 10.1007/s11205-020-02384-x.

[6] Fattore, M. Non-aggregated Indicators of Environmental Sustainability. Silesian Statistical Review/Slaski Przeglad Statystyczny, 16(22), 7–22. (2018).

[7] Maggino, F., Alaimo, L. S. Complexity and wellbeing: measurement and analysis. In: Bruni, L., Smerilli, A., De Rosa D. (eds.) A Modern Guide to the Economics of Happiness, pp. 113–128. Edward Elgar Publishing, Northampton (2021).

[8] Sen, A. Commodities and Capabilities. Oxford University Press, Oxford. (1999).

The SDGs System: a longitudinal analysis through PLS-PM

Il sistema degli SDGs: un'analisi longitudinale attraverso il PLS-PM

Cataldo Rosanna, Grassia Maria Gabriella, Antonucci Laura

Abstract The Sustainable Development Agenda [24] emphasizes measurement and monitoring progress of the Sustainable Development Goal (SDG) targets, stressing the need for "a data revolution for sustainable development to improve the quality of statistics and information available to citizens and governments". The main problem for researchers is to find appropriate tools to obtain a synthetic indicator able to synthesize these targets and monitor them over the time. The work focuses on using the Structural Equation Modeling and especially Higher Order Partial Least Squares Path Modeling as a valuable way to analyze longitudinal data of SDGs. The paper contributes to the European Community countries-analysis of SDG reporting by performing a longitudinal analysis over the 20-year period encompassing 2000 to 2019. Due to the difficulty of reporting on a paper a detailed analysis of all 17 SDGs, we focus only on social dimension.

Abstract *L'Agenda per lo Sviluppo Sostenibile [24] sottolinea la misurazione il monitoraggio dei progressi degli obiettivi di Sviluppo Sostenibile (Sustainable Development Goal (SDG)), evidenziando la necessità di "una rivoluzione di dati per migliorare la qualità delle statistiche e delle informazioni a disposizione di cittadini e governi". Il problema principale per i ricercatori è trovare strumenti adeguati per ottenere un'indicatore capace di sintetizzare questi target e monitorarli nel tempo. Il presente lavoro si focalizza sull'uso di Modelli ad Equazioni Strutturali e in modo particolare dei modelli gerarchici Partial Least Squares Path Modeling come uno strumento prezioso per analizzare dati longitudinali degli SDGs. Il lavoro si basa*

Cataldo Rosanna
Department of Social Sciences, University of Naples "Federico II", Vico Monte della Pietà, 1, 80138 Naples, Italy, e-mail: rosanna.cataldo2@unina.it

Grassia Maria Gabriella
Department of Social Sciences, University of Naples "Federico II", Vico Monte della Pietà, 1, 80138 Naples, Italy e-mail: mariagabriella.grassia@unina.it

Antonucci Laura
Department of Clinical and Experimental Medicine, University of Foggia, Via Luigi Pinto 1, 71100 Foggia, Italy e-mail: laura.antonucci@unifg.it

sull'analisi degli stati membri dell'Unione Europea eseguendo un'analisi longitudinale su un periodo di 20 anni compreso tra il 2000 e il 2019. A causa della difficoltà di riportare nel documento un'analisi dettagliata di tutti i 17 SDGs, ci siamo focalizzati solo sull'area sociale.

Key words: SDGs, Composite indicator, PLS-PM, longitudinal analysis

1 Introduction

Sustainable development has been at the heart of European policy for a long time, firmly anchored in the European Treaties. The 2030 agenda for Sustainable Development and its 17 Sustainable Development Goals (SDGs), adopted by the UN General Assembly in 2015, have given a new impetus to global efforts to achieve sustainable development. In the last years, the interest towards understanding and measuring the phenomena manifests in numerous researches and publications [3]; [20]; [9]; [6], aiming to review and compare the synthetic indices developed to measure sustainable development. Cataldo et al. [3] in a recent paper proposed Partial Least Squares Path Modeling (PLS-PM) as a method for studying SDGs indicators demonstrating how PLS-PM could help you to define the framework for SDGs indicators in order to provide a better measure of this complex multidimensional social phenomenon. This study can be considered an advancement of that work as it demonstrates how the PLS-PM, that worked with cross-sectional data in Cataldo et al. [3], is very useful in longitudinal data. In this work the statistical aim is to investigate the evolution of the effects between constructs over time and to test them.

According to Banati et al. [2], today "there is growing recognition of the potential of longitudinal research to contribute evidence for policy, insofar as it facilitates understanding of the dynamic nature of developmental trajectories and of the diverse processes that shape outcomes over time". Their paper highlight "how longitudinal data can be a resource for understanding the drivers underpinning SDG indicators and could provide an assessment of the timing of development windows, and related interventions to maximize the impact of interventions". Based on these considerations, in this work we want to demonstrate how PLS-PM can help researchers and funders to analyze the SDGs through a longitudinal analysis. To perform such an analysis, it is necessary to take into account the most important turning points in the evolution of sustainability: 1) the SDGs were born at the United Nations Conference on Sustainable Development in Rio de Janeiro in 2012 (the objective was to produce a set of universal goals that meet the urgent environmental, political and economic challenges facing our world) [23] and 2) the SDGs replace the Millennium Development Goals (MDGs), which started a global effort in 2000 to tackle the indignity of poverty (the MDGs established measurable, universally-agreed objectives for tackling extreme poverty and hunger, preventing deadly diseases, and expanding primary education to all children, among other development priorities) [22]. In this paper, we take into account only some SGDs, due to the difficulty of

reporting a detailed analysis of all 17 SDGs, in particular, we will focus on the first six Goals belonging to the social dimensions of SDGs, according to the three-way holistic framework (social, environment and economic area). The indicators at different points in time are used to create the exogenous and endogenous constructs at the different points in time in the PLS-PM [17]; [12]; [18]. The main research objective in this case is to investigate the evolution of the effects between constructs over time.

2 Longitudinal data analysis with PLS-PM approach

The model on longitudinal data can be approached from several perspectives, and the model can be constructed as a Structural Equation Model (SEM). According to Baltes and Nesselroade [1], SEM is a valuable way to analyze longitudinal data because it is both flexible and usefulfor answering common research questions. The explicit invocation of latent variables (LVs) afforded by the SEM makes this framework the one most commonly used to implement and analyze longitudinal data [16]; [19]. Recently, Roemer [17] has proposed using the component-based approach to SEM-PLS-PM in a longitudinal study [7]; [21]. In accordance with Roemer [17], we posit that PLS path modeling is highly appropriate for an analysis of the development and change in constructs in longitudinal studies, since it offers three favorable methodological characteristics. First, constructs often need to be predicted in evolutionary models [12]; [18]. Secondly, model complexity quickly increases when development and change need to be analyzed in longitudinal studies. This is due to the larger number of constructs that are measured at different points in time and the respective effects between those constructs [12]. PLS-PM is well suited to dealing with such complex models [8]; [27]. Thirdly, sample sizes can become quite small in longitudinal studies [13]. PLS path modeling is particularly appropriate in such cases [11]; [14]. There are many referenced review papers, in the literature, on the PLS approach to SEM [5], [10] and [21]. Recently, Lauro et al. [14] have presented some current developments in PLS-PM for the treatment of non-metric data, hierarchical data, longitudinal data and multi-block data.

3 Data and Analysis

The official data derive from the database of United Nation “Sustainable Development Goals” [1], they were mined in January 2021. The analysis was developed with reference to the European Community countries, including United Kingdom (left on 31 January 2020) over the 20-year period encompassing 2000 to 2019. Four thresholds were considered: 2000 (t_0), 2005 (t_1), 2010 (t_2) and 2015 (t_3); even 2019 (last

[1] https://unstats.un.org/sdgs/indicators/database/

year available) wanted to be included in the analysis but lacked data for almost all indicators, the last full year available was 2015. Considering that each indicators of the considered Goals have different units and values, for comparison purposes all units are first normalized to a value between 0 and 1, where 0 was assigned to the least wellbeing while 1 was the value assigned to the most wellbeing country for each indicators. Some variables are not considered because they are not available in the four periods considered and are not available for all countries. The scheme of theoretical model is shown in Fig. 1. The constructs have been created based on the

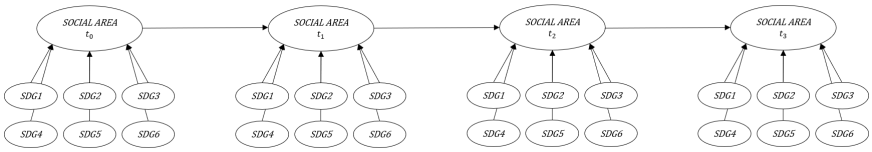


Fig. 1 Theoretical model

indicators at the three points in time. The study focuses on a formative measurement model and the XLSTAT software [2] was used for all the data processing and the PLS-PM. Any missing data was handled by using the NIPALS algorithm [26], while the path weighting scheme was chosen [10] for the model and it was estimated with a maximum of 1,000 iterations. The model, shown in Fig. 1, is a hierarchical model: the block "Social area" in the different time period is the Higher Order Construct (HOC) related to its concrete subdimensions (Lower-Order Components (LOCs)) represented by the six Goals. Different approaches have been developed and proposed in the literature [15]; [25] and [4]. In this work the HOCs have been estimated with the Mixed Two Step Approach [4]: in the first step the indicators of LOCs have been used as indicators of HOCs, and, after running the PLS-PM algorithm, the resulted scores of the blocks are used as indicators of the HOCs, and the PLS-PM algorithm is performed again.

Table 1 reports the main indices to test the overall model quality: R^2 coefficient, the Redundancy index, the Average variance Extracted (AVE) and the Goodness of Fit (GoF) indices. The R^2 coefficients show that the endogenous LV at time t is better predicted by the explanatory LV at the previous time *t-i*, while the values of the redundancy index are appreciably higher for all blocks (the value of 0,50 indicates a sufficient degree of construct validity). The prediction performance of the PLS-PM reflects the high quality of the constructs.

To test the significance of the path coefficients, the bootstrapping procedure was run [10]. Table 2 shows the effects of HOC block at t_1 to t_2 and t_2 to t_3 are positive and highly significant at p-value$<$ 0,001. Only one effect is not significant (t_0 to t_1).

[2] XLSTAT software Copyright © 2017 Addinsoft

Table 1 Overall model quality

Construct	R^2	Redundancies	AVE	GoF
Social Area (t_0)		0,703	0,705	
Social Area (t_1)	0,768	0,659	0,762	0,803
Social Area (t_2)	0,838	0,714	0,716	
Social Area (t_3)	0,821	0,737	0,737	

Table 2 Results of test of significance of the effects over time

Time	Effect	Path Coefficients	Standard Error	t-values	p-values
t_0/t_1	$SocialAreat_0 \rightarrow SocialAreat_1$	0,340	0,217	0,639	0,528
t_1/t_2	$SocialAreat_1 \rightarrow SocialAreat_2$	0,293	0,064	4,578	0,000
t_2/t_3	$SocialAreat_2 \rightarrow SocialAreat_3$	0,358	0,080	4,475	0,000

4 Conclusion

The aim of the work was to use the SEMs and especially Higher-Order PLS-PM as a valuable way to analyze longitudinal data of SDGs. In this analysis we take into account only the first six SGDs belonging to the social dimension of SDGs, due to the difficulty of reporting a detailed analysis of all 17 SDGs. The main research objective was to study the evolution of the effects between constructs of social dimension over time. The overall model quality indices reflect the high quality of the constructs and the path coefficients over time, after 2005, are positive and significant, effects related to the most important turning point in the evolution of sustainability, the Conference on SDG in Rio de Janiero in 2012, which had a very strong impact on the measurement of the indicators of the 17 SDGs. Only the first effect, from 2000 to 2005, results not significant, highlighting the fact that probably the MDGs, which started a global effort in 2000, did not have a strong social impact before 2005.

References

1. Baltes, P. B. and Nesselroade, J. R.: The developmental analysis of individual differences on multiple measures. Academic press. (1973)
2. Banati, P., & Oyugi, J.: Longitudinal Research for Sustainable Development. Zeitschrift für Psychologie. (2019)
3. Cataldo, R., Crocetta, C., Grassia, M. G., Lauro, N. C., Marino, M., & Voytsekhovska, V.: Methodological PLS-PM framework for SDGs system. Social Indicators Research, 1-23. (2020)
4. Cataldo, R., Grassia, M.G., Lauro, N.C. and Marino, M.:Developments in Higher-Order PLS-PM for the building of a system of Composite Indicators. *Quality & Quantity*, 51 (2), 657-674 (2017)

5. Chin, W. W.: The Partial Least Squares Approach to Structural Equation Modeling. Marcoulides G. A. Editor, *Modern Business Research Methods*, Mahwah, NJ: Lawrence Erlbaum Associates, 295 - 336 (1998)
6. De Smedt, M. Giovannini, E. and Radermacher, V.: Measuring sustainability, For Good Measure Advancing Research on Well-being Metrics Beyond GDP: Advancing Research on Well-being Metrics Beyond GDP, OECD Publishing, 241 (2018)
7. Esposito, Vinzi, V., Chin, W. W., Henseler, J., Wang, H.: *Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications*, Springer, Berlin, Heidelberg, New York (2010)
8. Fornell, C. and Cha, J.:Advanced Methods of Marketing Research, ed. RP Bagozzi, Blackwell, Cambridge. (1994)
9. Gan, X., Fernandez, I. C., Guo, J., Wilson, M., Zhao, Y., Zhou, B. and Wu, J.: When to use what: Methods for weighting and aggregating sustainability indicators. Ecological indicators, Elsevier, 81, 491–502 (2017)
10. Hair, J.F., Hult, G.T.M., Ringle, C.M., and Sarstedt, M.: A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). 2nd ed. Thousand Oaks. CA: Sage (2017)
11. Henseler, J., Ringle, C.M. and Sinkovics, R.R.:The use of partial least squares path modeling in international marketing. *New challenges to international marketing*, 277–319, Emerald Group Publishing Limited (2009)
12. Johnson, M.D., Herrmann, A. and Huber, F.:The evolution of loyalty intentions, *Journal of marketing*, 70 (2), 122-132, SAGE Publications Sage CA: Los Angeles, CA. (2006)
13. Jones, E., Sundaram, S.and Chin, W.:Factors leading to sales force automation use: A longitudinal analysis. *Journal of personal selling & sales management*, 22 (3), 145-156, Taylor & Francis. (2002)
14. Lauro, N.C., Grassia, M.G. and Cataldo, R.:Model based composite indicators: New developments in partial least squares-path modeling for the building of different types of composite indicators. *Social Indicators Research*, 135 (2), 421–455, Springer (2018)
15. Lohmöller, J. B.: Latent Variable Path Modeling with Partial Least Squares. *Springer Science & Business Media* (2013)
16. McArdle, J.J.: Dynamic but structural equation modeling of repeated measures data. *Handbook of multivariate experimental psychology*, 561–614. (1988)
17. Roemer, E.:A tutorial on the use of PLS path modeling in longitudinal studies, *Industrial Management & Data Systems*, 116 (9), 1901-1921,Emerald Group Publishing Limited. (2016)
18. Shea, C.M. and Howell, J.M.:Efficacy-performance spirals: An empirical test, *Journal of Management*, 26 (4), 791-812, Sage Publications Sage CA: Thousand Oaks, CA (2000)
19. Stoel, R.D., van den Wittenboer, G. and Hox, J.:Methodological issues in the application of the latent growth curve model, *Recent developments on structural equation models*, 241–261, Springer (2004)
20. Stiglitz, Joseph E and Sen, Amartya and Fitoussi, J-P: Report by the commission on the measurement of economic performance and social progress. (2017)
21. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y. M., Lauro, C. N.: PLS Path Modeling. *Computational Statistics and Data Analysis*, 48 (1), 159 - 205 (2005)
22. United Nations: Millennium Development Goals Report 2009. United Nations Publications (2009)
23. United Nations: United Nations Conference on Sustainable Development, Rio+20. (2015)
24. United Nations: Transforming our world: The 2030 agenda for sustainable development. New York: United Nations, Department of Economic and Social Affairs. (2015)
25. Wetzels, M., Odekerken-Schröder, G., van Oppen, C.: Using pls path modeling for assessing hierarchical construct models: guidelines and empirical illustration. *MIS Quarterly*, 33 (1), 177 - 195 (2009)
26. Wold, H.: Path models with latent variables: The NIPALS approach. In Quantitative sociology (pp. 307-357). Academic Press (1975)
27. Wold, H.:Partial least squares. S. Kotz and NL Johnson (Eds.), Encyclopedia of statistical sciences (vol. 6), Wiley, New York (1985)

3.2 Changes in the life course and social inequality

Heterogeneous Income Dynamics: Unemployment Consequences in Germany and the US

Dinamiche di reddito ineguali: le conseguenze della disoccupazione in Germania e negli Stati Uniti

Raffaele Grotti

Abstract This paper studies income trajectories after unemployment and their stratification by education in Germany and the United States. In the specific, this paper investigates how the labour market and the household shape income trajectories and buffer income losses following unemployment, and how this varies across educational levels, between sexes and in a comparative perspective. Empirical analyses are based on SOEP and PSID data and employ distributed fixed-effects models. Results show that institutions play a considerable role in shaping the consequences of unemployment but with varying intensity across groups and countries.

Abstract *Questo articolo studia le traiettorie di reddito associate alla disoccupazione e la loro stratificazione per livello di istruzione in Germania e negli Stati Uniti. Inoltre, l'articolo indaga come il mercato del lavoro e la famiglia siano in grado di modellare le traiettorie di reddito e contenere le perdite di reddito a seguito della disoccupazione, e come questo varia tra livelli di istruzione, tra uomini e donne e in una prospettiva comparativa. Le analisi empiriche si basano su dati SOEP e PSID e utilizzano modelli a effetti fissi. I risultati mostrano come mercato del lavoro e famiglia svolgano un ruolo importante nel plasmare le conseguenze della disoccupazione, anche se con differente intensità tra gruppi e paesi.*

Key words: income dynamics, unemployment, household, distributed fixed-effects

[1] Raffaele Grotti, Dept. of Sociology and Social Research, University of Trento, email: raffaele.grotti@unitn.it

1 Introduction

This paper studies the unemployment consequences for individual income trajectories and the extent to which the household shape such trajectories in Germany and the United States. The current paper aims to expand existing research by studying how income trajectories vary between individuals with different levels of education. The experience of a critical life event may trigger processes of increasing inequality over the life-course if the least resourceful individuals experience larger negative consequences. Therefore, my first research question is: *Do income trajectories after job loss differ across educational levels?*

The interaction between events and social stratification, moreover, takes place embedded in context, i.e. country. Differences between countries may affect (1) the labour income trajectories coming with job loss; (2) and the capacity of the household to buffer income losses. This raises further research questions: *To what extent does the household shape income trajectories? Does its role differ across educational levels? How does it operate in different countries?*

Existing literature agrees that unemployment has substantial negative consequences on earnings at the time the job is lost and potentially in the subsequent years (Gangl 2006). Reemployment is the main mechanism for compensating income losses. It's buffering capacity depends on the labour market structure and the standardization and reliability of worker's educational qualifications that the educational system provides. Given countries' differences in these institutional aspects, in Germany high-educated workers will experience faster re-entry and higher wages in the new job compared to low-educated workers. This stratified pattern should be less evident in the US.

A second mechanism that can buffer the consequences of job loss operates at the household level, namely the pooling of incomes from the partner and other household members. Income pooling should be stronger for women, as men usually can provide a larger amount of income. Moreover, when the man loses the job, income pooling should be stronger in the US as compared to Germany, given the higher labour market participation and intensity of women in the US. In addition, because of assortative mating, particularly educational homogamy, the capacity of the household to compensate for income losses via partner's income should increase with education in both countries (Grotti and Scherer, 2016).

2 Data and methods

I use data from the Socio-Economic Panel for Germany and the Panel Study of Income Dynamics for the US. I focus on the period from 1984 to 2015 and select individuals from 25 to 54 years. Analysis are separate for men and women and by education.

The consequences of job loss are studied for two income concepts: *individual labour earning* (*labour income* for simplicity); and *equivalent pre-government household income* (*household income*). Income losses are presented in relative terms.

The household buffer is measured in terms of percentage points: namely the difference between the percentage loss in labour income and the percentage loss in household income. I define the job loss event as the transition from a spell of employment to a spell of unemployment that lasts at least 3 months during the year.

The model – the 'distributed' fixed-effect
I model income trajectories in the years around job loss using distributed fixed-effects models (Dougherty 2006). My fixed-effects income equation can be written as

$$(1) \quad y_{it} - \bar{y}_i = \sum_j \beta_j \left(X_{jit} - \bar{X}_{ji}\right) + \gamma(UNEMP_{it} - \overline{UNEMP_i}) + (u_i - u_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

where y_{it} is a measure of income for individual i at time t, X_{jit} is a vector of time-varying covariates other than the employment status, $UNEMP_{it}$ stands for the (un)employment status while u_i and ε_{it} are respectively the individual-specific (that via demeaning disappears) and idiosyncratic error terms. i, j and t index over individuals, time-varying covariates, and time periods, respectively. In the distributed specification, the dummy variable $UNEMP_{it}$ in eq. (1) is substituted with a set of dummy variables $UNEMP_{pit}$ as in eq. (2), where p is the number of years before unemployment if negative, and the number of years after unemployment if positive, while s represents the maximum horizon in years backward (-5) and forward (+5) from the time to unemployment. Observations in which individuals are observed more than 5 years prior or after unemployment are coded as -5 and $+5$, respectively.

$$(2) \quad \sum_{p=-s}^{s} UNEMP_{pit} = UNEMP_{it}$$

In line with this model specification, the reference category ($s = -5$) also includes those who never experience unemployment. Results are presented in two-year intervals. In addition to the distributed employment status, the models control for partnership status, number of children younger than 14, age, age squared, and year.

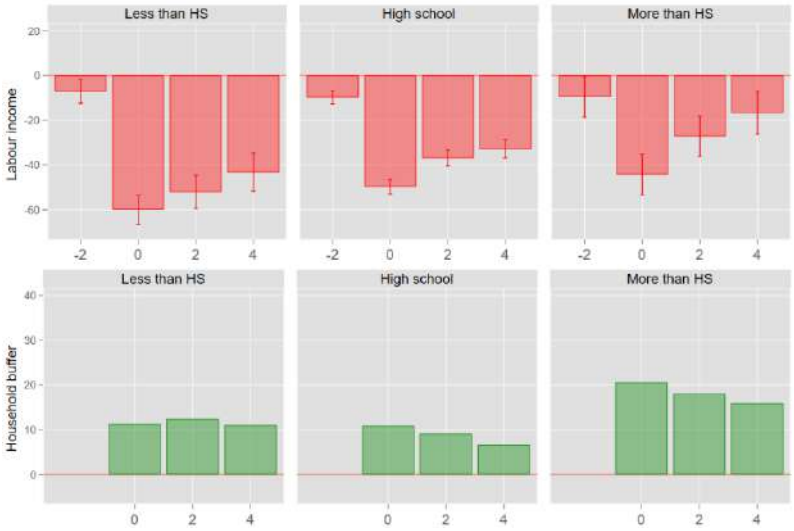
3 Results

Germany
I present income trajectories separately for levels of education and at different points in time. The first row in Figure 1 reports income trajectories in men's labour income in terms of percentage income losses with respect to 5 years before job loss. In the year of job loss, German men with less than high school experience an income reduction of 60%. Those with high school education lose 50% of income while those with more than high school lose 44%. The low educated also face more difficulties in recovering from their income losses later on, at both two and four years after the event. Results confirm my first expectation: a larger penalty in the after the event for the low educated compared to the high educated.

Results for German women (Figure 2) largely resemble those for German men: the higher the level of education, the lower the income losses. This is especially true

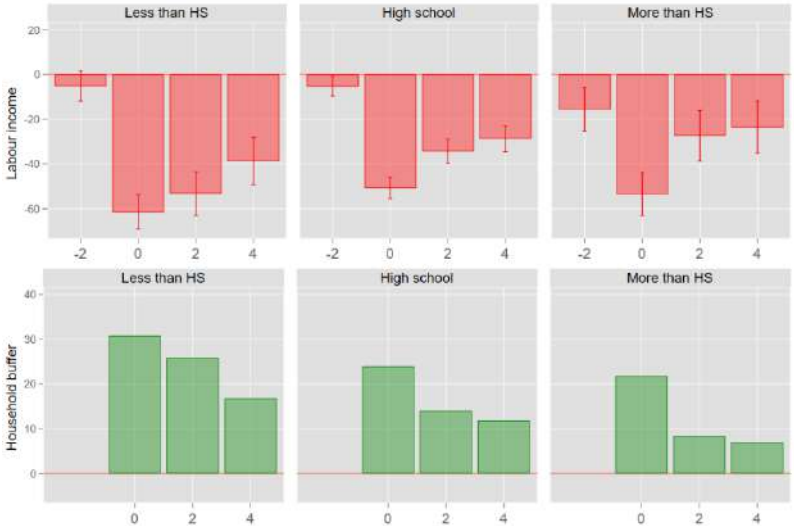
in the years after the event.

Figure 1: Estimated income trajectories for Individual Labour Income, and the Household buffer, by education. Germany, men



The second row of Figure 1 reports the 'household buffer'. For German men, the household redistributive capacity reduces income losses, especially among the high educated. German men, benefit from a smaller household buffer than women, which can be attributed to the lower employment of women – which less likely to provide additional income to the male partner.

Figure 2: Estimated income trajectories for Individual Labour Income, and the Household buffer, by education. Germany, women



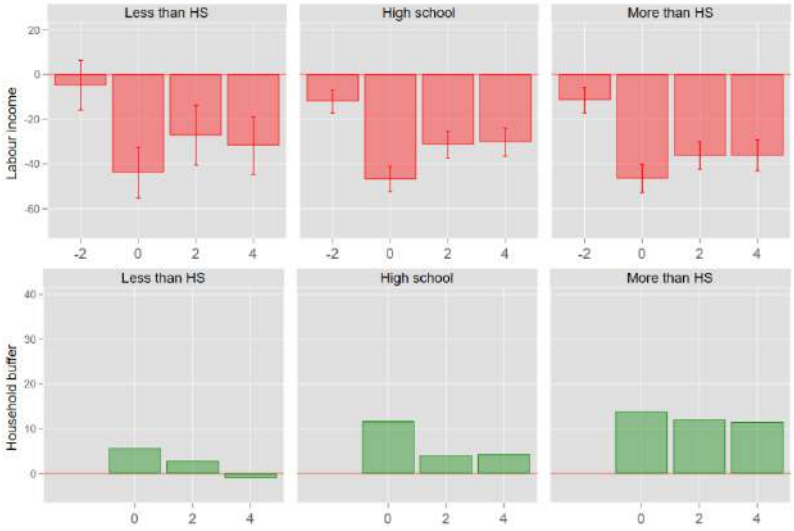
Differences across levels of education are in line with the expectation only for men: larger household buffer for the most educated. The pattern that we observe for

women could be partly explained by the largest share of total household income that high-educated women earn. In such situation, woman unemployment reduces household income strongly and the household will have a limited capacity to buffer the loss.

United States

I now turn my attention towards the US. At the time of the event, men lose between 44 (low-educated) and 47 (mid- and high-educated) percent of their labour income (Figure 3). In the following years, the losses decrease only slightly. As expected, in the US income losses after job loss do not vary significantly across levels of education.

Figure 3: Estimated income trajectories for Individual Labour Income, and the Household buffer, by education. United States, men



Income trajectories for American women, presented in Figure 4, are similar to those for men. However, we observe a larger penalty for the less educated groups in the year of job loss (55 percent), although differences are contained – by 3 and 7 percentage points compared to mid- and high educated respectively.

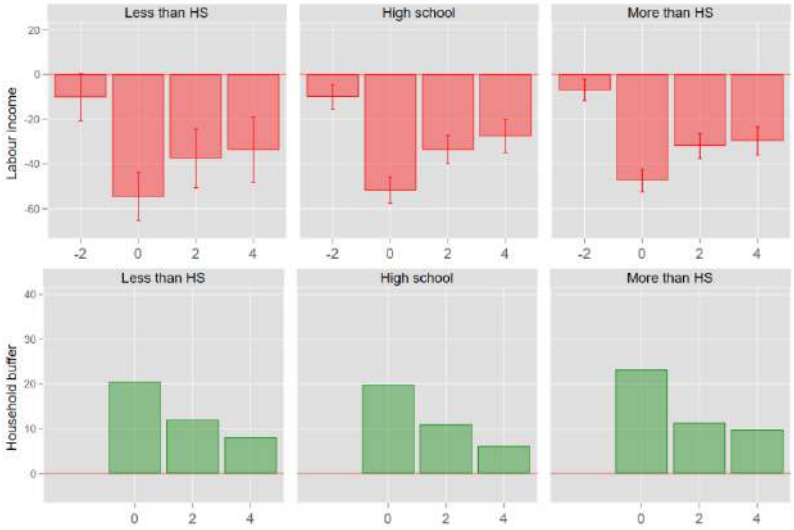
The household buffer reduces men's losses to a limited extent, with the most educated benefitting the most over the trajectory. The household role is considerably larger for women compared to men. However, for women I find only limited support to the expectation of a household buffer that increases with education.

4. Conclusions

This paper investigates the economic consequences of job loss in international comparison and among individuals with different educational levels. I find some evidence of the accumulation of inequalities over the life-course. Overall, at the time of the event we observed lower income losses in the US than in Germany, which might

be attributed to the higher dynamism of the US labour market. Education affects the consequences of job loss and exacerbates the disadvantage of the less educated, especially in Germany. The household substantially contributes in managing the consequences of unemployment. Women benefit the most from the support of other household members, mainly the partner, in both countries. Comparing men in the two countries, American men are not supported by their partner to a greater extent than German men, notwithstanding the higher labour market intensity of women in the US.

Figure 4: Estimated income trajectories for Individual Labour Income, and the Household buffer, by education. United States, women



To conclude, institutions in both countries play a considerable role in strengthening inequality and in (re)producing the system of socioeconomic stratification. These patterns are especially clear-cut for men. The market leads to an accumulation of disadvantages: the least educated have the lowest levels of income, experience the highest risk of unemployment, and suffer from the largest income losses with job loss. The market operates as an inequality 'booster', especially in Germany. The role of the household points toward the same direction by supporting the highest educated the most, with the exception of German women. Overall, my results show that the household plays a substantial role in shaping income trajectories of individuals. However, the extent to which it mitigates the income losses associated with unemployment varies according to several aspects, especially gender.

References

1. Dougherty, C. (2006). The marriage earnings premium as a distributed fixed effect. Journal of Human Resources, 41(2), 433-443.
2. Gangl, M. (2006). Scar effects of unemployment: An assessment of institutional complementarities. American Sociological Review, 71(6), 986-1013.
3. Grotti, R., & Scherer, S. (2016). Does gender equality increase economic inequality? Evidence from five countries. *Research in Social Stratification and Mobility*, 45, 13-26.

In-work poverty in Germany and in the US: The role of parity progression

Povertà da lavoro in Germania e negli Stati Uniti: il ruolo della transizione al primo figlio e delle nascite successive

Emanuela Struffolino and Zachary Van Winkle

Abstract This study contributes to the analysis of in-work poverty by analysing the short- and mid-term associations between parity transitions and the in-work poverty risk across the life course. Longitudinal data from the US and Germany are applied to between-within random effects linear regression models to estimate how the risk of in-work poverty increases initially following the transition to a first, second, and third child as well as how the initial increase in the risk of in-work poverty changes in the following six years. By comparing these two countries and tracking how the in-work poverty risk changes at the different steps of the household income generative process, we gain insights on whether institutional arrangements can ameliorate the association between parity transitions and the risk of in-work poverty.

Abstract *Questo studio contribuisce all'analisi della povertà da lavoro (*in-work poverty*) considerando l'associazione tra la transizione al primo figlio (e le nascite successive) e il rischio di* in-work poverty. *Applicando modelli di regressione lineare a effetti misti a dati longitudinali per gli Stati Uniti e la Germania, stimiamo il rischio di* in-work poverty *dopo la transizione a un primo, secondo e terzo figlio. Consideriamo, inoltre, le dinamiche di tale rischio nei sei anni successivi a ciascuna transizione. Confrontando Stati Uniti e Germania, osserviamo come cambia il rischio di* in-work poverty *in corrispondenza delle fasi del processo di generazione del reddito familiare: mostriamo, dunque, se e in che misura diverse configurazioni istituzionali moderano l'associazione tra la transizione alla nascita di un figlio e il rischio di* in-work poverty.

Key words: parenthood, poverty, cross-national comparison, life course, welfare transfers

[1] Emanuela Struffolino, Humboldt-Universität zu Berlin; emauela.struffolino@hu-berlin.de
Zachary Van Winkle, SciencePo; zachary.vanwinkle@sciencespo.fr

1 Introduction

In-work poverty is an increasing concern in most Western affluent democracies. The working poor are employed individuals who live in households whose income is below the poverty threshold. Most research has studied the association between structural characteristics (education and social class) as well as ascriptive characteristics (gender and race) with in-work poverty (e.g., Lohmann and Marx, 2018). However, family demographic processes, such as the transition to parenthood, as well as marriage and divorce, are tightly intertwined with labour market dynamics (Van Winkle & Struffolino, 2018). With a few exceptions, the literature has mainly relied on cross-sectional analyses and to date no comparative research has studied how the in-work poverty risk changes in correspondence of family demographic transitions and in the years to follow.

We aim at filling this gap by focusing on the short- and medium-term associations between parity progression and in-work poverty. We concentrate on parity progression, specifically the transition to first, second, and third child, for at two reasons. First, although it is acknowledged that household compositions has an influence on the probability to be employed and live in poverty (see Polizzi, Struffolino and Van Winkle [2020] and Crettaz [2013] for a review), there is no research on how changes in household composition due to parity progression affect the risk of in-work poverty and whether those vary by the number of children. Second, higher parity progression might not only increase households' risk of entering in-work poverty, but the number of young children exposed to households at risk of social exclusion.

We focus on the United States and Germany as ideal typical representatives of liberal and corporatist-conservative welfare states. The German labour market is highly regulated and German family policy is characterized by generous income support for families with children and long parental leave. In contrast, US family policy is residual without parental leave schemes and limited mostly to income tax credits and targeted relief for poor families with children. Comparing these two countries allows us to gain initial insights on whether institutional arrangements can ameliorate the association between parity transitions and the risk of in-work poverty across the life course. To achieve this, we will adopt the common Eurostat indicator for in-work poverty (being a worker whose household income is below the 60% of the median) and track how the in-work poverty risk changes at the different steps of the household income generative process: from market income to equivalized household income before and then after taxes and transfers.

2 Data and methods

US Panel Study of Income Dynamics (PSID 1970-2015) and the German Socio-Economic Panel (SOEP 1984-2017) included in the Cross-National Equivalent File (CNEF) are used. PSID and SOEP are both nationally representative household

panels. PSID sampled approximately 18,000 individuals within 5,000 households in 1968 and continued to collect economic, sociological and demographic information annually until 1997, since then on a biennial basis. SOEP began in 1984 with a sample of roughly 12,000 respondents living in West Germany and added a sample of approximately 5,000 East Germans in 1990.

We constructed three household samples to analyze how the risk of poverty changes following the transition 1) to parenthood, 2) to a second child, and 3) to a third child. Each household sample included households that (a) were observed to make a specific transition, for example to parenthood in the case of the first sample, and (b) households that did not make that transition, for example households that remain childless in the case of the first sample. All samples were restricted to single and couple households with a combined work intensity of at least 1,040 hours in the previous year, which corresponds with one adult working full-time for at least 26 weeks or working part-time for a full year. Households without an adult between age 18 and 50 were also excluded. The transition to parenthood sample was restricted a) to households that reported the birth of a first child or b) households without children that were not observed to transition to parenthood as a control. We included all observations before and after the transition to parenthood, retaining observations (household-years) even when households had additional children: this was consistent with our goal of estimating the association between the transition to parenthood and in-work poverty in the years following for all households. We constructed our samples to study the transition to a second and to a third child in a similar manner.

Outcome variables. In accordance with the EUROSTAT individuals with net equivalized household incomes under 60% of the annual median were considered to be in relative poverty. We estimated annual medians using the entire samples weighted to be nationally representative for the given year. We used this threshold to estimate changes in the in-work poverty risk for four different income types: 1) individual market income, pre-taxes pre-transfers equivalized household income (OECD scale), 2) pre-taxes pre-transfers equivalized market household income (OECD scale), 3) pre-taxes pre-transfers equivalized market household income including all additional sources of income, and 4) post-taxes post-transfers equivalized household income. Using four different income variables but one poverty threshold to define poverty status allows us assess how the consideration of needs, additional non-market income, and taxes and transfers change households' risk of entering poverty following the birth of a first, second, and third child.

Independent variables. To examine both short- and medium-term changes following the transition to a first, second, and third child, we included both a binary and a continuous indicator. For the analyses on the transition to parenthood, our binary indicator took the value of one when households had transitioned to have one child and zero when they were childless. The continuous indicator counted the number of years following the transition to first child and was zero while the household was childless and in the year of the transition. We constructed the binary and continuous indicators for the analyses on the transition to a second and third child in a similar manner. When these variables were simultaneously included in the regression models, the binary indicator captured the initial change following the

transition to a first, second, or third child, and the continuous indicator captured changes after the year of the respective transition. We used a quadratic specification of years after each transition, because it is more parsimonious than a non-parametric specification with categorical duration variables: given our sample size a categorical specification would be inefficient and reduce our capability to make comparisons across income types, time and country simultaneously.

Analytical strategy. We used between-within random effects linear regression models (Sjölander et al., 2013), also known as hybrid random effects regression models (Allison, 2009), with observation years nested in households to estimate changes in the probability of in-work poverty. To consistently estimate the within-person effects of parity transitions controlled for all time-constant covariates and unobserved characteristics, while simultaneously estimating between-person effects, all time-varying covariates entered in the model twice: one as household specific mean (between-effect); and one as deviation from the household mean (within-effect). We adjusted our models for the following variables referring to the focal person (highest earning member of the household): years of education, marital status, occupational group, gender, age. We further include in the models indicators for average years of observation and number of earners. For the United States we additionally controlled for race, and for Germany for East/West.. We also include the percentage distance from the poverty threshold in the year prior to childbirth to account for baseline differentials in the probability of in-work poverty that depend on household incomes before childbirth. Results are presented as changes in coefficients and predicted probabilities.

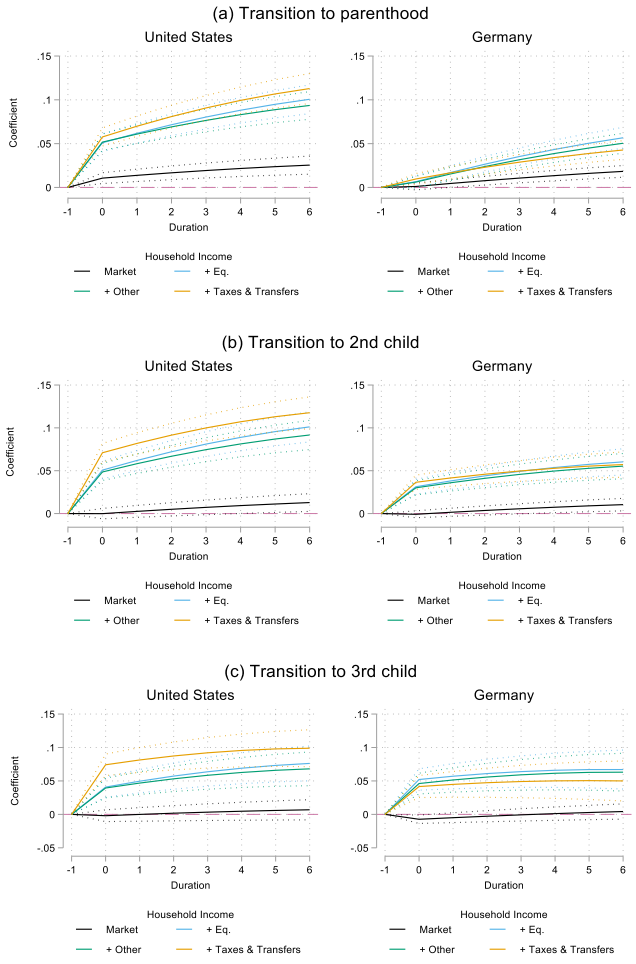
3 Main results

Figure 1 displays the results for changes in the in-work poverty risk after the transition to parenthood (panel a), the second child (panel b), and the third child (panel c) by income type in the US and Germany. Duration (x-axis) is equal to 0 in correspondence of the year when the child was born, so that the value of the coefficient is interpreted in changes in the in-work poverty risk compare to the previous year. In the US, the transition to parenthood is associated with an increase of 2 percentage points in the risk of being among the working poor based on market income: the risk significantly increases to 5 percentage points once equivalization is applied. Accounting for transfers and taxes does not alleviate the increase in the in-work poverty risk associated with parenthood. Most importantly, the risk of in-work poverty does not decrease in the following years but rather increases up to 10 percentage points by the sixth year after childbirth.

The figure for Germany differs to a large extent: the transition to parenthood is not associated with a significantly higher risk of in-work poverty when looking at market income. A small increase in the in-work poverty risk of around 1 percentage point emerges after equivalization. Similarly to the US, taxes and transfers to not mitigate the (admittedly small) effect and the in-work poverty risk increases over time, reaching 5 percentage points in the six years following the birth of the first

child. The picture for the transition to second and third child in the US mirror the one for the transition to parenthood, both in terms of size of the effect and of the role of transfers and taxes. On the contrary, for Germany the transition to the second and the third child are associated with an increase in the in-work poverty risk that is 4 and 5 times higher respectively compare to the transition to the first child, with not significant changes over the following six years. Also in these cases, taxes and transfers do not decrease the in-work poverty risk when applied to the household equivalized income.

Figure 1: Parity transitions and changes in in-work poverty risk by income type



Source: PSID (1970-2015) and SOEP (1984-2017). X-axis: duration from 1 year before to 6 years after childbirth.

4 Discussion and next steps

The causes of poverty identified in the literature are remarkably consistent and that the most important factor is a change in employment. Following the transition to parenthood, household resources will decrease if one earner, typically the mother, withdraws from the labor market or reduces their work intensity. Regardless of mothers' work intensity, household incomes may decrease if women's earnings decrease following childbirth. As children grow older and enter (formal) education, mothers are likely to re-enter the labor market, but changes in labor market supply and employment quality that (at least one of the) parents experienced have also long term consequences, as intermittent or reduced labor supply sum up to what can be conceptualized as the "career costs" of children (Adda et al., 2017). Our results seem to support the notion that increased needs are not off-set by market income nor state taxes and transfers.

Interestingly these considerations apply to both the US and Germany both before and after taxes and transfers, although the two contexts differ in terms of family policy and labor marker deregulation. One explanation for this unexpected finding relies on the fact that although 60% is the threshold commonly used to identify the poor in the working population, it might be too high to single out the effect of welfare in mitigating the in-work poverty risk after childbirth. In other words, the households whose income is below the 60% of the median are likely to be highly heterogeneous in terms of the actual access to welfare support in both contexts. For example, in the US individuals and households may qualify for accessing Supplemental Nutrition Assistance Program (SNAP) if they earn a gross monthly income that is 130% (or less) of the federal poverty level, which is based on an extremely low (compare to the relative one we use here) absolute threshold. These results hint at the fact that welfare transfers for the working poor represent a relief only for those at the very bottom of the income distribution and fail to address less extreme economic need, which nevertheless expose households with children to persistent vulnerability over time.

5 References

1. Adda, J., Dustmann, C., & Stevens, K. (2017). The Career Costs of Children. Journal of Political Economy, 125(2), 293–337. https://doi.org/10.1086/690952
2. Allison, P. (2009). *Fixed Effects Regression Models*. Sage Publications.
3. Crettaz, E. (2013). A state-of-the-art review of working poverty in advanced economies: Theoretical models, measurement issues and risk groups. *Journal of European Social Policy*, *23*(4), 347–362. https://doi.org/10.1177/0958928713507470
4. Lohmann, H., Marx, I. (2018). *Handbook on In-Work Poverty*. Edward Elgar Publishing.
5. Polizzi, A., Struffolino, E., Van Winkle, Z. (2020). Family Demographic Processes and In-Work Poverty: A Systematic Review. SocArXiv, August 4, https://doi.org/10.31235/osf.io/zncaq
6. Sjölander, A., Lichtenstein, P., Larsson, H., Pawitan, Y. (2013). Between-within models for survival analysis. *Statistics in Medicine*, *32*(18), 3067–3076. https://doi.org/10.1002/sim.5767
7. Van Winkle, Z., Struffolino, E. (2018). When working isn't enough: Family demographic processes and in-work poverty across the life course in the United States. *Demographic Research*, *39*(12), 365–380.

Parenthood, education and social stratification. An analysis of female occupational careers in Italy

Maternità, istruzione e stratificazione sociale. Un'analisi delle carriere occupazionali delle donne in Italia

Gabriele Ballarino and Stefano Cantalini

Abstract This paper studies the role of education and parenthood in shaping social inequality over the life course of Italian women. It asks a) if social inequalities already appear at young ages, and if they decrease, remain stable or increase over time; and b) if not only educational attainment, but also parenthood can contribute to shape these inequalities and their change over the life course. Preliminary findings, based on growth curve models estimated on Multipurpose Survey data (2009), show that gross differences between social classes in terms of labour market participation increase in the first years after completion of studies and then remain stable, whereas those in terms of occupational success are already visible at young ages and increase over the life course. Education appears to be the main factor explaining these differences, whereas parenthood primarily contributes to inequalities via different effects on female career according to social origin.

Abstract *Questo lavoro si focalizza sul ruolo dell'istruzione e della maternità nello sviluppo delle disuguaglianze sociali lungo il corso di vita delle donne italiane. Ci si chiede se: a) le disuguaglianze sociali si formano già in giovane età e se diminuiscono, persistono o aumentano lungo la carriera; b) se l'istruzione e la maternità contribuiscono a plasmare tali disuguaglianze e il loro cambiamento nel corso di vita. I primi risultati, basati su modelli growth curve stimati su dati Multiscopo (2009), mostrano che le differenze secondo le origini sociali aumentano nei primi anni post-uscita dal sistema scolastico e poi rimangono stabili se si considera la partecipazione al mercato del lavoro, mentre sono già visibili in giovane età e aumentano nel corso di vita se si considera il successo occupazionale.*

[1] Gabriele Ballarino, University of Milan; email: gabriele.ballarino@unimi.it
Stefano Cantalini, University of Milan; email: stefano.cantalini@unimi.it

L'istruzione è il fattore principale alla base di queste differenze, mentre la maternità contribuisce all'aumento delle disuguaglianze per via delle diverse conseguenze che essa esercita sulle carriere femminili a seconda delle origini sociali.

Key words: parenthood, education, social origin, career, growth curve

1 Introduction and background

Research on social stratification and mobility has extensively studied the relationship between social origins, education and occupational destinations (Blau and Duncan 1967). A vast literature has thus shown that family socio-economic background is substantially associated to occupational achievement, even when educational attainment is controlled for (see Bernardi and Ballarino 2016). However, analyses of intergenerational mobility have primarily looked cross-sectionally at snapshot measures of occupational destinations (e.g., first job, occupation at the moment of the interview, etc.). Only recently, thanks to the greater availability of panel data, research has started to study these issues from a dynamic perspective (see Ballarino et al. 2020), analysing the whole process of career development.

Despite this growing attention towards the analysis of intergenerational mobility from a dynamic perspective, longitudinal studies on women are still scarce (Härkönen et al. 2016). In fact, focusing on the female population requires to enlarge the theoretical and analytical framework and integrate social stratification and mobility research with the sociology of the family, focusing, among other things, on possible penalties related to family dynamics. Female careers are indeed strongly affected by family dynamics such as marriage and parenthood (see Cantalini 2020 for a review), which are key events in the processes of family formation and transition to adulthood. These issues have been seldom considered in their interaction with social inequalities, despite timing and propensities of entering a union or having children differ according to social background and career penalties and premia related to family formation can change across social groups. In this respect, the question is how much of the association between social origin and social destination among women is explained by factors related to the family sphere, and whether the latter are as important as education in accounting for this association.

This paper aims at answering these research questions, by studying the role of education and parenthood in shaping social inequality over the life course of Italian women. By means of a dynamic analysis of the association between social origin and social destination, it first asks if social inequalities among women appear already at young ages and if these inequalities decrease, remain stable or increase over time. Moreover, it investigates if not only educational attainment, but also parenthood can contribute to shape inequalities and their change over the life course.

2 Data and methods

Analyses were based on data from the Multipurpose Survey – Family and Social Actors (2009), a sample survey conducted by the Italian Statistical Institute (ISTAT). Retrospective longitudinal information was collected on a sample of about 8,000 families and 44,000 individuals, allowing us to reconstruct educational, career and family histories and to create a longitudinal yearly dataset. The analytical sample included 8,856 women born 1930-1980, who completed their studies at least twenty years before. Among these women, 5,867 had at least one employment episode. We followed them in the first twenty years after the conclusion of the educational career, truncating observations at age 15 and censoring at age 54.

We focused on two dependent variables: a) labour market participation, operationalised with a dummy equal to one if the individual was employed; b) occupational attainment, measured with the Standard International Occupational Prestige Score (SIOPS, Treiman 1977). Of course, the first variable was measured among the whole sample, whereas the second referred only to those women who had at least one employment episode.[2] Studying both variables was crucial if the focus are women, and it allowed to indirectly take into account the differential selection into employment.

The independent variable was social class of origin, constructed with the dominance principle and operationalised with a modified version of the Cobalti and Schizzerotto (1994) class scheme. We distinguished five social classes: service class (Ser), white collars (WhC), urban petit bourgeoisie (UPB), urban working class (UWC) and agricultural classes (Agr).

The main control variables, which allow studying the role of different factors on the social origin gap, referred to education and parenthood. Educational attainment was considered as a time-constant variable and measured using the most detailed categorization available in the data: illiterate or without education, primary, lower secondary, upper secondary, tertiary, and post-tertiary educated. Parenthood was measured in two ways. First, number of children was used, a time-varying variable with the following categories: childless, one child, two children or more. This variable was also interacted with marital status (single, married, divorced or widowed) in order to better control for the family situation over the life course. Second, we focused on the time elapsed from parenthood, entered as a set of dummy variables to capture the effect of years after first birth (year of first birth; one year; two years; 3-4 years; 5-7 years; and 8 years or more after parenthood).

Moreover, we controlled for years since completion of studies (our time axis, entered in yearly dummy variables) and for an interaction between geographical residence (North-west, North-east, Centre, South and Islands) and period dummies.

We estimated growth curve models with random effects (Halaby 2003), which have been used to study career progression in the recent social stratification and

[2] When SIOPS was studied, episodes of unemployment or inactivity were dropped from the sample. For an analysis including these episodes (i.e. career breaks) in the empirical strategy, see Cantalini (2020). Moreover, results on SIOPS are confirmed if alternative specifications of the time axis are used, e.g. years of effective work experience.

mobility literature (e.g., Härkönen et al. 2016; Ballarino et al. 2020). More specifically, we estimated group specific growth curves (Brüderl et al. 2019) by interacting social class of origin and years since completion of studies, in order to study how social inequalities evolve over the life course.

Three models were estimated, separately for each dependent variable: the first only controlled for an interaction between geographical residence and calendar year (model 1); the second included educational attainment (model 2); the third also controlled for an interaction between marital status and number of children (model 3). Additionally, a fourth model with full controls – with the exception of number of children, excluded because of collinearity – was estimated, including also an interaction between social class of origin and time since first birth. This model allowed us to study if parenthood contributes to social inequalities not only because of compositional effects (i.e. different parenthood probabilities across individuals from different social origins), but also as a result of differential career consequences of motherhood according to social origin.

3 Preliminary findings

Preliminary analyses show that differences between social classes in terms of labour market participation increase in the first years after completion of studies, reaching their peak around nine years after the end of school, when the probability of being employed for women from the UWC (agricultural classes) is 14.3 (14.6) percentage points (p.p.) lower compared to women from the service class (fig. 1, mod 1). After that, the employment gap decreases in the long run among women from the middle and agricultural classes – although it remains substantially relevant for the latter (9.0 p.p., twenty years after finishing school) –, whereas it persists without noticeable changes among those from the UWC.

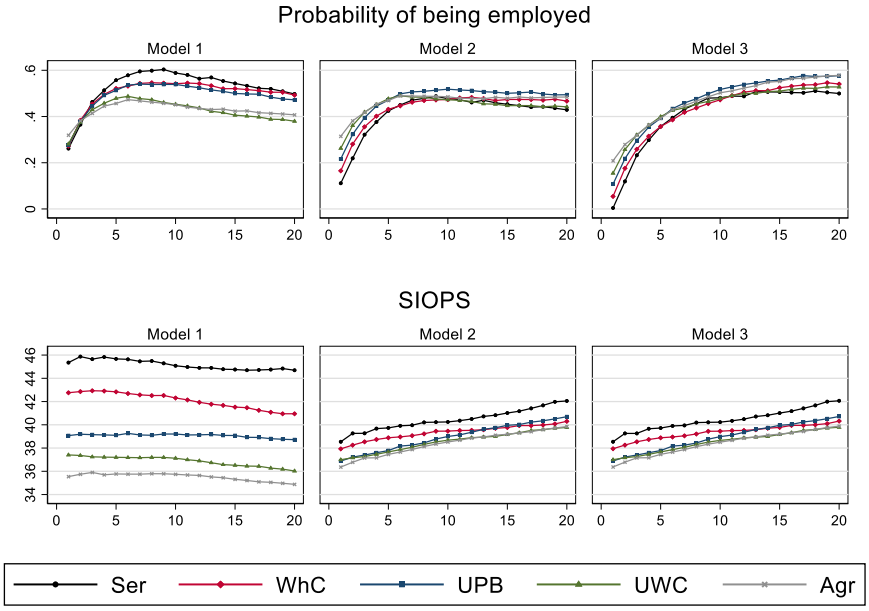
Gross differences by social origin – which appear already at the beginning of the career – increase if occupational prestige is considered, since the SIOPS of women from the service class remains substantially stable over the life course, while the SIOPS of women from other classes (with the only exception of UPB) starts to decrease around ten years after the end of school (fig. 1, mod. 1).

Education appears to be the main factor explaining the association between social origin and social destination. In fact, the social origin gap, net of education, in labour market participation is virtually non-existent across the whole life course, or it even changes direction (mod. 2). The gap in occupational prestige decreases as well if controlled for education, both at the beginning of the career as well as in the long run (mod. 2). Consistently with a pattern of cumulative advantage, however, it slightly increases over the career: the SIOPS of women from the service class improves more than the one of women from the lower classes over the life course, pointing to a gap of 2.28 (2.18) points in the SIOPS scale at the end of our observational window if children of the UWC (agricultural classes) are considered.

Parenthood (in its interaction with marital status) seems to account for the association between social origin and social destination to a much lower extent

(mod. 3). This primarily occurs because there are no huge differences in the propensities of having one or more children across social classes in the observational window. However, parenthood still contributes to increasing social inequalities over the life course as a result of different career trajectories after the birth of a child for women with different social background. Indeed, the probability of being employed strongly decreases in the year of motherhood for women from UWC, and it further diminishes in the following years (fig. 2). On the contrary, although women from the service class decrease their labour market attachment immediately after motherhood, in their case the penalty remains substantially stable over the life course.

Figure 1: Social class of origin, probability of being employed and SIOPS, by years since completion of studies. Linear (probability) models with random effects. Predicted probabilities

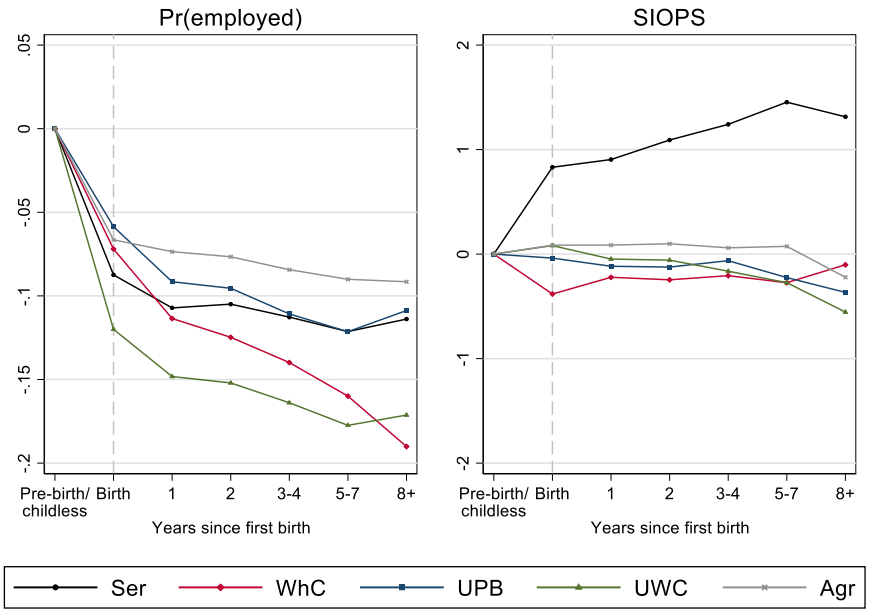


Source: Multipurpose Survey – Family and Social actors (2009)

Social inequalities slightly increase after parenthood also if occupational achievement is considered, not because of a motherhood penalty among the lower classes – whose SIOPS is not affected by the birth of the first child –, but rather because of a (small) premium for the children of the service class.

In conclusion, career differences by social origin among women slightly increase over the life course and are primarily driven by education, via inequality of educational opportunities. Parenthood does not contribute to these differences – and the changes therein – because of compositional effects, rather because of different career trajectories after motherhood, with smaller penalties (or premia) among women from the service class and higher penalties among children of the lower classes (UWC *in primis*), especially if labour market participation is considered.

Figure 2: Parenthood effects on the probability of being employed and SIOPS, by social class of origin. Linear (probability) models with random effects. Average marginal effects



Source: Multipurpose Survey – Family and Social actors (2009)

References

1. Ballarino, G., Cantalini, S. and Panichella, N. (2020). Social origin and compensation patterns over the occupational career in Italy. *Acta Sociologica*, doi:10.1177/0001699320920917.
2. Bernardi, F., Ballarino, G. (2016, eds.). *Education, Occupations and Social Stratification. A Comparative Analysis of the Transmission of Socio-Economic Inequalities*. Cheltenham: Elgar.
3. Blau PM and Duncan OD (1967) *The American Occupational Structure*. New York: Free Press.
4. Brüderl, J., Kratz, F., & Bauer, G. (2019). Life course research with panel data: an analysis of the reproduction of social inequality. *Advances in Life Course Research*, 41, 100247.
5. Cantalini, S. (2020). *Famiglia e disuguaglianza: matrimonio, fecondità e posizione sociale nell'Italia contemporanea*. Milano: Franco Angeli.
6. Halaby CN (2003) Panel models for the analysis of change and growth in life course studies. In: Mortimer J and Shanahan MJ (eds) *Handbook of the Life Course*. Hingham: Kluwer Academic, 503–527.
7. Härkönen, J., Manzoni, A., Bihagen, E. (2016). Gender inequalities in occupational prestige across the working life: An analysis of the careers of West Germans and Swedes born from the 1920s to the 1970s. *Advances in Life Course Research*, 29: 41-51.

3.3 Composition in the Data Science Era

Can we Ignore the Compositional Nature of Compositional Data by using Deep Learning Aproaches?

Possiamo ignorare la natura composizionale dei dati composizionali usando gli approcci di deep learning?

Matthias Templ

Abstract Care must be taken not to simply apply multivariate data analysis methods to compositional data. For example, one can show that correlations are biased to be negative, and almost all statistical methods result in biased estimates when applied to compositional data. One way out is to analyze data methods from compositional data analysis, i.e. by carrying out a log-ratio analysis. This contribution has its focus on settings where only the prediction and classification error is important rather than an interpretation of results. In this setting it is well-known that classification and prediction errors are smaller with a log-ratio approach using traditional machine learning methods. However, is this also true when training a neural network who may learn the inner relationships between parts of a whole also without representing the data in log-ratios? This contribution give an indication on this matter using one real data set from chemical measurements on beers.

Abstract Bisogna fare attenzione a non applicare semplicemente i metodi standard di analisi statistica multivariata ai dati composizionali. Per esempio, si può dimostrare che le correlazioni sono distorte in quanto necessariamente negative, e che quasi tutti i metodi statistici portano a stime distorte quando vengono applicati ai dati composizionali. Una possibile soluzione consiste nel ricorrere all'approccio logratio, basato su logaritmi di rapporti delle componenti. Questo contributo si concentra su un'impostazione in cui il focus dell'analisi è tenere sotto controllo l'errore di previsione e di classificazione piuttosto che l'interpretazione dei risultati. In questo contesto, è noto che gli errori di classificazione e di previsione sono più piccoli con un approccio logratio utilizzando metodi tradizionali di deep learning. Tuttavia, questo è vero anche quando si addestra una rete neurale, che può imparare le relazioni interne tra le parti di un insieme, anche senza rappresentare i dati sottoforma di logratio? Questo contributo fornisce un'indicazione su questo tema utilizzando un set di dati reali relativo a misurazioni chimiche sulle birre.

Matthias Templ
Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Rosenstrasse 3, 8400 Winterthur, Switzerland, e-mail: matthias.templ@zhaw.ch

Key words: compositional data analysis, artificial neural networks, prediction error

1 Compositional Data

Compositional data appears to be multivariate observations of a whole consisting of only parts above zero. Care has to be taken not to simply apply multivariate data analysis methods as per usual. For compositional data only the relative information of the observations is of interest. Consider a D-part composition $x = [x_1, \ldots, x_D]'$ with strictly positive parts $x_1, ..., x_D$. The same relative information is contained in x_i/x_j and $(ax_i)/(ax_j)$ for any non-zero scalar value a. The composition can be re-expressed as proportions, $x^* = ax$ by setting $a = 1/\sum x_i$. The composition x^* belongs to the $(D-1)$-standard simplex defined by

$$\left\{ x^* = \left[x_1^* \ \ldots \ x_D^* \right]' \mid x_i^* > 0, \sum_{i=1}^{D} x_i^* = 1 \right\} \quad .$$

In order to apply standard multivariate data analysis methods, the compositions need to be represented in the Euclidean space, for example, by using isometric log-ratio transformations, discussed in detail in Filzmoser et al. [2018].

The class of isometric logratio (ilr) coordinates aims to form an orthonormal basis in a hyperplane and express the composition in it. The resulting vector $\mathbf{z}$ is in $\mathbb{R}^{D-1}$, i.e. the isometric log-ratio transformtion maps the data from the simplex to the Euclidean vector space.

One particular choice of a basis leads to

$$\text{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$$

with

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}} \quad \text{, for} \quad j = 1, \ldots, D-1. \tag{1}$$

These ilr coordinates are referred as *pivot* (logratio) *coordinates* [Filzmoser et al., 2018]. Such a choice has also a primary importance for the coordinate system as a whole. z_j summarizes now all relative information (logratios) about x_j, and can thus be interpreted as the relative dominance of x_j within the given composition. These special pivot coordinates are used for model-based imputation of missing values or rounded zeros, but also generally often in the regression context. Finally, $z_j = 0$ indicates a balanced state between x_j and an average behavior of the other parts in the given composition.

2 Artificial Neural Networks

A neural network is just a non-linear statistical model [Hastie et al., 2009], which is based on a transformed (with an activation function) weighted linear combinations of the sample values. Each neuron is formed by transformed weights and thus a neuron hold some information on each variable and a set of observations. Each neuron has an activation, depending on how the input information looks like. A layer in a deep neural network is a collection of neurons in one step. There are three types of layers: the input layer, the hidden layers, and the output layer. The goal of training a network is to find the optimal weights for each connection between the neurons of two layers. After setting initially all weights randomly, a loss function of the network is defined. The output of the loss function is a single number judging the quality of the neural network. To lower the value of the loss function, an adaptive moment estimation called Adam [Kingma and Ba, 2014, Ruder, 2016] is used (other methods can be selected), which is a stochastic gradient descend method that uses adaptive learning rates for each parameter of the algorithm. With this gradient descend optimization all the weights are optimized to reach the next local minimum resulting in the most rapid decrease. This is causing the most rapid decrease in the loss function. The (stochastic) gradient used to adjust these weights is computed with back propagation, whereby the weights gets updated. Choosing a proper activation function, a deep neural network is able to find non-linear relationships between a target variable and predictors.

Next to the choice of the activation function a lot of parameters must be chosen. In brackets is our choice: the loss (mean squared error) and evaluation (absolute error) metrics, the number of epochs (500), a patient value (50), the number of layers (10), the number of neurons in each layer (1000, 900, ..., 100), possible dropout (10% in the first 5 layers) are the most important parameters to choose adequately to prevent over- or underfitting.

3 Data and Results

To demonstrate if one can ignore the compositional nature in a classification context, a real-world dataset on chemical compounds of beer had been selected.

The 48 beers of the beer data set [Varmuza et al., 2002] had been measured before and after beer aging, resulting in 96 measurements. The chemical measurements were taken using a gas spectrometer. It should be noted that beer experts can recognize a beer brand based on its constant, fresh flavor. Beer aging results in the production of stale flavors, regardless it is heat-induced or inadequate storage; thus, the original flavor is lost. A good classifier should recognize whether a measurement is from an aged beer, because the chemical composition differs [Templ and Templ, 2020].

Figure 1 represents an extended analysis of Templ and Templ [2020] using artificial deep neural networks. The following methods were compared: naive bayes,

linear discriminant analysis (lda), k nearest neighbor classification (knn), generalized linear models with family multinomial (glm), random forest with parameter tuning (cforest) and aritificial deep neural networks (ann). The dataset was put into classification untreated (*no*), column normalized to mean 0 and variance 1 (*normalized*), logarithmized (*logarithm*), logarithmized and normalized (*norm + log*), brought to constant row sum (*percentages*), brought to constant row sum and normalized (*percentages_norm*), displayed in centered and pivot coordinates. The missclassification was determined with 10 times repeated 10-fold cross validation, but for the artificial neural network a cut-off was made due to high computation times and a simplified cross validation with 5 times repeated random assignment in 75% training data and the rest as test data was used and the average reported.

It is shown that in principle a classifier applied to pivot coordinates gives better results, i.e. compositional data analysis results in smaller misclassification rates than normalizing or logarithmizing only. Random forests and artificial neural networks clearly represent non-linear methods. While for random forests a pivot log-ratio transformation helps as well, the results for logarithmized and normalized data are best for ann, followed by pivot coordinates. The presentation in constant row sums (e.g. as percentages) have negative influence on almost all methods, but especially the aritifical deep neural network cannot deal with such constraints.

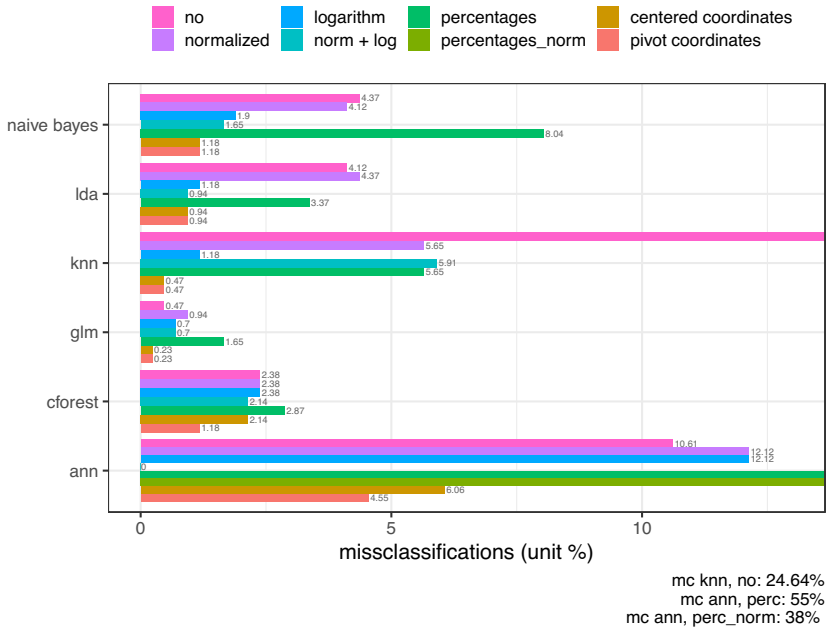


Fig. 1 Missclassification rate (in %) for different machine and deep learning methods, transformations and normalizations, for the beer data set.

4 Discussion and Conclusion

In the context of results of Templ [2020], which looked at the performance of imputing rounded zeros with artificial neural networks (similar context with compositional consideration and without it), and Lubbe et al. [2021], which looked at the performance of imputing rounded zeros from high-dimensional bionomic data, the following can be stated.

Templ [2020] and Lubbe et al. [2021] showed that a compositional treatment of data is important even for artificial neural networks and improves the results. The neural network learns better when the data is represented in pivot coordinates. Applying a neural network to data without taking the special nature of compositional data into account leads to worse results. Thus, the neural network does not learn all non-linear (compositional) relationships, although enough layers and neurons have been chosen and an optimal parameter setting has been set.

For the classification of the beer data set, the results are not black/white but a tendency is visible. We recommend a compositional pre-treatment of the data before estimating an artificial neural network. The tendency is that it gives better results and the neural network has an easier time learning the non-linear (compositional) relationships. And it can go really wrong when working with untransformed representation of the data in percentages (or generally with constant row-sums). For classification problems where a prediction error or misclassification rate is in the foreground, a compositional approach is therefore also advantageous for an artificial neural network.

This finding thus provides the result that - although an artificial deep neural network can learn non-linear dependencies - taking into account the compositional nature of the data by using appropriate logarithmic representations of the data is benefitial also before applying such non-linear methods. The results can improve significantly and the stochastic gradient algorithms used to optimise the network perform better in log-ratio coordinates.

Future work includes testing the methods on various other data sets and comparing them in simulation studies.

References

P. Filzmoser, K. Hron, and M. Templ. *Applied Compositional Data Analysis*. Springer International Publishing, 2018. ISBN 9783319964225. doi: 10.1007/978-3-319-96422-5.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2009. ISBN 978-0-387-84857-0.

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

S. Lubbe, M. Templ, and P. Filzmoser. Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems*, page 13, 2021. accepted for publication.

S. Ruder. An overview of gradient descent optimization algorithms, 2016. URL http://arxiv.org/abs/1609.04747. arXiv: 1609.04747.

M. Templ. Artificial neural networks to impute rounded zeros in compositional data, 2020.

M. Templ and B. Templ. Analysis of chemical compounds in beverages- guidance for establishing a compositional analysis. *Food Chemistry*, 325:1–7, 2020. doi: 10.1016/j.foodchem.2020.126755.

Kurt Varmuza, Ingrid Steiner, Thomas Glinsner, and Helmut Klein. Chemometric evaluation of concentration profiles from compounds relevant in beer ageing. *European Food Research and Technology*, 215(3):235–239, 2002. doi: 10.1007/s00217-002-0539-5.

Principal balances for three-way compositions

Bilanciamenti principali per composizioni a tre vie

Violetta Simonacci

Abstract Orthonormal balances resulting from a sequential binary partition (SBP) are one of the preferred tools for transforming compositional data in real space coordinates.The interpretability of this approach, however, greatly depends on the relevance of the SBP. SBPs can be chosen with the help of expert knowledge or with data-oriented methods, such as Principal Balances analysis. This results in an SBP whose balances maximize the explained variance in a subsequent manner. Principal balances can be calculated in an exact way or in an approximate fashion by using methods based on PCA for compositional data. In this work a method for the approximation of principal balances in the more complex case of three-way compositions is proposed. Here the additional difficulty given by the introduction of third mode variability is dealt with. In particular an algorithm based on the Tucker3 model is used which allows to keep the variability of the third dimension separate in the definition of principal balances.

Abstract *I bilanciamenti ortonormali risultanti da una partizione binaria sequenziale (SBP), sono uno dei metodi più usati per la trasformazione dei dati composizionali in coordinate reali. L'interpretabilità di questo approccio, tuttavia, dipende in gran parte dalla rilevanza dell'SBP. Una SBP può essere scelta con l'ausilio di conoscenze specialistiche o con metodi basati sui dati, come l'analisi dei bilanciamenti principali. Ciò si traduce in una SBP i cui bilanciamenti massimizzano in maniera sequenziale la varianza spiegata. I bilanciamenti principali possono essere calcolati in modo esatto o approssimativo utilizzando metodi basati sull'analisi delle componenti principali. In questo lavoro viene proposto un metodo per l'approssimazione dei bilanciamenti principali nel caso più complesso di composizioni a tre vie. Qui viene affrontata l'ulteriore difficoltà data dall'introduzione della variabilità del terzo modo. In particolare viene utilizzato un algoritmo basato sul modello Tucker3 che permette di mantenere separata la variabilità della terza dimensione nella definizione dei bilanciamenti principali.*

Violetta Simonacci
University of Naples Federico II - Department of Social Sciences ,Vico Monte della Pietà, 1 - 80138 Napoli e-mail: violetta.simonacci@unina.it

Key words: CoDa, Principal Balances, SBP, three-way data, Tucker3

1 Introduction and preliminary concepts

Compositional data (CoDa) are vectors of positive elements referred to as compositional parts because they describe the proportions of a whole. For this reason, these vectors carry relative information and are characterized by a negative covariance bias. An evident consequence of this constraint is that CoDa's geometry is non-Euclidean. Compositions are forced in a subspace of the real space, called simplex, therefore direct application of conventional statistical tools is inappropriate.

A set of I CoDa vectors comprising the same J-parts can be arranged in two-way matrices $\mathbf{V}(I \times J)$. If the same compositions are recorded at different K occasions (locations, time-points, etc) the K two-way matrices obtained can be considered as the frontal slices of a three-way compositional profile data array $\underline{\mathbf{V}}(I \times J \times K)$ and identified as $\mathbf{V}_k$. These three-way compositions present the multiple difficulty of having to deal with both a constrained structure and three-fold variability.

For two-way data, in order to overcome the problem of a bounded sample space CoDA are typically projected onto the unconstrained real space by expressing them in logarithms of ratios between parts. Several types of such transformations have been proposed. Specifically, Aitchison suggested three transformations [1, 2].

(i) The pairwise log-ratio (*plr*), which simply considers all the possible ratios among parts and has the disadvantage of greatly increasing data size. (ii) The additive log-ratio (*alr*) which considers the ratios of all component with a reference element and has the disadvantage of providing a non-isometric mapping of the original vector. (iii) The centered log-ratio (*clr*) which provides an isometric mapping by considering the ratios between the elements of a compositional vector and its geometric mean. The *clr* transformation has, however two major shortcomings: it yields a singular covariance matrix and does not provide an orthonormal basis. So far, due to its simplicity this has been the preferred method for most three-way applications where complex tools such as the Candecomp/Parafac and the Tucker3 models have been used to study the data accounting for three-way variability [3, 13].

Nonetheless, another transformation introduced by Egozcue *et al.* [4] ensures much better defined properties and could simplify interpretation under certain circumstances: the isometric log-ratio (*ilr*) transformation. This alternative provides an isometric isomorphism and allows to express CoDa with respect to an orthonormal basis.

A way to find *ilr*-coordinates for a D-part composition is to obtain a set of $D-1$ orthonormal balances with a sequential binary partition (SBP) [5] which divides the composition into a sequence of non-overlapping groups all characterized by the contrast of two sub-compositions. In this manner the total variance is also decomposed into the sum of the variances explained by each balance. The associated orthonormal basis in the simplex is composed of the corresponding balancing elements.

The ease of interpretation of *ilr*-coordinates obtained in such a fashion strongly relies on the capability of finding a meaningful SBP. This can be done manually with the help of expert knowledge, if available. Alternatively to manual selection of a proper SBP, Principal Balance Analysis (PBA) can be used instead. This methods is based on data exploration and aims to identify a set of principal balances (PB) which successively maximize the explained variance of the data.

The problem with the exact computation of PBs is that it becomes unfeasible as the number of parts increases [10, 6, 9, 11]. This is because the number of all possible SBPs to consider becomes hard to manage for compositions with a large number of parts. As shown in [10], for a J-part composition, the number of possible SBP is $\frac{J!(J-1)!}{2^{J-1}}$ thus approaches which provide an acceptable approximation of PBA search may be a better option than exact estimation. One of this approaches is the Maximum explained Variance hierarchical balances (MV) approach which uses CoDa-PCA, i.e. PCA performed on *clr*-coordinates, as a starting point.

For three-way composition the selection of an adequate SBP in terms of principal balances becomes even more challenging because studying K occasions means having to account for different variability structures. Each frontal slice represents its owns dataset with a given variability which may differ from other occasion especially if the data does not present close-to-perfect trilinearity. Thus, the best SBP for the $k = 1$ slice may not be the best one for the $k = 2$ slice. For this reason a method similar to the PBA approach modified as to allow the simultaneous modeling of the diffferent structure of the K matrices could be a good solution for a feasible approximation of PBs in a three-way setting.

In this work a three-way PBA method is proposed. The structure of the estimating algorithm is quite similar to that of MV, however instead of using Coda-PCA as a starting point CoDa-Tucker3 is used instead. The procedure is detailed in Section 2 while is Section 3 initial findings and a conclusive discussion are presented.

2 Methods

Balances are generally described by the following generic notation:

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{g_m(\mathbf{v}_+)}{g_m(\mathbf{v}_-)}, \tag{1}$$

where $\mathbf{v}_+$ and $\mathbf{v}_-$ are two non-overlapping sub-compositions of a complete $J-$part composition $\mathbf{v}$; r and s are the number of parts in $\mathbf{v}_+$ and $\mathbf{v}_-$ respectively, and g_m indicates the geometric mean of corresponding parts.

Two-way PBA looks for orthonormal PBs which maximize explained variability in a subsequent manner, so for a set of I compositions of J-parts arranged in a $\mathbf{V}(I \times J)$ matrix it yields $J-1$ Principal Balances obtained as the projection of the data on the corresponding balancing elements. A new matrix of *ilr*-coordinates $\mathbf{Y}(I \times (J-1))$ is obtained. This search can be carried out by considering all possible sets of

PBs, however, this is highly demanding from a computationally stand point, thus, sub-optimal algorithms which provide an adequate estimation can be used instead, like the MV algorithm.

For three-way compositions, estimation of PBs is an even more challenging task. Let us consider an array of compositions $\underline{\mathbf{V}}(I \times J \times K)$ with frontal slices $\mathbf{V}_k(I \times J)$. The selection of optimal Principal Balances for each $\mathbf{V}_k(I \times J)$ may differ as the variability structure of each slice is different. Thus, the first step would be to find a feasible compromise. One solution could be to simply avaraging all slices and then carrying out a standard PBA. As well-known in three-way analysis, however, averaging has always the great disadvantage of merging modes variability [8]. In this perspective, a way to obtain one estimate for all slices through proper modeling can provide a more reliable solution.

To study the variability structure of multiple populations simultaneously the Tucker3 model [14] can be used, which is in essence a three-way principal component analysis. This model yields four sets of parameters: (i) three loading matrices $\mathbf{A}_{(I\times P)}$, $\mathbf{B}_{(J\times Q)}$ and $\mathbf{C}_{(Q\times R)}$ where P, Q and R are the number of components extracted for the first, second and third mode respectively; (ii) and a tridimensional core array $\underline{\mathbf{G}}(P \times Q \times R)$ with generic element g_{pqr} representing the strength of the connection between each triad of components pqr.

In a compositional setting the Tucker3 model can be easily carried out on an array of *clr*-transformed data $\underline{\mathbf{Z}}(I \times J \times K)$, much like CoDa-PCA, without any particular issue (see [7]). Using the element-wise notation, the Coda-Tucker3 model can be written as:

$$z_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr}(a_{ip} b_{jq} c_{kr}) + e_{pqr}, \tag{2}$$

where z_{ijk} is the generic element of the array of *clr*-coordinates, e_{pqr} is the generic element of the array of residual $\mathbf{E}(I \times J \times K)$ and a_{ip}, b_{jq} and c_{kr} are the generic elements of the first, second and third mode loading matrices.

The standard algorithm returns orthonormal loading matrices even though the model is characterized only by subspace uniqueness, thus infinite bases could be identified. The advantage of using the loading matrix $\mathbf{B}$ for identifying a reliable SBP is that each mode subspace is estimated and can be read separately, in this way the influence of occasion variability is minimized.

At this point, just as in the MV algorithm, the first component is used to identify the first principal balance by contrasting parts with positive loadings and parts with negative loadings as in eq. 1. Afterwards, checks for variability changes given by moving each positive part to the negative grouping is performed and then the PB is stored. In a similar fashion the remaining PBs are found by performing the same steps on the largest sub-composition for a total of $J-1$ PBs.

3 Preliminary results and discussion

As a preliminary test of the presented methodolgy a three-way data-set on student satisfacion has been considered. The dataset is given by the scores assigned by the avarage student of 10 degree programs ($I = 10$) to 18 items (J=18) which can be intepreted as parts of a composition, since they represent different aspects of total student satisfacion [12]. Scores were recorded for 6 consecutive years from 2012 to 2017 ($K = 6$). In brief the considered parts of the compositiona are: 1 General Workload (G_W), 2 General Organization (G_O), 3 Individual Time (I_T), 4 Starting knowledge (S_K), 5 Topic (TOP), 6 Work load (W_L), 7 Teaching material (T_M), 8 Other activities (O_A), 9 Exam procedures (E_P), 10 Teacher timetable (R_H), 11 Teacher Motivation (MOT), 12 Teacher Clarity (CLA), 13 Teacher Availability (AVA), 14 Teacher Helpfulness (HEL), 15 Classroom (CLS), 16 Equipment (EQU), 17 Interest in the topic (INT), 18 Overall Satisfaction (SAT). For more details on the data-set set see also [15].

The Tucker3 methodology applied to the described data yielded the SBP presented in Table1 where the parts included in the $\mathbf{v}_+$ and the $\mathbf{v}_-$ sub-compositions of the corresponding principal balance are identified with the "+" and "-" symbols respectively. Initial findings show that the PBs obtained with the proposed three-way-PBA are different from those simply obtained with a standard PBA (exact or approximated) on the average matrix. Thus, a much more reliable SBP which considers the three-way nature of the data is ensured.

Table 1 SBP yielded by the three-way PBA

	G_W	G_O	I_T	S_K	TOP	W_L	T_M	O_A	E_P	R_H	MOT	CLA	AVA	HEL	CLS	EQU	INT	SAT
1	-	-	-	-	+	-	-	+	+	+	+	+	+	+	-	-	+	+
2					-			-	-	+	+	+	+	+			+	-
3	-	-	-	-		-	-								+	+		
4										+	-	-	-	+			-	
5	+	+	+	-		-	-											
6					-			+	+									-
7											-	-	+				+	
8				+		+	-											
9	-	-	+															
10	+	-																
11				+		-												
12													+				-	
13											+	-						
14					+													-
15								+	-									
16										+				-				
17															+	-		

Interpretation is coherent with the *clr* log-contrasts obtain with a PARAFAC model, however they appear much easier and straight forward to interpret and to further analyze. The first PB, for example, which explains over 55% of total variability shows the contrast between elements which evaluate the professional characteristics of the teacher against all other aspects of satisfaction (student readiness, facilities, course origanization, etc...). To further assess the methodology presented, a full

comparison with standard PBA on the average matrix with different estimation approaches will be developed. Additional steps also may be included in the estimation routine, such as rotations. scaling or proper representations, in order to improve results readability.

References

1. Aitchison, J.: The statistical analysis of compositional data. J. R. STAT. SOC. B. **44**(2), 139–160 (1982)
2. Aitchison, J.: The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability, Chapman & Hall Ltd., London (1986) (Reprinted in 2003 with additional material by The Blackburn Press)
3. Bergeron-Boucher, M.P., Simonacci, V., Oeppen, J., Gallo, M.: Coherent modeling and forecasting of mortality patterns for subpopulations using multiway analysis of compositions: an application to canadian provinces and territories. N. Am. Actuar. J. **22**(1), 92–118 (2018)
4. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric logratio transformations for compositional data analysis. Math. Geol. **35**, 279–300 (2003)
5. Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. Math. Geol. **37**(7), 795–828 (2005)
6. Filzmoser, P., Hron, K., Templ, M.: Methods for High-Dimensional Compositional Data. In Applied Compositional Data Analysis (pp. 207-225). Springer, Cham (2018)
7. Gallo, M., Simonacci, V.: A procedure for the three-mode analysis of compositions. Electron. J. Appl. Stat. Anal. **6**(2), 202–210 (2013)
8. Kroonenberg, P.M.: Applied multiway data analysis (volume 702). John Wiley & Sons (2008)
9. Martín-Fernández, J. A., Pawlowsky-Glahn, V., Egozcue, J. J., Tolosona-Delgado, R.: Advances in principal balances for compositional data. Math. Geosci, 50(3), 273-298 (2018)
10. Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana Delgado, R.: Principal balances. Proceedings of the 4th International Workshop on Compositional Data Analysis (2011)
11. Quinn, T.P.: Visualizing balances of compositional data: A new alternative to balance dendrograms. F1000Research **7** (2018)
12. Simonacci, V., Gallo, M.: Statistical tools for student evaluation of academic educational quality. Qual. Quant. **51**(2), 565–579 (2017)
13. Simonacci, V., Gallo, M.: Detecting public social spending patterns in Italy using a three-way relative variation approach. Soc. Indic. Res. **146**(1-2), 205–219 (2019)
14. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
15. VALMON: Gruppo di Ricerca sulla Valutazione ed il Monitoraggio delle Politiche e dei Servizi dell'Università degli Studi di Firenze. Copyright 2006 - VALMON s.r.l, Firenze (2018) Available at `https://valmon.disia.unifi.it/sisvaldidat/unifi`

Robust Regression for Compositional Data and its Application in the Context of SDG

Regressione Robusta per Dati Composizionali e un'Applicazione nel Contesto degli SDG

Valentin Todorov and Fatemah Alqallaf

Abstract In the course of the Agenda 2030 of Sustainable Development, agreed by all UN Member States in 2015, 17 Sustainable Development Goals (SDGs) with 169 specific targets have been identified to addresses the balance between the aspiration for a better life with the limitations imposed by nature. These measure different aspects of the economic, social and environmental development within countries and their mutual relationships has been the topic of numerous studies. We are interested in one especially important question, namely, how the size of different sectors of the manufacturing industry influences the well-being of population. In this case, as in many practical situations of data analysis of social, economic and technical data, the explanatory variables describe the relative contributions of the components on the whole and the sum of the variables (parts) is not important, i.e. we are dealing with compositional data. As it is often the case, outliers might have influence on the results of the analysis, and to cope with this we apply robust regression using MM-type estimates. The inference is performed through estimating the the distribution of the parameters using fast and robust bootstrap.

Abstract *Nel corso dell'Agenda 2030 per lo sviluppo sostenibile, approvata da tutti gli Stati membri delle Nazioni Unite nel 2015, sono stati identificati 17 obiettivi di sviluppo sostenibile (SDG) con 169 obiettivi specifici per affrontare l'equilibrio tra l'aspirazione a una vita migliore e i limiti imposti dalle natura. Questi obiettivi misurano diversi aspetti dello sviluppo economico, sociale e ambientale all'interno dei paesi e le loro interconnessioni sono state oggetto di numerosi studi. Siamo interessati a una questione di particolare importanza, ovvero come le dimensioni dei diversi settori dell'industria manifatturiera influenzano il benessere della popolazione. In questo caso, come in molte situazioni pratiche di analisi dei dati sociali, economici e tecnici, le variabili esplicative descrivono il contributo relativo di ciascuna componente rispetto al loro insieme e la somma delle variabili (parti) non è*

Valentin Todorov
UNIDO, Vienna, Austria, e-mail: valentin@todorov.at

Fatemah Alqallaf
Kuwait University, Kuwait, Kuwait, e-mail: fatma.alqallaf@ku.edu.kw

rilevante, in altri termini si tratta di dati composizionali. Come spesso accade, i valori anomali potrebbero avere un'influenza sui risultati dell'analisi, pertanto per far fronte a questa criticità applichiamo una regressione robusta utilizzando stime di tipo MM. L'inferenza viene eseguita stimando la distribuzione dei parametri attraverso l'algoritmo fast-robust bootstrap.

Key words: Industrialization, SDG, robust regression, compositional data

1 Introduction

The Sustainable Development Goals (SDGs) are the core of the 2030 Agenda for Sustainable Development and drive the global, regional and national development in the future. SDG 9 calls for building resilient infrastructure, promoting sustainable industrialization and fostering innovation and at the same time through the SDG interlinking promotes advancements in a large number of social, economic and environmental goals. The manufacturing sector and related service industries provide jobs, generate incomes, thus reducing poverty, and allows for rapid and sustained increases in living standards. A number of UNIDO reports search for evidence on how closely industrial development is linked to people's living conditions and the quality of their lives [see 9, and the references therein]. The level of industrialization, often measured by the manufacturing value added (MVA) per capita, is highly correlated with many social indicators. The higher a country's industrial development, the more resources are available for human development. Although industrialization contributes to the universal objective of economic growth, its impact differs depending on the country's stage of development. One very important question that can help reveal some specifics of development is how the size of different sectors of the manufacturing industry influences the well-being of population. To answer this substantial question, the manufacturing industry is divided according to the technology intensity in the sector into the three groups: *low technology*, *medium-low technology manufacturing* (referred to as medium-technology), and *medium-high and high technology manufacturing* (referred to as high-technology). Then the influence of the value added in these groups (relative to the total manufacturing value added) on social indicators related to different SDG goals, is analyzed. Alternatively, instead of using manufacturing value added, the country's exports, disaggregated into four categories could serve as proxy for measuring the industrialization intensity as shown in [7].

2 Compositional Data

Compositional data were defined traditionally as multivariate data with positive values that sum up to a constant (1 or 100 per cent or any other constant), i.e. constrained data [1]. Nowadays this definition is generalized in more practical terms to any set of multivariate observations with strictly positive components where relative rather than absolute information is relevant for the analysis. Often this property of the data is ignored, although linear regression models are only reasonable if the covariates $\mathbf{x} = (x_1, \ldots, x_D)'$ carry absolute information [3].

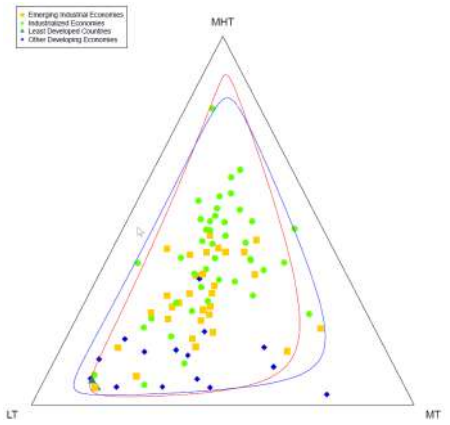


Figure 1 A ternary diagram displaying the proportion of the three technology groups: *low technology (LT)*, *medium-low technology (MT)*, and *medium-high and high technology (MHT)* which sum to 100%. The backtransformed Mahalanobis distance tolerance ellipses, classical (blue) and robust (red), identify a number of outliers (the countries falling out of the red ellipse).

Compositional data are described by the so-called *Aitchison* geometry which is isomorphic to a $D-1$ Euclidean space. The isomorphism between the two spaces is called *isometric logratio (ilr) transformation* and can be given as an orthonormal basis

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \log \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, \ldots, D-1. \tag{1}$$

All relevant information about the compositional part x_1 is given by z_1 and a permutation of the other parts $x_2, \ldots, x_D$ does not change the value of z_1 [3]. To represent the information about the other compositional parts $x_l, l = 1, \ldots, D$, a different orthonormal basis can be constructed by replacing $(x_1, \ldots, x_D)$ with $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D) := (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$:

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \log \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1. \tag{2}$$

With the transformed, $D-1$ dimensional, covariates $\mathbf{z}^{(l)}$, linear regression models can be used to analyze the data. The parameter estimates and inference statistics

obtained with the model for $\mathbf{z}^{(l)}$ are only interpretable for the parameter of $z_1^{(l)}$. To get information about all compositional parts, D models have to be fit to all possible transformations $\mathbf{z}^{(l)}, l = 1, \ldots, D$. In our analysis, the value added is split into 3 parts, resulting in 3 regression models for 3 different orthonormal bases. Figure 1 shows a ternary diagram with proportional representations of the three explanatory variables. The backtransformed Mahalanobis distance tolerance ellipses, which were computed using the Minimum Covariance Determinant (MCD) estimator [2] identify a number of outliers (the countries falling out of the red ellipse). Due to potential outlying data points, ordinary least squares regression would not give reliable results.

3 Robust Regression with Compositional Data

The approach for linear modeling of compositional data proposed by Hron *et al.* [3] is based on classical linear regression and thus will be sensitive to the presence of outliers in the data set. To cope wit this, we propose to use robust regression which is the means to analyze data when outlying data points are present. For this analysis, robust MM-type estimates for linear regression [10], will be used. The MM estimator has a high efficiency under the linear regression model with normally distributed errors. Because it is initialized at the high breakdown point S-estimates, it is also highly robust to outliers (see e.g. Chapter 5 in [4]). The computation will be performed with the R function `lmrob` from the **robustbase** package [5].

The regression models for the three different orthonormal bases are of the form

$$y = \beta_0 + \beta_1^{(l)} z_1^{(l)} + \beta_2^{(l)} z_2^{(l)} + \varepsilon \quad l = 1,2,3 \tag{3}$$

where y is the value of a social indicator and the $z_i^{(l)}$ are the isometric logratios of the share in total value added with different orthonormal bases l.

In order to get appropriate point-wise confidence intervals for the parameters $\beta_1^{(l)}$ without assuming a certain data distribution, the distribution of the parameters was estimated using fast and robust bootstrap [6].

4 Example: Life Expectancy at Birth

To illustrate the approach we consider the example described in the introduction. The data comes from different international statistical databases. Data on value added by industry was obtained from the UNIDO INDSTAT database available at stat.unido.org which comprises industrial statistics for all 22 divisions of the manufacturing sector in 174 countries. Several variables, including the value added, are

available from 1963 to 2018, but only data for 2018 has been used for this example. To reduce the complexity of the analysis, instead of the 22 divisions, a derived classification into three technology groups *low technology*, *medium-low technology manufacturing*, and *medium-high and high technology manufacturing*, as defined in [8, p. 244], was used. The final value added, aggregated for the three technology groups, was available for 91 countries. Data on overall human development were taken from the *Human Development Report* published by United Nations Development Programme http://hdr.undp.org/en.

Life expectancy at birth is a key indicator for SDG 3 (Ensure healthy lives and promote well-being for all at all ages). For an appropriate analysis of this relation, the compositional structure of the data must be taken into consideration. Furthermore, to control the influence of outlying data points, robust regression estimates have to be employed and fast and robust bootstrap yields reliable inference for these estimates. The value of the parameter $\beta_1^{(l)}\, l = 1,2,3$ indicates how much the response

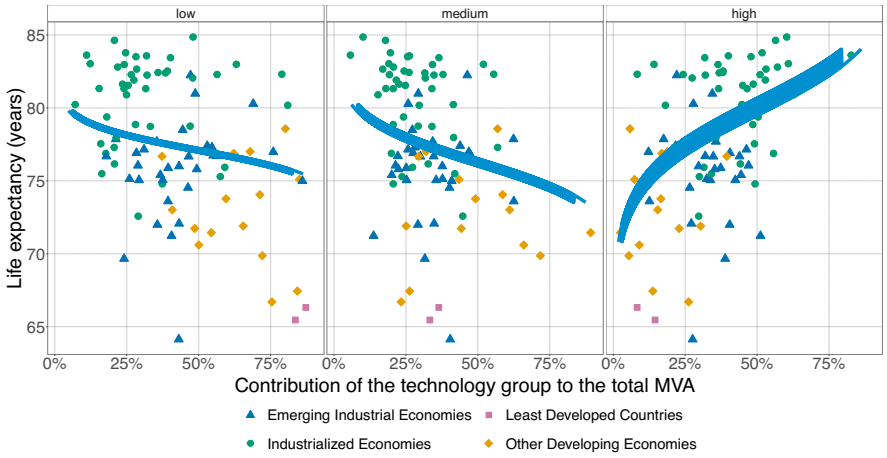


Figure 2 The life expectancy at birth indicator as a function of the contribution of each of the three technology groups to the total MVA.

variable changes on average by a unit change of the logratio between the contribution of the sector of interest and an average of the contributions of the remaining sectors. While the relative size of the low- and medium-technology sector has a negative effect on the life expectancy, a larger high-technology sector results in a higher life expectancy. The most pronounced influence has the contribution of the high-technology industry. The estimated coefficient is high and significantly positive (3.05814; p-value is 4.8e-07) (it lies between 1.94 and 4.18 with 95% certainty, and is therefore significantly positive). The values of $\beta_1^{(1)}$ and $\beta_1^{(2)}$ are negative, so the larger the share of the low- or medium-technology industry in total MVA is, the

lower the life expectancy. If using the standard model, the high-technology sector does not have a significant influence on the life expectancy anymore, in comparison to the very high influence in the model with the *ilr*-transformed value added.

5 Summary and Conclusions

Well-being is strongly influenced by the relative contribution of different manufacturing industries to the total manufacturing value added. To build appropriate regression models, it is crucial to be aware of the compositional nature of the data and for the estimates and inference to be resistant to outlying data points. Taking into account all these aspects, the resulting regression models for different social indicators and the structure of manufacturing support the statement. A large contribution of the high-technology manufacturing industries helps to significantly enhance human development.

References

[1] Aitchison, J.: The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability, Chapman & Hall Ltd., London (UK) (1986)
[2] Filzmoser, P., Hron, K.: Outlier detection for compositional data using robust methods. Math. Geosci. **40**, 233–248 (2008)
[3] Hron, K., Filzmoser, P., Thompson, K.: Linear regression with compositional explanatory variables. J. Appl. Stat.**39**, 1115–1128 (2012)
[4] Maronna, R. A., Martin, D., Yohai, V., Salibian-Barrera, M.: Robust Statistics: Theory and Methods (with R): Second edition. John Wiley & Sons, New York (2019)
[5] Maechler, M., Rousseeuw, P. J., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., di Palma, M. A.: **robustbase**: Basic Robust Statistics. http://CRAN.R-project.org/package=robustbase, R package version 0.93-7 (2020)
[6] Salibian-Barrera, M., Zamar, R.: Bootrapping robust estimates of regression. Ann. Stat. **30** 556–582 (2002)
[7] Todorov, V.: Monitoring robust estimates for compositional data. Austrian J. Stat. **50** 16–37 (2021)
[8] UNIDO: Industrial statistics: Guidelines and methodology. Report, Vienna (2010)
[9] UNIDO: How industrial development matters to the well-being of the population. Report, Vienna, http://stat.unido.org (2020)
[10] Yohai, V. J.: High breakdown-point and high efficiency robust estimates for regression. Ann. Stat. **15** 642–656 (1987)

3.4 Evaluation of undercoverage for censuses and administrative data

Spatially balanced indirect sampling to estimate the coverage of the agricultural census

Campionamento indiretto spazialmente bilanciato per la stima di copertura del censimento dell'agricoltura

Federica Piersimoni and Francesco Pantalone and Roberto Benedetti

Abstract Coverage error in census has become an important statistical issue. In this paper we address the design of the coverage survey through the use of spatially balanced sampling designs and the employment of indirect sampling framework. Spatially balanced sampling exploits the spatial component of the target population, while indirect sampling is taken into account since a frame linked to the target population is assumed to be available. Some proposals are presented and their efficiency investigated by means of Monte Carlo simulations.

Abstract *L'errore di copertura nei censimenti è diventato un importante problema statistico. In questo articolo studiamo il campionamento dell'indagine di copertura attraverso l'uso di campionamenti spazialmente bilanciati e il framework del campionamento indiretto. Campionamenti spazialmente bilanciati catturano la componente spaziale della popolazione di interesse, mentre il campionamento indiretto è preso in considerazione siccome ipotizziamo la disponibilità di un frame collegato alla popolazione di interesse. Alcune proposte sono presentate e la loro efficienza investigata per mezzo di simulazioni Monte Carlo.*

Key words: Agricultural Census, Coverage Survey, Indirect Sampling, Spatially Balanced Sampling.

Federica Piersimoni
Directorate for Methodology and Statistical Process Design, Istat, Rome, Italy, e-mail: piersimo@istat.it

Francesco Pantalone
Dept. of Economics, University of Perugia, Italy, e-mail: francesco.pantalone@studenti.unipg.it

Roberto Benedetti
Dept. of Economic Studies, "G. d'Annunzio" University, Pescara, Italy, e-mail: benedett@unich.it

1 Introduction

Coverage error in census is due to omissions or duplications of statistical units in the census enumeration. This has become an important statistical issue. For example, the U.S. Bureau of the Census has been sued in federal court more than 50 times regarding the completeness of the 1980 census. In order to obtain an estimate of the coverage error additional information is necessarily needed. Indeed, an estimate of the coverage error cannot be obtained from the census data themselves, and usually an independent sample survey is obtained over the same target population U. This independent survey is often called *coverage survey (CS)* and can either precedes (pre-enumeration survey) or follows (post-enumeration survey) the census. Once the data from the coverage survey are obtained, a model is employed in order to estimate the under/over-count. In agricultural census, the CS is usually employed for the selection of areas, and coverage rate of the agricultural holdings are of interest, i.e. the ratio between the number of farms pointed out over the Census and the number of really existing farms.

In 2010, ISTAT performed the agricultural census with the aim of enumerate all the agricultural holdings in the country. Afterwards, ISTAT carried out the CS aimed at providing a measurement of the degree of coverage of the Census with respect to the population of farms through an areal sample where the final sampling units were about 1,500 cadastral maps extracted from the Land Registry Office [ISTAT, 2013]. It was designed with a two-stage sampling. In the first stage, municipalities were stratified according to their provinces. For each stratum, a number of municipalities were selected with probability proportional to the number of agricultural holdings they had. In the second-stage, portion of maps stored in the cadastre were selected with equal inclusion probabilities from each municipality selected in the first stage. Note that the sampling units are the cadastral maps (territorial unit into which each municipality is divided, and each map is divided into continuous parcels) of the Land Registry.

We address the designing of the CS for agricultural census through spatially balanced sampling, and we consider the CS in the context of indirect sampling. The use of the former is motivated by the strong spatial component of the problem. Indeed, units distributed over a region of interest tend to be similar since they are influenced by the same set of factors, which is especially true in agricultural surveys, where units close together are influenced by the same soil fertility, weather, pollution, and other spatial factors. In order to exploit this feature, spatially balanced sampling designs select sample well spread over the region of interest. For a review, see [Benedetti et al., 2015]. The framework of indirect sampling [Lavallée, 2007] is taken into consideration, since in our proposal we sample portion of cadaster in order to select holdings. Indeed, we do not sample from a frame of the target population (of holdings), but we sample from a frame (of portion of cadaster) linked to the target population. The paper is organized as follows. In Section 2 we introduce the theoretical framework, while in Section 3 we discuss some proposals for the CS and corresponding efficiency is investigated by means of a Monte Carlo simulation. Finally, Section 4 provides conclusions and future research.

2 Theoretical framework

Suppose a finite population $U_B = \{1,\dots,N\}$ is of interest. We denote with $\mathscr{S}$ the set of all possible subset of U_B. A sample without replacement is an element $s \in \mathscr{S}$, and a sampling design is a probability distribution on $\mathscr{S}$ such that $p(s) \geq 0$ and $\sum_{s\in\mathscr{S}} p(s) = 1$. The probability of selecting unit i is called first-order inclusion probability and given by $\pi_i = \sum_{s\ni i} p(s)$, while the probability of selecting unit i and j in the same sample is called second-order inclusion probability and given by $\pi_{ij} = \sum_{s\supset\{i,j\}} p(s)$. In the *design-based* approach the variable of interest y is considered deterministic, and we can estimate the total of y, $t_y = \sum_{i=1}^{N} y_i$, through the Horvitz-Thompson estimator [Horvitz and Thompson, 1952] $t_{y,HT} = \sum_{i\in s} \frac{y_i}{\pi_i}$, which is unbiased with respect to the design when $\pi_i > 0 \forall i \in U$.

Standard way to proceed in survey sampling requires a frame for the target population U_B, from which a sample of units is selected by means of a sampling design $p(s)$. Unfortunately, in some circumstances no frame is available for U_B. We may have a frame for another population U_A, which is linked to the target population U_B. Indirect sampling consists on select a sample s_A from U_A and exploit the linkage with U_B in order to produce the desired estimate. The major challenge is to assign a selection probability or an estimation weight to the units of the population U_B.

The generalised weight share method (GSWM) [Lavallée, 1995] is a generalisation of the weight share method [Ernst, 1986] and produces an estimation weight for each surveyed unit of the population U_B, through an average of the sampling weights of the population U_A. Suppose the population target U_B is composed by M^B units and divided into N clusters, where the i-th cluster has M_i^B units, and the linked population U^A is composed by M^A units. We select a sample s^A of m_A units from U_A by means of a sampling design with first order inclusion probabilities $\pi_j^A > 0$ $\forall j \in U^A$. Moreover, we suppose there exists a relationship between units j of population U^A and units k of cluster i of the population U^B, and we indicate this relationship through an indicator variable $l_{j,ik}$, which is equal to 1 when a link exists between $j \in U^A$ and unit $ik \in U^B$, and equal to 0 otherwise. The total of links between the unit $j \in U^A$ and the units k of cluster i of population U^B is given by $L_j^A = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik}$, whereas the total of links for any unit k of a cluster i of population U^B is given by $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$. Note that in order to use the GSWM, we must satisfy $L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik} > 0$. Starting from the sample s^A, for each unit $j \in s^A$ we identify the units ik of U^B that have a non-zero link with j, and for each of those units we assume we can set up the list of M_i^B units of cluster i containing this unit. Therefore, each cluster i represents within itself a population U_i^B where $U^B = \bigcup_{i=1}^{N} U_i^B$. Then, let Ω^B be the set of n clusters identified by the units $j \in s^A$, that is $\Omega^B = \{i \in U^B | \exists j \in s^A \text{ and } L_{i,j} > 0\}$ with $L_{j,i} = \sum_{k=1}^{M_i^B} l_{j,ik}$.

We can now survey all the units k of cluster $i \in \Omega^B$. In particular, we record the variable of interest y_{ik} and the number of links L_{ik}^B. We suppose that, for the target population U^B, the parameter of interest is the total $Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$, which

can be estimated by $\hat{Y}^B \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik}$, where n is the number of clusters surveyed and w_{ik} is the weight assigned to unit k of cluster i. These weights are obtained by the GWSM, in such a way the final weights are calculated according to a weighted method within each surveyed cluster. Indeed, for the weight w_{ik}, the method starts with the computation of an initial weight, which is the inverse of the inclusion probability of the unit selected from U^A and with link with the cluster i (if the unit does not have a link, the initial weight is equal to zero). Then, the final weight is obtained as the ratio of the sum of the initial weights for the cluster over the total number of links in the cluster, and it is assigned to all units in the cluster.

In this work we investigate the use of spatially balanced sampling designs for the selection of units from the population U_A. Spatially balanced sampling designs select samples well-spread over the population of interest. Technically, a well spread sample has a number of selected units on every part of the study region close to what is expected on average [Grafström and Lundström, 2013]. The idea is that a spread sample could capture the spatial heterogeneity of the population, which in turn could improve the efficiency of estimates compared to the efficiency of estimates achieved by data obtained from non-spatial sampling designs. For more details about efficiency and extensive simulations, see [Benedetti et al., 2017].

3 Proposals and simulations

In this section we discuss some proposals for the sampling design to employ in the CS, and we present a set of simulations where the efficiency of these proposals are investigated. An artificial population that mimics the situation that ISTAT faced in 2010 is generated through the following steps.

1. A regular 30×30 grid that represents the cadastral maps is generated, and the municipalities limits are generated by aggregation of the cadastral maps by means of a clustering algorithm based on a Minimum Spanning Tree [Assunção et al., 2006], which is used in 50 contiguous groups (clusters) and with a number of prefixed cadastral maps between 12 and 38.
2. Two populations that represent clustered and sparse configuration of holdings, respectively, are generated. Toward this end, two series of 1,000 points are generated according to the Neyman-Scott process with Cauchy cluster kernel [Waagepetersen, 2007] and are overlaid on the grid. These points represent the center position of the farms. The intensity of the Poisson process is equal to 10, and the mean number of units per cluster is 100. Two different scale parameters 0.05 and 0.1 are used for the two series, respectively, where the former is used for the clustered population while the latter for the sparse population.
3. A size is assigned at each farm according to a negative exponential distribution of parameter $\beta = [0,2]$. The sizes are approximated to the upper integer.
4. A variable $[0,1]$ (censused/not censused) is generated through a Markov Chain Monte Carlo (MCMC) algorithm such that the probability of being 1 is simulta-

neously inversely proportional to the size and proportional to the frequencies of 1 in the neighborhood (spatial dependence). The frequencies are fixed to 850 and 150, for the 0 and 1, respectively.

5. Three sets of links are generated with the population of the cadastral maps (i.e. in which cadastral maps the company has land in use). After removing the cadastral map where the business center is located, links are generated (size 1) with probability inversely proportional to the distance between the farm and the cadastral map, raised to a control parameter set equal to 2, 3, 10 to ensure that these links are more or less probable increasing the distance.

Simple Random Sampling (SRS), Local Pivotal Method (LPM) [Grafström, 2012], two-stage Simple Random Sampling (2S_SRS), and two-stage Local Pivotal Method (2S_LPM) are investigated as proposals for the CS. In order to evaluate the efficiency of these proposals in terms of Root Mean Squared Error (RMSE) of the HT estimator of the total, a Monte Carlo simulation is performed, with $M = 10,000$ replicated samples of cadastral maps of size $n = \{9, 24, 45\}$ selected from the aforementioned populations, in different scenarios characterized by different sets of links (L^2, L^3, and L^{10}) , which are generated as explained in the previous step 5. In the two-stage sampling designs, the first-stage is employed to select municipalities, and the second-stage selects cadastral maps. Therefore, for $n = 9$ three municipalities and three cadastral maps per municipality are selected, for $n = 24$ six municipalities and four cadastral maps per municipality are selected, and for $n = 45$ nine municipalities and five cadastral maps per municipality are selected. From the cadastral maps, every agricultural holding belonging to them is then selected. The HT estimator is employed, where the sample weights are obtained by the GSWM, and the proposals are compared in terms of the Monte Carlo RMSE. In particular, we compare SRS with LPM (both single-stage sampling), 2S_SRS with 2S_LPM (both two-stage sampling), 2S_SRS with SRS, and 2S_LPM with LPM (the last two comparisons are performed for investigation of single-stage against two-stage sampling). We report the results in Table 1. In case of single-stage sampling, LPM obtains a lower RMSE (Table 1a), while for the two-stage sampling design the best result is obtained when LPM is employed in both stages (Table 1b). Tables 1c and 1d report the performance of single-stage against two-stage sampling. The results suggest that when possible, a one-stage design should be employed. However, many situations may require a two-stage sampling, due to cost or administrative reasons, among many others. In this case, the spatial version of the two-stage sampling could be used.

4 Conclusion and future research

In this paper we focused on the CS of agricultural census. Indeed, we proposed the use of spatially balanced designs while the framework of indirect sampling is employed. Results show that the combination of these two methods allows to achieve good results in terms of RMSE, which is investigated by means of Monte Carlo sim-

ulation. We plan to extend this work in order to account for costs of the sampling designs and for variance estimation, and to extend simulations on real data as well.

	Sparse population			Cluster population		
n	L^2	L^3	L^{10}	L^2	L^3	L^{10}
9	0.988	0.990	0.985	0.929	0.919	0.928
24	0.959	0.939	0.942	0.878	0.858	0.855
45	0.935	0.903	0.894	0.838	0.817	0.806

(a) LPM vs SRS

	Sparse population			Cluster population		
n	L^2	L^3	L^{10}	L^2	L^3	L^{10}
9	0.954	0.939	0.927	0.909	0.899	0.894
24	0.950	0.935	0.920	0.880	0.872	0.878
45	0.938	0.925	0.915	0.883	0.867	0.870

(b) 2S LPM vs 2S SRS

	Sparse population			Cluster population		
n	L^2	L^3	L^{10}	L^2	L^3	L^{10}
9	1.229	1.258	1.256	1.156	1.157	1.181
24	1.293	1.316	1.330	1.238	1.260	1.273
45	1.340	1.371	1.385	1.266	1.311	1.317

(c) 2S SRS vs SRS

	Sparse population			Cluster population		
n	L^2	L^3	L^{10}	L^2	L^3	L^{10}
9	1.186	1.192	1.183	1.131	1.133	1.138
24	1.280	1.310	1.299	1.240	1.280	1.308
45	1.344	1.406	1.416	1.335	1.392	1.423

(d) 2S LPM vs LPM

Table 1: Relative Root Mean Squared Error (rRMSE) of the HT estimator for the indicated cases.

References

[Assunção et al., 2006] Assunção, R. M., Neves, M. C., Câmara, G., and da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811.

[Benedetti et al., 2017] Benedetti, R., Piersimoni, F., and Postiglione, P. (2017). Spatially balanced sampling: a review and a reappraisal. *International Statistical Review*, 85(3):439–454.

[Benedetti et al., 2015] Benedetti, R., Piersimoni, F., Postiglione, P., et al. (2015). *Sampling spatial units for agricultural surveys*. Springer.

[Ernst, 1986] Ernst, L. R. (1986). *Weighting issues for longitudinal household and family estimates*. US Bureau of the Census.

[Grafström, 2012] Grafström, A. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–521.

[Grafström and Lundström, 2013] Grafström, A. and Lundström, N. L. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, 3(1):36–41.

[Horvitz and Thompson, 1952] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

[ISTAT, 2013] ISTAT (2013). Atti del 6 censimento generale dellagricoltura, la valutazione della qualità, fascicolo 5. Technical report, ISTAT.

[Lavallée, 1995] Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21(1):25–32.

[Lavallée, 2007] Lavallée, P. (2007). *Indirect Sampling*. Springer New York.

[Waagepetersen, 2007] Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous neyman–scott processes. *Biometrics*, 63(1):252–258.

Next Census in Israel: Strategy, Estimation and Evaluation

Progettazione del Censimento in Israele. Strategia, Stima e Valutazione

Danny Pfeffermann

Abstract In this short article I describe our proposed plans for running Israel's next census, due to start towards the end of 2021. The census will combine a sample survey with administrative data and I consider four main issues: Sample selection and data collection; Sample estimation, accounting for non-ignorable nonresponse; Combination of the sample estimates with available administrative data to form the final census estimates in small statistical areas; Census evaluation.

Abstract *In questo breve articolo descrivo i nostri piani per il prossimo ciclo del Censimento in Israele, che si avvierà a fine 2021, e che integrerà una indagine campionaria e dati amministrativi. Mi soffermo su quattro aspetti principali: disegno campionario e raccolta dati; stima, tenendo conto delle mancate risposte; integrazione delle stime da indagine con dati disponibili da fonti amministrative per la produzione delle stime censuarie in piccoli domini; valutazione del Censimento.*

Key words: Census evaluation; Composite estimator, NMAR nonresponse

1 Sample selection, data collection and estimation

Israel has a fairly accurate Central Population Register (CPR); almost perfect at the country level. However, the CPR is much less accurate for small statistical areas, with an average enumeration error of 13% and a 95 percentile of 40%. The country is divided into about 3,500 statistical areas, and census information is required for every area. In this article, I consider count estimation, generally considered as the main target of any census. The main reason for the inaccuracy in the CPR at the area level is that people moving in or out of areas, often report late their change of address, sometimes because they possess an address of interest (tax benefits, school area, parking, etc.).

For our 2021 census, we plan to combine information from a probabilistic sample, taken from an improved CPR (integrating the CPR with other available administrative data files, hereafter the ICPR), with the information available from the ICPR itself. The sample will collect geo-demographic information on all members of the household on census day, as well as socio-economic information.

Sample selection: The sample will be selected by use of a stratified, two-stage cluster sample, with many strata defined by the statistical areas, size of administrative families (AF, persons registered in the ICPR as living in the same address), marital status and age of the mother in the AF. The clusters are the AF, and the ultimate sampling units are the adult persons within them. Each sampled person will be asked to provide information about all

[1] Danny Pfeffermann, Central Bureau of Statistic, Israel; email: msdanny@cbs.gov.il

the persons residing in the household, including children. The questionnaire is designed electronically in such a way that every information known from administrative flies, like for example about education, will be added automatically, thus reducing response burden.

Data collection: The data will be collected using three separate modes. First, all the sampled units will be encouraged to respond via the internet. Non-respondents will be asked to respond via the telephone. Persons not responding by either one of the two modes, will be approached at home for a personal interview. The Bedouin population will only be approached for personal interviews, because of difficulties to obtain sufficiently accurate addresses otherwise.

Sample estimation: Denote by N the (known) number of residents in the country on census day and by N_i the unknown number of residents in area i. Let $P_i = N_i / N$ denote the true proportion of residents in area i and $\hat{P}_{i,HT} = \sum_{(i,j)\in S} I_{j\in i} w_{ij} z_{ij} / \sum_{(i,j)\in S} w_{ij}$ the corresponding sample estimator, where $I_{j\in i} = 1(0)$ if unit j resided (did not reside) in area i on census day, z_{ij} represents the number of residents reported by sampled unit (i,j) as belonging to her or his household, $\pi_{ij} = \Pr[(i,j)\in S]$ is the sample inclusion probability of unit (i,j) and $w_{ij} = (1/\pi_{ij})$ is the corresponding base sampling weight. (Some modifications are needed if two or more sampled units report on the same household). Assuming complete response (see below), the direct, Horwitz-Thompson ratio estimator for the count of area i is then,

$$\hat{N}_{i,HT} = N \times \hat{P}_{i,HT}\,. \tag{1.1}$$

The estimator (1.1) can be improved by calibrating the base sampling weights, $\{w_{ij}\}$ to make estimators of auxiliary variables $\mathbf{x}$. match their known population totals from the ICPR, resulting in the familiar Generalized Regression estimator (GREG, Deville and Särndal, 1992). The "standard" calibrated weights take the form,

$$\boldsymbol{c} = \boldsymbol{w} + BX(X'BX)^{-1}(\boldsymbol{T} - X'\boldsymbol{w})\,, \tag{1.2}$$

where $\boldsymbol{c}$ and $\boldsymbol{w}$ are the vectors of the calibrated- and base sampling weights of order n, T is the vector of known totals, X is the "design matrix" of the vectors $\mathbf{x}_k$ observed for the sampled units and $B = Diag(\boldsymbol{w})$. Clearly, $X'\boldsymbol{c} = \boldsymbol{T}$.

Accounting for non-ignorable nonresponse: So far we assumed complete response but in practice, nonresponse exists in every survey and we expect to encounter it in the census sample as well. In particular, the nonresponse may depend on the target variable of interest, known as not missing at random (NMAR), in which case the use of calibration alone, although possibly reducing the effect of nonresponse, does not solve the problem. As well known, not accounting for NMAR nonresponse may result in highly biased estimators.

To deal with this problem, we plan to follow a methodology developed in Sverchkov and Pfeffermann (2018). The basic idea is to fit a model for the observed data, assume a model for the response probabilities as a function of appropriate covariates known for the responding units and of the target variable of interest, and then using the combined model for estimating the response probabilities. Having estimated the response probabilities, we view the response as an additional, ultimate sampling process and modify the original base sampling weights $w_{ij} = (1/\pi_{ij})$ as $a_{ij} = w_{ij} q_{ij}$, where by denoting $R_{ij} = 1(0)$ if sampled unit (i,j) responds (does not respond), $q_{ij} = 1/\hat{\Pr}(R_{ij} = 1)$. In the final stage the weights a_{ij} are calibrated and the GREG estimates are computed as in (1.1) and (1.2) when assuming complete response.

In what follows I define plausible models for estimating the counts. Define the target outcome as $y_{ij} = 1(0)$ if the ICPR address of unit (i,j) is correct (not correct). Assume the following two-level model for the observed data,

$$\Pr_\gamma(\mathbf{x}_{ij}, u_i; \beta) = \Pr(y_{ij} \mid x_{ij}, u_i, R_{ij} = 1; \beta) = \frac{\exp(\beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i)}{1 + \exp(\beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i)};\ u_i \sqcap N(0, \sigma_u^2)\,. \tag{1.3}$$

Note that the outcome model refers to the observed data and hence can be estimated and tested by standard small area estimation (SAE) methods.

The response probabilities are also assumed to follow a logistic model,

$$P_R(y_{ij},x_{ij};\gamma)=Pr(R_{ij}=1\,|\,y_{ij},\mathbf{x}_{ij};\gamma)=\frac{\exp(\gamma_0+\mathbf{x}'_{ij}\boldsymbol{\gamma}_{\mathbf{x}}+\gamma_y y_{ij})}{1+\exp(\gamma_0+\mathbf{x}'_{ij}\boldsymbol{\gamma}_{\mathbf{x}}+\gamma_y y_{ij})}. \tag{1.4}$$

Notice that the response probabilities are allowed to depend on the outcome, thus accounting for NMAR nonresponse. To simplify the notation, we assume the same covariates $\mathbf{x}_{ij}$ in (1.3) and (1.4), but in practice they may differ.

Let $R^c=\{(i,k);(i,k)\in S,R_{ik}=0\}$. If the missing outcomes were actually observed, the vector $\boldsymbol{\gamma}'=(\boldsymbol{\gamma}'_{\mathbf{x}},\gamma_y)$ could be estimated by solving the likelihood equations,

$$\sum_{(i,j)\in R}\frac{\partial\log p_r(y_{ij},x_{ij};\gamma)}{\partial\gamma}+\sum_{(i,\mathrm{k})\in R^c}\frac{\partial\log[1-p_r(y_{ik},x_{ik};\gamma)]}{\partial\gamma}=0. \tag{1.5}$$

However, since the missing data are practically unknown, Sverchkov and Pfeffermann (2018) apply the missing information principle (MIP, Orchard and Woodbury, 1972) and replace the likelihood (1.5) by its expectation with respect to the distribution of the missing outcomes, given all the observed data. The latter distribution is obtained from the distribution fitted to the observed values, using the relationship between the two distributions for given covariates and response probabilities, as developed in Sverchkov and Pfeffermann (2004). See Sverchkov and Pfeffermann (2018) and Pfeffermann, Ben-Hur and Blum (2019) for numerical illustrations, using simulated and real data collected in Israel's 2008 census. Pfeffermann and Sverchkov (2019) extend the methodology for the case of several target outcomes, assuming multivariate outcome models and allowing for the different response patterns appearing in the actual data.

Combining the sample estimates with the ICPR: The Greg estimator with the calibrated weights defined above does not use the area counts in the ICPR. Although not fully accurate, the ICPR counts provide valuable information about the true counts, which should not be ignored.

For our next census, we consider the use of composite estimators of the form,

$$\hat{N}_i^{Com}=A_i\hat{N}_i^{GES}+(1-A_i)K_i, \tag{1.6}$$

where K_i is the ICPR count. Notice that $\hat{N}_i^{GES}$ has a variance, denoted hereafter by $\sigma^2_{GES,i}$ but (approximately) no bias, while K_i has a bias (measurement error). Thus,

$$MSE(\hat{N}_i^{Com})=A_i^2\sigma^2_{GES,i}+(1-A_i)^2(K_i-N_i)^2, \tag{1.7}$$

and it is minimized when

$$A_i=(K_i-N_i)^2/[(K_i-N_i)^2+\sigma^2_{GES,i}]. \tag{1.8}$$

In practice, N_i and $\sigma^2_{GES,i}$ are unknown, and we plan to replace them in (1.8) and (1.6) by their sample estimates, yielding the empirical composite estimator,

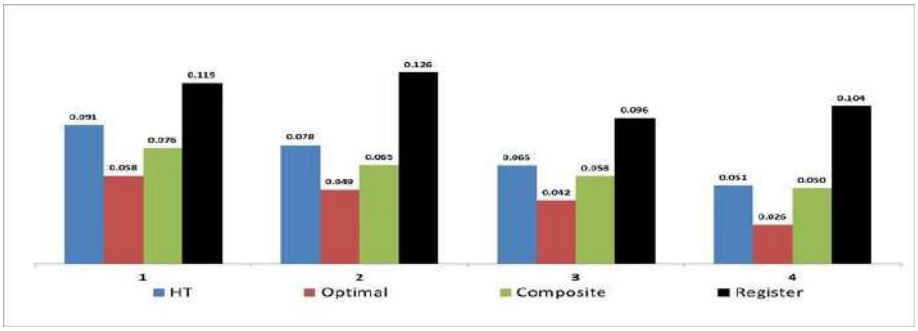
$$\hat{N}^{Com}_{EMP,i}=\hat{A}_i\hat{N}_i^{GES}+(1-\hat{A}_i)K_i;\ \hat{A}_i=\frac{(K_i-\hat{N}_i^{GES})^2}{(K_i-\hat{N}_i^{GES})^2+\hat{\sigma}^2_{GES,i}}. \tag{1.9}$$

In order to illustrate the performance of the estimate (1.9), we use data from the 2008 census in Israel. The 2008 census consisted of an area sample of about 1,100,000 individuals (*U* sample) for estimating the register *U*ndercount, and a telephone sample of people registered in each area, for estimating the register *O*ver-count (*O* sample). We consider the addresses in the *U* sample as the correct addresses. We sampled 10% of the individuals listed in the 2008 ICPR from each statistical area. For each area we calculated the HT estimator, the "optimal" estimator obtained by substituting (1.8) in(1.6) and the empirical composite estimator (1.9). For all three estimators, we used the HT estimator without calibration. We repeated the sampling 1,000 times.

Figure 1 shows the averages of the relative root mean square errors of the three estimators over the 1,000 samples within the four quarters of the size of the statistical areas

(SA). As can be seen, the theoretical optimal estimator is clearly the best for all the areas and the HT estimator is the worst, although its performance improves as the sample sizes within the SA increase. Except in the case of the large sample sizes, the empirical composite estimator performs better than the HT estimator. Finally, notice the inaccuracy of the ICPR counts, which indicates that in 2008, one could not rely solely on them.

Figure 1: Average relative root mean square error of estimators by quarters of SA size.



2 Census evaluation

A first, natural evaluation will be to compare the ICPR values, K_i, with the sample estimates, $\hat{N}_i^{(GES)}$. Although as stated before, the K_i's are "biased", we expect that a confidence interval (C.I.) based on $\hat{N}_i^{(GES)}$ will cover most of them. Thus, we plan to compute for each area the standard C.I. $\hat{N}_i^{(GES)} \pm 2Std(\hat{N}_i^{(GES)})$ and consider areas where the C.I does not cover K_i for further examination (e.g., add personal interviews in those areas). We consider this procedure as a Post Enumeration Survey, which we won't be able to carry out because of budget constraints.

In what follows I discuss several "bureau evaluations", which we plan to apply. As becomes evident below, these kinds of evaluations are "cost free". The basic idea is to attribute to each area many inherent indices of credibility. For each index, it will be clear what would be considered as a satisfying census estimate, as opposed to a potential error in the estimate. The indices target the sampling, data collection, and estimation phases, with each phase having its own indices. In particular, the indices of the sampling stage may enable replacing samples in areas where the random sample selection fails to provide appropriate samples.

The indices of the data collection and estimation processes, examine internal incoherence between survey estimates. The basic idea is that the larger the area, the more reliable is its estimate. This is because of (i)- Larger sample sizes in larger areas and (ii)- Fewer anchoring (identifying the correct address) problems in larger areas (re-locations are more likely to be within large areas). In particular, when there is a large time gap between the census date and the sample response date, recollection problems are likely to be less significant in larger areas. Therefore, we shall compare the sum of census estimates of small areas with known estimates of a larger area (locality) in which they are nested. A contradiction may indicate problems in the census estimate for at least one of the SA's.

Evaluation indices for the sampling process: The probability sampling design permits comparing the census sample estimates at the national level with their corresponding known values; for example, between estimates of counts by age and gender, for which the true values are known from the ICPR. The comparison should employ the initial sample

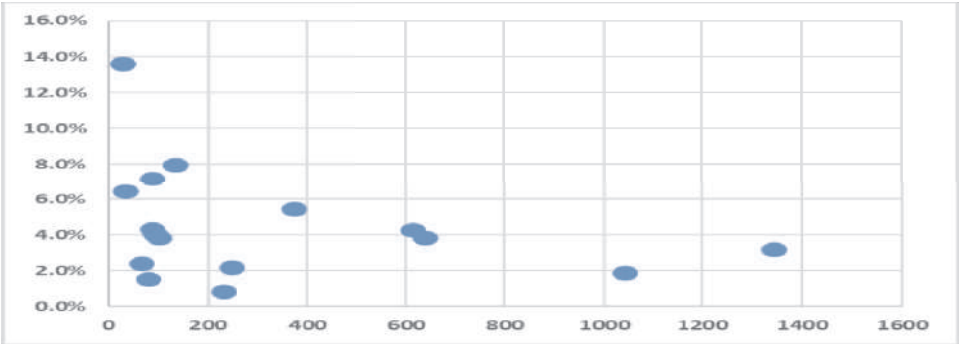
estimates before calibration. In Israel, we have a very reliable administrative education file to which the census sample estimates may also be compared.

So far, I only considered comparisons to known population values, but we may also compare the final census estimates for large localities to estimates obtained from other surveys, for which the latter estimates are sufficiently reliable. An example of such a survey is the labour force survey (LFS), which collects detailed information about demographic and many other variables, which are also investigated in the census. The LFS is a monthly survey of dwellings with high response rates. Moreover, it is a field survey where interviewers visit the sampled dwellings, so that the addresses of the inhabitants found in these dwellings are measured without error. The LFS estimates are not accurate for small SAs and in some months there are no sampled units from these areas, but they are sufficiently accurate for large localities. Nonetheless, any comparison between the census and the LFS estimates should account for the MSE of both.

For sufficiently large localities L, one could estimate the size of the locality, N_L, from the LFS in the following way. Estimate, $\hat{P}_{L|R}$ - the proportion of individuals living in L out of those registered as living in L, and $\hat{P}_{R|L}$ - the proportion of individuals registered in L out of those living in L. Denote by K_L the ICPR of the locality. Thence, $\hat{N}_L = K_L \hat{P}_{L|R} / \hat{P}_{R|L}$ is a consistent estimator of N_L.

Figure 2 compares the LFS estimates with the census sample estimates as obtained in a recent Census Rehearsal without personal interviews of non-respondents and with no nonresponse adjustments, for 16 localities. Notice that the larger the LFS sample size in the locality, the larger is the locality and the more accurate are the LFS and the census sample estimates. Consequently, for the large localities, the differences are relatively small.

Figure 2: Percent difference between census rehearsal sample and LFS estimates, 16 localities



Evaluation indices for the data collection process: Relevant indices for this process at the area level include, **a-** Percent non-response in the area, **b-** Distribution of non-response across demographic categories, **c-** Percent of anchoring success, **d-** Estimated percent of recorded immigrants found in the area, **e-** Zero cases.

Evaluation indices for the estimation process: As stated before, the estimates for large geographical areas are expected to be more reliable than for the small areas. We plan therefore to apply a 'top down' evaluation procedure. The basic premise of this procedure is that the sum of estimates of totals over all the SAs nested in a large area (locality) L, should agree with the estimate of the total in L. Denoting the sampled units residing in SA i by j, and the sampled units residing in the locality L by k, ideally,

$$\sum_{i \in L} \sum_{j \in i} a_{ij}^{(sa)} z_{ij} = \sum_{k \in L} a_k^{(L)} z_k , \qquad (2.1)$$

where $a_{ij}^{(sa)}$ is the weight of unit j in SA i after adjusting for nonresponse and $a_k^{(L)}$ is the adjusted weight of unit k in the Locality L. A significant difference between the left- and right hand side of (2.1) can be attributed to many more anchoring problems at the SA level,

which can be viewed as another facet of nonresponse. In such cases, the weights $a_{ij}^{(sa)}$ should be modified to satisfy the constraint (2.1). A modification often used in SAE, known as pro-rata benchmarking or calibration is,

$$a_{ij,cal}^{(sa)} = a_{ij}^{(sa)} \frac{\sum_{k\in L} a_k^{(L)} z_k}{\sum_{i\in L}\sum_{j\in i} a_{ij}^{(sa)} z_{ij}}. \tag{2.2}$$

The use of the weights (2.2) guarantees agreement of the SA estimates in a locality L with the corresponding locality estimate.

Rather than calibrating the weights by use of the target outcome (the counts in our case), we can calibrate them by use of other variables, known for every sampled unit in the locality. Let X be such a variable. A ratio type estimate of N_i, the size of SA i, calibrated with respect to X is,

$$\begin{aligned}\hat{N}_i &= \sum_{j\in i} a_{ij}^{(sa)} z_{ij} + [\sum_{k\in L} a_k^{(L)} x_k - \sum_{i\in L}\sum_{j\in i} a_{ij}^{(sa)} x_{ij}] \frac{\sum_{j\in i} a_{ij}^{(sa)} z_{ij}}{\sum_{i\in L}\sum_{j\in i} a_{ij}^{(sa)} x_{ij}} \\ &= \sum_{j\in i} a_{ij}^{(sa)} z_{ij} \frac{\sum_{k\in L} a_k^{(L)} x_k}{\sum_{i\in L}\sum_{j\in i} a_{ij}^{(sa)} x_{ij}}.\end{aligned} \tag{2.3}$$

Note that for $x_{ij} \equiv z_{ij}$ $(x_k \equiv z_k)$, the weights used in (2.3) reduce to the weights in (2.2).

For evaluation purposes, it is advisable to study the magnitude of the correction terms $[\sum_{k\in L} a_k^{(L)} x_k - \sum_{i\in L}\sum_{j\in i} a_{ij}^{(sa)} x_{ij}] \frac{\sum_{j\in i} a_{ij}^{(sa)} z_{ij}}{\sum_{i\in L}\sum_{j\in i} a_{ij}^{(sa)} x_{ij}}$, and check whether they are in a plausible range. This kind of evaluation will assess the consistency between the separate SA estimates and the locality estimate before calibration, for a range of calibration variables X. Standard calculations show that the correction terms are of order $N_L \times CV_{XL} \times O_p(n_L^{-1/2})$, where N_L is the locality size, CV_{XL} is the coefficient of variation of X in the locality and n_L is the sample size. Thus, we may divide each correction term by the estimate, $\hat{N}_L \times \hat{C}V_X \times n_L^{-1/2}$, and order the results. The larger the ratio, the more questionable is the area estimate.

We conclude that the use of reliable estimates of the locality size may be useful for evaluating the SA size estimates.

References

1. Deville, J.C. and Särndal, C.E.: Calibration Estimators in Survey Sampling. In: Journal of the American Statistical Association, 87, 376-382 (1992).
2. Orchard, T. and Woodbury, M.A.: A missing information principle: theory and application. In: Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 697-715 (1972).
3. Pfeffermann, D., Ben-Hur, D. and Blum, L.: Planning the next Census for Israel. In: Statistics in Transition, 20, 433-443 (2019).
4. Pfeffermann, D. and Sverchkov, M.: Multivariate small area estimation under nonignorable nonresponse. In: Statistical theory and related fields, 3, 213-223 (2019).
5. Sverchkov, M. and Pfeffermann, D.: Prediction of finite Population Totals Based on the Sample Distribution. In: Survey Methodology, 30, 79-92 (2004).
6. Sverchkov, M. and Pfeffermann, D.: Small area estimation under informative sampling and not missing at random nonresponse. In: Journal of the Royal Statistical Society, Series A. 181, 981-1008 (2018).

Administrative data for population counts estimations in Italian Population Census

I dati amministrativi per le stime dei conteggi di popolazione nel Censimento della popolazione italiano

A. Bernardini, A. Chieppa, N. Cibella, F. Solari[1]

Abstract. In 2018 ISTAT adopted the Permanent Population Census, that is a new census strategy that combines the Base Population Register, administrative archives and data collected with annual sample surveys to obtain counts for the usual resident population for each territorial unit, taking into account coverage errors in the Base Population Register. In this paper, the coverage errors estimation procedure used for Census waves 2018 and 2019 is described. Administrative data resulted to be useful to detect possible over/undercoverage errors of the registers and they are expected to be exploited at a greater degree in next Census rounds, when the reduction of costs dedicated to surveys and the minimization of response burden are relevant.

Abstract. *Dal 2018 l'ISTAT ha avviato il Censimento Permanente, un nuovo sistema censuario, finalizzato ad integrare il Registro Base degli Individui, i dati amministrativi e i dati provenienti dalle indagini campionarie per ottenere i conteggi della popolazione abitualmente residente, ripulendo i conteggi del Censimento dagli errori di copertura che interessano il Registro Base degli Individui. Nel lavoro viene presentata la procedura di stima degli errori di copertura per le ondate censuarie 2018 e 2019. L'uso dei dati amministrativi è risultato essere utile per individuare eventuali errori di sovra/sotto copertura dei registri e si prevede un loro maggiore utilizzo nelle prossime tornate censuarie, considerata la riduzione dei costi dedicati alle indagini e l'obiettivo di minimizzazione del 'disturbo statistico' sui rispondenti.*

[1] Antonella Bernardini, ISTAT; email: anbernar@istat.it
Angela Chieppa, ISTAT; email: chieppa@istat.it
Nicoletta Cibella, ISTAT; email: cibella@istat.it
Fabrizio Solari, ISTAT; email: solari@istat.it

Key words: administrative data, coverage errors, permanent population census

1 Producing Census results combining data from statistical registers and surveys

Administrative data are nowadays an unavoidable resource for official statistics, since they allow considerable reduction in costs, both in economical terms and in respondent burden. Statistical registers are the solution many National Statistical Institutes have chosen to manage and to supervise the integration between data coming from different sources, including administrative archives not originally designed for statistical purposes [2]. In 2016, Italian National Institute of Statistics, ISTAT, adopted a so-called 'modernization programme' involving a statistical production based on an Integrated System of Statistical Registers, combining administrative and surveys data. The Permanent Census of Population and Housing means both the new census strategy and the informative system designed to accomplish the traditional Census goals in the framework of the 'modernized' statistical production, based on statistical registers with surveys data used to feed registers and to integrate missing, incomplete or insufficient quality information.

The main register at the core of the Permanent Census is the Population Base Register (PBR), whose main administrative sources are the Local Population Registers of each Italian Municipality. Others registers involved in Population Census production are: Statistical Base Register of Addresses/Micro-zones and thematic registers on education and employment; moreover, a thematic database (register) called AIDA, Integrated Archive of Usual Resident Population, has been set up, since 2015, to exploit the administrative sources and to find relevant patterns useful for estimation of population counts, improving the quality of municipal population registers and PBR [1].

Two sample surveys are conducted annually to assess the quality of registers and complete them for Census purposes: the Area survey is especially designed for measuring coverage errors of the PBR, through the enumerations of all households living in the addresses/zones sampled from the Addresses Register; the List survey is primarily designed for thematic integration of registers, with a sample of households filling in a questionnaire with variables not included in registers (or included, but with insufficient quality): since the sample is extracted from the PBR, List survey's results are crucial also in detecting overcoverage error in that register.

All Census results are, in this new framework, an estimation that combines data from both registers and surveys, through a proper statistical model. The primary output of the Census consists in the Population counts, i.e. total amounts of usual

resident population for each Municipality: the estimation process, in this case, is based on PBR data corrected with survey results and, possibly, integration with AIDA.

The coverage errors estimation procedure adopted for the production of the population counts for the first two waves of Permanent Census is described in the following paragraph. This first experience showed that, considering the quality levels of surveys and the PBR, it is necessary to use additional administrative sources, via AIDA register, that could be translated into "signs of life": the final paragraph addresses current investigation and future scenario centred on better exploiting administrative signs of life to correct PBR.

2 The estimation process

For Census waves 2018 and 2019, survey data are used to correct PBR data under a Dual System Estimation framework aiming at estimating coverage errors of the register. In the traditional census a Post Enumeration Survey (PES) is usually used to measure the census undercoverage errors, the census considered as the first capture and the PES the second capture. In the PPHC the PBR represents the first capture while the annual sample surveys. Furthermore, differently from a typical PES, aimed at measuring only undercoverage, in the PPHC design the second capture is able to measure both undercoverage and overcoverage of the first capture, i.e. the PBR.

For each domain of interest, given by the cross-classification of municipality i and Italian or foreign nationality status j, the parameter of interest $corr_{ij}$ is the correction factor to be applied to each individual in the PBR in order to correct the PBR for overcoverage and undercoverage:

$$corr_{ij} = \frac{1-p_{ij,over}}{1-p_{ij,under}}.$$

Here, $p_{ij,under}$ is the undercoverage probability defined as 1 minus the ratio between the newly enumerated individuals and the total number of individuals enumerated (i.e. individuals not expected according to the PBR), while $p_{ij,over}$ is the overcoverage probability given by the ratio between individuals not enumerated in the survey and individuals expected to be enumerated according to the PBR.

The Areal survey is used for measuring the undercoverage error of the PBR. Unfortunately, since not all the municipalities carried out all the Areal survey phases with required level of quality, Areal survey could not have been used for the estimation of the overcoverage probability. Therefore, the List survey is used, together with information on ‘administrative signs of life’ from AIDA in order to measure the overcoverage error of the PBR. To this aim, the subset of ‘potential overcoverage’

individuals (individuals still present in the municipality according to the PBR and not found on the field), is 'cleaned' by means of 'signs of life' in the municipality tracked down in AIDA. Non respondents to the List survey are thus 'recovered' if they show strong (i.e. at least 8-12 months) "signs of life" in the same municipality where they are recorded in the PBR. On the contrary, individuals with no 'signs of life' in the municipality (or with 'signs of life' of intensity below 8 months) will be confirmed as actual register overcoverage. The 'signs of life' considered for this purpose are the following: public, private employee or self-employed indicator, retirement pension beneficiary indicator, school attendance (including pre-primary), unemployment assistance or basic income beneficiary indicator, indicator of being fiscally dependent household member from an individual with strong signs of life.

For sampled municipalities direct estimates calibrated of over/undercoverage probabilities for each domain have been calculated. Then, indirect estimates are computed using small areas estimation models defined separately for overcoverage and undercoverage, in order to reduce direct estimates' variability for sampled municipalities and to calculate estimates for non-sampled municipalities. In details, [3] model is applied to the direct estimates in order to obtain indirect estimates for all the domain of interest. Generalized variance function (see [7]) is used to obtain smoothed variances of the direct estimates.

For each estimation domain and separately for overcoverage and undercoverage, the average of 2018 and 2019 correctors is calculated, weighted with the respective demographic sizes. The estimate of the 2018-2019 average correction factor is, therefore, obtained as the ratio between the weighted averages, between 2019 and 2019, of the estimates overcoverage and undercoverage, weighted with the corresponding population size.

3 The future scenario

For census waves 2018 and 2019 the crucial steps in estimating the correctors of the population register consisted in the use of the dual systems approach and of the small area estimation methods [8,9]. The population counts, corrected for over/under coverage at individual profiles level were published by ISTAT at the end of 2020. In this first experience the administrative sources were used to correct the undercoverage that can affect the census surveys, but the use of administrative data could help also for evaluating the overall quality of the base population register in terms of over/under coverage errors. so determining a reduction of costs dedicated to surveys and the minimization of response burden. To this aim, ISTAT is currently conducting a feasibility study in order to estimate the population counts for different profiles (e.g. citizenship, age groups) through the use of administrative data as primary source; this approach could be especially relevant for the scenario post 2021 also because the COVID pandemic emergency could cause a decrease of the survey response rates and

the new habits related to place in which people decide to live or work could increase misclassification of place of usual residence. Richness in information of administrative sources has to be exploited to evaluate the presence in Italy and detect patterns useful for the estimation of people usually living in each municipality, improving the quality of data existing in the population register and also resulting more informative of the sample surveys for some specific subpopulation [4,5,6]. The estimate of the percentage of persons who usually reside in a municipality compared to those who are included in the register could be derived by properly processing administrative data to extract 'signs of life' of individuals, that could be used to predict if people are actually usually living in a specific municipality. This predicted cases could be used to measure and then correct the coverage errors in the PBR.

References

1. Bernardini. A., Cibella. N., Falorsi. S., Fasulo. A., Gallo, G. (2019). Empirical evidence for population counting: the combined use of administrative sources and survey data, ESS Workshop on the use of administrative data and social statistics, Valencia, 4-5 June, 2019, https://ec.europa.eu/eurostat/cros/system/files/gerardo-gallo_empirical-evidence-population-counting_istat.pdf.
2. Chieppa, A., et al. (2018). Knowledge discovery for inferring the usually resident population from administrative registers. In Mathematical Population Studies, http://www.tandfonline.com/loi/gmps20.
3. Fay, R.E. Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data, Journal of the American Statistical Association, 74, 269-277.
4. Pfeffermann D. (2015). Methodological issues and challenges in the production of Official Statistics, Journal of Survey Statistics and Methodology, 3, 425–483.
5. UNECE (2018), Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses.
6. UNECE (2020), New frontiers for censuses beyond 2020, https://unece.org/statistics/publications/new-frontiers-censuses-beyond-2020.
7. Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling, Journal of the American Statistical Association, 82(398), 499-50.
8. Wolter K.M. (1986), Some Coverage Error Models for Census Data, Journal of the American Statistical Association, 81, 338-346.
9. Zhang, L., Dunne, J. (2017). Trimmed dual system estimation. In D. Bohning, P. G. M. van der Heijden, and J. Bunge (eds.), Capture Recapture Methods for the Social and Medical Sciences, Chapman & Hall/CRC Press, Boca Raton, FL, USA, 239-259.

LFS non response indicators for population register overcoverage estimation

Indicatori di mancato contatto dell'indagine sulle Forze lavoro per la valutazione della sovracopertura del registro della popolazione

Loredana Di Consiglio, Stefano Falorsi[1]

Abstract The statistical register of individuals that is built on administrative sources is the basis for the population counts of the Italian population census. As the register may be affected by over and under coverage, the population census produces a statistical correction to account for its coverage, exploiting the two different surveys that compose census master sample. In this work, we study the feasibility of integrating the LFS in this process, for the estimation of overcoverage, analysing the non response indicators of the survey (specifically non-contacts).

Abstract *Il registro base degli individui costruito a partire da fonti amministrative rappresenta la base delle statistiche sul conteggio di popolazione. Tale registro però può essere affetto da errori di sovra e sotto copertura, pertanto il censimento della popolazione ne prevede una correzione statistica basata sulle due componenti di indagine del master sample. In questo lavoro si analizza il possibile contributo dell'indagine Forze lavoro per la stima della sovracopertura del registro, a partire dagli indicatori di mancata risposta dell'indagine (in particolare quelli relativi ai mancati contatti).*

Key words: population counts, population census, register overcoverage, social surveys

1 Introduction

One of the main objectives of the Census of Population and Housing (CP) is the production of population counts; in the current framework these are obtained through

[1] Loredana Di Consiglio, Istat; diconsig@istat.it
Stefano Falorsi, Istat; email: stfalors@istat

the Population Register (Base Register of Individuals, Registro Base degli individui - RBI), properly adjusted to account for over- and under-coverage. Over-coverage refers to the inclusion in RBI of individuals who are not present and do not usually live in the municipality; under-coverage refers to the non-inclusion in RBI of individuals present and usually living in the municipality. They are estimated from CP Master Sample (CP-MS) List (L) and Area (A) surveys, supplemented with administrative information, by applying a small area estimation method to guarantee reliable estimates at municipality level.

On the other hand, other repeated social surveys currently produce a set of standardized non-response indicators, some of which could be exploited in this context. The largest of these surveys is the Labour Force Survey (LFS), characterized by a high quality profile and producing a long time series of quarterly non-response indicators since 2004. The LFS sampling design is similar to the CP-MS List survey: a large number of municipalities overlap between the two samples and the LFS household sample is selected from an administrative frame that is the main source of RBI. The goal of this work is to analyse LFS non-response indicators and to assess the usability of LFS data to contribute to the evaluation of RBI overcoverage.

2 LFS data collection indicators for the over-coverage estimation

The Labour force survey (LFS) is the largest social survey conducted by Istat, apart from the CP-MS surveys; both are based on a two stages sampling, municipalities represent primary sampling units, while households are the final sampling units.

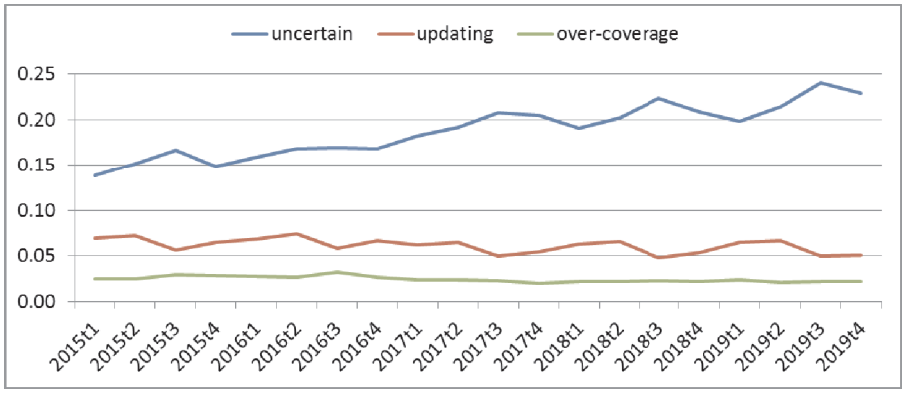
For the purposes of this work it is important to underline that most of the municipalities selected as primary sampling units for the LFS are also included in the sample of the L survey of CP-MS, only the smallest municipalities, which are subject to yearly rotation in LFS, are not included with probability equal to one in CP-MS. It is worth noting that LFS sample is selected from the administrative lists of population taken by the municipalities, which are the main sources (99.97%) of RBI from which the L survey of CP-MS is selected.

Each household is interviewed four times over a 15 months period. The first interview of each household is conducted with the CAPI technique, while subsequent interviews are conducted with the CATI technique (except for families without telephone or with a foreign family head). In general, the interview is conducted in the week following the reference week or, less frequently, in the three following weeks. Interviews are conducted by professional interviewers, directly trained by Istat, most of them are long-term workers in the LFS data collection. The theoretical quarterly sample is composed by almost 18,000 households to be interviewed for the first wave, about 71,000 each year.

Each time a household is contacted, the outcome of the contact is recorded: in case of non-response the interviewer has to record the reason, choosing it among a large set of items. Some of these items are not related to the willingness of the

household to participate in the survey, but refer to the quality of the frame from which the sample is selected (in relation to the reference population). In particular, a couple of items refer to houses that are vacant, for example, holiday properties or second homes, or to persons living in a collective household; these items are considered as signals of over-coverage of the frame. On the other hand other items are related to updating of the frame, i.e. they refer to non-contact because the household moved to another municipality or abroad or due to death of the member. These cases depend on the time lag between the sample selection and the data collection and may be solved ex-post by linking to the updated frame. It is worth noting that in many other cases, indeed the majority, the non-response depends on the impossibility of contacting the family, due to several reasons, the more frequent are the absence from home for many hours every day, or for a longer period. We cannot exclude that among them there are other over-coverage signals, that we are not able to distinguish. In Figure 1 LFS quarterly household non-response indicators are shown (years 2015-2019).

Figure 1: LFS non-response indicators (quarterly data years 2015-2019)



We may observe that the "uncertain" and "updating" indicators show a seasonal pattern: the "uncertain" indicator is always higher in the third quarter, when many families are away from home because of summer holidays; the "updating" indicator is always lower in the third quarter because starting from this quarter the sample of households is selected from an updated frame. On the other hand, the "over-coverage" indicator, as expected, is rather stable over time.

In this work the main focus regards signals of over-coverage of the frame coming from LFS data collection. The goal is analysing the quality of the over-coverage estimates produced from LFS, in terms of stability over time, coherence, and predictive power. Finally, the aim is to assess the usability of LFS over-coverage rates to contribute to the evaluation of RBI over-coverage.

We examined the "over-coverage" indicator estimated at the municipality level, exploring the relationship with certain characteristics of the municipalities taken from the register of the territorial units. Through a simple regression model it results that the following characteristics are significant: being a holiday destination, coastal,

having a low degree of urbanization, 20-40 minutes are needed to reach the closest centre providing main services. Similar conclusions have been derived in the over-coverage estimation conducted on Census data.

The LFS over-coverage indicator was initially estimated at the household level, as we described above, exploiting household non-response indicators; however as the final objective of the over-coverage estimation is the production of population counts, thus the evaluation of the over-coverage should be conducted at the individual level. To this aim the household non-response indicators have been assigned to the individuals belonging to the household, as they result in the administrative frame (this represents the over_cov_fam rate). To take account also of the over-coverage of individuals in responding households, we compared the composition of the household in the administrative frame with the individuals belonging to the *de-facto* household as it was collected by the survey (in this way we estimated the over_cov_ind_rate). Jointly, the two components of over-covered individuals in non-responding and responding households represent the total over-coverage estimate (over_cov_tot). We may observe that through this approach we are able to evaluate over-coverage of individuals in responding and non responding households, while under-coverage may be evaluated only in responding households.

In Table 1 the individual level components of the over-coverage are shown by NUTS1 and NUTS2 regions. Moreover, the share coming from the household/individual component of the over-coverage is added.

Table 1: LFS estimated over-coverage rates and share of family/individual components (year 2019)

	over_cov fam	over_cov ind	over_cov tot	% fam	% ind
Italy	1.6	2.3	3.9	40.3	59.7
North	0.7	1.6	2.3	28.9	71.1
Centre	1.8	1.8	3.6	49.0	51.0
South	2.7	3.5	6.2	43.1	56.9
Piemonte	0.6	1.9	2.5	23.5	76.5
Valle d'Aosta	0.9	1.4	2.2	38.7	61.3
Lombardia	0.6	1.3	1.9	29.4	70.6
Trentino Alto Adige	0.6	1.5	2.1	29.3	70.7
Veneto	0.6	2.1	2.7	20.9	79.1
Friuli Venezia Giulia	0.7	2.8	3.5	20.5	79.5
Liguria	2.3	1.7	4.0	58.0	42.0
Emilia Romagna	0.5	1.3	1.8	29.4	70.6
Toscana	1.9	1.8	3.8	51.7	48.3
Umbria	0.7	1.0	1.8	41.3	58.7
Marche	0.5	1.9	2.4	20.7	79.3
Lazio	2.1	1.9	4.1	52.3	47.7
Abruzzo	1.1	2.9	4.0	28.3	71.7
Molise	4.7	3.9	8.6	54.3	45.7
Campania	1.9	3.6	5.5	34.4	65.6
Puglia	1.0	3.9	4.9	20.6	79.4
Basilicata	3.2	3.7	6.9	46.9	53.1
Calabria	3.8	4.2	8.0	47.5	52.5
Sicilia	4.8	3.2	8.0	59.9	40.1
Sardegna	2.7	3.1	5.8	46.4	53.6

We can observe that the LFS estimated over-overage rates are rather different between North, Centre and South and between regions, they are generally higher in the South, the household component has higher variability compared to the individual one. The share of the two components (household/individual) is rather different, the household component seems to be more influent than the individual component on the total rate.

We could interpret the over_cov_fam and the over_cov_tot rates as lower and upper bound of the over-coverage at the municipality level, respectively: the over_cov_ind is indeed an overestimation of the over-coverage of the individuals in the interviewed households, because the individuals may still live in another house in the same municipality (this does not represent over-coverage at the municipality level). As regards the over_cov_fam, the interviewer verifies the presence of the household on a different address in the same municipality. To further analyse this phenomenon, with the aim of obtaining a more precise estimation of the over-coverage from LFS data, it could be useful integrating LFS data on over-coverage with administrative signals on the presence of the individuals on the territory, coming from administrative sources on work and education (and other). This deepening is currently ongoing. Moreover these administrative signals about the presence of the individuals on the territory will be useful also to analyse the "uncertain" component of households non-response composed by households for which it was not possible to make any contact and we do not dispose of any information (see Fig. 1).

In Table 2 correlations between yearly LFS estimated over-overage rates at the municipality level are shown, for the years 2014-2019. Correlations are high, confirming that the over-coverage is rather stable over time (see Fig. 1).

Table 2: Correlations between LFS estimated over-overage rates at the municipality level (years 2014-2019 and multiannual averages)

	2014	2015	2016	2017	2018	2019	2014-2016	2017-2019	2014-2019
2014	1.00	0.75	0.70	0.60	0.60	0.60	0.89	0.66	0.81
2015	0.75	1.00	0.78	0.71	0.64	0.65	0.93	0.72	0.87
2016	0.70	0.78	1.00	0.76	0.74	0.69	0.91	0.79	0.90
2017	0.60	0.71	0.76	1.00	0.78	0.71	0.76	0.91	0.88
2018	0.60	0.64	0.74	0.78	1.00	0.80	0.73	0.94	0.88
2019	0.60	0.65	0.69	0.71	0.80	1.00	0.71	0.91	0.86
2014-2016	0.89	0.93	0.91	0.76	0.73	0.71	1.00	0.80	0.94
2017-2019	0.66	0.72	0.79	0.91	0.94	0.91	0.80	1.00	0.95
2014-2019	0.81	0.87	0.90	0.88	0.88	0.86	0.94	0.95	1.00

The stability of the LFS estimated over-coverage rates allow us to consider in the analysis also multiannual rates, in order to reduce the variability for the smallest municipalities and the concentration of the distribution on the zero, in particular for the household component (while in 1-year LFS sample about 50% of the municipalities has over_cov_fam=0, the share is about 30% in a 3-years sample and less than 20% in a 6-years sample). It is worth noting that over the 6-years period

2014-2019 the dimension of the theoretical LFS sample (for the first wave) is about 420,000 households, close to the dimension of the CP-MS Area survey.

In Table 3 the mean and median of the three over-coverage rates (family, individual and total) at the municipality level in 2019 are shown.

Table 3: Mean and median of LFS estimated over-coverage rates (year 2019)

	over_cov fam	over_cov ind	over_cov tot
mean	1.9	2.5	4.5
median	0.0	2.2	3.5

The goal of this work was to analyse LFS non-response indicators and to assess their usability to contribute to the evaluation of the statistical register coverage. The preliminary results of these analyses encourage the possibility to exploit the over-coverage rates estimated by the LFS to this aim, in an integrated framework with the CP-MS surveys.

References

1. UNECE, 2018, Guidelines on the use of registers and administrative data for population and housing censuses
2. Stefano Falorsi, 2017, Census and Social Surveys Integrated System. UNECE working paper 23
3. ISTAT, 2006, La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione. Metodi e norme 32
4. Bergamasco, S., Gazzelloni, S., Quattrociocchi, L., Ranaldi, R., Toma, A., 2004, New Strategies to Improve Quality of ISTAT new CAPI/CATI Labour Force Survey. European Conference on Quality and Methodology in Official Statistics
5. Giuliani, G., Grassia, M. G., Quattrociocchi, L., Ranaldi, R., 2004, New Methods for Measuring Quality Indicators of ISTAT's New CAPI/CATI Labour Force Survey. European Conference on Quality and Methodology in Official Statistics

3.5 Excesses and rare events in complex systems

Space-time extreme rainfall simulation under a geostatistical approach

Simulazione spazio-temporale di precipitazioni estreme tramite un approccio geo-statistico

Gianmarco Callegher, Carlo Gaetan, Noemie Le Carrer, Ilaria Prosdocimi

Abstract In this work we illustrate an approach to simulate extreme events with high resolution in space. First we model spatio-temporal variability in the marginal distributions with a flexible semi-parametric specification. Then the Gaussian copula is used to model locally in time and space the extremal dependence. The methods are showcased with an application to daily precipitations in the Venice lagoon catchment.

Abstract *In questo lavoro illustriamo un approccio per simulare eventi con alta risoluzione nello spazio. Per prima cosa modelliamo la variabilità spazio-temporale delle distribuzioni marginali con un approccio semi-parametrico. Quindi la copula gaussiana viene utilizzata per modellare localmente la dipendenza estrema nel tempo e nello spazio. Come esempio mostriamo un'applicazione alle precipitazioni nella laguna di Venezia.*

Key words: Copula, Peak-over-threshold, quantiles, rainfall, Venice lagoon

1 Introduction

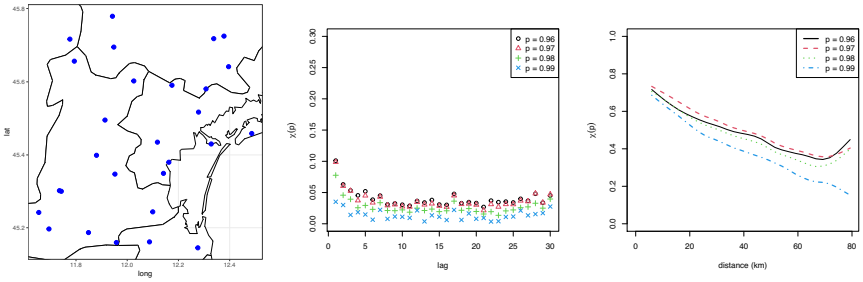
Observational studies have found that extreme precipitation can have heavy-tailed behaviour, i.e. the tail of the distribution of the magnitude of extreme events decays slower than an exponential. In the literature, we can find examples in which climate models of sufficiently high resolution may be capable of simulating precipitation extremes of comparable intensity to observed extremes. However it is not clear that they simulate daily intensities that are as heavy-tailed as observed, nor is it clear that they do so given the different scales in the observations at distinct points and simulated grid-box values. Moreover averaging in space and time smooths the tail

Gianmarco Callegher, Carlo Gaetan, Noemie Le Carrer, Ilaria Prosdocimi
Ca' Foscari University of Venice, DAIS (Italy)
e-mail: [giammarco.callegher,gaetan,noemie.lecarrer,ilaria.prosdocimi]@unive.it

behaviour recorded at weather stations, reducing the usefulness of simulated outputs for impact studies.

In this note we present a two-stage framework in which we couple models for extreme values and geostatistical space-time models. In the first stage, described in Section 2, we focus on the tail of the distribution of the rainfall amount by means of the the so-called peaks-over-threshold (POT) approach and we capture marginal spatio-temporal variation using regression splines (Youngman, 2019). Moreover, we use a Gaussian copula model to capture the short-range spatial and time dependence of the observed data (Section 3). This will allow to simulate extreme events which are consistent with the observed local variability in space and in time.

As a motivating example we consider daily rainfall records from long-term gauging stations in the Venice lagoon catchment from 1956 to 2018. The 28 locations of the stations are plotted in Figure 1-(a).



(a) (b) (c)

Fig. 1 (a) Locations of the stations in Venice lagoon catchment. Empirical estimates of $\chi(p)$ for pairs of observations at increasing temporal lags in (b) and spatial distances (in kilometers) in (c).

2 Extreme value model for a single site and time

We denote with $X(s,t)$ the daily rainfall accumulation at location s and time t. We consider a fixed high threshold $u(s,t)$, and we look at the distribution of the exceedances $(X(s,t)-u(s,t))$, conditional on $X(s,t)$ being larger than $u(s,t)$. Extreme value theory argues that it is possible to approximate this conditional distribution by a Generalized Pareto (GP) distribution. More precisely the distribution of threshold exceedances $Y(s,t) = (X(s,t)-u(s,t))$, given that $X(s,t) > u(s,t)$ has cumulative distribution function (cdf)

$$GPD(x;\xi,\sigma,u) = 1 - \left(1+\xi\frac{x-u}{\sigma}\right)^{-1/\xi}, \tag{1}$$

where $\sigma > 0$ and $\xi \in \mathbf{R}$ are the scale and shape parameter of the distribution for $\{x > u : \left(1 + \xi \frac{x-u}{\sigma}\right) > 0\}$. The threshold can either be chosen or estimated, but must be sufficiently high that the GPD assumption is valid.

For simulations to represent rainfall at given locations and time periods, the rate, i.e. the probability of exceeding the threshold must be taken into account: we denote this probability as $\zeta(s,t) = \Pr(Y(s,t) > u(s,t))$. For now, we assume that the threshold is known and we further simplify the modeling procedure by making the assumption that we are interested in a rate of exceedances that is constant at every site and at every time step, i.e. $\zeta(s,t) = \zeta$.

Then the unconditional distribution for $X(s,t)$ is defined as

$$F(x;\xi(s,t),\sigma(s,t),u(s,t)) = \begin{cases} 1-\zeta+\zeta\left(1+\xi(s,t)\frac{x-u(s,t)}{\sigma(s,t)}\right)^{-1/\xi(s,t)} & x > u(s,t), \\ 1-\zeta & x \leq u(s,t) \end{cases} \tag{2}$$

For the scale parameter $\sigma(s,t)$ we adopt an additive form of the log-link function

$$\log\sigma(s,t) = \beta^{\sigma} + f_1^{\sigma}(lon(s),lat(s)) + f_2^{\sigma}(t) \tag{3}$$

Here f_1^{σ} is thin plate regression spline where $lon(s)$ and $lat(s)$ represent longitude and latitude and f_2^{σ} is a cyclic cubic regression spline of period 365.25 to account for the leap years. Under this setup (3) can be written as

$$\log\sigma(s,t) = \beta^{\sigma} + \sum_{k=1}^{b_1}\beta_{1,k}B_{0,k}(lon(s),lat(s)) + \sum_{k=1}^{b_2}\beta_{2,k}B_{1,k}(t)$$

where $B_{k,i}(\cdot)$ are basis functions and $\beta_{i,k}$ the coefficient multiplying the spline basis. A similar specification is adopted for the shape parameter, namely

$$\xi(s,t) = \beta^{\xi} + f_1^{\xi}(lon(s),lat(s)) + f_2^{\xi}(t) \tag{4}$$

The model (2), (3) and (4) with parameters in the spline forms can be fitted using an approach that maximizes an independence likelihood (Chandler and Bate, 2007). More precisely, let $x(s_j,t)$ be realizations of $X(s_j,t)$ for s_j, $j = 1,\ldots,n$ sites and $t = \ldots,T$ times. By pretending that the observations are independent, the independence likelihood of the model (1) takes the form

$$L(\theta) = \prod_{j=1}^{n}\prod_{t=1}^{T}\frac{1}{\sigma(s_j,t)}\left(1+\xi(s_j,t)\frac{x(s_j,t)-u(s_j,t)}{\sigma(s_j,t)}\right)^{-1/\xi(s_j,t)-1} \tag{5}$$

where θ contains the unknown parameters in (3) and (4).

Maximization of (5) requires the knowledge of the space-time varying threshold $u(s)$. We follow Northrop and Jonathan (2011) and we estimate it by quantile regression (Koenker and Bassett, 1978) assuming that the threshold $u(s,t) = u_1(s) + u_2(t)$

can be splitted in two components: one ($u_1(s)$) which depends on the geographical coordinates and the other ($u_2(t)$) on the season. The effect of the season is modelled by a harmonic regression term. We have

$$\begin{aligned} u(s,t) &= u_1(s) + u_2(t) \\ &= \delta_0 + \sum_{i=1}^{d_0} \delta_{0,i} B_{1i}(lon(s), lat(s)) + \sum_{k=1}^{d_1} \delta_{1,k} \cos(\omega_k t) + \sum_{k=1}^{d_1} \delta_{2,k} \sin(\omega_k t) \end{aligned}$$

where $\omega_k = 2\pi k/365.25$.

3 Copula and extremal dependence

In the previous section we have described how to specify a model for the distribution of the extreme rainfall in one site s and at time t. However, this model can only reproduce the variability of the data at a low resolution (i.e at the point scale), and we need a model for efficient simulations of high-resolution extreme events. For this reason we couple the marginal model with a model for the local variation on space and time under a copula approach (Joe, 2014). It can be shown that every continuous multivariate distribution can be represented in terms of a copula which couples the univariate marginal distributions. More precisely, for a n-variate cdf $F(x_1,\ldots,x_n) := \Pr(X_1 \leq x_1,\ldots,X_n \leq x_n)$ with i-th univariate margin $F_i(x_i) := \Pr(X_i \leq x_i)$, the copula associated with F is a cdf function $C_n : [0,1]^n \to [0,1]$ with $\mathscr{U}(0,1)$ margins that satisfies $F_n(x_1,\ldots,x_n) = C_n(F_1(x_1),\ldots,F_n(x_n))$. Note that the copula does not depend on the marginal distributions. For this reason, it is possible to characterize the extremal dependence through the copula function and distinguish between asymptotic independence and asymptotic dependence (Coles et al, 1999). Formally, let X_1 and X_2 be continuous random variables with distribution functions F_1 and F_2, respectively, and let

$$\chi(p) = \Pr(F_2(X_2) > p | F_1(X_1) > p) = \frac{1-2p-C_2(p,p)}{1-p}, \qquad 0 \leq p < 1. \quad (6)$$

X_1 and X_2 are then said to be asymptotically independent if the limit $\chi := \lim_{p\to 1^-} \chi(p)$ is zero and asymptotically dependent if $\chi > 0$. Broadly speaking, under asymptotic independence the conditional probability of observing an exceedance in one variable given that the other variable has produced an exceedance converges to 0 as the threshold increases.

Copula based on Gaussian process can represent pairs of random variable which are asymptotically independent (Bortot et al, 2000). They play an important role since they can accommodate a variety of spatio-temporal dependence.

Assuming that the estimated marginal model (2,3,4) is the "true" generating model, we calculate uniformly distributed residuals on $[1-\zeta, 1]$:

$$R^*(s,t) = 1 - \zeta \left[1 + \xi(s,t) \frac{X(s,t) - u(s,t)}{\sigma(s,t)} \right]^{-1/\xi(s,t)}, \qquad \text{if } X(s,t) > u(s,t)$$

Figure 1 displays estimates of $\chi(p)$ for probabilities $p = 0.96, 0.97, 0.98, 0.99$ for pairs $R^*(s,t), R^*(s,t+h)$ with only temporal lag, and for pairs $R^*(s,t), R^*(s',t)$ with only spatial lag. The curves for spatial lags are the result of a smoothing procedure. These plots support the assumption of asymptotic independence at all positive distances and at all positive temporal lags.

Finally, the $R^*(s,t)$ random variable is transformed on a Gaussian scale by $R(s,t) = \Phi^{-1}(V(s,t))$ where $\Phi^{-1}(u)$ is the inverse of the cumulative distribution function of a standardized Gaussian random variable. We model $R(s,t)$ as a space-time zero mean Gaussian process with $\rho(s,s',t,t',\phi) = \text{cor}(R(s,t), R(s',t'))$, a correlation function that depends on an unknown parameter ϕ. Since for large data sets the evaluation of the censored likelihood becomes unfeasible the correlation can be estimated by maximizing the censored composite log-likelihood (see Bacro et al, 2020, for an example).

4 Results

Simulations are based on a stationary isotropic separable covariance function. The limited size of the area under analysis and daily temporal lags suggest that they will not have any impact on the recorded values. We consider an exponential-exponential separable correlation function, $\rho(s,s',t,t',\phi_1,\phi_2) = \exp(-\|s-s'\|/\phi_1) \times \exp(-|t-t'\|/\phi_2)$, $\phi_1, \phi_2 > 0$. The resulting estimates are $\hat{\phi}_1 = 127.41$ and $\hat{\phi}_2 = 1.04$, respectively. The low value of $\hat{\phi}_2$ indicate weak temporal dependence in the rainfall phenomena. As expected, an high spatial dependence is estimated. A fifty-years simulation is performed over a 700 evenly-spaced points grid. This results in a distance of $\simeq 2.43$ Km between two neighbouring locations. Figure 2 shows three different randomly selected events. In each row we report the spatial pattern of the day d, over an year, during which we simulated the maximum-precipitation in one site day, with the previous $(d-1)$ and following $d+1$ days. Extreme events can occur in different seasons, but their magnitude is strongly time-dependent by construction.

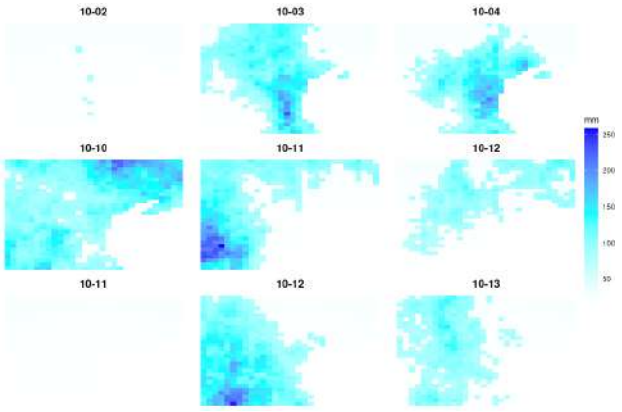


Fig. 2 Three examples of extreme precipitations simulations in three different period of the year. The central plot represents the day with highest precipitations

Acknowledgements

The research presented was funded by CORILA - research program "VENEZIA 2021". The authors thank ARPAV for providing the rainfall data.

References

Bacro JN, Gaetan C, Opitz T, Toulemonde G (2020) Hierarchical space-time modeling of exceedances with an application to rainfall data. Journal of the American Statistical Association 115:555–569

Bortot P, Coles S, Tawn J (2000) The multivariate gaussian tail model: An application to oceanographic data. Journal of the Royal Statistical Society Series C 49:31–49

Chandler RE, Bate S (2007) Inference for clustered data using the independence loglikelihood. Biometrika 94:167–183

Coles S, Heffernan J, Tawn JA (1999) Dependence measures for extreme value analyses. Extremes 2:339–365

Joe H (2014) Dependence Modeling with Copulas. Chapman and Hall/CRC, Boca Raton, FL

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33 – 50

Northrop PJ, Jonathan P (2011) Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. Environmetrics 22:799–809

Youngman BD (2019) Generalized additive models for exceedances of high thresholds with an application to return level estimation for u.s. wind gusts. Journal of the American Statistical Association 114:1865–1879

3.6 Hierarchical forecasting and forecast combination

Density calibration with consistent scoring functions

Roberto Casarin and and Francesco Ravazzolo

Abstract This contribution studies a calibration approach for predictive densities based on generalized scoring rules. We consider a set of simulated experiments in order to study the effectiveness of the method.

Questo lavoro studia un approccio di calibrazione per densità previsive basato su regole di scoring generalizzate. Considera una serie di esperimenti simulati per studiare l'efficacia del metodo.

Key words: Density calibration, Predictive distributions, Scoring rules.

1 Introduction

When multiple forecasts are available from different models or sources it is possible to combine these in order to make use of all relevant information on the variable to be predicted and, as a consequence, to produce better forecasts. This is particular important when working with large database and selection of relevant information *a priori* is not an easy task. Early papers on forecasting with model combinations are [1], who considered air passenger data, and [17] who introduced a distribution which includes the predictions from two experts (or models). This latter distribution is essentially a weighted average of the posterior distributions of two models and is similar to the result of a Bayesian Model Averaging (BMA) procedure, see [15]. [14] extend the BMA framework by introducing a method for obtaining probabilistic forecasts from ensembles in the form of predictive densities and [12] extend it to Bayesian predictive synthesis.

[3] deal with the combination of predictions from different forecasting models using descriptive regression. [9] extend this and propose to combine forecasts with unrestricted regression coefficients as weights. [18] generalize the problem to a state space with weights that are assumed to follow a random walk process. [11] propose robust time-varying weights and account for both model and parameter uncertainty

Roberto Casarin
University Ca' Foscari of Venice e-mail: r.casarin@unive.it

Francesco Ravazzolo
Free University of Bozen-Bolzano, BI Norwegian Business School and RCEA e-mail: Francesco.Ravazzolo@unibz.it

in model averaging. [13] derive time-varying weights in dynamic model averaging, and speed up computations by applying forgetting factors in the recursive Kalman filter updating.

Combination weights that depend on (optimal) score functions have also been studied. [10] introduce the Kullback-Leibler divergence as a unified measure for the evaluation and suggest weights that maximize such a distance, see also [6] for a comprehensive discussion on how such weights are robust to model incompleteness, that is the true model is not included in the model set. [8] recommend strictly proper scoring rules. [4] develops a general method that can deal with most of issues discussed above, including time-variation in combination weights, learning from past performance, model incompleteness, correlations among weights and joint combined predictions of several variables.

Finally, the last aspect relates to calibration and combinations. [16] and [7] introduces the idea of recalibration density forecasts when the density is not well-calibrated. They introduce a monotone non-decreasing map via a Beta distribution to achieve it. [2] generalize to Beta mixtures, allowing for more flexibility in calibrating and combinations in presence of fat tails, skewness and multiple-modes.

This paper extends the density calibration literature and proposes to apply consistent scoring functions when calibrating models. We follow [5] and consider three different consistent scoring functions. These functions are minimized to compute the parameters of a beta calibration scheme. We study in simulation exercises the effectiveness of the method.

The structure of the paper is organized as follows. Section 2 presents the optimal calibration method. Section 3 provides numerical examples and directions for future research.

2 Optimal calibration

Consider the forecast distribution F_1 from a predictive model and F the distribution of Y, the variable to forecast. One can consider the following map

$$(\theta,\xi) \mapsto D(\theta,\xi) = \mathbb{E}_{\mathbb{Q},\xi}(S_{\alpha,\theta}(X,Y)) \tag{1}$$

where X is a point forecast from F_1, α a quantile level, $\theta \in \Theta \subset \mathbb{R}$ a threshold parameter and $\xi \in \Xi \subset \mathbb{R}^k$ a combination/calibration parameter vector. If the parameter ξ is indexing a family of continuous distributions $H_{\xi,F_1}(X) = (G_\xi \circ F_1)(X)$, with $x \mapsto G_\xi(x) \in (0,1)$ $\xi \in \Xi$ a sequence of non-decreasing functions with $G(0) = 0$ and $G(1) = 1$, then we obtain a calibration scheme. Our optimal calibration scheme can be defined as

$$\theta \mapsto \inf_{\xi\in\Xi} D(\theta,\xi) \tag{2}$$

In this study we follow [2] and assume $G_\xi(x)$ is the cumulative distribution function (cdf) of a beta distribution $B(x;\mu\phi,(1-\mu)\phi)$ with parameters $\xi = (\mu\phi,(1-\mu)\phi)$.

It follows that the calibrated density is $h_{\xi,F_1}(X) = (g_\xi \circ F_1)(X) f_1(X)$ where f_1 and g_ξ are the probability density functions of H_ξ and F_1. We denote with $\xi(\theta)$ the solution of Eq. (2).

The scoring function $S_{\alpha,\theta}(X,Y)$ can assume different forms. Following [5], we investigate three different consistent specifications. The first one is an elementary weighted average over elementary or extreme scores:

$$S_\theta(X,Y) = (Y-\theta)_+ - (Y-\theta)_+ - \mathbb{I}(x>\theta)(Y-X) \tag{3}$$

with $(t)_+ = max(t,0)$ and $\mathbb{I}(A)$ the indicator of the event A.

The second specification refers to a quantile consistent scoring function:

$$\begin{aligned} S^q_{\alpha,\theta}(X,Y) &= \{\mathbb{I}(Y<X)-\alpha\}\{\mathbb{I}(\theta<X)-\mathbb{I}(\theta<Y)\} \\ &= \begin{cases} 1-\alpha, & Y \le \theta < X \\ \alpha, & X \le \theta < Y \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

The third specification relies to an expectile consistent representation:

$$\begin{aligned} S^e_{\alpha,\theta}(X,Y) &= |\mathbb{I}(Y<X)-\alpha|\{(Y-\theta)_+ - (X-\theta)_+ - (Y-X)\mathbb{I}(\theta<X)\} \\ &= \begin{cases} (1-\alpha)|Y-\theta|, & Y \le \theta < X \\ \alpha|Y-\theta|, & X \le \theta < Y \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

We apply model in (2), with a beta calibration scheme and scoring functions in (3)-(5).

3 Numerical Illustration

The assume $Y_i \sim \mathscr{G}a(1,1)$ $i = 1,\dots,N$, where $\mathscr{G}a(a,b)$ denotes a gamma distribution with mean ab. The misspecified model is alternatively in the same distribution family $\mathscr{G}a(2,1)$ or in the lognormal distribution family $\Lambda(2,2)$. The score function $\mathbb{E}_{\mathbb{Q},\xi}(S_{\alpha,\theta}(X,Y))$ is evaluated on the data

$$\widehat{D(\theta,\xi)}_N = \frac{1}{N}\mathbb{E}_{\mathbb{Q},\xi}(S_{\alpha,\theta}(X,Y_i)) == \frac{1}{N}\sum_{i=1}^{N}\int S_{\alpha,\theta}(x,Y_i)h_{\xi,F}(x)dx \tag{6}$$

and the expectation approximated with M Monte Carlo samples from the predictive distribution, that is

$$\widehat{D(\theta,\xi)}_{M,N} = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N} S_{\alpha,\theta}(X_j,Y_i) \xrightarrow[M\to\infty]{as} \widehat{D(\theta,\xi)}_N \tag{7}$$

which converges to the empirical scoring function by the SLLN. Samples from the calibrated predictive distribution are obtained by inverse cdf methods, that is we first generate U from the standard uniform $\mathscr{U}(0,1)$ and then $X = F^{-1}(G_\xi^{-1}(U))$ where F^{-1} and G_ξ^{-1} are the inverse cdf of the misspecified model and of the beta distribution. The validity of the method follows from $P(X \leq x) = P(F^{-1}(G_\xi^{-1}(U)) < x) = P(G_\xi^{-1}(U)) < F(x)) = P(U < G_\xi(F(x))) = G_\xi(F(x)))$. In the numerical experiments we set $N = 100$ and $M = 1000$.

Fig. 1 reports the Murphy's diagrams for three scoring functions in (3)-(4)-(5) of the true model, the misspecified gamma models and the optimal calibration scheme. Results indicate that the calibrated forecasts is closer to the true model in all cases.

References

1. Barnard, G. A. (1963). New methods of quality control. *Journal of the Royal Statistical Society, Series A*, **126**, 255–259.
2. Bassetti, F., Casarin, R., and Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association*, **113(522)**, 675–685.
3. Bates, J. and Granger, C. (1969). The combination of forecasts. *Operations Research Quarterly*, **20(4)**, 451–468.
4. Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, **177**, 213–232.
5. Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78(3)**, 505–562.
6. Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, **164**, 130–141.
7. Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747–1782.
8. Gneiting, T. G. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102(477)**, 359–378.
9. Granger, C. W. J. and Ramanathan, R. (1984). Improved Methods of Combining Forecasts. *Journal of Forecasting*, **3**, 197–204.
10. Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, **23**, 1–13.
11. Hoogerheide, L., Kleijn, R., Ravazzolo, R., van Dijk, H. K., and Verbeek, M. (2010). Forecast Accuracy and Economic Gains from Bayesian Model Averaging using Time Varying Weights. *Journal of Forecasting*, **29(1-2)**, 251–269.
12. McAlinn, K. and West, M. (2018). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, forthcoming.
13. Raftery, A., Karny, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, **52**, 52–66.
14. Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, **133**, 1155–1174.
15. Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92(437)**, 179–91.
16. Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72(1)**, 71–91.

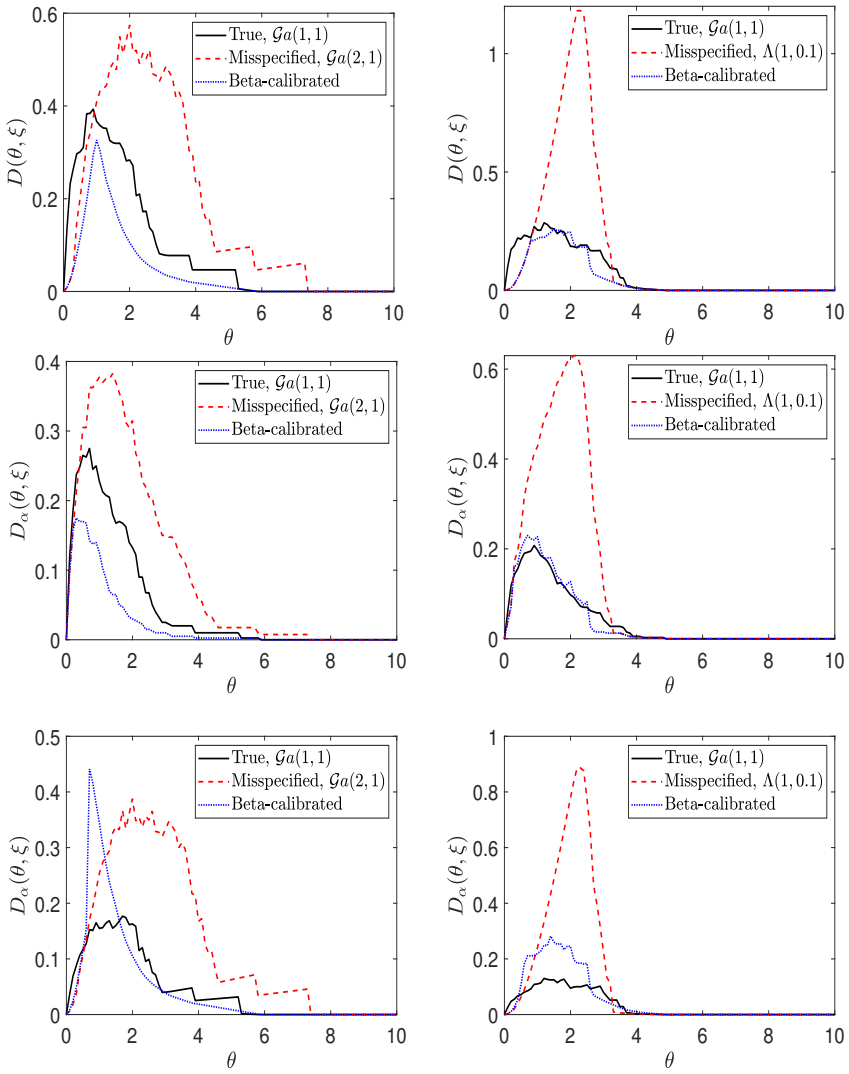


Fig. 1 Murphy's diagrams for score functions $S_\theta(X,Y)$ (first line), $S^q_{\alpha,\theta}(X,Y)$ (second line) and $S^e_{\alpha,\theta}(X,Y)$ (third line) with $\alpha = 0.25$. The misspecified models are in the Gamma (left) and lognormal (right) families of distributions. In each plot: the perfect (black solid), misspecified (red dotted) and calibrated (blue dashed) forecasters, where the calibrated forecaster is obtained by applying a beta calibration function to the misspecified model.

17. Roberts, H. V. (1965). Probabilistic prediction. *Journal of American Statistical Association*, **60**,50–62.
18. Terui, N. and van Dijk, H. K. (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, **18**, 421–438.

Forecasting combination of hierarchical time series: a novel method with an application to CoVid-19

Combinazione di Previsioni per serie storiche gerarchiche: proposta di un metodo e sua applicazione a dati relativi al CoViD-19

Livio Fenga

Abstract Multiple, hierarchically organized time series are routinely submitted to the forecaster upon request to provide estimates of their future values, regardless the level occupied in the hierarchy. In this paper, a novel method for the prediction of hierarchically structured time series will be presented. The idea is to enhance the quality of the predictions obtained using a technique of the type forecast reconciliation, by applying this procedure to a set of optimally combined predictions, generated by different statistical models. The goodness of the proposed method will be evaluated using the official time series related to the number of people tested positive to the SARS-CoV-2 in each of the Italian regions, between February 24^{th} 2020 and August 31^{th} 2020.

Abstract *Serie storiche multiple, gerarchicamente organizzate, sono spesso usate per la previsione di aggregati intermedi o totale ma anche delle singole serie componenti all'interno della struttura gerarchica di riferimento. In questo paper viene proposto un nuovo metodo per condurre l'esercizio previsivo in un tale contesto. L'idea è quella di migliorare la qualità delle previsioni ottenute usando la tecnica della riconciliazione ad un insieme di previsioni, generate da differenti modelli statistici e combinate secondo diversi approcci. La bontà del metodo viene discussa sulla base delle serie storiche ufficiali disponibili sul fenomeno CoViD per il periodo 24 Febbraio – 31 Agosto 2020*

Key words: ARIMA model, ARFIMA model, ETS model, forecast reconciliation, forecast combination,model uncertainty, SARS-CoV-2 , Theta method

Italian National Institute of Statistics
Via Cesare Balbo, 16 00184 Roma; e-mail: livio.fenga@istat.it

1 Introduction

In many applications, it is often the case that accurate forecasts are needed for time series showing an inherent hierarchical structure. The estimation of the future demand of domestic tourism usually follows a geographical proximity criterion, based on which the related time series are organized (and predicted) according to homogeneous groups. Sometimes, emergency situations require close monitoring of the spread of a disease not only at a national but also at a regional level, e.g. in order to set up more appropriate countermeasures for elderly and chronically ill people. These are all cases where a single line of hierarchy generates the overall structure of the data which therefore is referred to as "hierarchical time series".

2 Hierarchical cross-sectional reconciliation: the chosen method

This paper focuses on structures of the type summation constrained, in the sense that the underlying hierarchic structure of a given $m-$dimensional time series x_t, arises by summing up the bottom-level series into the higher ones. Figure 1 is an example of such a structure, under the condition that the constraints $x_t = x_{a,t} + x_{b,t}$, $x_{a,t} = x_{aa,t} + x_{ab,t} + x_{ac,t}$ and $x_{b,t} = x_{ba,t} + x_{bb,t}$ are all satisfied.

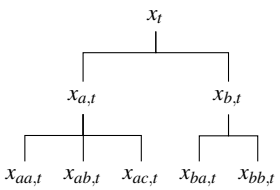


Fig. 1: A two-level hierarchical structure

Formally, we have that the observed data x_t – as well as their estimated future values, defined as x_h; $h = 1,2,\ldots,H$, with H the prediction horizon – lie in the summation-coherent subspace $\{\mathscr{U}\}$; $\forall t = 1,2,\ldots,T$ and $\forall h = 1,2,\ldots,H$. The prediction step subscript h has been omitted in Figure 1, for the sake of a better readability. In total, this hierarchy contains $m = 8$ time series, $n = 5$ of which are the lowest level time series, which therefore constitute the highest level of disaggregation. The observed series $x_t \in \mathbb{R}^m$ can be broken down as follows: $x_t = [u_t^T, b_t^T]^T$, where $b_t^T \in \mathbb{R}^n$ and $u_t^T \in \mathbb{R}^{m-n}$ respectively contain the data pertaining to the bottom and upper series. Therefore, according this representation, the structure of Figure 1 (omitting the subscript t) can be broken down as follows: $[u_t^T, b_t^T]^T \equiv [x, x_a, x_b, x_{aa}, x_{ab}, x_{ac}, x_{ba}, x_{bb}]^T$, $u_t^T \equiv [x_a, x_b]^T$ and $b_t^T \equiv [x_{aa}, x_{ab}, x_{ac}, x_{ba}, x_{bb}]^T$. The hierarchic structure – satisfying $x \subset \{\mathscr{U}\}$ – is induced by the summing matrix $\mathscr{S}$ of

dimension $m \times n$ such that $x = \mathscr{S} b_t$. Formally: $x \subset \{\mathscr{U}\} \iff x_t = \mathscr{S} b_t$ (the symbol $\iff$ replacing the locution "in and only if"). The $\mathscr{S}$ matrix for the hierarchy in Figure 1 is as follows:

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Using the symbols $\tilde{}$ and $\hat{}$ respectively to refer to the case of coherent and base (generally non coherent) forecasts, the reconciliated forecast $h-$ step ahead can be expressed as proposed by Hyndman et al. (2011), i.e.

$$\tilde{x}(h) = \mathscr{S} P \hat{x}(h), \tag{1}$$

for some appropriately chosen matrix $P \in \mathbb{R}^{m \times n}$. Assuming unbiased base forecasts, the best (in the sense of minimum sum of variances) linear unbiased revised forecasts are given by Equation 1 with

$$P = (\mathscr{S}' W^{-1} \mathscr{S})^{-1} \mathscr{S}' W^{-1} \tag{2}$$

and thus (see Taieb et al. (2017), Theorem 1)

$$\tilde{x}(h) = \mathscr{S} (\mathscr{S}' W^{-1} \mathscr{S})^{-1} \mathscr{S}' W^{-1} \hat{x}(h), \tag{3}$$

where $\mathscr{S}$ is as above defined and $\hat{x}(h)$ and $\tilde{x}(h)$; $h = 1, 2, \ldots, H$ represent respectively the set of H predictions independently generated and the ones made coherent. Finally, W is the positive definite covariance matrix of the base forecast errors, i.e. $\hat{e}_t(h) = \hat{x}_t(h) - x_t(h)$, so that $W(h) = \mathbb{E}[\hat{e}_t(h) - \hat{e}'_t(h)]$. As shown by Wickramasuriya et al. (2019), matrix W appears in the equation for the estimation of the error variance of the reconciled forecasts, i.e.

$$V(h) = Var[x(T+h) - \tilde{x}(h)] = \mathscr{S} P W(h) P' \mathscr{S}', \tag{4}$$

whose diagonal elements are the variances of the forecast errors. Their minimization can thus be performed in terms of the trace of $V(h)$ and given by Equation 2 (therefore, this method is called Minimum Trace Estimator). Unfortunately, as proved by the same Authors, W is not identifiable, therefore, in the empirical section, the workaround proposed by them will be adopted. In essence, it is assumed $W_h = k(h)\text{diag}\hat{W}_1$; $\forall h$ and assuming $k(h) > 0$ and denoting with W_1 the forecast errors covariance matrix estimated at horizon $h = 1$ – i.e. $\hat{W}_1 = \frac{1}{T} \sum_1^T e_t e_t{}'$ – and with K is an unknown constant depending on the time horizon h.

2.1 The Forecast combination methods adopted

The first method considered in this paper is of the type simple average, which assigns equal weights to all predictors, i.e. $w^{sa} = \frac{1}{N}$ and thus the combined forecast is

$$f = f_t' w^{sa}.$$

In the second method chosen, the forecast combination weights $w^{ols} = (w_1, w_2, \ldots, w_n)$, along with the intercept b, are computed using ordinary least squares (OLS) regression (Granger (1980)), i.e.

$$f = b + f_t' w^{ols}. \tag{5}$$

The third method applied – of the type Least Absolute Deviation (LAD) – is a modification of the OLS method, and it is expressed as in Equation 5, replacing the superscript *ols* with *lad*. Finally, a modification of the method proposed by Newbold and Granger (1974), built upon an earlier methodology of Bates and Granger (1969), is our fourth approach. Let $\sum$ be the positive definite matrix of the mean squared prediction errors (MSPE) of f_t and g is an $N \times 1$ vector of $(1, 1, \ldots, 1)'$ their method relies on a constrained minimization of the MSPE under the normalizing condition $g'w = 1$. The resulting combination of weights is

$$w^{ng} = \frac{\sum^{-1} g}{g' \sum^{-1} g},$$

so that the combined forecast is

$$f = f_t' w^{ng}. \tag{6}$$

However, unlike the original method, the variant employed here follows the proposal by Hsiao and Wan (2014), which does not impose the prior restriction that the matrix $\sum$ is diagonal.

3 The proposed method

Let us indicate with the symbols $\mathscr{R}$ and $|\cdot|$ respectively a suitable reconciliation method and the cardinality function (assuming the number of elements in a given set to be finite, $|\cdot|$ simply returns the number of the elements belonging to that set). Let the symbol **ncol** identify the function which, applied to a given matrix, returns its number of columns and $\mathscr{M} \equiv \{\mu_1, \mu_2, \ldots, \mu_M\}$ and $\mathscr{D} \equiv \{\delta_1, \delta_2, \ldots, \delta_D\}$ respectively the set of $|\mathscr{M}| = M$ prediction models and the set $|\mathscr{D}| = D$ of forecast combination methods entertained, both arbitrarily chosen. Once applied to the time series of interest x_t; $t = 1, 2, \ldots, T$, each model $\{\mu_j \in \mathscr{M};\ j = 1, 2, \ldots M\}$ generates a set, called $\mathscr{F}^H$, made up with M H–step ahead predictions, i.e.: $\{\mathscr{F}^H(\mu_j);\ j = 1, 2, \ldots M\}$. Each of the elements of this set is a base forecasts, in the sense that

it is generated by individually applying a given statistical model μ_j to the observed time series without any attempt of reconciliation. Each of these M elements in $\mathscr{F}$ (the M forecast vectors) is individually reconciled through the reconciliation procedure $\mathscr{R}$, i.e. $\left\{\mathscr{R}(\mathscr{F}(\mu_j);\ j=1,2,\dots M)\right\}$ (the superscript h is omitted for brevity). At this point, the resulting set $\left\{\mathscr{P}(\mu_j);\ j=1,2,\dots M\right\}$ of M model–dependent reconciled forecasts (first optimization) is optimally combined by applying each method in the set $\mathscr{D}$ to any possible combination (without repetition) of order $\{k=1,2,\dots,M\}$ to the set $\mathscr{P}$ (second optimization). The resulting set $\mathscr{Z}$ – with cardinality $(|\mathscr{D}| * \sum_{k=1}^{|\mathscr{P}|}\binom{M}{k})$ – contains all the possible combinations – $\forall k$-order – of the model-dependent reconciled forecasts. The third optimization step is carried out by applying to $\mathscr{Z}$ a suitable loss function, here denoted with the symbol $\mathscr{L}(\cdot)$. The optimal vector of forecasts is thus the element $z^* \in \mathscr{Z}$ minimizing this function, i.e. $\mathbf{z}^* = \min \mathscr{L}(\mathscr{Z})$. This optimality condition is expressed as

$$\mathbf{z}^* = f(\mu^*, \delta^*), \tag{7}$$

being the arguments of f respectively the "best" forecasting model and forecast combination technique. This last step, by ruling out the less performing combination method(s), has been introduced in order to reduce the overall uncertainty level of the analysis. In fact, suppose that the original set $\mathscr{D}$ reduces to $\mathscr{D}'$ – being clearly $|\mathscr{D}'| < |\mathscr{D}|$ – the additional amounts of undesired fluctuations and noise – which one can reasonably expect as a consequence of the employment of one (more) under-performing combination method(s) – are avoided. Finally, the model bias β^* is empirically estimated using the in–sample residuals generated by employing the winners techniques μ^* and δ^*, according to an optimal choice made on a set of suitable central tendency functions (fourth optimization).

4 Empirical study

In this section the goodness of the proposed method will be evaluated using the official time series related to the number of people tested positive to the SARS-CoV-2 in each of the Italian regions, between February 24^{th} 2020 and October 7^{th} 2020. The whole data set – issued by the Italian National Institute of Health – are publicly and freely available at the web address *https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni*. The data, sampled at a daily frequency, are stored in a matrix called $\mathscr{O}$ of dimension 227×2, where 2 are the Italian regions reported here. The performances of the method – expressed in terms of the cost function $\mathscr{L}$, i.e. the RMSFE – are summarized in Table 1 where the components statistical models are reported in the column "Single models".

Table 1: *Performances of the method for two Italian regions. Outcomes of the winner models and of the single statistical models. See text for details*

Region	Winner combination	$\mathscr{L}^*$	$\mathscr{E}^*$	β^*	β^{out}	Single models	$\mathscr{L}$
Piemonte	$LAD-BET$	135.6	${}^a\tau$	≈ 0	54.09	$ARFIMA$	1050.8
						ETS	1065.8
						θ	1271.8
						$ARIMA$	683.6
Val d'Aosta	$SA-BT$	8.6	${}^a\tau$	≈ 0	-2.25	$ARFIMA$	60.7
						ETS	177.5
						θ	30.5
						$ARIMA$	29.5

References

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468.

Granger, C. W. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of econometrics*, 14(2):227–238.

Hsiao, C. and Wan, S. K. (2014). Is there an optimal forecast combination? *Journal of Econometrics*, 178:294–309.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational statistics & data analysis*, 55(9):2579–2589.

Newbold, P. and Granger, C. W. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)*, 137(2):131–146.

Taieb, S. B., Taylor, J. W., and Hyndman, R. J. (2017). Coherent probabilistic forecasts for hierarchical time series. In *International Conference on Machine Learning*, pages 3348–3357. PMLR.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819.

3.7 Household surveys for policy analysis

Did the policy responses to COVID-19 protect Italian households' incomes? Evidence from survey and administrative data

Le policies in risposta al COVID-19 sono state efficaci nel sostenere i redditi delle famiglie italiane? Evidenze da dati amministrativi e di survey

Abstract This paper addresses the economic impact of the COVID-19 pandemic by providing timely and accurate information on Italian households' income distribution, inequality, poverty and liquidity risk, and assessing the effects of policy responses during 2020. By building a unique and wide database with the latest survey, tax and administrative data at individual and firm level, and by using the micro-simulation model *Taxben-DF* from the Italian Department of Finance, the analysis nowcasts the income loss due to the economic shutdown since March 2020 and simulates most of the measures adopted by the Government from March to December 2020. Results suggest that policy measures in response to the first pandemic wave have been effective in keeping inequalities stable. However, when considering the effects on the whole 2020, measures do not appear to ensure that inequalities and poverty risk will return to pre-COVID levels.

Abstract Questo lavoro intende analizzare gli effetti economici della pandemia da COVID-19 utilizzando informazioni tempestive e accurate per aggiornare la distribuzione dei redditi delle famiglie italiane durante la crisi, e calcolando gli indicatori della disuguaglianza, della povertà e del rischio di liquidità per valutare l'efficacia degli interventi del Governo durante il 2020. Attraverso microsimulazioni elaborate con il modello *Taxben-DF* del Dipartimento delle Finanze e la costruzione di un ampio database, unico nel suo genere, che sfrutta i più recenti dati di survey, fiscali e amministrativi a livello individuale e d'impresa, l'analisi utilizza metodi di *nowcasting* per stimare le perdite di reddito subite a causa del blocco delle attività produttive a partire da marzo 2020 e simula le principali misure adottate dal Governo tra marzo e dicembre 2020. I risultati mostrano che, durante la prima ondata pandemica, le misure sono state efficaci nel mantenere stabile il livello delle disuguaglianze. Tuttavia, considerando l'effetto sull'intero 2020, gli interventi di sostegno ai redditi non sembrano sufficienti a riportare ai livelli pre-crisi disuguaglianza e rischio di povertà.

Key words: COVID-19; inequalities; administrative and survey data; micro-simulation

1 Introduction

The COVID-19 pandemic risks exacerbating existing and new inequalities. The pandemic is having particularly adverse effects on younger workers, women and people that are more vulnerable and affected by longstanding socio-economic inequalities. Pandemics have been shown to increase new inequalities, in the form of job losses, income reductions, or exposure to health risks. As a result, income inequality is likely to rise further over the medium term, unless policies succeed in breaking this trend.

Since the beginning of the COVID-19 crisis, a fast-growing literature has underlined that the most vulnerable social groups have borne the costs of the

pandemic disproportionately, since more likely they lose their job or experience a drop in economic activity during the "Great lockdown".

Less educated and low-skilled workers, younger or unstable employees and self-employed (including those employed in the "black economy") are concentrated in the sectors more affected by the shutdown, or also could work from home less likely (Blundell et al., (2020); Benzeval et al., (2020)). On these groups the consequences of the pandemic were more severe. As a secondary effect, the lockdown also reduced the size of irregular economy (by chance falling in the sectors most hit by the shutdown), reducing fiscal revenues less than expected as well. In addition, households of the top quintile, which were used to spend a third of their income on services restricted by the shutdown, experienced unexpected savings capacity (Crawford et al., (2020)), reinvigorating such dynamics of inequality.

At the same time, new forms of disparity arose along various key dimensions: the *stay-at-home* policy charged a disproportionate burden on women increasing gender inequality (Alon et al., (2020); Andrew et al., (2020a); Del Boca et al., (2020)). Schools' closure amplified education inequalities with strong advantage in favour of children from richest families (Coe et al., (2020); Andrew at al., (2020b)). Men, the elder, and those with lower income were more vulnerable to the infection and faced the highest health risks and fatality ratio (Blundell et al., (2020)). In this context, the level of disposable income before and during COVID-19 is the crucial driver in exacerbating or mitigating the whole inequality dynamics.

Policies aimed at protecting those most directly hit by the crisis, either through automatic stabilisation (e.g. unemployment benefits) or through discretionary measures (e.g. income subsidies or non-refundable grants) adopted to alleviate the impact of COVID-19 on household's incomes and inequalities.

Within this framework, a key question in academic and policy current debates concerns the extent to which policy measures in response to the COVID-19 outbreak have been effective in preserving households' incomes and fencing off an increase in inequality (Gallo and Raitano, (2020); Brunori et al., (2020); Fiorio and Figari, (2020); Clark et al., (2020)).

This paper contributes to the current debate by addressing the economic impact of the COVID-19 pandemic on Italian households' income distribution, inequality, poverty and liquidity risk, and by assessing the mitigation effects of policy responses during 2020.

It is worth noting that tracking household disposable-income inequality during COVID-19 poses several methodological challenges.

First, income-related inequality is one of the most important key variables in social science notwithstanding the difficulty to measure it accurately. In Italy, income data are available from many sources, including household surveys, tax records and administrative data from government program schemes providing transfer payments. Each source has important strengths and major limitations when used alone. Surveys provide rich social and demographic information, but underreport certain type of incomes. Tax records are consistently and accurately collected, resulting in highly reliable data covering a large number of observations, but are only available for those obliged to fill tax forms, missing low-income

households. Finally, administrative data from government programs provide details for safety schemes that are not always captured by other sources but, being collected only for administrative purposes, contain a limited range of variables.

Second, the asymmetric and heterogeneous shock induced by the pandemic is not captured by the available data sources. The lack of up-to-date information on the labour market and on the differential impacts across the population constrain the scale and direction of recent changes in the income distribution, which in turn does not allow timely, effective policy analysis and hinders the efforts to target income support measures. In fact, Personal Income Tax (PIT) returns and representative surveys data on population's incomes and living conditions are usually available two years later from the year in which incomes have been earned.

To overcome this data delay and shortcomings, innovative nowcasting methods are needed to assess the distributional implications of the COVID-19 crisis in light of any differences with pre-pandemic incomes. These would require to adjust and align micro data with the recent changes in labour market and income variations, by taking into account more timely and high frequency data.

2 Data and research design

Matching survey and tax and administrative data at individual level and integrating them with new real-time data sources would yield the most complete and accurate estimates of income and inequality, and would provide timely and sufficiently robust evidence on policies evaluation even in the uncertain scenario of the COVID-19 pandemic.

To this aim, we build a wide and detailed dataset that merges the individual data from the Italian EU-SILC survey for the year 2018 - which contains information on incomes, skills, education level, and employment conditions for a large and representative sample of Italian households - and administrative micro-data drawn from PIT returns of 2018, properly extended to account for the households' income changes due to the pandemic.

More specifically, we included:

- individual data from firms' balance sheets of 2019, the electronic invoicing and periodic VAT returns of 2020 (at firm level) to simulate trends in turnover and costs for goods and services and, thus, derive the income variation experienced by self-employed workers and entrepreneurs;
- administrative firm-level data by the Italian National Social Security Institute (INPS) on the personnel costs actually paid in each firm, suited to define the recipients of the wage-supplementation scheme for working-time reduction *(Cassa Integrazione Guadagni,* CIG) and the employees' income changes.

Relying on this comprehensive dataset, we used the micro-simulation model *Taxben-DF* developed by the Italian Department of Finance to evaluate the impact

of the policies undertaken in response to the pandemic and to assess their heterogeneous distributional impact across taxpayers' groups, regions, and sectors.

Our analysis shows how an approach that combines microsimulation and nowcasting is suited to provide real-time information, leveraging in particular on more frequent updates of firm-level data to input current individual changes in disposable income, so allowing to tailor policy responses to support households through the systemic shocks.

To assess the potential impact of policy measures adopted in the wake of the COVID-19 crisis on household income, poverty and inequality in 2020, we estimated a range of monthly indicators to produce short-term forecasts of the turnover, costs of goods and services, and personnel costs for each combination of 6-digit NACE and Italian region. These indicators have been developed by the Department of Finance and SOSE S.p.a to nowcast the individuals' income variation by differentiating between the incomes of self-employed and entrepreneurs (using the data from periodic VAT returns) and the income of employees (using the data on personnel costs by INPS).

We build two alternative scenarios in which the differences of the households' equivalised disposable income levels are compared to the hypothetical scenario in which COVID-19 has not occurred (**baseline scenario**):

- the **counterfactual scenario** allows to gauge what would have been the economic effect of the pandemic on households' incomes if no discretionary policy measures had been taken to dampen the impact of COVID-19;
- the **real scenario** simulates the impact of both the pandemic and the government policies on income distribution.

In the counterfactual scenario, the monthly indicators of income loss of individuals for each month between March and December 2020 are applied to the pre-Covid individuals' incomes from 2018 PIT returns. In this scenario, we assume that without postponing layoffs and extending the CIG between March and December 2020, employees would have lost their entire salary for a number of months equal to the number of estimated months in which they received the CIG.

In the real scenario, to the income loss caused by the economic shutdown since March 2020, we add the simulation of most of the measures designed to support workers' incomes and to protect firms' jobs during COVID-19.

In detail, we simulate the following combination of policies:

- For employees: the wage-supplementation scheme for working-time reduction "CIG" for a maximum of 10 months between March and December 2020 and suspending layoffs for the full period.
- For self-employed workers and entrepreneurs: the lump-sum benefit of 600 euros for all self-employed workers and entrepreneurs; the lump-sum benefit of 1,000 euros for self-employed professionals; the non-refundable grant aid and the tax credit for rents of non-residential buildings for self-employed workers and entrepreneurs.

- For all households that did not receive the above-mentioned benefits and were not receiving the Citizenship Income, the Emergency Income was assigned, integrating the dataset with open-data by INPS.

3 Main results

To illustrate the disposable income distribution and the dynamics of income inequality under the different scenarios, we focus on particular measures, notably the distribution of net income loss after aids received and the evolution of some inequality and poverty indicators: the interquintile ratio, the Gini index and the risk of poverty indicator.

Simulations based on the policy responses to the first pandemic wave showed that Government income support schemes, combined with existing progressive features of the tax benefit system, have been effective to dampen the increase of income inequalities and the share of individuals at risk of poverty between March and May 2020, especially among the poorest households. Nonetheless, policies produced heterogeneous effects across different types of workers with employees that were finally, less compensated than the self-employed and more likely to incur in a liquidity risk. Nevertheless, the comparison does not take into account the advantage of employees keeping their jobs because of firing blockage.

Results over the whole 2020 do not appear to ensure that inequalities and poverty risk will return to pre-COVID levels. In fact, inequality showed a positive trend since 2017, and, without the pandemic, the interquintile ratio would have reduced to 5.8 (Figure 1). Due to the pandemic, the ratio actually increased by 0.3 points in the real scenario, falling back to its 2017 level, and it increased by 0.9 points in the "no policy scenario".

Nonetheless, public policies adopted soon after the beginning of the pandemic have mitigated the raise of inequality produced by the COVID-19 pandemic and mainly compensated the individuals at the bottom 20 percent of income distribution. However, the Gini index indicates a less marked increase in income inequality, which remains substantially stable with respect to 2019, thanks to the measures adopted. This suggests that the pandemic has driven incomes towards a polarization between the poorest and the richest, more than to a spread of inequality over the whole population.

The first quintile dynamics are also captured by the poverty risk ratio (Figure 2). Despite policies have not completely sterilised the increase in the percentage of families in relative poverty, the CIG to employees, as well as grants and transfers to self-employed and firms, limited the severe increase in poverty that households would have experienced (+1 p.p.), saving more than 90,000 households from poverty risk.

Figure 1: Interquintale ratio – year 2017- 2020

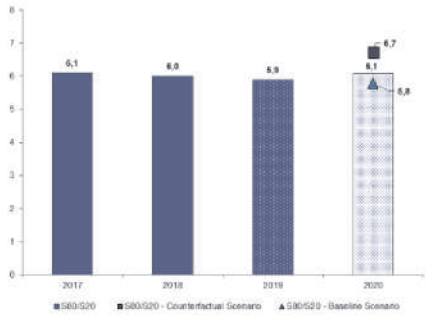
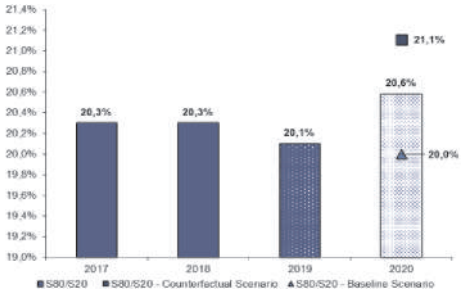


Figure 2: % of individuals exposed to poverty risk – year 2017- 2020



Source: 2017-2019 Istat, EU-Silc Survey; 2019 Istat estimates under a macroeconomic approach; 2020: MEF-DF estimates under the Taxben-DF micro-simulation model. For the 2020, the column shows the data of the real scenario in which the economic effects of both the pandemic and the Government measures are considered.

References

1. Alon, T., Doepke, M., Olmstead-Rumsey, J., Tertilt, M. (2020): The impact of COVID-19 on gender equality, Covid Economics: Vetted and Real-Time Papers, no. 4, pp. 62–85.
2. Andrew, A., Cattan, S., Costa Dias, M., Farquharson, C., Kraftman, L., Krutikova, S., Phimister, A. and Sevilla, A. (2020a): How are mothers and fathers balancing work and family under lockdown?, Institute for Fiscal Studies (IFS), Briefing Note no. BN290.
3. Andrew, A., Cattan, S., Costa Dias, M., Farquharson, C., Kraftman, L., Krutikova, S., Phimister, A. and Sevilla, A. (2020b): Learning during the lockdown: real-time data on children's experiences during home learning, Institute for Fiscal Studies (IFS), Briefing Note no. BN288.
4. Benzeval, M., Burton, J., Crossley, T., Fisher, P., Jäckle, A., Low, H. and Read, B. (2020): Understanding Society COVID-19 Survey, April Briefing Note: the economic effects', Institute for Social and Economic Research (ISER), Working Paper no. 10/2020.
5. Blundell, R., Costa Dias, M., Joyce, R., Xu, X. (2020): COVID-19 and Inequalities. Fiscal Studies, 41(2), 291-319.
6. Brunori, P., Maitino, M. L., Ravagli, L., Sciclone, N. (2020): Distant and Unequal: Lockdown and Inequalities in Italy. DISEI, Università degli Studi di Firenze.
7. Coe, R., Weidmann, B., Coleman, R., Kay, J. (2020): Impact of school closures on the attainment gap: rapid evidence assessment.
8. Clark, A. E., D'Ambrosio, C., Lepinteur, A. (2020): The Fall in Income Inequality during COVID-19 in Five European Countries (No. 565). ECINEQ, Society for the Study of Economic Inequality.
9. Crawford, R., Davenport, A., Joyce, R., Levell, P. (2020): Household spending and coronavirus, Institute for Fiscal Studies (IFS), Briefing Note no. BN279 (https://www.ifs. org.uk/public)
10. Del Boca, D., Oggero, N., Profeta, P., Rossi, M. (2020): Women's and men's work, housework and childcare, before and during COVID-19. Review of Economics of the Household, 18(4), 1001-1017.
11. Figari, F., Fiorio, C. V. (2020): Welfare resilience in the immediate aftermath of the covid-19 outbreak in Italy (No. EM6/20). EUROMOD Working Paper.
12. Gallo, G., Raitano, M. (2020): SOS incomes: Simulated effects of COVID-19 and emergency benefits on individual and household income distribution in Italy (No. 566). ECINEQ, Society for the Study of Economic Inequality.
13. Ministry of Economy and Finance – Department of Finance (2020) Nota Tematica n. 3 and n. 5: L'impatto del Covid-19 e degli interventi del Governo sulla situazione socio-economica delle famiglie italiane nei primi tre mesi della pandemia. L'impatto del Covid sul fabbisogno di liquidità delle imprese.

3.8 Learning analytics methods and applications

Open-Source Automated Test Assembly: the Challenges of Large-Sized Models

L'assemblaggio dei Test Automatizzato e a Sorgente Aperto: le Sfide dei Modelli di Grandi Dimensioni

Giada Spaccapanico Proietti

Abstract Commonly, in large-scale educational assessments, several optimal parallel test forms are assembled through automated test assembly (ATA) softwares. However, it is not rare that ATA programs are not able to produce satisfying solutions. Specifically, this work focuses on the practical concerns which arise using open-source tools and it offers a mixed algorithm based on the sequential strategies introduced in Spaccapanico P. et al. (2020) to identify the constraints which make the ATA model infeasible and to incrementally relax or further restrict the model to find a satisfying solution in feasible amount of time. An application on the TIMSS item bank shows the capabilities of the method in a real-world scenario.

Abstract *Generalmente, nelle valutazioni di apprendimento su larga scala, una moltitudine di test paralleli viene prodotta attraverso l'uso di software di assemblaggio automatizzato (ATA). Tuttavia, non sempre i programmi per l'ATA sono in grado di produrre soluzioni soddisfacenti. Nello specifico, questo lavoro si concentra su problemi pratici derivanti dall'uso di applicazioni a sorgente aperto e offre un algoritmo misto basato sulle strategie sequenziali introdotte in Spaccapanico P. et al. (2020) per identificare i vincoli che rendono il modello ATA irrisolvibile e ottenere una soluzione soddisfacente in tempi ragionevoli. Un'applicazione sui dati TIMSS dimostra l'efficacia del metodo in un caso reale.*

Key words: automated test assembly; infeasibility; open-source; large-scale assessment; psychometrics; TIMSS

Giada Spaccapanico Proietti
Department of Statistical Sciences, University of Bologna, Via delle Belle Arti, 41, 40126 Bologna, Italy, e-mail: giada.spaccapanico2@unibo.it

1 Introduction

In recent years, along with the advent of innovations in item banking systems and in cloud testing services, paper-and-pencil tests are being progressively replaced by computer-based tests. The burdensome task of choosing items by hand is making the way to more sophisticated selection techniques, as well. Among the mentioned upgrades in test forms generation, automated test assembly (ATA) is the most widely applied. In practice, by ATA, the selection of items and the subsequent test construction are automatically performed by a software.

Ideal use cases of ATA are large-scale standardized assessments since the administration in multiple sessions and locations and security concerns, such as cheating and leaking of information require to assemble several test forms. Moreover, especially for high-stake tests, it is essential that the tests follow fairness principles, i.e., they must be parallel with respect to their statistical and content-related properties [4]. In addition, they must achieve the highest of precision of ability measurement to obtain the features of validity and reliability [5]. Once specified, the mentioned requirements must be translated into a formal language through constraints or objectives of a mixed-integer linear programming (MILP) model [2]. Then, a program, called *solver*, takes the model as an input and tries to obtain the best solution available. However, despite the attractive user interfaces and usability of ATA softwares, in concrete circumstances, the large-sized models put a strain on the open-source solvers, and it is not uncommon that the desired results are not achieved.

Therefore, the aim of this article is to provide a workflow to unravel the intricacy of a non promptly solvable ATA model, adopting an approach which blends the additive and subtractive strategies introduced in [7]. The work focuses on the MAXIMIN paradigm for the assembly of parallel test forms within the item response theory (IRT) framework [8].

2 The MAXIMIN ATA Model

Within the IRT and MILP frameworks, a common example of ATA model employes the MAXIMIN objective [2]. This objective is widely used in practice, since it allows to find the tests which have the maximum Fisher test information function (TIF), i.e. the minimum expected ability estimation error. Formally, the TIF is the sum of the Fisher information functions (IIFs) of the items selected to be in the test. For example, for the unidimensional 3-parameter logistic (3PL) model [8] with dichotomous responses, the IIF of item i at ability θ, $I_i(\theta)$ is equal to $a_i^2 \frac{1-P_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta)-c_i}{1-c_i}\right]^2$, where the parameters a_i and c_i represent the discrimination and the pseudo-guessing parameters of item i, respectively. Moreover, $P_i(\theta)$ is the probability of a correct answer to item i for an examinee with ability level θ and it is equal to $c_i + (1-c_i)\frac{\exp(a_i(\theta-b_i))}{1+\exp(a_i(\theta-b_i))}$, where b_i is the difficulty of item i.

Thus, given a set of optimization variables $x_{it} \in \{0,1\}$, where $i = 1,\ldots,I$ are the indices of the items in the item bank and $t = 1,\ldots,T$ are the indices of the test forms to be assembled, the MAXIMIN ATA model takes the following form:

$$\begin{aligned}
&\text{maximize} && y && \text{(objective function)} && \text{(1a)}\\
&\text{subject to} && \sum_{i=1}^{I} I_i(\theta)x_{it} - y \geq 0, && \forall i,t, && \text{(TIF constraints)} && \text{(1b)}\\
& && \sum_{i=1}^{I} v_{itm}x_{it} \leq b_{tm}, && \forall m,t && \text{(other constraints)} && \text{(1c)}
\end{aligned}$$

where v_{itm} are the coefficients and b_{tm} is the lower bound for defining the m-th constraint for test t, such as the test length, the number of items with a certain content feature, item use, and the overlap within each of the other tests. Furthermore, by the linear inequalities (1c) we can also impose friend sets or enemy sets. If at least one combination of decision variables which satisfies the constraints (1b) and (1c) exists, the model is said to be feasible, otherwise it is infeasible.

3 The Challenges of Large-Sized Models

In large-scale assessments, in order to ensure a safe, valid and fair administration of the tests, the forms must fulfill a complex set of requirements. Subsequently, the item bank must be rich enough to comply with the imposed conditions. It entails having an ATA model with several constraints and decision variables, decreasing the chance of obtaining a satisfying solution from the software. Nevertheless, the imposed constraints may collide with each other and the inspection and identification of the conflicts are usually very intricate tasks. Furthermore, another consideration for the practitioners is the choice of the software to optimize the ATA model. For example, the open-source packages `xxIRT` [6] and `ATA.jl`[1], written in `R` and `Julia`, repectively, are available. The mentioned packages wrap open-source solvers[2] which are widely recognized to not be the best-performing MILP solvers available, especially if they are compared with their commercial alternatives, such as `CPLEX` or `Gurobi`.

In [7], the authors presented a detailed list of sources of infeasibility and size growth of ATA models, together with the formalization of two pragmatic strategies to overcome these issues, called additive and subtractive. By these strategies, the constraints are added, relaxed or deleted from the model in a way analogous to the forward and backward stepwise selection used for statistical purposes. To increase the effectiveness of the mentioned approaches in detecting conflicts between constraints and to provide the best solution given the initial infeasible test specifications,

[1] https://github.com/giadasp/ATA.jl

[2] Nonetheless, `ATA.jl` supports also commercial solvers.

our proposal is to blend the two algorithms and develop a *mixed* alternative. First of all, as in the original paper, the constraints must be sorted in order of priority: from the most to the least important. Then, some backup plans must be prepared for the constraints that can be relaxed. The mixed algorithm starts in the same way as the additive model. If the solver cannot find a solution at the end of an additive step, the subtractive algorithm is implemented. So, the last constraint added, A, is relaxed. Then, if the model is feasible, the algorithm continues with the additive mechanism, otherwise, A is relaxed until a predefined limit, and the model is solved again. If A has reached its most relaxed version and the model is still infeasible, a *repair* phase starts, and the previously added constraints, except A, are relaxed one-by-one starting from the least important, using a modified subtractive algorithm. If relaxing some further constraint B, the model is feasible, B is kept as relaxed, and the subsequent constraints are restored using the additive technique. On the other hand, if the feasibility cannot be reached before arriving to the essential model, A must be deleted since it is incompatible with the other more significant requirements. The additive algorithm continues from the latest feasible obtained model, until the next infeasibility is found. Note that, in this case, A is still imposed in its relaxed version.

By this algorithm, incompatibilities between constraints can be better detected, namely, it is clear that, constraint A is in conflict with constraint B. The following pseudo algorithm formalizes the mixed strategy.

Algorithm 1: Mixed algorithm

$accept = FALSE$; $model = \{\}$; $repair = FALSE$; $m = 1; n_m = 1\ \forall m$
while *!accept* **do**
  $model = model \cup c_{mn_m}$ (additive); $solve(model)$
  if *model is feasible* **then**
    $solution = get\_solution(model)$
    if *!repair* **then**
      $n^*_m = n_m\ \forall m$ (save current optimal state)
      $m = m + 1$ (next constraint)
    else
      (conflict resolved, $c_{m(n_m-1)}$ was in conflict with A)
      $m = A + 1$; $repair = FALSE$
  else
    $model = model \setminus c_{mn_m}$ (subtractive)
    if $n_m < N_m$(can c_m be further relaxed?) **then**
      $n_m = n_m + 1$ (relaxation)
    else
      if *!repair* **then**
        (conflict found, m is a troublesome constraint)
        $A = m$; $repair = TRUE$
      else
        $n_{m'} = n^*_{m'}\ \forall m' \neq A$ (restore to previous optimal state)
        $m = m - 1$ (previous constraint)
  if *solution is satisfying* **then** $accept = TRUE$

where the constraints and related relaxations c_{mn_m}, $m = 1,\ldots,M, n_m = 1,\ldots,N_m$ are sorted by strictness, i.e. c_{m1} is the m-th original constraint c_m and c_{mN_m} is its most relaxed version that may correspond to deletion. For example, the lower and upper bound of a constraint can be decreased or increased to enlarge or narrow the space of possible values from a categorical or quantitative variable. In extreme cases, if a constraint is not absolutely necessary for the final testing purpose, it can be fully relaxed, i.e., deleted.

4 Application to TIMSS Data

An ATA model of the type (1a) is solved using the mixed strategy and the item bank coming from the Trends in International Mathematics and Science Study (TIMSS) international database [9]. TIMSS is a large-scale standardized student assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA). The project evaluates the skills in mathematics and science of pupils coming from 39 countries every four years. In particular, 325 science items have been calibrated following a unidimensional 3PL model using the dichotomous response matrices of the 2015 and 2019 survey data on Italian 8th grade students. Polytomous items, items with a derived score and items with a calibrated discrimination or difficulty out of the range [-3.0, 3.0] and with a guessing parameter higher or equal to 0.5, are excluded. The items are grouped in friend sets and they are categorized by content and cognitive domain, item type and cycle in which they have been created. For the ATA model, the specifications for the assembly have been set following the TIMSS 2019 objectives described in [9]. Specifically, we assembled 14 test forms with 45 items, and we maximized the TIFs at $\theta = 0$. It follows, in round parenthesis the level of priority of each imposed constraint, where I is the highest and VI is the lowest. The first class of constraints is enforced for security purposes: the items must be in at most 2 different tests (I), the majority of items in a test must be created in the last 2 assessments (II; cycles 6 and 7), and maximum overlap between test forms is 15 items (VI). Moreover, for ensuring the content validity, test forms must have the same distribution of content and cognitive domains (III), between 30 and 40 multiple choice items (IV), and at least 3 items for each combination of content/cognitive domain (V). The backup plan for the item use is to relax the upper bound to 3 for specific problematic items. For the other constraints, we allow a relaxation of the lower and upper bounds of -1 and $+1$, respectively. We choose the software `ATA.jl`[3] and the `Cbc` solver with a time limit of 500 seconds to perform each step of the mixed algorithm.

Thus, the ATA process starts with the empty model and the item use specification. Progressively, the requirements with priorities I, II, and III are added. The bounds imposed about physics items creates infeasibility also in their relaxed version, so the repair phase starts. The feasibility is restored when the constraint on earth science is relaxed, as well. Then, the constraints on the cognitive domains are added to the model. The requirement on reasoning items creates infeasibility, so it is relaxed. Unfortunately, this backup plan does not fix the infeasibility, so another repair phase starts. The relaxation of the earlier imposed constraints do not help, so the item use for reasoning items is increased to 3. At the end, the constraints about about the combinations of cognitive and content domain (IV) and multiple choice items (V) are inserted into the model. In particular, the items in earth science are not enough to provide at least three items in applying and reasoning, so the lower bounds for the related constraints are decreased to 2. The maximum overlap (VI) is not added to the model since the latest solution already satisfies the requirement.

[3] Code, detailed specifications and results are available at https://github.com/giadasp/SIS2021.

The results obtained in each step of the algorithm reveals that, given all the previously imposed specifications, the requirement over the content domain "physics" disagrees with the bounds enforced on the earth science items. Moreover, the deficiency of reasoning items required to increase their maximum item use, possibly raising the items circulation and consequently compromising the security of the test. Furthermore, the relaxation on the item use contributed to a 11.5% increment of the minimum TIF across the tests revealing that this deficiency caused a worsening in the measurement precision of the tests.

5 Conclusion

In this article, we proposed a mixed strategy to deal with large-sized ATA model in an open-source software framework which blends the two pragmatic algorithm provided by [7]. The mentioned approach should give a deeper insight into the conflicts between requirements and offers, as a final solution, the set of tests which requires the softest relaxation of the constraints. By means of a real data example on the 2015/2019 TIMSS item bank on science, the effectiveness of the mixed algorithm has been tested. However, this study has some limitations that should be improved in the future, starting from the applications to other item banks and the extension of the approach to other ATA models, such as non-parallel test assembly or other commonly used objective functions.

References

1. Wightman, L.F.: Practical Issues in Computerized Test Assembly. *Appl. Psychol. Meas.*, **22**, 292–302 (1998) doi:10.1177/01466216980223009.
2. Van der Linden, W.J.: Linear Models for Optimal Test Design; Springer: New York, NY, USA (2005)
3. Report INVALSI CBT 2018–Aspetti Metodologici. Available at https://invalsi-areaprove.cineca.it/docs/2019/Parte_I_capitolo_2_aspetti_metodologici_CBT_2018.pdf.
4. Samejima, F.: Weakly Parallel Tests in Latent Trait Theory with some Criticisms of Classical Test Theory. Psychometrika, **42**, 193–198 (1997). doi:10.1007/BF02294048
5. American Educational Research Association; American Psychological Association; National Council on Measurement Education. Standards for Educational and Psychological Testing; American Educational Research Association: Washington, DC, USA (2014)
6. Luo, X.: xxIRT: Item Response Theory and Computer-Based Testing in R (2019) Available at https://CRAN.R-project.org/package=xxIRT.
7. Spaccapanico Proietti, G.; Matteucci, M.; Mignani, S.: Automated Test Assembly for Large-Scale Standardized Assessments: Practical Issues and Possible Solutions. Psych, **2**, 315–337 (2020) doi:10.3390/psych2040024
8. Hambleton, R.K.; Swaminathan, H.; Rogers, J.H.: Fundamentals of Item Response Theory; Sage: Newbury Park, CA, USA, **2** (1991)
9. Martin, M.O.; von Davier, M.; Mullis.: Methods and Procedures: TIMSS 2019 Technical Report. Retrieved from Boston College, Available at https://timssandpirls.bc.edu/timss2019/methods. I. V. S. (Eds.) (2020).

How Much Tutoring Activities May Improve Academic Careers of At-Risk Students? An Evaluation Study

Valutare l'Efficacia delle Attività di Tutorato per gli Studenti a Rischio Dropout

Marta Cannistra, Tommaso Agasisti, Anna Maria Paganoni and Chiara Masci

Abstract Through the development of models able to predict who will be the students with the highest probability of dropping out, our aim is to understand whether and how the tutoring courses already active will benefit students in difficulty. Through an experimental setting and the adoption of different statistical methodologies, such as Regression Discontinuity Design and the Propensity Score Matching, we answer the following research questions: (a) are tutoring activities attractive for at-risk students? (b) are tutoring activities effective for at-risk students? (c) how much a nudging communication may influence academic careers of at-risk students? Generally, emerges that tutoring activities are more effective than attractive, and a nudging communication is not enough to improve academic career of at-risk students.

Abstract *Attraverso lo sviluppo di modelli statistici in grado di prevedere quali saranno gli studenti con la maggiore probabilità di abbandono scolastico, il nostro obiettivo è capire se e come i corsi di tutoraggio già attivi andranno a beneficio degli studenti in difficoltà. Grazie a un contesto sperimentale e all'adozione di diverse metodologie statistiche quali Regression Discontinuity Design e Propensity Score Matching, rispondiamo alle seguenti domande di ricerca: (a) le attività di tutoraggio sono attraenti per gli studenti a rischio? (b) le attività di tutoraggio sono efficaci per gli studenti a rischio? (c) quanto una comunicazione sollecita può influenzare le carriere accademiche degli studenti a rischio? In generale, emerge che le attività di tutorato siano più efficaci che attrattive, e come una comunicazione "soft" non sia abbastanza per migliorare le carriere degli studenti a rischio.*

Key words: early warning system, remedial intervention, propensity score matching

---

Cannistra, M., Politecnico di Milano; e-mail: marta.cannistra@polimi.it
Agasisti, T., Politecnico di Milano; e-mail: tommaso.agasisti@polimi.it
Paganoni, A., Politecnico di Milano; e-mail: anna.paganoni@polimi.it
Masci, C., Politecnico di Milano; e-mail: chiaramasci@polimi.it

1 Introduction

The discussion about the use of analytics for predicting students' performance and accompany remedial programs stem from the traditional attention to the serious problem of dropout. On one hand, empirical studies define and estimate dropout rates with ever-increasing precision and examine the factors associated with dropout of individual students. On the other hand, papers, articles and reports describe the efforts and interventions to prevent students from leaving schools and universities [3, 4, 2, 6]. In fact, simply identifying at-risk students does not alleviate the risk these students face. To make EWSs impactful to prevent students from dropping out, educational institutions must tailor intervention and prevention efforts based on the data [7]. Indeed, we can consider the two research streams as sequential: the outputs produced by the analyses of dropouts functioning as the key information source when setting the remedial interventions. We define this two-steps process as Early Warning System (EWS). Commonly, the use of EWS is related to diverse fields of applications where detection is important – as, for example, military attacks, conflict prevention, economical/banking crisis, environment disasters/hazards, human and animal epidemics, and so on. In the educational domain, an EWS consists of a set of procedures and instruments for (i) early detection of indicators of students at risk of dropping out and, in a second moment, (ii) the implementation of appropriate interventions to make them stay in school [5]. Early warning indicators are used for early intervention with students to help them get back on track and meet major educational milestones. Consequently, the second step of EWS needs to take into consideration that at-risk students are not a homogenous group, therefore policy makers need to design specific interventions to efficiently target them but, most important, evaluate them. In this vein, it must be emphasized that identifying students at risk of dropping out by using an EWS is only the first step in addressing the issue of school dropout. Evidence-based education specifically supports decision-making processes [8]. The scientific revolution in education will only take hold and produce its desired impacts if research in fact begins to focus on replicable programs and practices central to education policy and teaching [9]. In this vein, the contributions given by experimental settings to evaluate remedial interventions in education is crucial. Only understanding the main drivers of academic success (or retention) is possible to help and support students stay on-track. The proposed research aims at investigating the effectiveness of a tutoring program in a technical university for at-risk students. The analysis is computed after the prediction of dropout probabilities of freshman students, with Machine Learning techniques, which produces the dropout probabilities for every first-year student. Then the university contacts the most at-risk ones with a personalized communication which warns out about their academic career, suggesting to enrol to tutoring activities. Then, through a quasi-experimental approach we evaluate the effect of such activities on at-risk students. To set out our argument, the paper is structured as follows: Section 2 presents the experimental setting, then in Section 3 data and methodology are explained, which brings to Section 4 dedicated to the main results, while the last section resumes the main findings together with some final considerations.

2 Experimental Design

If we consider the information in our possession, about 500 students leave our university every year about half of these drop out before the end of the first year. Through the development of statistical models and analysing the data of the last years on the careers of the students of the university, we are able to predict the dropout risk of students with high accuracy. It is worth to note, that the prediction is made mainly for first-year students at the end of their first semester at university. The ”live” prediction is made and for each student enrolled in the first year (the population of interest), we get a dropout probability, from 0% (no risk of dropout) to 100% (sure to dropout). Then, it is the task of the research group to understand if and how the already active tutoring courses will benefit students in difficulty. This part will be carried out by adopting an experimental methodology, to answer to the following research question: Tutoring is effective in enhancing students' academic career at risk of dropout? Students who are associated with a dropout risk of more than 80% are contacted by university to fill in a short questionnaire (six questions) regarding the perception of their academic career. In addition, to those with a risk of dropout of more than 90%, a short message will also be attached. The message will aim to remind the student the possibility of enrolling in different remedial education interventions, mainly tutoring activities offered by the university. In this experimental design, we adopt a Propensity Score Matching methodology to assess the effectiveness of being enrolled to tutoring activities for at-risk students.

3 Data and Methods

3.1 Data

The data used in our analysis comes from the Information Technology (IT) system of the university, which collects both dynamic and static data about enrolled students. The former ones are the so-called ”digital prints” left in correspondence to some key administrative facts, such as register at exams' sessions, accept or retake grades or pay university's fees. Static data comprises all the information that administrative office registers at the moment of enrolment, such as citizenship, gender or date/place of birth, previous school performance or the university admission test score.

3.2 Methodology

The main methodology adopted to understand whether tutoring activities are effective in improving students' academic performance is the Propensity Score Matching. This method allows to set a quasi-experimental approach to observe the effect

of the treatment status (i.e. the enrolment to tutoring activities) on the academic performance (i.e. GPA and credits obtained). Indeed, randomization accounts for both observed and unobserved characteristics of study participants, and the analyst can be certain that any differences between the two groups on outcomes is the result of the intervention [1]. In the social sciences, such randomization is often logistically impossible and, even if it is possible, it is often ethically questionable. Quasi-experimental methods such as those used in this study provide researchers with analytic tools to mimic randomization, accounting for both observable and unobservable differences between treatment (tutoring's enrolment) and control groups. Specifically, propensity scores are calculated using logistic regression (see eq. 1), wherein the outcome is the log-odds that a student will participate to tutoring activities represents an intercept term (β_0), and β_1 is a vector of regression coefficients corresponding to the individual student characteristics.

$$\log(p/p-1) = \beta_0 + \beta_1 * X_1 \tag{1}$$

Each student's predicted probability (p), or propensity, to enrol to tutoring that resulted from this equation forms the basis for matching tutoring participants with non-participants. Once students are matched according to mostly similar propensity scores, a final linear regression is computed to capture the effects of participating to tutoring on the academic careers (see eq. 2).

$$Y_1 = \alpha_0 + \alpha_1 * Z_1 + \varepsilon_1 \tag{2}$$

In particular, Y_1 represents the outcomes of the academic career at the end of the second semester: GPA, the weighted average grade of the passed exams, and the credits obtained. While Z_1 represents the student's level covariates, ranging from demographic information (such as age or gender) to previous studies (such high school track) and residency (whether the student lives alone or with family).

4 Results

Thanks to the adopted methodology, the Propensity Score Matching, we are able to catch the effect of being enrolled to tutoring activities on students' academic performance for at-risk careers (more than 80% of dropout probability). Once having matched students with similar features, they are paired according to if they receive the treatment (treatment is frequency at tutoring activities). The results of the linear regressions showed in Table 1 mainly highlight how important is the frequency to tutoring in improving students' performance. In fact, both GPA and credit are influenced by the dummy variable $freq\_Tutoring$ which takes 1 whether the student follows at least one lecture of tutoring. The final grade obtained at high school represents another key feature positive associated to higher performance. Generally, tutoring activities allows at-risk student having better results to their counterpart, who do not frequent them.

Table 1 Linear Regression models to evaluate the effectiveness of the frequency at tutoring activities, after matching similar students with propensity score matching.

	Dependent variable:	
	GPA II semester	Credits II semester
	(1)	(2)
freq_Tutoring	2.690*** (0.706)	5.801*** (1.569)
stud_birth_yyyy	0.350 (5.015)	30.349*** (11.147)
stud_genderM	−0.340 (0.868)	−2.207 (1.930)
stud_citizenship	2.976* (1.679)	−0.521 (3.732)
stud_admission_score	4.744 (2.909)	9.415 (6.467)
stud_career_degree_changed	−0.761 (0.942)	−0.513 (2.093)
highschool_grade	8.104*** (1.590)	25.189*** (3.534)
previousStudiesOthers	1.046 (2.053)	−4.349 (4.562)
previousStudiesScientifica	0.429 (1.474)	3.116 (3.277)
previousStudiesStraniera	−0.981 (4.227)	−12.318 (9.396)
previousStudiesTecnica	−1.785 (1.652)	−1.179 (3.671)
income_bracket_normalized_on4fascia bassa	0.248 (1.285)	4.261 (2.855)
income_bracket_normalized_on4fascia media	−0.239 (0.984)	1.537 (2.188)
income_bracket_normalized_on4LS	−1.708* (0.884)	−4.010** (1.965)
originsForeigner	−0.120 (4.223)	6.450 (9.386)
originsMilanese	−0.999 (0.952)	−2.882 (2.116)
originsOff-site student	1.771 (2.097)	−8.703* (4.661)
Constant	8.533 (5.285)	−22.137* (11.747)
Observations	448	448
R^2	0.156	0.245
Adjusted R^2	0.123	0.215
Residual Std. Error (df = 430)	7.378	16.398
F Statistic (df = 17; 430)	4.671***	8.210***
Note:		*p<0.1; **p<0.05; ***p<0.01

5 Conclusions

The importance of detecting and monitoring at-risk students is only the first step to help and support them in staying on track, not leaving university. In fact, increasing retention rates at higher education institutions represent a key today's challenge. In doing this, we experiment and analyse the effect of tutoring activities freely offered by the university to support students with learning difficulties. Thanks to a quasi-experimental approach, the Propensity Score Matching, we see how the enrolment to these activities are effective in improving students' academic performance, in terms of GPA and credits. This results is profoundly important since around one third of students leave their academic career. Hence, putting in place effective remedial interventions represent a key action in order to stop this phenomenon.

References

[1] J. D. Angrist and J.-S. Pischke. *Mastering metrics: The path from cause to effect*. Princeton Univ. Press, 2014.

[2] R. Balfanz, L. Herzog, and D. J. Mac Iver. Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychol.*, 42(4):223–235, 2007.

[3] A. L. Caison. Determinants of systematic retention: Implications for improving retention practice in higher education. *J. of College Student retent.: Res., Theory & Practice*, 6(4):425–441, 2005.

[4] E. Ghignoni. Family background and university dropouts during the crisis: the case of italy. *High. Education*, 73(1):127–151, 2017.

[5] J. B. Heppen and S. B. Therriault. Developing early warning system to identify potential high school dropouts. issue brief. *Natl. High School Cent.*, 2008.

[6] M. A. Mac Iver. The challenge of improving urban high school graduation outcomes: Findings from a randomized study of dropout prevention efforts. *J. of Education for Students Placed at Risk (JESPAR)*, 16(3):167–184, 2011.

[7] L. Pinkus. Using early-warning data to improve graduation rates: Closing cracks in the education system. *Washington, DC: Alliance for Excell. Education*, 2008.

[8] R. Pring and G. Thomas. *Evidence-based practice in education*. McGraw-Hill Education (UK), 2004.

[9] R. E. Slavin. Evidence-based education policies: Transforming educational practice and research. *Educational res.*, 31(7):15–21, 2002.

Composite–based Segmentation Trees to Model Learners' Performance

Alberi di segmentazione composite–based per la stima dei risultati di apprendimento

Cristina Davino and Giuseppe Lamberti

Abstract In this paper, we explore whether the performance of students attending Massive Open Online Courses and its main drivers, learning and engagement, can be differentiated according to external stratification variables represented by socio–demographic characteristics and type of course. Performance and related main drivers of students attending platform FedericaX, the "Federica WebLearning" center of the University of Naples Federico II, are used. Their relationship are analyzed using partial least squares structural equation modeling, while the potential effect of the stratification variables are explored through a Pathmox analysis.

Abstract *L'obiettivo del lavoro è esplorare la relazione tra la performance degli studenti dei MOOC (Massive Open Online Courses) e i suoi principali driver, l'apprendimento ed il coinvolgimento, evidenziando possibili differenze in base a variabili di stratificazione esterne quali le caratteristiche socio–demografiche e il tipo di corso. Il lavoro è basato sull'analisi dei dati relativi alla partecipazione degli studenti che frequentano due MOOC disponibili sulla piattaforma FedericaX dell'Università di Napoli Federico II. La metodologia di riferimento è rappresentata dai modelli ad equazioni strutturali di tipo composite–based e da un approccio basato sulla segmentazione ad alberi per confrontare tali modelli osservati su partizioni del campione osservato.*

Key words: massive open on-line courses, performance, partial least squares structural equation modeling, heterogeneity, pathmox analysis

---

Cristina Davino
Department of Economics and Statistics, University of Naples Federico II, Naples, Italy, e-mail: cristina.davino@unina.it

Giuseppe Lamberti
Department of Business, Universitat Autonoma de Barcelona, Barcelona, Spain e-mail: giuseppe.lamberti@uab.cat

1 Introduction

Predicting students' performance is one of the main challenges in learning analytics. This aim becomes even more critical in the case of the analysis of data collected through Massive Open Online Courses (MOOCs). Due to the peculiarity of being massive, MOOCs are attended by very heterogeneous students with respect to socio–demographic characteristics as well as cultural and educational backgrounds. In this framework, learning analytics focusing on learners' behavioral patterns aims to study the effects on student's performance and cannot disregard the students' heterogeneity who decide to attend the MOOC.

This contribution extends the proposal of Carannante *et al.* [2] where a structural equation model, in the framework of the composite–based approach [11] [6] [7], is proposed to measure the main factors affecting students' performance in MOOCs. Exploiting the conceptualization of performance and its main drivers (learning and engagement), this paper carries out a deeper analysis to assess whether the structural model shows significant differences according to the characteristics of the students or considering different types of courses. In essence, we aim to handle a possible heterogeneity in modeling students'performance according to their learning and engagement behavior.

Heterogeneity is a major issue in statistical modeling since a model could be valid for the whole population but not for possible partitions. In particular, in the framework of composite–based modeling, also known as Partial Least Squares Structural Equation Modeling (PLS–SEM), several contributes have been proposed to handle both observed [4] and unobserved heterogeneity [5]. We exploit here the former approach analysing if the structure of relationships in a PLS–SEM differs according a set of stratifying or segmentation variables not included into the model. In particular, the Pathmox analysis [8] is adopted. It applies the principles of binary segmentation to produce a tree with different models in each node according to the set of stratifying variables.

The analysis refers to two courses in Political Science offered on the Platform FedericaX, the EdX MOOCs platform of the "Federica WebLearning" Centre at University of Naples Federico II, and it is carried out on the tracking logs of users' actions.

2 Data Description and Global Model

It is a matter of fact that the digitalization of education allows tracking all the students' actions from event logs (also called clickstream), participation to forums and discussions, and test sessions.

Data presented in this paper refer to 3578 students who attended two courses in Political Science on the FedericaX platform. Each course was offered in two versions: an instructor–paced version and a self–paced version. The instructor–paced is strictly scheduled, with specific dates for assignments, course materials, exams,

and a deadline for learners to complete the course and get a certification. Usually, this modality is integrated into an in–site course delivered in blended mode. For the self–paced version, all of the course materials are available as soon as the course starts, assignments and exams do not have due dates, and therefore a learner can progress through the course at its own speed and pass grade in the course, even without completing all of the course materials.

Learning and engagement, the two considered drivers of students' performance, have been measured by a set of 14 indicators. In particular *learning* considers the number of actions undertaken in order to acquire knowledge. In contrast, *engagement* represents the emotional perspective, i.e., quantify the way the user approaches the course. Both of them are complex constructs that requires second–order constructs to be measured: *frequency–based activities*, *time–based activities*, *interaction* for the *learning* construct and *regularity* and *no procrastination* for the *engagement* one. The outcome dimension, *performance*, is measured through the rate of correct answers to a set of questions provided by the instructor. For a complete description of the indicators, please refer to the original paper of Carannante *et al.* [2]. In this paper we consider only a subset of the original data: students registered to the course but not active were excluded from the analysis, and some indicators (average backward, average time spent on video, first delay and rate of return) have been discarded due to low reliability.

The relationship among *performance*, *engagement*, and *learning* is explored considering the structural model provided in Figure 1 (left–hand side) and validated in [2]. The reference framework is represented by PLS–SEM, a multivariate approach for latent variables estimation. PLS–SEM allows multiple blocks of observed variables (or indicators) to be analyzed, with each block playing the role of a latent variable; a linear relationship is assumed between blocks. Two models are computed: the measurement (or outer) model, that relates observed variables to latent variables, and the structural (or inner) model, that calculates the strength and direction of relationships among the latent variables.

The proposed model considers reflective LVs, and therefore unidimensionality and internal consistency were checked revealing satisfactory results. The global model results (estimated on the whole sample) are consistent with the results in [2]. For the sake of brevity, results related to the measurement part of the model are not presented. Estimates of the structural model, reported in Figure 1 (left–hand side), confirm that *learning* is globally the main driver for performance (coefficient equal to 0.563, p–value$<$0.001), while *engagement* has a lower impact (coefficient equal to 0.261, p–value$<$0.001). The R^2 is 0.648, which ensures an adequate predictive power of the model.

3 Modeling Heterogeneity

The model presented in Section 2 can be refined by analyzing any differences among subgroups in the considered sample. At this regard, the Pathmox analysis [8] allows

to identify different segments of the sample holding a different set of relationships among constructs. The method applies the principles of binary segmentation to produce a tree with different models in each of the obtained nodes. The algorithm starts by fitting a global model on all the data, thereby defining the root of the tree, and then identifies the models with the greatest differences in each child node in a procedure that is repeated iteratively. The candidate splitting variables are represented by stratification or segmentation variables, not included into the model. As in each binary segmentation procedure [1], these variables are previously transformed in all possible dichotomous variables. The available data is recursively partitioned to identify the iterations whose segmentation variable yields the most significant difference after comparing two PLS–SEM models of child nodes. To this end, Pathmox uses a statistical test based on Fisher's F–test for equality in regression models [10] adapted to compare two PLS–SEM models. The approach results in a valid alternative when the context of the research is explorative. It is useful for detecting heterogeneity in case of several segmentation variables, and it is difficult to determine which of these variables are responsible for differences.

The segmentation variables considered in this paper as possible sources of heterogeneity are: the course's version offered on the FedericaX platform (instructor–paced = 26.91%; self–paced = 73.09%), gender (female = 23.87%, male = 40.55%; prefer not to say = 35.58%), age (14–18 years = 4.22%, 19–25 years = 21.32%, 26–32 years = 48.69%; 33–45 years = 15.57%, 46–60 years = 6.96%, >60 years = 3.24%), and country (Africa = 7.57%, Americas = 26.10%, Asia = 31.11%, Europe = 21.66%, Oceania = 13.56%).

The Pathmox results are presented in Figure 1 (right–hand side) and Table 1. The root node corresponds to the global model estimated on the whole sample (n=3578), while the four leaf nodes identify four different Local Models (LMs), respectively labeled LM4, LM5, LM6, and LM7. Pathmox produced a total number of four nodes

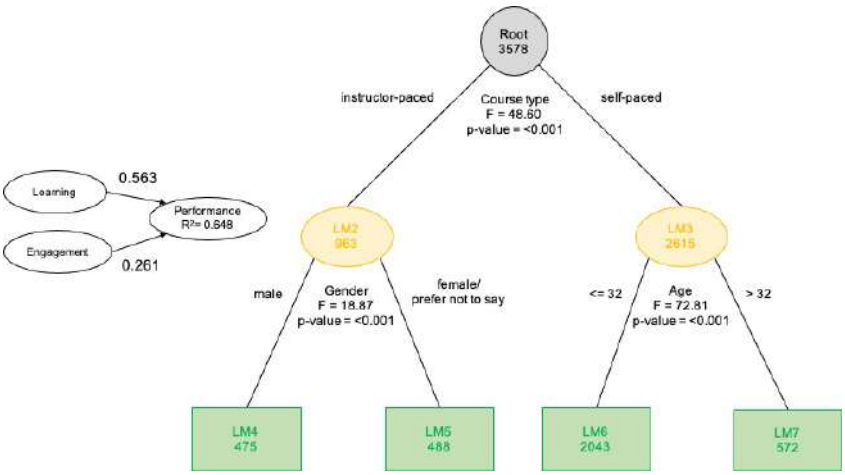


Fig. 1 Global model estimation (left–hand side) and Pathmox segmentation tree (right–hand side)

as we limited the tree depth to two (i.e., four terminal nodes as maximum) for the interpretation.

The algorithm selects *course type* as the variable with the greatest discriminant power ($p<0.001$), distinguishing between students that attended *instructor–paced* courses and students that attended *self–paced* courses. The students involved in the *instructor–paced* courses are further differentiated by the variable *gender* ($p<0.001$), defining two groups, *male* (LM4) against *female and others* (LM5). The students involved in *self–paced* courses are differentiated instead by the variable *age* ($p<0.001$), classifying the students as younger than 32 (LM6) and older than 32 (LM7). The PLS–SEM models associated with the four groups are presented in Table 1. Results show that, unlike evidence obtained for the whole sample, *engagement* has not a significant impact on the *male* students involved in *instructor–paced* courses (LM4), while for the *female and not declared gender* (LM5), *engagement* and *learning* are both crucial in defining their *performance*. Moreover, in this model, the effect difference between *learning* and *engagement* is lower than the global model (0.069 vs. 0.302). Younger students involved in *self–paced* courses (LM6) present effects similar to the global model. Finally, the older than 32 (LM7) students involved in *self–paced* courses define their *performance* mainly by *engagement* rather than by *learning*.

Summarizing, Pathmox has identified *course type, gender*, and *age* as the three variables with the highest discriminant power. It reveals that the global model masks different behaviors according to the detected sources of heterogeneity. The identified subgroups of students present difference in how they define their *performance*, particularly for the *male instructor–based* courses where *performance* depends only on *learning*. In contrast, students *older than 32 attending self–based* courses associate the *performance* principally to the *engagement*.

Table 1 Local model results

	Learning	**Engagement**	R^2	**N**
Global model	0.563	0.261	0.648	3578
LM 4: Male students involved in the instructor–paced course	0.927	-0.082^{NS}	0.728	475
LM 5: Female students involved in the instructor–paced course	0.444	0.375	0.644	488
LM 6: Younger students involves in self–based course	0.562	0.276	0.671	2043
LM 7: Older students involves in self–based course	0.312	0.527	0.651	572

^{NS} non–significant

Further developments of the study will regard the use of multi–group non parametric tests to evaluate between–segment differences in path coefficients [9]. Moreover, an exploration of the impact of learning and engagement on the whole distribution of performance could highlight different effects in low, middle, or high performances [3].

References

1. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. California: Chapman & Hall (1984)
2. Carannante, M., Davino, C., Vistocco, D.: Modelling students' performance in MOOCs: a multivariate approach. Stud. High. Educ. (2020)
3. Davino, C., and Vinzi, V.E.: Quantile composite-based path modeling. Adv Data Anal Classif **10**, 491–520 (2016)
4. Chin, W.W., Dibbern,J.: An Introduction to a Permutation Based Procedure for Multi-Group PLS Analysis: Results of Tests of Differences on Simulated Data and a Cross Cultural Analysis of the Sourcing of Information System Services Between Germany and the USA. In: Esposito Vinzi V., Chin W., Henseler J., Wang H. (eds) Handbook of Partial Least Squares. Springer Handbooks of Computational Statistics. Berlin Heidelberg: Springer (2010)
5. Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., Tenenhaus, M.: REBUS-PLS: A response based procedure for detecting unit segments in PLS path modelling. Appl Stoch Models Bus Ind **28**(5) (2008)
6. Esposito Vinzi, V., Chin, W.W. , Henseler, J. , Wang, H.: Handbook of Partial Least Squares: Concepts, Methods and Applications. Berlin Heidelberg: Springer (2010)
7. Hair Jr, J. F., Hult, G. T. M., Ringle, C., Sarstedt, M.: A primer on partial least squares structural equation modeling (PLS-SEM). Sage publications (2016)
8. Lamberti, G., Aluja, T.B., Sanchez, G.: The pathmox approach for PLS path modeling segmentation. Appl Stoch Models Bus Ind. **32**, 453–468 (2016)
9. Lamberti, G.: Hybrid multigroup partial least squares structural equation modelling: an application to bank employee satisfaction and loyalty. Qual. Quant. (2021).
10. Lebart, L., Morineau, A., and Féenelon, J.P.: Traitement des donneés statistiques. Paris: Dunod (1979)
11. Wold, H.: Partial least squares. In S. Kotz e N. Johnson. (Eds.) Encyclopedia of Statistical Sciences. John Wiley & Sons (1985)

Test-taking Effort in INVALSI Assessments
L'Impegno nello Svolgimento delle Prove INVALSI

Chiara Sacco

Abstract A topic of great interest is the students' test-taking behaviour in large scale assessment, and, particularly, the estimation of the students' engagement level and of the impact of unmotivated test-taking on test performance. The present study pursued two goals in order to better understand the students' behaviour during the INVALSI Math tests. In a first step, the identification of item responses that are not reliable indicators of test-takers competence level has been performed computing a non-effortful indicator exploiting item response time. In a second step, the effects of student and item characteristics on disengagement have been investigated using explanatory item response models.

Abstract *Un argomento di grande interesse è il comportamento degli studenti nei test di valutazione su larga scala e, in particolare, la stima del livello di impegno degli studenti e dell'impatto della scarsa motivazione sul punteggio. Per comprendere meglio il comportamento degli studenti durante le prove INVALSI di Matematica, lo studio si propone di raggiungere due obiettivi. Dapprima, al fine di identificare le risposte ai singoli item che non sono degli indicatori affidabili del livello di competenza degli studenti, è stato stimato un indicatore di non-effort basato sul tempo di risposta. Successivamente, attraverso l'uso di explanatory item response models sono stati studiati gli effetti delle caratteristiche degli studenti e degli item sulla motivazione degli studenti.*

Key words: test-taking disengagement, response time thresholds, low stakes assessment, student effects, item effects

1 Introduction

[1] Chiara Sacco, INVALSI; chiara.sacco@invalsi.it

In the last years, there has been a growing interest in the study of the impact of student's disengagement on test performance and on the validity of the resulting test score in the framework of large scale assessment. National and international large-scale assessments are widely used to assess and to monitor students' competences and they are also used as a benchmark from policymakers for planning educational reforms. The validity of the test score depends in part on the assumption that the students are motivated and behave with effort during the test event [3, 9]. These tests often are low stakes, which means that the tests have no personal consequences for students, yet the scores are used to compare performance across schools and to make inferences about student learning and educational quality. Several researchers showed that low effort tends to bias student test score downward [6, 15]. It has been widely demonstrated that test-taking motivation is correlated with students' performance and this suggests that performance depends not only on the student ability and knowledge, but also on motivational and emotional aspects. [1] assessed the negative impact of boredom on test-taking efforts and, based on the framework of the expectancy-value theory, demonstrated that the engagement is positively associated to the student expectancy of being able of solving an item. [4] showed how disengaged responses can occur for several reasons at the student level and item level. Analysing the differences in student engagement and the factors that affect the test-taking motivation could help to shed light to understand which test scores truly reflect individual differences in terms of competence.

The diffusion of computer-based tests has opened the way to new measures of student effort: thus far, it has been shown that the response time at item-level is a good indicator of student engagement [10, 11].

The aim of this study is to identify the non-effortful behaviour during the INVALSI test, using the time spent by each student to respond to each item. We first addressed the question of whether disengaged responses occurs during INVALSI Math tests, estimating the effort indicator for each response and examining the prevalence by grade. Secondly, the effects of student and item characteristics on disengagement have been investigated using explanatory item response models.

2 Indicators of Test-taking Engagement

Several methods have been proposed in the literature to assess the test takers engagement. The time spent on a given item is a good indicator of the behaviour engaged by the test-taker during the test. It is reasonable to suppose that very rapid responses are indicative of a non-effortful behaviour of the test-taker [8, 14]. The assumption under this approach is that an item engaged in a solution behaviour fashion implicates taking at least a certain minimum amount of time to read fully and understand the test instructions, process the content and finally give a response. In this framework, a non-effortful behaviour means taking less time: this behaviour is termed *rapid-guessing behaviour* when the test-taker chooses a response too rapidly, and *rapid omit behaviour* when the test-taker viewed the item but left it

quickly without answering [13]. On the other hand, when the test-taker engages the item in an effortful manner, the behaviour is termed *solution behaviour*.

Several methods have been applied in the literature for the identification of the response time threshold for each item to distinguish between effortful and non-effortful responses. In the present study, the item-specific response-time thresholds are identified using the normative threshold approach.

3 Methods

3.1 Sample

National Institute for the Evaluation of the System of Education and Training (INVALSI) conducts every year large-scale survey assessments in Italy to monitor students' skills. In this work, the Mathematics standardized tests administered by INVALSI at the 8th (last year of the lower secondary school) and 10th (second year of the higher secondary school) grade students in the school year 2017-2018 have been exploited. In addition to the student answer, for each item and each student we used the INVALSI data collected during the computer-based tests: the response time (i.e. the total time spent on the item by the student) and the number of attempts (i.e. the number of times that the item has been viewed by the student).

A data cleaning procedure has been applied to remove students with multiple accesses to the test platform and remove the students with at least one item not reached. Thus, the exploited sample for the identification of the non-effortful responses is composed of 503063 out of 508033 students for grade 8 and of 454162 out of 455254 students for grade 10. For investigating the factors that affect the disengagement in INVALSI test, the analysis has been performed on the representative sample of randomly selected classes where tests are administered in presence of an external examiner who supervises and controls the test administration procedure. The representative sample is composed of 26828 and 39152 students for grade 8 and 10, respectively.

3.2 INVALSI Test Design

The INVALSI test design assumed 90 minutes of testing time for the mathematics assessment of students in grade 8 and 10. The platform is designed in order to allow the test takers to: answer items in any order they chose; flag items for possible later review; review and possibly change answers any time they want; skip items without an answer. The test is composed of multiple test booklets that contain different sets of items. Each student was administered a single booklet, whereas each

item can be contained in different booklets, thus the same item can be administered in different positions. A total of 206 and 143 items were administrated during the INVALSI mathematics tests of grade 8 and 10, respectively. The administrated items can be classified into two macro-categories: close-ended items and open-ended items.

3.3 Effortful Indicator

According to the literature, we exploited item-specific response time thresholds to distinguish between non-effortful and effortful responses. Since the design of the INVALSI tests implicates that the same item can be observed at different positions, and it is well known that the test fatigue could lead to observe a larger threshold [5], the normative threshold method for the identification of the rapid guessing and the rapid omit responses has been computed for each item within each booklet. This is an adapted version of [13] normative threshold method for the INVALSI test design.

Let non-effortful behaviour indicator for the response of the student *i* at item *j* in booklet *b* be equal to one if the response of the student *i* at item *j* in booklet *b* is classified as rapid guessing or rapid omit and zero otherwise.

Validation of the indicator has been computed by checking three different points: the item responses classified as non-effortful should be correct at a rate lower than those classified as solution behaviour and close to the rate expected from the random responding [13]. The number of actions should be lower than those classified as solution behaviour [7]. In this study as a proxy of the number of actions, the number of attempts has been used.

3.4 Data Analysis

To test the effects of the student and item characteristics on the non-effortful behaviour we applied explanatory item response models, using the generalized linear mixed modelling (GLMM) framework [2, 5]. The model explains the logit for the probability of having a non-effortful behaviour for student i and item j with the effects of M student covariates and H item covariates. Differences in test-taking engagement were explained by the following student-level variables: gender, late-enrolled status (i.e. students enrolled at least one year after the age of 6 or repeated one or more years), primary language spoken at home (that assume values 1 if the spoken language is different from the test language), parental educational attainment. Furthermore, to investigate the item effect on test-taking engagement the following set of item-level variables were included in the model: item position, item difficulty and item dimension (that is categorized in the following classes: knowing, problem-solving, arguing). Four model were estimated to address the second research aim: Model 0 is an empty model, Model 1 includes the student characteristics as predictors, Model 2 include only the item characteristics and

Model 3 is the full model with all student and item characteristics. All analyses were performed using the R environment. The lme4 package was exploited for the estimation of the explanatory item response model.

4 Results

4.1 Indicator Validation Results

The first validation hypothesis was supported by the data; for all items, solution behaviours, in both grade 8 and grade 10, were found to be correct about 50% of the time, whereas the accuracy of the non-effortful responses was approximately 19% and 15% for grade 8 and grade 10, respectively. The second validation hypothesis was that the non-effortful responses should be correct at a rate close to that expected from the random responding. The proportion of correct responses of non-effortful responses ranged from 0.1 for the open-ended items to 0.26 for the close-ended items in grade 8, and from 0.04 for the open-ended items to 0.23 for the close-ended items in grade 10. Assuming that most of the close-ended items are four-option multiple choice and for the open-ended random responses are expected to be rarely correct, the observed values result to be very close to the accuracy rate expected from random responding. The data supported the third validation hypothesis, too. The average number of attempts results to be close to two for all the types of items in solution behaviour and close to one in case of non-effortful responses, in correspondence of both grades.

4.2 Math Disengagement

The results from Model 1, of both grade 8 and grade 10, showed that non-effortful behaviour was higher among students late-enrolled and among males. Interestingly, the spoken language showed differential effects on disengagement: it had a significant positive effect on grade 10 students but a not significant effect for grade 8. Parental educational attainment exhibited a significant negative effect: test-takers with parents with higher educational attainment were less disengaged. In correspondence of both scholastic grades, Model 2 showed that item position and item difficulty have a significant positive effect on disengagement, whereas there was no significant effect of item dimension. All these findings were confirmed in the full Model 3.

Focusing on the models of 8th-grade students, the variance of the random intercept of the student decreased by 4.25% from Model 0 to Model 3, whereas the variance in the random item intercept decreased by 40.44%. The variance of the

random person intercept was slightly lower for mathematics in grade 8 than for mathematics in grade 10. For the tests administered in grade 10, the variance of the random intercept of the student decreased by 5.90% from Model 0 to Model 3, whereas the variance in the random item intercept decreased by 53.31%.

5 Conclusions

The present study measures the extent of the non-effortful responding in the large-scale assessment INVALSI, providing an effort indicator based on the response time able to monitor the student behaviour over the entire test event. Moreover, a model-based approach has been used to identify the factors related to disengagement. Future studies will investigate the impact of individual emotional characteristics on the response time and the student's engagement.

References

1. Asseburg, R., Frey, A.: Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. Psychol Test Assess Model, 55(1), 92–104. (2013)
2. De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. The estimation of item response models with the lmer function from the lme4 package in R. J Stat Softw, 39(12), 1-28 (2011)
3. Eklöf, H. Skill and will: Test-taking motivation and assessment quality. Assess Educ, 17(4), 345-356 (2010)
4. Finn, B. Measuring motivation in low-stakes assessments. ETS Research Report Series, 2015(2), 1-17 (2015)
5. Goldhammer, F., Martens, T., & Lüdtke, O. Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. Large Scale Assess Educ, 5(1), 1-25 (2017)
6. Rios, J. A., Guo, H., Mao, L., & Liu, O. L. Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not?. Int J Test, 17(1), 74-104 (2017)
7. Sahin, F., & Colvin, K. F. Enhancing response time thresholds with response behaviors for detecting disengaged examinees. Large Scale Assess Educ, 8, 1-24 (2020)
8. Schipnike, D. L. Assessing speededness in computer-based tests using item response times. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco, CA. (1995)
9. Wise, S. L. Effort analysis: Individual score validation of achievement test data. Appl Meas Educ, 28(3), 237-252 (2015)
10. Wise, S. L. Rapid-guessing behavior: Its identification, interpretation, and implications. Educ Meas, 36(4), 52-61 (2017)
11. Wise, S. L. Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. Educ Inq, 10(1), 21-33 (2019)
12. Wise, S. L., & DeMars, C. E. Low examinee effort in low-stakes assessment: Problems and potential solutions. Educ Assess, 10(1), 1-17 (2005)
13. Wise, S. L., & Gao, L. A general approach to measuring test-taking effort on computer-based tests. Appl Meas Educ, 30(4), 343-354 (2017)
14. Wise, S. L., & Kong, X. Response time effort: A new measure of examinee motivation in computer-based tests. Appl Meas Educ, 18(2), 163–183 (2005)

3.9 Light methods for hard problems

Fast Divide-and-Conquer Strategies to Solve Spatial Big Data Problems

Strategie Veloci di Divide-et-Impera per la Risoluzione di Problemi con Dati Spaziali Massivi

Michele Peruzzi

Abstract Massive geolocated datasets are increasingly common in many scientific fields and industry. While latent Gaussian Process (GP) models are frequently chosen to perform statistical analyses and to quantify the associated uncertainties, they scale poorly to growing data sizes due to the intricate spatial dependencies. A simple class of methods for scalable computations involve partitioning scheme or tessellations of the spatial domain; independence or conditional independence assumption across such domain partitons give rise to different models. When a directed acyclic graph (DAG) is used to characterize conditional independence across domain partitions, one obtains *meshed* GPs which can be constructed to greatly reduce the number of expensive matrix operations. In this article, we show that similar strategies can be used for independent partitions methods. We consider satellite imaging data to compare several methods.

Abstract *I modelli con processi gaussiani latenti (GP) che vengono spesso scelti per eseguire analisi statistiche con dati geolocalizzati e quantificarne le incertezze si adattano male al crescere della dimensione dei dati a causa delle complesse dipendenze spaziali. Quando è un grafo aciclico diretto (DAG) a caratterizzare le dipendenze tra regioni spaziali, si ottiene un GP* a maglia *(MGP) che può essere costruito in modo da ridurre il numero di operazioni costose di algebra lineare. In questo articolo si mostra che anche un modello a partizioni indipendenti può essere costruito in modo simile. I metodi considerati sono messi a confronto su un dataset costruito da immagini satellitari.*

Key words: Spatial, big data, Bayesian regression, latent Gaussian, domain partitioning.

Michele Peruzzi
Dept. of Statistical Science, Duke University e-mail: michele.peruzzi@duke.edu

1 Introduction

When modeling the relationships between several spatially referenced outcomes and corresponding predictors, it is convenient to consider a linear spatial regression model. Let $\mathscr{D} \subset \Re^d$ denote the spatial domain and $\boldsymbol{\ell} \in \mathscr{D}$ be a generic spatial location. At $\boldsymbol{\ell}$ we have

$$\boldsymbol{y}(\boldsymbol{\ell}) = \boldsymbol{X}(\boldsymbol{\ell})^\top \boldsymbol{\beta} + \boldsymbol{w}(\boldsymbol{\ell}) + \boldsymbol{\varepsilon}(\boldsymbol{\ell}) \,, \tag{1}$$

where $\boldsymbol{y}(\boldsymbol{\ell}) = (y_1(\boldsymbol{\ell}), y_2(\boldsymbol{\ell}), \ldots, y_q(\boldsymbol{\ell}))^\top$ is the $q \times 1$ vector of outcomes at location $\boldsymbol{\ell}$, $\boldsymbol{X}(\boldsymbol{\ell})^\top = \text{blockdiag}\{\boldsymbol{x}_i(\boldsymbol{\ell})^\top\}_{i=1}^q$ is a $q \times p$ matrix of spatially referenced predictors with each $\boldsymbol{x}_i(\boldsymbol{\ell})$ being $p_i \times 1$ vector of predictors corresponding to $y_i(\boldsymbol{\ell})$ at location $\boldsymbol{\ell}$ and $p = \sum_{i=1}^q p_i$, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, $\boldsymbol{w}(\boldsymbol{\ell})$ and $\boldsymbol{\varepsilon}(\boldsymbol{\ell})$ are $q \times 1$ vectors of spatial random effects and random noise with elements $w_i(\boldsymbol{\ell})$ and $\varepsilon_i(\boldsymbol{\ell})$, respectively, for $i = 1, 2, \ldots, q$ such that $\boldsymbol{\varepsilon}(\boldsymbol{\ell}) \sim N(0, \boldsymbol{D})$ with a diagonal $\boldsymbol{D} = \text{diag}(\tau_1^2, \ldots, \tau_q^2)$, i.i.d. for all $\boldsymbol{\ell}$.

Dependence across spatial locations arises by modeling $\{\boldsymbol{w}(\boldsymbol{\ell}) : \boldsymbol{\ell} \in \mathscr{D}\}$ as a q-variate multivariate Gaussian process (GP), i.e. $\boldsymbol{w}(\boldsymbol{\ell}) \sim GP(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\theta}}(\cdot,\cdot))$, where $\boldsymbol{C}_{\boldsymbol{\theta}}(\cdot,\cdot)$ is a $q \times q$ matrix-valued *cross-covariance* function (see e.g. [4]) indexed by unknown parameters $\boldsymbol{\theta}$ which we may suppress to simplify notation. This means that $\boldsymbol{C}(\boldsymbol{\ell}, \boldsymbol{\ell}') = [\text{cov}\{w_i(\boldsymbol{\ell}), w_j(\boldsymbol{\ell}')\}]$ is the $q \times q$ matrix with (i,j)th element given by the covariance between $w_i(\boldsymbol{\ell})$ and $w_j(\boldsymbol{\ell}')$. Over any finite set $\mathscr{L} \subset \mathscr{D}$ the GP leads to $\boldsymbol{w}_{\mathscr{L}} \sim N(0, \boldsymbol{C}_{\mathscr{L}})$, where $\boldsymbol{w}_{\mathscr{L}}$ is the $qn_{\mathscr{L}} \times 1$ column vector and $\boldsymbol{C}_{\mathscr{L}}$ is the $qn_{\mathscr{L}} \times qn_{\mathscr{L}}$ block matrix with the $q \times q$ matrix $\boldsymbol{C}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_j)$ as its (i,j) block for $i, j = 1, \ldots, n_{\mathscr{L}}$. We focus here on the case $q = 1$ for simplicity.

Let $\mathscr{N} = \{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_n\} \subset \mathscr{D}$ be the set of observed locations. We construct $\boldsymbol{y} = [y(\boldsymbol{\ell}_1)^\top, \ldots, y(\boldsymbol{\ell}_n)^\top]^\top$ as the $n \times 1$ vector of $y(\boldsymbol{\ell}_i)$'s over the n locations, analogously define $\boldsymbol{w}$ and $\boldsymbol{\varepsilon}$, and let $\boldsymbol{X} = [\boldsymbol{X}(\boldsymbol{\ell}_1) : \ldots : \boldsymbol{X}(\boldsymbol{\ell}_n)]^\top$. The regression model (1) can be written as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w} + \boldsymbol{\varepsilon}$. Spatial dependence is highlighted by noting that marginally $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{C}_{\mathscr{N}} + \boldsymbol{D}_n)$, in which the dense $\boldsymbol{C}_{\mathscr{N}}$ alludes to the inadequacy of models assuming independent residuals in this context. The resulting Bayesian hierarchical model is constructed from the above specifications as

$$p(\boldsymbol{\beta}, \boldsymbol{w}, \tau^2, \boldsymbol{\theta} \,|\, \boldsymbol{y}) \propto p(\tau^2, \boldsymbol{\theta}) \times N(\boldsymbol{\beta}; \boldsymbol{m}_\beta, \boldsymbol{V}_\beta) \times N(\boldsymbol{w}\,; \boldsymbol{0}, \boldsymbol{C}_{\mathscr{N}}) \times N(\boldsymbol{y}\,; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}, \boldsymbol{D}_n) \,, \tag{2}$$

where $\boldsymbol{D}_n = \text{blockdiag}(\{\boldsymbol{D}\}_{i=1}^n)$. Bayesian inference typically proceeds by sampling from the posterior distribution in (2) via Markov-chain Monte Carlo (MCMC) algorithms. Unfortunately, a major roadblock arises when $\boldsymbol{\theta}$ is unknown: $\boldsymbol{C}_{\mathscr{N}}$ is a dense $n \times n$ matrix, and this implies that evaluating $N(\boldsymbol{w}; \boldsymbol{0}, \boldsymbol{C}_{\mathscr{N}})$ has a computational complexity of $O(n^3)$. The cubic cost associated to GPs has been addressed in a large literature which includes low-rank methods. We refer the reader to [11], [1], and [5] for in-depth reviews and comparisons of many scalable geostatistical methods.

2 Domain partitioning for scalable computations

Spatial domain partitioning may be leveraged for scalable computations: the spatial domain $\mathscr{D}$ is split into M mutually exclusive regions, i.e. $\mathscr{D} = \cup_{i=1}^{M} \mathscr{D}_i$ and $\mathscr{D}_i \cap \mathscr{D}_{i^*} = \emptyset$. We focus here on axis-parallel tesselations (APT), which subdivide each of the d coordinate axes into intervals—the partitions are the cartesian product of intervals.

Once a partition of the domain has been defined, perhaps recursively [6, 7], one can assume that if $\boldsymbol{\ell} \in \mathscr{D}_i$ and $\boldsymbol{\ell}' \in \mathscr{D}_{i^*}$ with $i \neq i^*$, then the latent process is such that $\boldsymbol{w}(\boldsymbol{\ell}) \perp\!\!\!\perp \boldsymbol{w}(\boldsymbol{\ell}')$. This *block-independence* assumption leads to a *new* spatial process $\widetilde{\boldsymbol{w}}$ which is associated to a block-diagonal covariance matrix $\widetilde{\boldsymbol{C}}_{\mathscr{N}}$ which replaces the off-diagonal blocks of $\boldsymbol{C}_{\mathscr{N}}$ with zeros. Evaluation of the new process density proceeds swiftly; by taking $\mathscr{N}_i = \mathscr{D}_i \cap \mathscr{N}$, one has $p(\widetilde{\boldsymbol{w}} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{M} p(\boldsymbol{w}_i \,|\, \boldsymbol{\theta})$ where $\boldsymbol{w}_i$ is the $|\mathscr{N}_i| \times 1$ vector collecting realizations of $\boldsymbol{w}$ at $\mathscr{N}_i$. We refer to this scheme as the *naïve* Divide & Conquer (*nD&C*) method; in the literature, it appears in [10] and as the "Partition" method in [5]. The computational complexity of *nD&C* is $O(MS^3)$ and driven by the cost of computing $p(\widetilde{\boldsymbol{w}} \,|\, \boldsymbol{\theta})$ assuming $|\mathscr{N}_i| \approx S$ for all $i = 1, \ldots, M$. If we further assume that S is constant for growing data size n, we have $M = n/S$ so the asymptotic complexity is $O(nS^2)$. Sampling $\boldsymbol{w}$ has the same cost.

If spatial independence assumptions are overly restrictive, one can consider a *Meshed* GP (MGP) [8, 9], in which independence across partitions is replaced by a fixed directed acyclic graph (DAG). *Conditional* independence across partitions amounts to taking $p(\widetilde{\boldsymbol{w}} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{M} p(\boldsymbol{w}_i \,|\, \boldsymbol{w}_{[i]}, \boldsymbol{\theta})$—we are now conditioning $\boldsymbol{w}$ on $\boldsymbol{w}_{[i]}$, which is the realization of the latent process $\boldsymbol{w}$ at the collection of locations corresponding to parents of node i in the DAG $\mathscr{G}$, denoted by $[i]$. A valid standalone spatial process can be created by assuming that $\boldsymbol{w}(\boldsymbol{\ell}) \perp\!\!\!\perp \boldsymbol{w}(\boldsymbol{\ell}') \,|\, \boldsymbol{w}_{\mathscr{N}}$ for any pair of locations $\boldsymbol{\ell}, \boldsymbol{\ell} \notin \mathscr{N}$. If $\mathscr{G}$ is sparse, the induced precision matrix $\widetilde{\boldsymbol{C}}_{\mathscr{N}}^{-1}$ is block sparse, leading to advantages in computations for collapsed MCMC methods. Here, we focus on the non-marginalized posterior (2); algorithmic complexity is driven by $p(\boldsymbol{w}_i \,|\, \boldsymbol{w}_{[i]} \boldsymbol{\theta}) = N(\boldsymbol{w}_i; \boldsymbol{H}_i \boldsymbol{w}_{[i]}, \boldsymbol{R}_i)$, where $\boldsymbol{H}_i = \boldsymbol{C}_{i,[i]} \boldsymbol{C}_{[i]}^{-1}$ and $\boldsymbol{R}_i = \boldsymbol{C}_i - \boldsymbol{H}_i \boldsymbol{C}_{[i],i}$. Taking $\mathscr{G}$ as the cubic mesh of [8], $\boldsymbol{C}_{[i]}$ is of size $2S \times 2S$ or less, whereas $\boldsymbol{R}_i$ is of size $S \times S$. Therefore, the complexity for spatial domains of dimension $d = 2$ is $O(9MS^3)$; this is asymptotically the same as with the *nD&C* method as it involves a fixed additional cost for the computation of $\boldsymbol{C}_{[i]}$. In practice, this leads to improved predictions and estimation of $\boldsymbol{\theta}$ for a relatively small additional computational cost.

3 Grid-caching to trim computational costs

DAG-based methods such as [2, 8, 7] all involve a reference set $\mathscr{M}$ of spatial locations (sometimes referred to as the knots). Knots do not necessarily coincide with observed locations. Methods based on sparse DAGs lead to sparse Gaussian precision matrices and one may choose $\mathscr{N} = \mathscr{M}$. However, setting $\mathscr{M} \supset \mathscr{N}$ may be computationally faster when data locations are on a partly observed lattice–such as

pixels of a satellite image, some of which missing due to cloud cover [8]. The DAG $\mathscr{G}$ can be made to mimick the patterns in $\mathscr{M}$, resulting in extreme reductions in the number of unique $\boldsymbol{C}_{[i]}$ and $\boldsymbol{R}_i$ which need to be computed. With *caching*, rather than having to compute $9M$ system solvers of the MGP above, one only computes 4, regardless of data size. [9] extended these ideas to completely irregular data. To clarify, suppose the covariance function $\boldsymbol{C}(\cdot,\cdot)$ is stationary, and let $\mathscr{M}$ be an equally-spaced grid of points in $\mathscr{D}$. Following an APT scheme, one obtains $\{\mathscr{M}_i\}_{i=1}^M$ where $\mathscr{M}_i = \mathscr{M} \cap \mathscr{D}_i$. The cubic DAG of [8] leads to most nodes i having exactly 2 parents, e.g. the "south" parent and the "west" parent. Then, there exist nodes $i^\#$ and i^* such that S_i is obtained by translating $\mathscr{M}_{i^\#}$ along the spatial domain, and similarly $S_{[i]}$ with $S_{[i^*]}$ (it may be $i^* = i^\#$). But then $\boldsymbol{C}_i = \boldsymbol{C}_{i^\#}$ and $\boldsymbol{C}_{[i]} = \boldsymbol{C}_{[i^*]}$; one computes the inverses for these *prototypes* and copies the result across the whole domain.

A *fast* D&C method (*fD&C*) can be built to allow caching. Suppose $\mathscr{N}$ are irregularly-spaced spatial locations and take a gridded set of knots $\mathscr{M}$ such that, after APT, $\mathscr{M}_i \overset{\rightarrow}{=} \mathscr{M}_{i'}$ for any pair $i, i' = 1, \ldots, M$, where $\overset{\rightarrow}{=}$ refers to equality up to translations along the domain, and $|\mathscr{M}_i| = S$. Then we let $\boldsymbol{w}(\boldsymbol{\ell}) \perp\!\!\!\perp \boldsymbol{w}(\boldsymbol{\ell}') \,|\, \boldsymbol{w}_{\mathscr{M}}$ for any pair of locations $\boldsymbol{\ell}, \boldsymbol{\ell} \notin \mathscr{M}$; these include the observed locations $\mathscr{N}$. The regression model at location $\boldsymbol{\ell} \in \mathscr{D}_i$ can now be written as $y(\boldsymbol{\ell}) = \boldsymbol{X}(\boldsymbol{\ell})^\top \boldsymbol{\beta} + \boldsymbol{H}_{\boldsymbol{\ell}} \boldsymbol{w}_{\mathscr{M}_i} + \boldsymbol{\nu}(\boldsymbol{\ell})$ where $\boldsymbol{\nu}(\boldsymbol{\ell}) \sim N(0, \boldsymbol{R}_{\boldsymbol{\ell}} + \tau^2)$, $\boldsymbol{H}_{\boldsymbol{\ell}} = \boldsymbol{C}_{\boldsymbol{\ell},i} \boldsymbol{C}_i^{-1}$, $\boldsymbol{R}_{\boldsymbol{\ell}} = \boldsymbol{C}_{\boldsymbol{\ell}} - \boldsymbol{C}_{\boldsymbol{\ell},i} \boldsymbol{C}_i^{-1} \boldsymbol{C}_{i,\boldsymbol{\ell}}$ and $\boldsymbol{C}_i$ is the covariance matrix at locations $\mathscr{M}_i$. In practice one only needs to compute a *single* Cholesky decomposition for a "prototype" $\boldsymbol{C}^*$ of size $S \times S$ rather than for $\boldsymbol{C}_i$ for all i; this implies that almost all the complexity of this model is due to sampling $\boldsymbol{w}$ – sampling inefficiencies can be reduced in this context [9].

4 Application: FPAR data in North Carolina

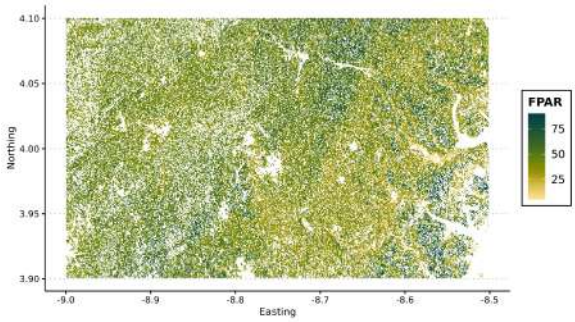
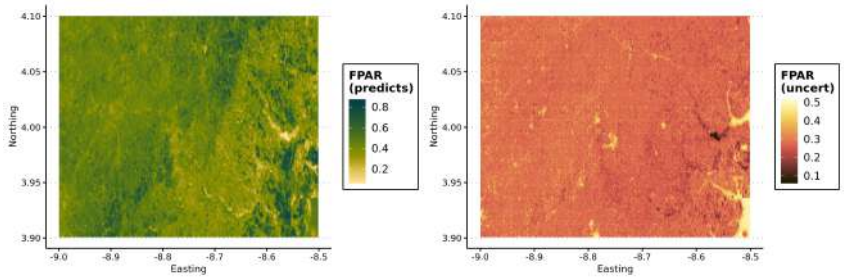


Fig. 1 Observed FPAR data from MODIS satellite imaging.

The fraction of photosynthetically active radiation (i.e. with wavelengths of 400-700 nm) absorbed by green vegetation (FPAR) is used for calculating several mea-

sures related to the biogeochemistry of vegetation. The MODIS *Aqua* and *Terra* satellites product MOD15A2H v.6 includes FPAR at a 500m grid of worldwide locations averaged over 8-day periods. Here, we consider a large dataset with $n = 109,610$ observed locations over a region in North Carolina, USA, in the 8-day time period ending on February 18th, 2020. We compare our *fD&C* to its naive counterpart, a MGP model and a NNGP model of the response (i.e. not involving latent random effects, see [3]). All methods run 5000 MCMC iterations and compute predictions at a set of 50,000 left-out locations in about 6 minutes; they all use the same covariance $\boldsymbol{C}(\ell,\ell') = \sigma^2 \exp\{-\phi\|\ell-\ell'\|\}$, and the same prior distributions, which are $\tau^2 \sim Inv.Gamma(2,1)$, $\sigma^2 \sim Inv.Gamma(2,1)$, $\phi \sim U(1/20,2000)$; computations ran on a workstation with a Ryzen 9 5950X CPU running on 10 threads. Figure 2 reports *fD&C* predictions with uncertainty *a posteriori*; Table 1 compares

Fig. 2 Predictions (left) and pointwise uncertainty band width (right) obtained from *fD&C*.



predictive performance of the models. All models perform very similarly; we highlight the improvement of our proposed *fD&C* strategy relative to its naïve counterpart. Thanks to caching, the *fD&C* method uses much coarser partitioning than *nD&C*, while performing all operations in the same timeframe.

Table 1 Predictive performance of the methods under considerations.

Method	RMSPE	MAPE	Coverage (95% nominal)	Time (min)
nD&C (60×34 partitioning)	0.0764	0.0558	93.93%	5.4
fD&C (34×17 partitioning)	0.0755	0.0552	93.92%	5.7
MGP (50×34 partitioning)	0.0746	**0.0545**	**93.98%**	5.9
NNGP ($m = 10$ neighbors)	**0.0745**	**0.0545**	93.65%	5.9

Acknowledgements The author received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506), and grant R01ES028804 of the United States National Institutes of Health (NIH).

References

[1] S. Banerjee. High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12 (2):583–614, 2017. doi:10.1214/17-BA1056R.

[2] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111:800–812, 2016. doi:10.1080/01621459.2015.1044091.

[3] A. O. Finley, A. Datta, B. D. Cook, D. C. Morton, H. E. Andersen, and S. Banerjee. Efficient Algorithms for Bayesian Nearest Neighbor Gaussian Processes. *Journal of Computational and Graphical Statistics*, 28:401–414, 2019. doi:10.1080/10618600.2018.1537924.

[4] M. G. Genton and W. Kleiber. Cross-Covariance Functions for Multivariate Geostatistics. *Statistical Science*, 30:147–163, 2015. doi:10.1214/14-STS487.

[5] M. J. Heaton, A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, D. W. Nychka, F. Sun, and A. Zammit-Mangion. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, Sep 2019. doi:10.1007/s13253-018-00348-w.

[6] M. Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112:201–214, 2017. doi:10.1080/01621459.2015.1123632.

[7] M. Peruzzi and D. B. Dunson. Spatial multivariate trees for big data Bayesian regression, 2020. arXiv:2012.00943.

[8] M. Peruzzi, S. Banerjee, and A. O. Finley. Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 2020. in press. doi:10.1080/01621459.2020.1833889.

[9] M. Peruzzi, S. Banerjee, D. B. Dunson, and A. O. Finley. Grid-Parametrize-Split (GriPS) for improved scalable inference in spatial big data analysis, 2021. arXiv:2101.03579.

[10] M. L. Stein. Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19, 2014. doi:doi:10.1016/j.spasta.2013.06.003.

[11] Y. Sun, B. Li, and M. Genton. Geostatistics for large datasets. In J. Montero, E. Porcu, and M. Schlather, editors, *Advances and Challenges in Space-time Modelling of Natural Events*, pages 55–77. Springer-Verlag, Berlin Heidelberg, 2011. doi:10.1007/978-3-642-17086-7.

Application of hierarchical matrices in spatial statistics

Applicazione di matrici gerarchiche nella statistica spaziale

Anastasiia Gorshechnikova and Carlo Gaetan

Abstract Large datasets with irregularly spatial (or spatio-temporal) locations are difficult to handle in many applications of Gaussian random fields such as maximum likelihood estimation (MLE) and prediction. We aim to approximate covariance functions in a format that facilitates the computation of MLE and prediction with very large datasets using a hierarchical matrix approach. We present a numerical study where we compare this approach with the covariance tapering method.

Abstract *Grandi dataset contenenti posizioni spazio-temporali disposte in maniera irregolare sono molto difficili da trattare in parecchie applicazioni dei campi aleatori gaussiani, quali la stima di massima verosimiglianza o la previsione. Il nostro obiettivo è di approssimare le funzioni di covarianza in un formato che faciliti il calcolo della stima di massima verosimiglianza e della previsione in caso di dataset molto grandi si basa sull'uso di matrici gerarchiche. Un esempio numerico in cui si confronta il metodo proposto con il 'tapering' viene presentato.*

Key words: computational methods, hierarchical matrices, large datasets, covariance matrices

1 Introduction

Large data sets are common in environmental sciences where data are often observed at a large number of spatial locations and at different temporal intervals. Therefore, computational and modeling challenges arise which were labeled by as

Anastasiia Gorshechnikova
Proximus, Boulevard du Roi Albert II 27 Brussels, Belgium, e-mail: agorshechnikova@gmail.com

Carlo Gaetan
Ca' Foscari University of Venice, Dorsoduro 3246 Venice, Italy, e-mail: gaetan@unive.it

"big N problem". The exact computation of the likelihood of a Gaussian Random Field (GRF) observed at N irregularly sited locations generally requires $O(N^3)$ floating point operations and $O(N^2)$ memory [3].

Consider a vector Z of N observations from a GRF $\{Z(x)\}$ defined over a domain indexed by x, where x denotes either a spatial $x : s \in \mathbb{R}^d$ or spatio-temporal domain of observations $x : (s,t) \in \mathbb{R}^d \times \mathbb{R}$. Without loss of generality we consider a zero-mean GRF. Considering parametric covariance function with the vector of the unknown p- dimensional parameters $\theta \in \Theta \subseteq \mathbb{R}^p$, the covariance function $c(x) := c(x;\theta)$ depends on unknown parameter θ. We make statistical inference with respect to θ based on the Gaussian log-likelihood

$$L(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|C_Z| - \frac{1}{2}Z^\top C_Z^{-1} Z \tag{1}$$

where N is the sample size and C_Z is the covariance matrix of Z. As can be seen from (1), to make an inference on the unknown parameter θ the exact computation of the log-likelihood requires a computation of the determinant of the covariance matrix $|C_Z|$ as well as its inverse C_Z^{-1} which both require $O(N^3)$ operations.

A similar computational burden is involved in evaluating the best linear unbiased prediction (BLUP), at an unobserved location x_0 defined as follows

$$Z(x_0) = c(x_0)^\top C_Z^{-1} Z, \tag{2}$$

where $c(x_0) = [c(x_0,x_1),\ldots,c(x_0,x_N)]'$ is covariance vector formed based on a new location x_0 and $C_Z = C(x_i,x_j)$.

A comparison of current methods to tackle this computational problem is contained in [3]. For instance, in the covariance tapering approach [4] the covariance matrices are multiplied element-wise by a sparse correlation matrix which results in another positive definite function with a compact support, i.e. $C_T = C_Z \circ T(\delta)$, where $T(\delta)$ is a compactly supported correlation function which is identically zero whenever $||s-s'|| \geq \delta$ with $s,s' \in \mathbb{R}^d$ and taper (or cut-off distance) δ. Therefore

$$L(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|C_T| - \frac{1}{2}Z^T C_T^{-1} Z \tag{3}$$

is the tapered likelihood, where $C_T = C_Z \circ T(\delta)$.

The covariance tapering method may not be effective in accounting for spatial dependence with long range dependence thereby sacrificing some precision. Also it is not straightforward how to choose the distance to taper off. In this work we present an approach based on the approximation of covariance functions by hierarchical matrices (or shortly $\mathscr{H}$-matrices). Focusing on the numerical analysis, the method of $\mathscr{H}$-matrix was exploited by [5] for MLE estimation. We extend this work by adapting the regularity conditions, performing kriging prediction on a simulated dataset and comparing this technique with covariance tapering in terms of both computational and statistical efficiencies.

2 Hierarchical matrices

The idea behind $\mathscr{H}$-matrices is to use a low-rank approximation of the blocks of a covariance matrix which are located far from the diagonal entries. To obtain the structure of a covariance matrix with the off-diagonal blocks $\tilde{C}^k_{\text{block}}$ approximated in a low-rank k format, an index $i \in I$ from the index set $I \subset \mathbb{N}$ is firstly assigned to each data location $x_i \in \mathbb{R}^d$. The hierarchical structure of a matrix is then obtained by partitioning the index set I into subsets or, equivalently, associated data locations x_i into clusters. This is required in order to obtain matrix blocks which further can be factored, such that a low-rank block $\tilde{C}^k_{\text{block}}$ is characterised by the rank $k << N$. These all are crucial steps required to compress data and perform matrix operations in a linear cost. We refer to [2] for the technical details of the $\mathscr{H}$-matrix method.

The matrix C resulting from a covariance function $c(\cdot)$ is not sparse. To find a data sparse representation of some blocks of the covariance matrix, their low-rank decomposition must be exploited. We refer to [1] for the description of analytical techniques to find a low-rank approximation of a block $\tilde{c}^k(x_i,x_j)$ of the covariance function $c(x_i,x_j)$. According to [2], to admit a low-rank representation it is necessary that the underlying functions satisfy so called 'asymptotic smoothness condition'.

We define a d-dimensional multi-index notation $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ of non-negative integers. For the multi-index $\alpha \in \mathbb{N}_0^d$ sum of the components or absolute value can be written as $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$ and higher-order partial derivatives as $\partial^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \ldots \partial_d^{\alpha_d}$, where $\partial_i^{\alpha_i} = \partial^{\alpha_i} / \partial x_i^{\alpha_i}$ of the dimension d. Let $X_i, X_j \subset \mathbb{R}^d$ be subsets such that the function $c(x_i,x_j)$ is defined and arbitrarily often differentiable for all spatial locations $x_i \in X_i$ and $x_j \in X_j$ with $x_i \neq x_j$ for $i, j = 1, \ldots, N$. Then the covariance function $c(x_i,x_j)$ is asymptotically smooth if there exist constants $p_1, p_2 \in \mathbb{R}^+$, such that for all multi-indices $\alpha \in N_0^d$, one has

$$|\partial_x^\alpha c(x_i,x_j)| \leq p_1 |\alpha|! p_2^{|\alpha|} (||x_i - x_j||)^{-|\alpha|} \tag{4}$$

for all $x_i \neq x_j$.

The factor $p_2^{|\alpha|}$ allows for a change of the growth behaviour. The derivatives tend to 0 as $||x_i - x_j|| \to \infty$. The condition (4) is required to guarantee a fast decay of the eigenvalues of the underlying function which leads to an effective low-rank approximation $\tilde{C}^k_{\text{block}}$ of specific blocks of C, so that the error $|c(x_i,x_j) - \tilde{c}(x_i,x_j)|$ of a low-rank approximation of $c(x_i,x_j)$ converges exponentially fast. At first sight, the condition (4) seems to be restrictive. However, this condition is satisfied by some classes of spatial covariance functions such as Matérn and spatio-temporal covariance functions, see [1] for the details.

We aim to approximate a covariance matrix by the $\mathscr{H}$-method and perform a fast approximated Cholesky decomposition. We denote the $\mathscr{H}$-matrix approximation of the covariance matrix by $\tilde{C}$ and approximation of the Cholesky factor by $\tilde{\Lambda}$, so that $\tilde{C} = \tilde{\Lambda}\tilde{\Lambda}^T$. To be able to perform approximate Cholesky decomposition, the positive definiteness property of $\tilde{C}$ should be preserved. With the approximation by $\mathscr{H}$-matrices, the error can propagate and perturb the eigenvalues of the resulting matrix.

If the smallest eigenvalue is close to the origin, the result of these operations might become indefinite. To tackle this problem, we follow the suggestion of [5] to add a nugget value to the diagonal of $\tilde{C}$, albeit sacrificing approximation accuracy for the sake of positive definiteness. The $\mathscr{H}$-approximation of the exact log-likelihood $L(\theta)$ is defined by $\tilde{L}(\theta,k)$ with the maximal rank k

$$\tilde{L}(\theta,k) = -\frac{N}{2}\log 2\pi - \sum_{i=1}^{N}\log\tilde{\lambda}_i - \frac{1}{2}U^{\top}U, \tag{5}$$

where $U^TU = Z^T(\tilde{\Lambda}\tilde{\Lambda}^T)^{-1}Z = Z^TCZ$ which is composed of the matrix-vector multiplications with a log-linear cost and $\tilde{\lambda}_i$ are diagonal elements of $\tilde{\Lambda}$, such that $\log\det(C) = \log\det\tilde{\Lambda}\tilde{\Lambda}^T = \log\det\left(\prod_{i=1}^{N}\tilde{\lambda}_i^2\right) = 2\sum_{i=1}^{N}\log\tilde{\lambda}_i$.

As with the likelihood in (5), we substitute C_Z in (2) by the approximated by $\mathscr{H}$- covariance $\tilde{C}_Z$. Then a simple kriging prediction for a location x_0 using the estimated covariance function with $\hat{\theta}$ in (5) is $\tilde{Z}(x_0) = \tilde{c}(x_0)^{\top}\tilde{C}_Z^{-1}Z$, where $\tilde{c}(x_0) = [\tilde{c}(x_0,x_1),\ldots,\tilde{c}(x_0,x_N)]'$ is the $\mathscr{H}$-matrix approximation of the corresponding covariance vector. We note that as in (5) it is also based on the matrix-vector multiplications which leads to the log-linear cost computation due to the $\mathscr{H}$-matrices.

3 Numerical results

For the covariance tapering approach distant pairs of observations are modelled using a compactly supported covariance function. With the $\mathscr{H}$-matrices, off-diagonal elements of C_Z are defined through the low-rank factors. Because of the similarity of both methods, the main purpose of this section is to compare their performance based on computational and statistical efficiency. With the covariance tapering the 'score' function for θ based on (3) is biased. Since (5) also entails a biased score function, i.e $\frac{1}{2}\left(Z^T\tilde{C}_Z^{-1}C_i\tilde{C}_Z^{-1}Z - \mathrm{tr}(\tilde{C}_Z^{-1}C_{iZ})\right)$, we use (5) with $\mathscr{H}$-covariance $\tilde{C}_Z$ and (3) with tapered covariance C_T.

The simulation study is performed with the increasing domain asymptotics setup on the randomly perturbed grid of spatial locations by constructing a regular grid with increments 0.03 over $W_k = [0,2^{(k+2)/2}]\times[0,2^{(k+2)/2}], \quad k = 0,\ldots,2.$ and perturbing the regular grid points by adding a uniform random value on $[-0.01,0.01]$. With this setup, each data location is at least 0.01 units distant from its neighbours.

For the different sample size of $N_k = \{2000,4000,8000\}$ points with $k = 0,\ldots,2$ chosen without replacement, we simulate $L = 100$ realizations of zero-mean GRF with Matérn covariance with the true parameters $\theta = (\sigma^2,\varphi,\nu,\tau^2) = (1,0.1,0.5,0.1)$, where σ^2 is the marginal variance and τ^2 is the nugget parameter. We fix the smoothness parameter $\nu = 0.5$ (exponential covariance function) and $\tau^2 = 0.1$ is added to the diagonal in order to preserve the positive definiteness property. In addition, we scale the distance in (4) by the range parameter φ. This adjustment resulted in a computational efficiency that is doubled compared to the standard condition.

To check the predictive performance with the increasing N_k, we divide the simulated data into a training dataset chosen at random and a validation dataset containing the remaining 10%, i.e $M = \{200, 400, 800\}$ observations respectively. As practical range we set $\varphi = 0.1$ due to consistency of φ over the spatial domain to increasing domain framework. Because we keep distance as fixed, increasing k and consequently the number N_k of observations, the percentage of nonzero elements in the resulting tapered covariance matrix decreases. By varying the practical range $\delta = \{0.15, 0.3, 0.5\}$, the percentage of non-zero elements p in the tapered covariance matrix increases. For the $\mathscr{H}$-matrices we control the compression ratio q which is defined as the ratio between the sizes of a compressed (hierarchical matrix) $\tilde{C}$ and original matrix C. The *h2lib* library[1] was exploited for application of $\mathscr{H}$-matrices); for covariance tapering method *spam*[6] was used.

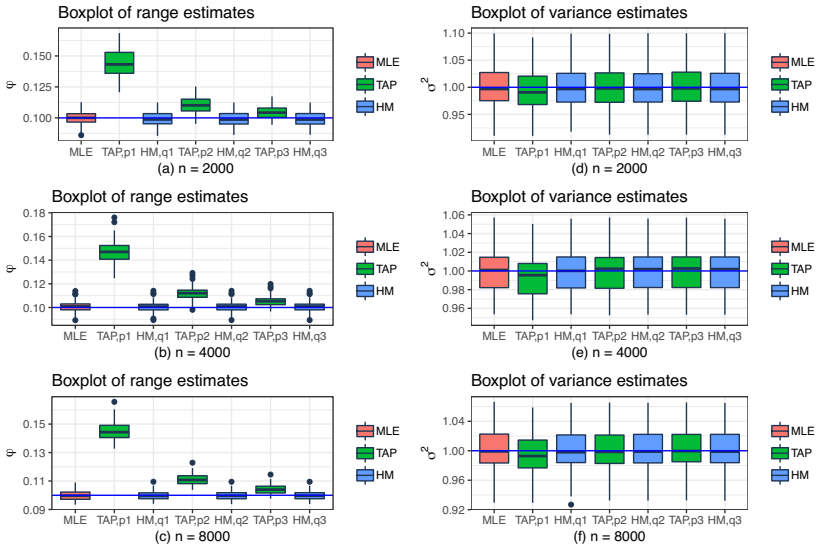


Fig. 1 Boxplots of sampled estimates (a)-(c) $\hat{\varphi}$ and (d)-(f) $\hat{\sigma}^2$ with the horizontal line of the true estimates $(\varphi = 0.1, \sigma^2 = 1)$ under the exact maximum likelihood estimation (MLE), covariance tapering (TAP) and $\mathscr{H}$-matrices (HM)

Figure 1 shows boxplots of the estimates of the φ and σ^2 parameters with the both methods, including the exact ML estimation. The horizontal line indicates the true values of the estimates $(\varphi = 0.1, \sigma^2 = 1)$. As taper δ decreases, the biases in the one-taper estimates increase. In contrast, we see negligible bias in the $\mathscr{H}$-matrices estimates. The difference in variance estimates with both methods is almost indistinguishable. In terms of computational efficiency, for example, for $n = 8000$ likelihood evaluation based on the $\mathscr{H}$-matrices with q_{max} required $t_{hm} = 2$ min compared to $t_{tap} = 7$ min by the covariance tapering approach with p_{max}. Thus, the application

[1] https://github.com/H2Lib/H2Lib, developed by Steffen Boerm and his group, Kiel, Germany

of $\mathscr{H}$-matrices approach for ML estimation results in computational efficiency as well as in a good statistical efficiency even with a small compression ratio q.

To compare the predictive performance of both methods we compute Root-Mean-Squared Prediction Error (RMSPE). The set of the predicted locations for each M is denoted as D_M^* with each new location $x_0 \in D_M^* \subset \mathbb{R}^d$. If $\tilde{Z}(x_0,l)$ denote the model-A predictor, where $Z(x_0,l)$ is the lth simulated process evaluated at a new location x_0 and A =TAP, HM, then the model-A predictor RMSPE for the lth simulation is $\text{RMSPE}_A(l) = \sqrt{\sum_{x_0 \in D_M^*} \left(\tilde{Z}(x_0,l) - Z(x_0,l)\right)^2}$, $l = 1,\ldots,L$. We then consider a measure of relative skill (RS), relative to HM, namely $\text{RS}(N) = \text{RMSPE}_{\text{HM}}(l)/\text{RMSPE}_{\text{TAP}}(l)$ for different $N = 2000, 4000, 8000$. As can be seen from the Figure 2 for different sample size N_k, density and correlation ratio p and q $RS(N) < 1$. Therefore, $\mathscr{H}$-matrices approach has a better predictive accuracy.

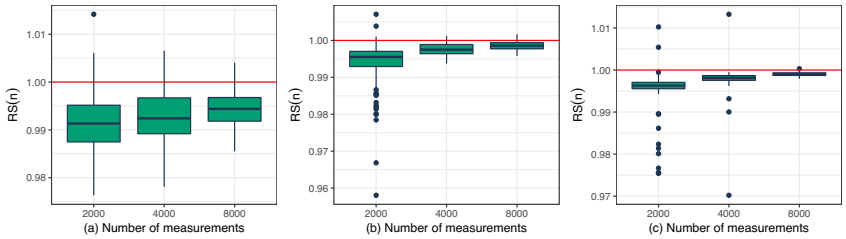


Fig. 2 Boxplots of RS(N) for (a) $N_1 = 2000 : p_1 = 0.2, q_1 = 0.3, N_2 = 4000 : p_1 = 0.15, q_1 = 0.25$, $N_3 = 8000 : p_1 = 0.1, q_1 = 0.2$, (b) $N_1 = 2000 : p_2 = 0.5, q_2 = 0.8, N_2 = 4000 : p_2 = 0.38, q_2 = 0.48$, $N_3 = 8000 : p_2 = 0.27, q_2 = 0.37$, (c) $N_1 = 2000 : p_3 = 1.5, q_3 = 1.9, N_2 = 4000 : p_3 = 1.33, q_3 = 1.56, N_3 = 8000 : p_3 = 1.12, q_3 = 1.31$

References

1. Gorshechnikova, A.: Likelihood Approximation and Prediction for Large Spatial and Spatio-temporal Datasets using H-matrix Approach. PhD-Thesis, University of Padua, Italy (2019)
2. Hackbusch, W.: Hierarchical matrices: algorithms and analysis. Vol. 49. Heidelberg: Springer (2015)
3. Heaton, M.J., Datta,A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Zammit-Mangion, A. et al.: A case study competition among methods for analyzing large spatial data. Journal of Agricultural, Biological and Environmental Statistics, 24(3), 398–425 (2019).
4. Kaufman, C.G., Nychka, D.W., Schervish, M.J.: Covariance tapering for likelihood-based estimation in large spatial data sets. Journal of the American Statistical Association 103.484, 1545–1555. (2008)
5. Litvinenko, A., Genton, M.G., Keyes, D.E., Sun, Y.: Likelihood approximation with hierarchical matrices for large spatial datasets. Computational Statistics & Data Analysis 137, 115–132 (2019)
6. Furrer, R., Sain, S.R.: spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. Computational Statistics & Data Analysis 137, 36, 1–25. (2010)

3.10 Management and statistics in search for a common ground (AIDEA)

Customer Segmentation: it's time to make a change

La segmentazione della clientela: è tempo di un cambiamento

Fabrizio Laurini and Beatrice Luceri and Sabrina Latusi

Abstract Companies routinely use customer segmentation, often tracking fidelity cards, and collect huge amount of information. Nowadays the $k-$means method is still widespread, having the merits of being available on many platforms and being computationally efficient even with big data. However, the $k-$means has still many drawbacks, and in this work we propose an algorithm which overcomes many limitations and it mixes the benefits of model-based clustering, efficient computational times and is resistant to outliers. We apply our algorithm to a real dataset using data by a non-food retail chain. We find interesting results profiling several hundred thousand customers, of which we know shopping habits and social-demographic indicators.

Abstract *Le aziende utilizzano abitualmente la segmentazione dei clienti, spesso monitorando le carte fedeltà, e raccolgono enormi quantità di informazioni. Al giorno d'oggi il metodo delle $k-$medie è ancora diffuso, avendo il merito di essere disponibile su molte piattaforme ed essere computazionalmente efficiente anche con grandi dati. Tuttavia, le $k-$medie hanno ancora molti svantaggi, e in questo lavoro proponiamo un algoritmo che supera molte limitazioni e mescola i benefici del clustering basato su modelli causa-effetto, con tempi di calcolo molto rapidi ed è resistente ad outliers. L'algoritmo viene testato sui dati di una catena di vendita al dettaglio non alimentare. Si sono trovati interessanti risultati di profilazione di diverse centinaia di migliaia di clienti, di cui si conoscono abitudini di shopping e indicatori socio-demografici.*

Fabrizio Laurini
Department of Economics and Management, University of Parma, Via J.F. Kennedy 6, 43125, Parma, Italy, e-mail: fabrizio.laurini@unipr.it

Beatrice Luceri
Department of Economics and Management, University of Parma, Via J.F. Kennedy 6, 43125, Parma, Italy, e-mail: beatrice.luceri@unipr.it

Sabrina Latusi
Department of Economics and Management, University of Parma, Via J.F. Kennedy 6, 43125, Parma, Italy, e-mail: sabrina.latusi@unipr.it

Key words: Clustering, Big data, Robustness

1 Introduction

Customer segmentation is the basis of marketing strategy and practice. Market segmentation is an essential step in the development of marketing strategies. Market segmentation (clustering or partitioning) identifies homogeneous groups of consumers called "segments". In practice, segmentation is commonly based on demographic variables (such as gender and age), consumers' psychographic characteristics (such as lifestyles, values, and attitudes), and geographic variables.

Segmentation of a market enables us to design different advertising campaigns for different segments, develop products with suitable characteristics for consumers in specific segments, propose a price differentiation along the defined segmentation, and design other marketing strategies.

The American Marketing Association defines segmentation as "the process of subdividing a market into distinct subsets of customers that behave in the same way or have similar needs. Each subset may conceivably be chosen as a market target to be reached with a distinct marketing strategy". Defining customer requirements by homogeneous groups is the first step of the Segmentation-Targeting-Positioning approach (STP, e.g., [2]). This sequential process starts with the extraction, profiling and description of segments (market segmentation), followed by the selection of a target segment on the basis of its attractiveness and product/brand competitive position (targeting), and finally the development of a positioning for different target segments to ensure that the product/brand is perceived as distinctly different from competing products/brands, and in line with segment needs (positioning). In sum, the STP process supports the company in developing marketing mix decisions for each segment.

The challenge for both scholars and practitioners is to identify the best criteria for segmenting the market and profiling the different segments assuring that they are: homogenous within, heterogeneous between, measurable, substantial, and operational (e.g., [1]). Homogeneity and heterogeneity conditions are often hindered by boundary data which are close to more than one segment and can be incorrectly assigned. This focal issue calls for the development of efficient segmentation models using advanced techniques to avoid segments overlapping and to support marketers to make the right evaluations and decisions.

The study aims to develop and test a new clustering algorithm able to overcome the limitations of the most widely used segmentation model, namely k-means clustering. The proposed clustering algorithm is applied to the customer data of a primary brick-and-click non-food retail chain operating in Italy. The data represent 24-month customer shopping data for eight product groups. The paper contributes to the existing marketing literature by developing a customer segmentation approach that is theoretical consistent and reliable.

2 Robust clustering for supervised problems

In regression model-based clustering, outliers and noise can be handled in different ways. For example, one approach is to represent them with one (or more) finite mixture model component(s) additional to those for the meaningful part of the data via maximum likelihood (ML). Alternatively, it is possible to rely on normally distributed variables and downweight the contribution of atypical observations using, e.g., M-estimation, updating the components of a Gaussian mixture in the M step of the EM algorithm.

Furthermore, it is important to note that cluster analysis is also not a well-defined problem from an applied viewpoint. There is nowadays a wide consensus about the fact that clustering techniques should always depend on the final data-analysis purpose, so that different goals require the use of different clustering approaches. This does not seem to be well understood by many practitioners. This holds in many applications of market research, where the construction of relevant clusters must be coupled with subject matter aims. We thus argue that clustering should not be seen as a fully automatic task providing just one single solution and that the user always has to play an active role in it. However, when decision support is needed, at the very least outliers must be accounted for, and their effect reduced to avoid any data-driven bias.

From a quite technical viewpoint, given a sample of observations $x_1, \ldots, x_n$ a widely used method in unsupervised learning is to assume multivariate normal components and to adopt a maximum likelihood approach for clustering purposes. With this idea in mind, well-known classification and mixture likelihood approaches can be followed. Unfortunately, it is well-known that the maximization of "log-likelihoods" without proper deviation from model's assumption create numerically unstable problems. Additionally, multivariate normality is mostly "a dream that turns into a nightmare": very often it is assumed for mathematical convenience, but equally often all diagnostics miserably reject that theoretical simplification.

Robust methods are designed to overcome these restrictions at the cost of being computationally expensive in terms of evaluation time. Nowadays this looks a rather minor issue.

To tackle this feature, we mix the some approaches adapting the "impartial trimming", consisting in removing from the dataset a fraction α of the "most outlying" data units, but constraining the group scatters to reduce the possibility of spurious solutions in the ML. This computationally efficient framework based on trimming and scatter constraints was introduced for multivariate clustering in the TCLUST method ([6]; [3]) and then extended to the regression setting with the TCLUST-REG of [4].

The peculiar feature of the TCLUST-REG is to provide clustering of units under a supervised learning algorithm. In other words, it is possible to give clustering under a model of type $Y = f(x_1, x_2, \ldots, x_p; \theta) + \varepsilon$, with f a function linear in the parameters p-dimensional vector of parameters θ and ε an error term (not restricted to be Gaussian). This is a convenient generalization to the k-means where casual-effect models are not usable. We implement a robust version of the TCLUST-REG

which is not affected by outliers and extreme values. There are several theoretical results supporting our methods [5]

3 Data and preliminary results

In the data analysed, kept anonymous for confidentiality, there are shopping tracks of 24 month sales of non-food items. The number of customers is approximately 470000. The average sale of each customer in the time period is the response variable Y from which the supervised clustering algorithm. The set of explanatory variables is given by the number of visits, the number of items bought per visit, the percentage value bought with promotion/sales, age and gender of the customer. After running the algorithm a fine-tuning is required to select the optimal number of clusters, according to a well-defined objective function. With our data we obtained five groups and with approximately 5% of customers identified as outliers and unallocated. The outliers in the data (approximately 25000 customers) are mostly characterized by occasional shops and low revenue for the chain. For clustering we use a method which includes constraints on the eigenvalues of the dispersion matrices, so avoiding thread shaped clusters. The pattern of clusters and outliers alters appreciably for low levels of trimming.

We want to remark that the identification of these outliers is fully automatic and not arbitrary, but comes as a by-product of an optimal model-based algorithm. A subset of 1000 customers and class allocation associated is displayed in Figure 1. The clusters projected in the 2 dimensional plan of the first principal components is sketched in Figure 2. Notice the elliptical shapes and little overlapping in clusters. The robust model-based clustering approaches, achieved by maximizing the trimmed likelihood functions is very useful for any practical implication, as when clusters are disentangled, better managerial decisions are possible.

In order to understand the main features of the customers that belong to each cluster, it is useful to analyse some of the means of the variables used for clustering. Broadly speaking it is possible to detect that in cluster 4, the customers spend a lot more and buy more articles compared to the average. Interesting to notice that these customers spend less on average for the articles they buy tend to buy on sales rather than full price even though they overall buy more compared to the other clusters. As regards for cluster 5, the customers spend the least compared to the others, even though in cluster 6 customers buy the least number of articles. This means that in cluster 5 people buy less expensive articles but more often, while in the cluster 6 they spend more buying less articles. The cluster where customers buy more in full price is the 2 and on average it is the group where it is spent the most after cluster 4. To conclude, it can be said that cluster 1 and 3 are the less valuable and less affiliated customers, because it they seldom buy and when they do, they do not spend much. On the contrary, cluster 4, followed by cluster 2, are the most valuable and loyal groups for the retailer.

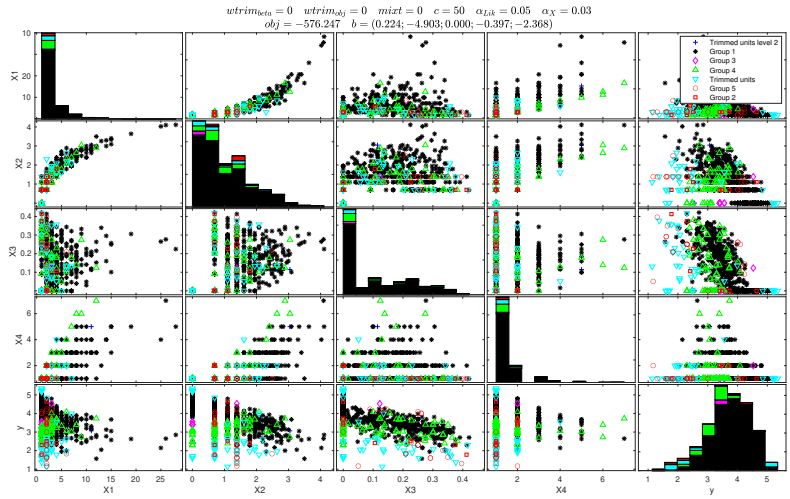


Fig. 1 Scatter plot of a subsample of approximately 1000 points of customers. The marginal histograms are for the whole set of customers

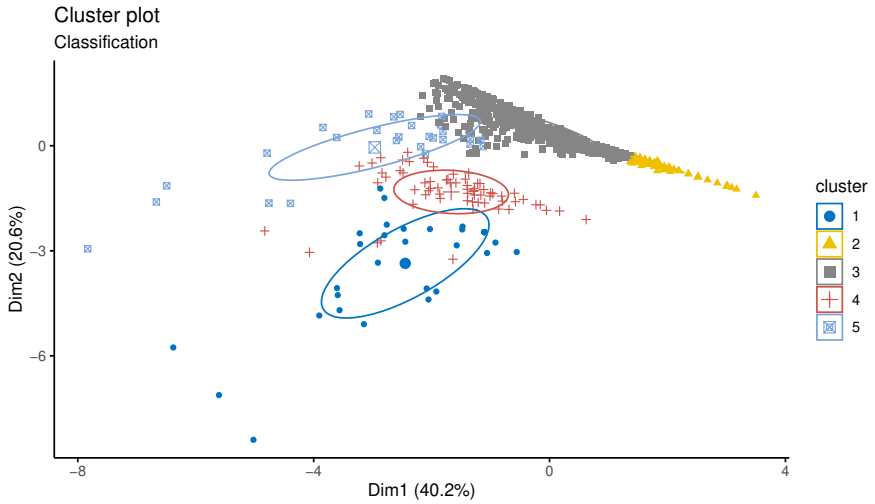


Fig. 2 Cluster shapes for the allocation with the 5 clusters and little overlapping. About 60% of the overall variance is captured in the first 2 principal components

4 Final comments

The paper offers a glimpse on the flexibility of a novel method for clustering which overcomes many restrictions of the popular k-means. The TCLUST-REG has many advantages: does not require elliptical distributions; allows for correlated input variables; does not give clusters of similar size and shape; causal-effect models can be introduces; is robust to outliers; the order of the inputs does not affect the final allocation. A numerically efficient implementation of the TCLUST-REG routine is available in the Matlab toolbox FSDA, whose first appearance is in [7].

There is further research to undertake. We have analysed a subset of all possible explanatory variables and we have not discussed the importance of each predictor in terms of the final classification. The proposed methodology is not limited to normal mixtures, and other mixtures where a scatter matrix Σ is included in the definition of the k component-specific multivariate distributions, may benefit from its use. Some examples are mixtures of t distributions or contaminated normal distributions. A very detailed discussion of business implications will be given.

The proposed methodology is very flexible because it can cope with very different types of data thanks to the flexibility that the tuning parameters entail. Although often the choices of the tuning parameters are motivated by the final clustering purposes, certain guidance about how to choose all of them is sometimes required.

References

1. Alon, I., Jaffe, E., Prange, C., Vianelli D.: Global marketing. Routledge, New York (2021)
2. Kotler, P., Keller, K.L.: Marketing management. Pearson Education, Harlow (2012)
3. García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar A. A general trimming approach to robust cluster analysis. Ann Stat **36** 1324–1345 (2008)
4. García-Escudero L.A., Gordaliza A., Mayo-Iscar A., San Martin R. Robust clusterwise linear regression through trimming. Comput Stat Data Anal **54** 3057–3069 (2010)
5. García-Escudero L.A., Mayo-Iscar A., Riani, M. Model-based clustering with determinant-and-shape constraint, Statistics and Computing, **30**, 1363–1380 (2020).
6. Gordaliza A. Best approximations to random variables based on trimming procedures. J Approx Theory **64** 162–180 (1991)
7. Riani M., Perrotta D., Torti F. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration Chem Intell Lab Sys **116** 17–32 (2012)

Multivariate prediction models: Altman's Z-Score and CNDCEC's sectoral indicators

Modelli multivariati di previsione: lo Z-Score di Altman e gli indicatori settoriali del CNDCEC

Alessandro Danovi, Alberto Falini, Massimo Postiglione

Abstract In line with EU Directive 2019/1023 on early warning provisions, the recent reform of Italian Bankruptcy Law introduced a new procedure ("*procedura di allerta*") for Italian companies. This procedure asks companies to monitor a selection of sectoral indexes that would allow the early detection of corporate crisis. The model applied sits between multivariate and univariate forecasting models. This paper aims to evaluate the applicability of a different multivariate approach, comparing the adopted model with one of the most successful multivariate insolvency forecasting models, that was proposed by E.I. Altman.

Abstract *Allineandosi alla Direttiva UE 1023/2019 in merito agli "early warning", la recente riforma della Legge Fallimentare italiana ha introdotto una nuova procedura (c.d. "procedura di allerta") per le imprese. Tale procedura impone alle società di monitorare una serie di indicatori settoriali finalizzati ad individuare in via preventiva la possibile crisi di impresa. In termini applicativi, il modello si colloca fra i modelli di previsione univariati e quelli multivariati. L'obiettivo dell'elaborato è valutare l'applicabilità di un approccio multivariato al sistema di allerta confrontandolo con uno dei modelli multivariati più di successo, ovvero quello proposto da Altman.*

Key words: prevision, multivariate, insolvency.

[1] Prof. Alessandro Danovi, Università degli Studi di Bergamo; alessandro.danovi@unibg.it.
Prof. Alberto Falini, Università degli Studi di Brescia; alberto.falini@unibs.it.
Dott. Massimo Postiglione, dottore magistrale in Management; massimoposti@icloud.com.

1 The insolvency prediction: multivariate and univariate models

The first studies of insolvency prediction models date back to the Twenties with the so-called univariate approach, which consists in a model that aims to establish the condition of the economic-financial equilibrium of a company using a system of accounting indicators in order to identify the most suitable variable to represent a firm's weaknesses. This method is characterised by the non-attributability of the results obtained to a single summary value. Therefore, there is no overall evaluation, typically represented by a score rating, although nothing prohibits the application of this approach through a system of several ratios investigated separately one from each other. In this regard, the contribution of Beaver (1966), who weighted the accuracy of some indicators in terms of prediction of a possible state of insolvency, is one of the most relevant[1].

However, the critical element of univariate models consists in the absence of a criterion in order to aggregate the results obtained (Varetto, 1999). To overcome this limitation, a defined multivariate approach was conceived. Its aim was to investigate qualitatively the meaning of the indicators through a single score, which has to be suitable to express an overall rating. The Author who probably contributed the most to the development of such a multivariate model was certainly E. I. Altman who, more than 50 years ago, created the so-called Z-Score (Altman, 1968). This model was improved and adapted several times to different companies (Altman, 2018) with the contribution of other illustrious researchers (Altman, Haldeman and Narayanan, 1977 and Altman, 1993)[2].

These studies were so successful that they captured the attention of financial intermediaries, who were eager to create a method able to accurately determine creditworthiness[3]. From a business point of view, the evolution of this research has been proved particularly successful in turnaround strategies, expanding the interest in the management on the company's state of insolvency (Chowdhury, 2002) whose foundations came from the contributions of multiple authors like Schendel, Patton and Riggs (1976), Guatri (1985), Weitzel and Jonsson (1989), Krueger and Willard (1991), Bhave (1994).

1.1 The Altman Z-Score (1993)

Altman (1993) has made a decisive contribution to the insolvency prediction theme. Among all the multivariate approaches offered by a lot of authors, this is

[1] The Author, analysing a sample made of hundreds unstable companies, concluded that there is a proven systematic difference between the values obtained by healthy companies compared to distressed companies.

[2] For a more complete story of Z-Score, see Altman, 2018. An interesting review of insolvency prediction methodology can be found in Bellovary, J. L., Giacomino, D., Akers, M. (2007).

[3] Creditworthiness is a judgment made by financial intermediaries regarding the probability of the debtor's default.

not probably the most known, but it is also one of the most successful in terms of practical applications (Chiaramonte, Croci e Poli, 2015). The formula is as follows:

$$Z = 0{,}717 \bullet X1 + 0{,}847 \bullet X2 + 3{,}107 \bullet X3 + 0{,}420 \bullet X4 + 0{,}998 \bullet X5$$

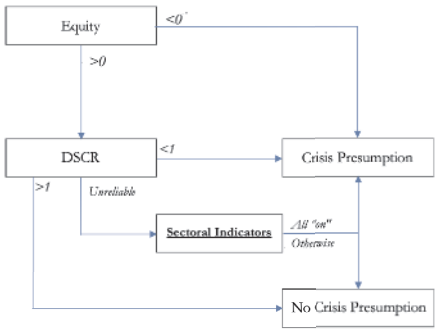
X1 = Net Working Capital on Invested Capital ratio;
X2 = Profit reserves on Invested Capital ratio;
X3 = EBIT on Invested Capital ratio;
X4 = Fairness at market value on book value of payables ratio;
X5 = Revenues on Invested Capital ratio.

Altman identifies a 1,23 threshold below which the firm will declare insolvency.

1.2 The CNDCEC Model (2019)

According to the Art. 13 of the Italian new Crisis and Insolvency Code, in 2019 the CNDCEC[1] was required to determine suitable indicators to reveal the impending state of crisis for a company[2]. The result is a set of seven ratios calculated on a sample of approximately one million financial statements in relation to the operating periods between 2010 and 2015. The model is based on a hierarchical scheme (fig. 1):

Figure 1: The Italian alert system



The approach is sequential and progressive and spreads out the analysis over several stages starting with a check on the value of equity. If it is negative, this can immediately lead to the presumption of a state of business crisis. On the other hand, if it is positive, a test is performed on the DSCR index, with three different possible outcomes:

[1] CNDCEC stands for "*Consiglio Nazionale dei Dottori Commercialisti ed Esperti Contabili*" which is the Italian National Council of Chartered Public Accountants.

[2] According to the new Code definition, crisis is a much broader concept of insolvency. In fact, the latter can be identified as the creditor's inability to fulfill his obligations due to expire. Crisis is defined in terms of likelihood of insolvency between six months.

1. DSCR > 1: eliminates the presumption of a state of crisis;
2. DSCR < 1: leads directly to the presumption of a state of crisis;
3. DSCR unreliable: involves a second level of analysis through sectoral indicators.

The use of sectoral indicators must take place in compliance with pre-established critical thresholds classified for ten types of businesses. There are 5 sectoral indices and their calculation method is as follow:

X1 = Short-term assets on short-term liabilities ratio;
X2 = Cash flow from operations on total assets ratio;
X3 = Financial expense (borrowing costs) on sales ratio;
X4 = Equity on debts ratio;
X5 = Social security debts on total assets ratio.

These ratios "turn on" when certain critical thresholds, identified by type of sector and reviewed periodically by the CNDCEC, are exceeded. When the company exceeds simultaneously all the limits set by these five indicators, the presumption of a state of crisis becomes concrete.

2 Research Methods

In order to evaluate the CNDCEC model, we tried to compare its results with Altman 1993 model by a deep empirical analysis using data from companies that were admitted to Extraordinary Administration, which is an Italian procedure to manage bankruptcy of large companies. As Prof. Altman suggested, all financial statements of the companies admitted to the Extraordinary Administration procedure from 2013 to 2019 were considered[1]. Therefore, both the variables of the Z-Score model and the sectoral alert thresholds of the CNDCEC model were calculated. In both cases, only the values obtained starting from the 4th year before the beginning of the Procedure were considered.

The results of the alert thresholds, in order to come to an overall rating like any multivariate model, were weighted for each year of analysis (T-period). The final score was reached as follows:

$$0.2 \cdot X1\,(T) + 0.2 \cdot X2\,(T) + 0.2 \cdot X3\,(T) + 0.2 \cdot X4\,(T) + 0.2 \cdot X5\,(T) = R\,(T)$$

It is emphasized that the choice to distribute the weights in an equitable way (20% for each variable) was adopted according to the model criteria, i.e. the positive outcome of any ratio (threshold not exceeded) is sufficient to exclude the

[1] These are large Italian companies that experienced full-blown distress manifestation before the current crisis brought by the COVID-19 pandemic, in order to not rely on data altered by non-ordinary systematic phenomena.

reliability of the state of crisis, attributing the same importance to each variable in order to identify the final result.

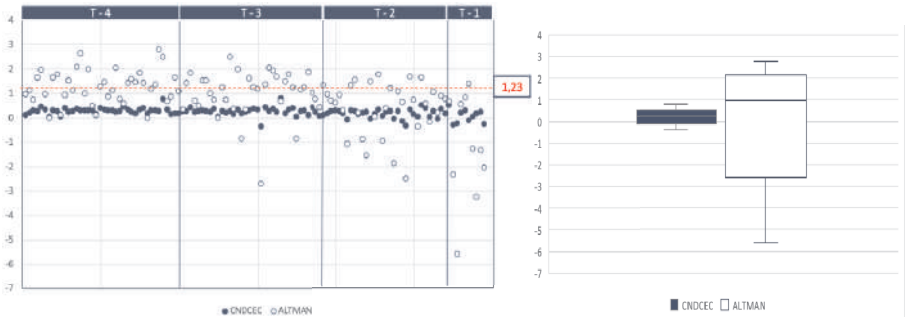
The results obtained from the application of the Z-Score Model and the proposed CNDCEC Multivariate Model (hereinafter also the "CNDCEC Model") were analyzed by elaborating:

1. The average;
2. The Standard Deviation
3. The normal values attributed to each result;
4. The normal distribution attributed (Gaussian) to each result as a function of the values obtained in the three previous items.

3 Results and conclusions

The comparison between the different results highlights a tendency of the CNDCEC model to report extremely homogeneous values, regardless of the analyzed T-period. On the other hand, the Altman model shows a decreasing trend that, getting closer to the start of the Procedure, identifies an increasing number of cases with lower values to the pre-established threshold Z=1,23 (fig. 2.*a.*). This should obviously be linked to the fact that Altman's indicators were calculated many years ago and not specifically for Italian companies.

Figure 2: Comparison between CNDCEC and Altman results (*a*) and relative Box-plots (*b*)

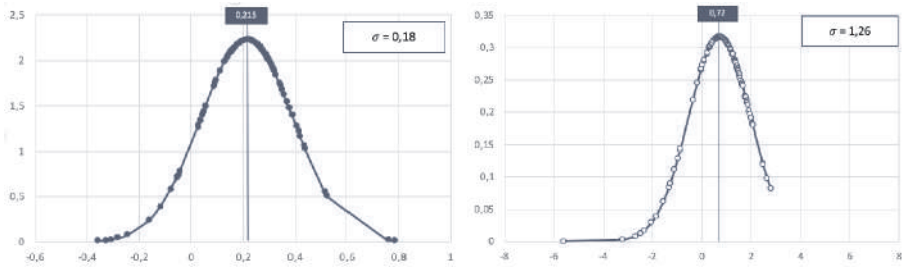


(a) *(b)*

The CNDCEC model reports values with a very low dispersion if compared with the Altman's model. This conclusion is confirmed by the box plot (fig. 2.*b.*), which represents the distance between the minimum value, the 1st quartile, the 2nd quartile (median), the 3rd quartile and the maximum value.

Analyzing both models in terms of results variability, the CNDCEC Model has a much higher accuracy than Altman's Model, whose standard deviation (1,26) results 7 times higher than the first one (0,18) (fig. 3). Therefore, investigating the normal distribution of the CNDCEC Model and assuming a threshold slightly above the average value of its results (0,215), it can identify all the financial statements of the companies that will be admitted to the Procedure, regardless of the forecast T-period, within a range of 4 years.

Figure 3: Gauss distribution for CNDCEC's Model (*a*) and Altman's Model (*b*)



(a) *(b)*

In conclusion, the applicability of a multivariate version of the CNDCEC model is potentially very effective if applied only to companies which are admitted to the Extraordinary Administration. However, as this cannot be considered an unbiased sample from several points of view (like the company size), we will continue this research investigating the CNDCEC model results related to its application to healthy companies, in order to establish whether this trend is an exclusive feature of companies admitted to the Procedure or not.

References

1. Altman, E. I. (1968). "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy". *The journal of finance* 23 (4), 589-609.
2. Altman, E. I., Haldeman, R. G., Narayanan, P. (1977). Zeta™ analysis a new model to identify bankruptcy risk of corporations". *Journal of banking & finance,* 1 (1), 29-54.
3. Altman, E. I. (1993). "Corporate Financial Distress and Bankruptcy". John Wiley & Sons, II Edition, New York.
4. Altman E. I (2018). "A fifty-year retrospective on credit risk models, the Altman Z-score family of models and their applications to financial markets and managerial strategies". *Journal of Credit Risk*, 14 (4), 1-34.
5. Beaver, W. H. (1966). "Financial ratios as predictors of failure". *Journal of accounting research*, 71-111.
6. Bellovary, J. L., Giacomino, D., Akers, M. (2007). "A review of bankruptcy prediction studies: 1930 to present". *Journal of Financial Education,* 33, 1-42.
7. Bhave, M. P. (1994). "A process model of entrepreneurial venture creation." *Journal of business venturing*, 9 (3), 223-242.
8. Chiaramonte, L., Croci, E., Poli, F. (2015). "Should we trust the Z-score? Evidence from the European Banking Industry." *Global Finance Journal* 28, 111-131.
9. Chowdhury, S. D. (2002). "Turnarounds: A stage theory perspective". *Canadian Journal of Administrative Sciences*, 19 (3), 249-266.
10. Guatri, L. (1985). "Le crisi d'impresa". *Rivista milanese di economia,* 13, 5-8.
11. Krueger, D. A., Willard, G. E. (1991). "Turnarounds: a process, not an event". *Academy of a management proceedings*, 26-30.
12. Schendel, D., Patton, G. R., Riggs, J. (1976). "Corporate turnaround strategies: a study of profit decline and recovery". *Journal of general management,* 3 (3), 3-11.
13. Weitzel, W., Jonsson, E. (1989). "Decline in organizations: a literature integration and extension". *Administrative science quarterly,* 91-109.
14. Varetto, F. (1999). "Metodi di previsione delle insolvenza: un'analisi comparata." *Il rischio Creditizio. Misura e Controllo*: 178-301.

Comparing Entrepreneurship and Perceived Quality of Life in the European Smart Cities: a "Posetic" Approach

Analisi comparativa dell'imprenditorialità e della qualità della vita percepita nelle smart city europee: l'approccio su Poset

Lara Penco, Enrico Ivaldi and Andrea Ciacci

Abstract This work deals with the issues of entrepreneurship and perceived well-being in European smart cities. The aim is to analyse the relationships between the different subjective well-being dimensions and the smart cities entrepreneurial performance. By using the non-aggregative quantitative method known as POSET, we compare the different European smart cities. This non-aggregative method allows us to realise comparisons based on the single elementary indicators that make up the various dimensions.

Abstract *Questo lavoro affronta i temi dell'imprenditorialità e del benessere percepito nelle smart city europee. L'obiettivo è quello di analizzare le relazioni tra le diverse dimensioni di benessere soggettivo e le performance imprenditoriali delle smart city. Utilizzando il metodo quantitativo non aggregativo noto come POSET, confrontiamo le diverse smart city europee. Questo metodo non aggregativo permette di realizzare confronti a partire dai singoli indicatori elementari che compongono le diverse dimensioni delle smart city. Attraverso diversi processi di computazione abbiamo ottenuto una misura sintetica di performance per le dimensioni costitutive delle differenti smart city.*

Keywords: POSET, Entrepreneurship, Well-being, Smart City, Europe.

1. Introduction

Recent research has started to analyse cities as an environment for entrepreneurship (Glaeser et al., 2014). Entrepreneurial activity (especially the Schumpeterian type) is more concentrated and clustered than manufacturing industries (Adler et al., 2019). Cities are an appropriate environment for entrepreneurship (Szerb et al., 2013), providing a relevant socio-economic and institutional context for the entrepreneurial

[1] Lara Penco, Department of Economics; lara.penco@economia.unige.it
Enrico Ivaldi, Department of Political Science; enrico.ivaldi@unige.it
Andrea Ciacci, Department of Economics; andrea.ciacci@edu.unige.it

ecosystem (Audretsch at al., 2015). The extant literature on the urban entrepreneurship is focused on the analysis of the characteristics of the urban ecosystem in terms of Regional System of Entrepreneurship (Szerb et al., 2013; Acs et al., 2014), while only few contributions are focused on the perceived quality of life at city level as a driver of entrepreneurial activities research (Penning; 1982; Florida et al., 2013; Audretsch & Beliski, 2017).
In this vein, the aim of this research is to address the following research questions (RQ): RQ1: *is the perceived urban quality of life a stimulus for entrepreneurship at EU level?* (RQ1) and *which quality of life Smart City dimension is more conducive to stimulate the urban entrepreneurship at EU level?* (RQ2).
In this work, we provide a concrete procedure to perform data analysis and to compute synthetic indexes for policy and decision making. Compared to mainstream evaluation approaches, the peculiarity is represented by the non-aggregative method of analysis and by the systematic use of Partially Ordered Set (POSET), a set of algebraic and combinatory tools designed to properly treat order relations (Fattore, 2017). "Posetic" approach is employed to conduct quantitative analysis on multidimensional systems (Carlsen & Bruggemann, 2014; Ivaldi et al., 2020a).

2. Methodology

Consistent with Penco et al. (2020) and Bruzzi et al. (2019), the sample comprises 43 EU cities involved in smart city policies. They were divided into three groups depending on their population amount (quartile allocation on the initial sample):
Large Cities (N=11): London, Paris, Madrid, Barcelona, Berlin, Rome, Athens, Warsaw, Manchester, Hamburg, and Budapest;
Medium-sized Cities (N=21): Lisbon, Munich, Vienna, Stuttgart, Amsterdam, Lille, Frankfurt, Prague, Brussels, Turin, Bucharest, Stockholm, Copenhagen, Dublin, Glasgow, Sofia, Helsinki, Bordeaux, Düsseldorf, Krakow, and Dresden;
Small Cities (N=11): Malmo, Zagreb, Cardiff, Vilnius, Karlsruhe, Riga, Bratislava, Tallinn, Luxembourg, Ljubljana, and Valletta.

2.1 Variables selection

Based on sub-indexes connected to smart city dimensions, we develop an innovative multidimensional index (**UHI**—Urban Happiness Index) to better explain the different dimensions/measures of the perceived (subjective) quality of life (Eurostat, 2015):

- ECO: Ease in finding a job; Satisfaction with the financial situation of your household; Difficulty paying bills at the end of the month (inverted value); Satisfaction with the efficiency of administrative services of your city;
- GOV: Satisfaction with the way your city spends its resources; Satisfaction with the public administration;
- PEO: Satisfaction with the integration of foreigners who live in your city; Satisfaction about the presence of foreigners in your city; Satisfaction with the credibility of individuals;
- WEL: Satisfaction with the healthcare services offered by doctors and hospitals; Satisfaction with the safety of your neighbourhood; Satisfaction with education and training in your city:

- CUL: Satisfaction with schools in your city; Satisfaction with cultural facilities; Satisfaction with cinemas in your city;
- ENV: Satisfaction about the presence of green spaces in your city; Satisfaction with the quality of air in your city; Satisfaction with the cleanliness in your city
- ICT&MOB: Satisfaction with the Internet access in your city; Satisfaction with the Internet access at home in your city; Satisfaction with public transport in your city.

ENT comprehends the following objective measures:

- STAR: Average monthly increase in the number of startups (Teleport[1], 2019); INV: Number of investors (Teleport, 2019); ACC: Accelerated startups (Gust[2], 2016 and Seed DB, 2016); UNI: History of highly successful digital companies (unicorns) (GP Bullhound, 2016; CB Insights, 2016). All normalized by the number of inhabitants.

2.2 From variables to sub-indexes and indexes

The methodology employed in this study is based on a formative approach, according to which the latent factor (entrepreneurship and well-being) depends on the indicators that "explain" the factor, and not vice versa (Diamantopoulos et al. 2008). This implies the variables are functional to the definition of the phenomenon (Maggino, 2017). Our analysis is based on the POSET method, a quantitative non-aggregative analytical approach used to compare different statistical units (Ivaldi et al., 2020a). POSET makes it possible to establish unequivocally whether it is appropriate to compare the statistical units of a distribution. Graphically, POSET is represented through a tool called Hasse diagram. It is composed by the following elements: the nodes, which correspond to profiles, each one identifying a statistical unit; a connector, or line, which unites the comparable elements, while it is absent if it is not possible to determine with certainty the prevalence of one statistical unit over the other. The existence of incomparability is a direct consequence of the complexity and multiformity of a given phenomenon.

According to Fattore (2017), after assigning the evaluation scores to the elements of a POSET, they can be linearly ordered. In other words, we can assign a score to each element of a finite POSET; the score represents the position of the elements in a "low-high" axis (Fattore et al., 2019). After having arranged all the linear extensions, we must calculate the height of each linear extension, defined as 1 plus the number of elements below x in the linear order; then, for each element, we compute the average height [0, 1] on its linear extensions, corresponding to the arithmetic mean of the heights of x in all linear extensions (*ibidem*).

3. Results and Discussion

3.1 Average height values

[1] https://teleport.org/ (data were collected in November 2019).

[2] Data have been obtained from Bannerjee et al. (2016). It refers to the following sources: (http://gust.com/accelerator_reports/2016/global; /https://www.seed-db.com/accelerators/view?acceleratorid=3012; GP Bullhound (2016): www.gpbullhound.com/wp-content/uploads/2016/06/GP-Bullhound-Research-European-Unicorns-2016-Survival-of-the-fittest.pdf; CB Insights2016: www.cbinsights.com/researchunicorncompanies).

The application of the average height allowed us to obtain synthetic performance measures of the different cities. We measured the performance for each city group. Higher values identify better situations and vice-versa.
Among the Large cities, London and Berlin exhibited the ENT best performance (average height=10.5). London is also characterized by a positive perception of quality of life (UHI=10.1). The Medium-sized cities that present the best ENT average height are Amsterdam (19.4), Copenhagen (18), and Dublin (17.9). Copenhagen and Dublin also rank positively in terms of UHI, recording an average height equal to 15.9 and 15.8, respectively. The best Small cities are Tallinn (10.8), Valletta (9.5), and Luxembourg (8.6); nevertheless, only Luxembourg presents a good average height in terms of UHI (7.1).

3.2 Correlation analysis

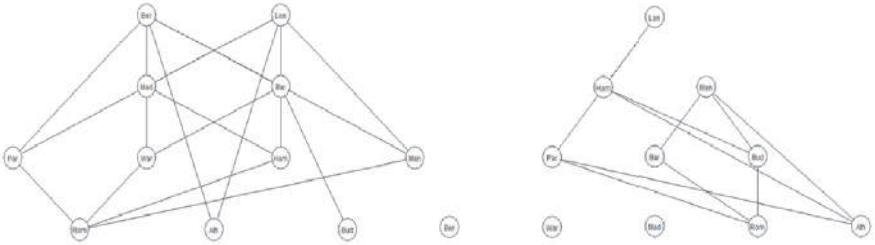
In order to answer to RQ1, a Spearman's correlation analysis was performed. The Spearman's index value of 0.09 reveals that there is no correlation between ENT and UHI for the entire sample (43 cities). Hence, we decided to distinctly analyze each group of cities. In Large cities, we found a significant Spearman's correlation between ENT and UHI. In this group we also found a positive relationship between entrepreneurship and the population (0.7), while the correlation between UHI and population was not significant. For Medium-sized and Small cities, all the relationships were not significant and were sometimes negative.
To identify which important quality of life dimensions trigger entrepreneurial activities (RQ2), we employed a Spearman's correlation matrix for each group of cities.
For Large cities, the correlation between ENT and WEL was highly significant and positive (0.7). A positive evaluation by the citizens of welfare, transport, and environment services can provide an "economic leverage," as seen empirically in Ivaldi et al. (2020b). There was a positive correlation between entrepreneurship and the perceived quality of life in terms of ECO (0.5), ENV (0.6), and ICT/MOB (0.5). The correlations of ENT with the other dimensions did not exist (i.e., CUL) or were not particularly significant (i.e., GOV and PEO). For Medium-sized cities the Spearman's correlation matrix did not identify any significant relationships, as well as for the Small cities.

3.3 The hierarchy for each group

To better understand ENT and UHI, we used the Hasse diagram. It helps to understand the city partial position for each sub-dimension by disposing the cities from the top (the best cities) to the lower level (the worst cities). In this paragraph, we only present the Hasse diagrams associated to Large Cities (figure 1 and 2).

Figure 1: big cities ENT Hasse diagram

Figure 2: big cities UHI Hasse diagram



4. Conclusions and implications

This study examines the relationship between the perceived quality of life and entrepreneurship at the urban level on a sample of 43 smart cities. The choice to analyze smart city dimensions derives from the fact that smart city policies are considered a mean of urban development and of quality of life creation (Dameri et al., 2019).

Correlation analysis demonstrates that only in Large cities entrepreneurship is positively related to the perceived quality of life (RQ1). For the Large cities, the dimensions that help to enhance the entrepreneurial attitude are the positive perception of welfare (WEL), economy (ECO), environment (ENV) and ICT and mobility dimensions (RQ2). Cities showing both high levels of ENT and UHI represent the ideal environment aimed at the creation of virtual circle for the entrepreneurial activity prosperity and citizen satisfaction. These cities have implemented explicit policies to earn this 'status'.

Our research advances knowledge to entrepreneurship literature by bringing together Regional System of Entrepreneurship (Szerb et al., 2013; Acs et al., 2014), quality of life and well-being at city level (Florida et al., 2013) and smart city (Nam & Pardo, 2010) and to develop synthetic indexes and test the relationship between the perceived quality of life and entrepreneurship. In addition, this contribution identifies the most important drivers of quality of life that stimulate innovative activities at the urban level. Moreover, it provides an innovative classification of EU smart cities.

In terms of practical implications, the contribution helps to explain to policy makers and city managers the importance of the perceived quality of life and the most important drivers for the creation of an attractive entrepreneurial environment. This study has some inherent limitations to be addressed by future research. First, the investigation is performed on EU cities alone. Second, our cross-sectional analysis leaves causality issues between UHI and ENT for future research. In addition, the number of variables and attributes that refer to each dimension may be expanded toward a better understanding of the determinants affecting the development of urban-level entrepreneurship.

References

1. Acs, Z.J., Autio, E., Szerb, L. (2014). National Systems of Entrepreneurship: Measurement Issues and Policy Implications. *Research Policy*, 43, 476–449. DOI: 10.1016/j.respol.2013.08.016.
2. Adler, P., Florida, R., King, K., Mellander, C. (2019). The city and high-tech startups: The spatial organization of Schumpeterian entrepreneurship. *Cities*, 87: 121-130. DOI: 10.1016/j.cities.2018.12.013.
3. Audretsch, D. B., Belitski, M. (2017). Entrepreneurial ecosystems in cities: establishing the framework conditions. *Journal of Technology Transfer*, 42 (5): 1030-1051. ISSN 1573-7047. DOI: 10.1007/s10961-016-9473-8.
4. Audretsch, D.B., Heger, D., Veith, T. (2015). Infrastructure and entrepreneurship. *Small Business Economics*, 44(2), 219-230. DOI: 10.1007/s11187-014-9600-6.
5. Bannerjee, J. Bone, Y. Finger, (2016), European Digital City Index – Methodology Report. Nesta Report, Article 978-1-84875-153-8.
6. Bruzzi, C., Ivaldi, E., Musso E., Penco, L. (2019). The Role of Knowledge City Features in Nurturing Entrepreneurship: Evidence from EU Cities. In M.N. Iftikhar, J. Justice, and D. Audretsch (eds.), *Urban Studies and Entrepreneurship*, The Urban Book Series. Cham: Springer. DOI: 10.1007/978-3-030-15164-5_4.
7. Carlsen, L., Bruggemann, R. (2014). The 'Failed State Index' Offers more than just a simple ranking. *Social Indicators Research*, 115, 525–530. DOI: 10.1007/s11205-012-9999-6.
8. Dameri, R. P., Benevolo, C., Veglianti, E., Li, Y. (2019). Understanding smart cities as a glocal strategy: A comparison between Italy and China. *Technological Forecasting and Social Change*, 142: 26-41. DOI: 10.1016/j.techfore.2018.07.025
9. Diamantopoulos, A., Riefler, P., Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, 61, 1203–1218. DOI: 10.1016/j.jbusres.2008.01.009.
10. Eurostat. (2015). *Quality of life in European Cities 2015. Regional and Urban Policy.* Luxembourg: Publications Office of the European Union, Belgium. DOI: 10.2776/870421.
11. Fattore M. (2017). Functionals and synthetic indicators over finite Posets. In M. Fattore and R. Bruggemann (eds.), *Partial order concepts in applied sciences*. pp. 71-86. Cham: Springer AG.
12. Fattore, M., Arcagni, A., Maggino, F. (2019). Optimal Scoring of Partially Ordered Data, with an Application to the Ranking of Smart Cities. In G. Arbia, S. Peluso, A. Pini, and G. Rivellini (eds.), SIS 2019 – *Smart Statistics for Smart Applications*. pp. 855-860. Milano: Pearson. Società Italiana di Statistica.
13. Florida, R., Mellander, C., Rentfrow, P.J. (2013). The happiness of cities. *Regional Studies*, 47(4), 613-627. DOI: 10.1080/00343404.2011.589830.
14. Glaeser, E. L., Ponzetto, G., Tobio, K. (2014). Cities, skills and regional change. *Regional Studies*, 48: 7–43. DOI: 10.1080/00343404.2012.674637.
15. Ivaldi, E.; Ciacci, A.; Soliani, R. (2020a) Urban deprivation in Argentina: A POSET analysis. *Pap. Reg. Sci.*, 99, 1723–1747. doi:10.1111/pirs.12555.
16. Ivaldi, E., Penco, L., Isola, G., Musso, E. (2020b). Smart Sustainable Cities and the Urban Knowledge-Based Economy: A NUTS3 Level Analysis. *Social Indicators Research*. DOI: 10.1007/s11205-020-02292-0.
17. Maggino, F. (2017). Developing indicators and managing the complexity. In F. Maggino (ed.), *Complexity in society: From indicators construction to their synthesis*. Pp. 87-114. Cham: Springer.
18. Nam, T., Pardo, T.A. (2011). *Conceptualizing smart city with dimensions of technology, people, and institutions.* In 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times. pp. 282-291. New York: ACM. DOI: 10.1145/2037556.2037602.
19. Penco, L., Ivaldi, E., Bruzzi, C., Musso, E. (2020). Knowledge-based urban environments and entrepreneurship: Inside EU cities. *Cities*, 96. DOI: 10.1016/j.cities.2019.102443.
20. Pennings, J.M. (1982). The urban quality of life and entrepreneurship. *Academy of Management Journal*, 25(1), 63-79. DOI: 10.5465/256024.
21. Szerb, L., Acs, Z., Autio, E., Ortega-Argiles, R., Komlosi, E. (2013). *REDI: The Regional Entrepreneurship and Development Index – Measuring regional entrepreneurship. Final Report*. European Commission, Directorate-General for Regional and Urban policy. REGIO DG 02 – Communication.

The Relationship between Business Economics and Statistics: Taking Stock and Ways Forward

Il Rapporto tra le Scienze Economico-Aziendali e le Scienze Statistiche: Verso Nuovi Modelli di Interazione?

Amedeo Pugliese

Abstract

The last two decades have witnessed a steady increase in terms of proximity between business economics and statistics. On the demand side, the process has been sparked by firms and institutions seeking out for graduates able to handle complex organizational problems through data analysis in support of decision-making. On the supply side, the relatively ease in accessing data prompted the quest for complex tools to address impending issues. Alongside with such closeness in education and learning, research in business economics has steadily shifted towards empirical analyses in lieu of more conceptual and interpretative research. I suggest that the next iteration in the relationship between the two disciplines should evolve towards a *problem-centric* approach whereby key societal, business, operational issues are tackled jointly, rather than in silos by the two disciplines. The benefits of this approach are discussed.

Abstract

Negli ultimi anni l'integrazione tra gli ambiti business e statistico è cresciuta esponenzialmente. Il processo è spinto da aziende e istituzioni alla ricerca di laureati in grado di gestire complesse problematiche organizzative attraverso l'analisi dei dati a supporto del processo decisionale. Dal lato dell'offerta, la relativa facilità di accesso ai dati ha spinto alla ricerca di strumenti per affrontare problemi complessi. Accanto a tale prossimitànell'ambito della formazione, la ricerca in economia aziendale si è spostata verso analisi empiriche al posto di una ricerca più concettuale e interpretativa. Suggerisco che la prossima iterazione nella relazione tra le due discipline dovrebbe evolversi verso un approccio incentrato sui problemi in base al quale le questioni chiave della società, degli affari e delle

[1] Amedeo Pugliese, Univeristà degli Studi di Padova &UniversitatPompeuFabra; email:amedeo.pugliese@unipd.it

operazioni vengono affrontate congiuntamente, piuttosto che in silos, dalle due discipline.

Key words: accounting, business, complexity, data analysis, education

1. Introduction

The last few years witnessed a growing proximity between the disciplines of business economics and statistics. The two fields, whilst autonomous and distinct in terms of objects of inquiry, methods and epistemology are becoming increasingly intertwined both in terms of research and scientific endeavours, as well as in the domain of education. Undoubtedly, the push factors are the availability of large amount of data that feed into the decision-making process of managers and line staff, spanning across multiple levels within organizations. Similarly, on the research side, the steady growth of empirical research has led the Italian community of researchers in business economics (both in AIDEA and SIDREA) to embrace different epistemology and start engaging with tools and instruments typically employed in the fields of statistics and econometrics.

In this short essay, I argue that whilst the integration is tangible, a lot more could be done both in terms of education and research. Specifically, the two disciplines should come closer together in trying to address important economic and social issues: the current health and economic crisis offers a clear and tangible example of an area in which both could offer a significant contribution, and incrementally so if a strong integration would be pursued, abstracting for the oftentimes harmful boundaries imposed by the *scientific sectors* (e.g.settori scientifico-disciplinari).

2. Taking Stock of Research in Business Economics

Since the introduction, in 2003, of nation-wide systems to assess research productivity on the basis of the 'quality' of published works (Bertoni, Brunello, Cecchi & Rocco, 2021), the Italian academy of scholars in business economics is facing growing challenges, possibly due to two reasons: (a) the call for a switch from an inductive, mainly descriptive and case-led approach to research,towards a deductive and positivistic approach, thus departing from a normative frame, largely dictated by the professional twist of the discipline; (b) next, the breadth of the scope of the discipline makes it hard converging towards agreed upon models in terms of how to conduct high quality research. A case in point is the heterogeneity in terms of objects and topics, spanning from capital market research to auditing, from public management to not-for-profit organizations, from corporate governance to behavioral accounting (Bond, Clout, Czernkowski, & Wright, 2021).

More pointedly, within the accounting domain – the one attracting the most of the researchers' attention within the 'business economics' sub-field, the split between financial and management accounting is even more pronounced because of the preferences for more quantitative or qualitative approaches, with different aims, targets and scope of applicability for firms and their managers.

Inadvertently, the exogenous shock imposed by the introduction of evaluation criteria and the parallel growth of empirical studies within the international arena have surely provided impetus to the discipline, yet with a noticeable downside: the quest for methodological rigour has been traded off with relevance of research questions and issues contextualized to our setting (Bertoni et al., 2021). This tendency is not solely observable in Italy, but it generalizes to several countries (Bond et al., 2021) and spurred discontent and breaches in terms of appreciating and assessing the progression of the field.

Notably, this is a pressing issue in doctoral programmes in 'business management' in general that tend to 'borrow' courses from statistics colleagues without offering or requiring a contextualized application of the tools taught. This cold fusion endangers both the perceived utility of statistics and inference – if fraught with lack of contextualization - as well as enhancing the distance in terms of the (perceived) usefulness of statistics in terms of addressing interesting questions.

3. Education: Where Integration is Real

An interesting example of a full-blown integration between the two disciplines is the education setting. Curricula in business administration, and even more in accounting and finance heavily rely on courses centered around statistical analysis and tools. The synergy is well appreciated since at least three decades at the national level and it appears to be growing even more in light of the latest calls from practice in terms of the skills and ability required to our graduands.

The accounting courses are a case in point in this regard, in two ways: the emergence of fair value accounting in evaluating financial instruments in companies' financial statements requires a 'probabilistic' approach in estimating the central possible value of an asset with associated probabilities of extreme values on the tails. Even more prominent is the use of business intelligence tools and a statistical approach in the auditing profession: sampling is a the routine approach when performing routine audits in companies engaging with hundreds of thousands of transactions.

At a more general level, the whole decision-processes entertained by managers is more recurrently based on data analytics to learn from past occurrences or making forecasts about future events: once again, the development of ad-hoc tools to support managers in their decision-making, without mastering the topic will offer significant opportunities in terms of post-graduate formation and training.

The latter is an interesting domain, yet ill-explored with few noticeable exceptions within the Italian landscape and surely one opportunity to further expand collaborations and synergies.

4.Towards Integration? A problem-centric Approach

Whilst a certain degree of integration has been reached in both the education and – somewhat less – in the research sphere, one area in which cooperation is still languishing is the so-called 'impact' to the real world. This is difficult to achieve because of the lack of a clear incentive system pushing for more (and truly) interdisciplinary work to tackle key problems requiring the contribution of both disciplines – equally weighted – to address them.

This is what I define the 'problem-centric' approach that recognizes a first and higher order problem that requires a heterogenous skill set to be properly addressed. As it oftentimes happens with crisis, it is timely moving as soon as possible in this direction (Rajgopal, 2020). Some key fundamental issues sparked by the health pandemic require more serious consideration and appreciation from researchers in both disciplines: two noticeable examples are the design of public policy interventions to restore economic viability of corporations adversely affected by the pandemic. Whilst several tools are potentially available to the policy makers, two key problems surface and require *integration*: predicting and simulating the effectiveness and efficacy of alternative tools and instruments – this is typical of the statistical sciences – as well as an embedded and first-hand knowledge of the corporations' financial structure, business models and how mangers and owners react differentially to the crisis. The latter is a typical object of analysis of business economics.

Overall, we need more analyses on the Italian context, empirical assessment on a large scale as well as on the outliers. We shall present ourselves as jointly trying to address these systemic and higher order problems. We shall try to set on a shared agenda, define the most urgent issues, split areas of competence and offer a synthesis with potential solutions and thus make impact on firms, entrepreneurs and the society at large. The need spans across multiple areas, from regulation to infrastructures, financing and policy support.

Needless to say, the current incentive system does not help much in moving in this direction. Career concerns drive young and promising researchers look for consensus within the discipline, chasing recognition from those who will be assessing their suitability to move up and ensuring progression. A clear step forward would be giving serious credit and recognition (also) to those who choose to venture in trying to be a bridge with sibling disciplines like business economics and co-operate in applying state-of-the-art methodology and approaches to tackle key and urging issues.

References

1. Bertoni, M., Brunello, G., Cecchi, D., & Rocco, L. 2021. Where do i stand? Assessing researchers' beliefs about their productivity. Journal of Economic Behavior & Organization 185: 61-80.
2. Bond, D., Clout, V. J., Czernkowski, R. M. J., & Wright, A. 2021. Research productivity of Australian accounting academics. Accounting & Finance, 61(1): 1081–1104.

Business and Statistics: towards an Integration?
3. Rajgopal, S. 2020. Integrating Practice into Accounting Research. Management Science. https://doi.org/10.1287/mnsc.2020.3590.
4. Wasserstein, R. L., & Lazar, N. A. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70(2): 129–133.

3.11 Mathematical methods and tools for finance and insurance (AMASES)

On the valuation of the initiation option in a GLWB variable annuity

Valutazione dell'opzione di inizio prelievi in una GLWB variable annuity

Anna Rita Bacinello and Pietro Millossovich

Abstract In this paper we focus on the initiation option featured in many Guaranteed Lifelong Withdrawal Benefit variable annuity contracts, granting their owner the right to decide the age at which lifetime withdrawals should begin. Such contracts have been successfully analysed using a PDE approach. The latter method is elegant, but becomes less viable when the valuation model is more involved and other guarantees are considered. As an alternative, we exploit the Least Squares Monte Carlo approach and model the interaction of the initiation option with lapses and other riders.

Abstract *In questo lavoro valutiamo un'opzione presente in molti contratti di variable annuity con prelievi garantiti per tutta la durata di vita dell'assicurato. Si tratta, precisamente, del diritto di decidere quando (e se) iniziare tali prelievi. Questa opzione è già stata studiata in letteratura tramite equazioni alle derivate parziali. Tuttavia questo metodo, seppur elegante, si rivela inadatto a cogliere aspetti modellistici non eccessivamente semplificati e, in particolare, a tener conto della presenza di altre garanzie o opzioni. Come alternativa, noi proponiamo di utilizzare l'approccio Least Squares Monte Carlo, che ci consente anche di modellare l'interazione tra l'opzione di inizio prelievi e i riscatti.*

Key words: GLWB, initiation option, surrender option, LSMC

Anna Rita Bacinello
Department of Business, Economics, Mathematics and Statistics 'Bruno de Finetti', University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy, e-mail: bacinel@units.it

Pietro Millossovich
Department of Business, Economics, Mathematics and Statistics 'Bruno de Finetti', University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy, e-mail: pietro.millossovich@deams.units.it
Faculty of Actuarial Science and Insurance, The Business School (formerly Cass), City, University of London, 106 Bunhill Row, London EC1Y 8TZ, UK, e-mail: pietro.millossovich.1@city.ac.uk

1 Introduction

In this paper we consider a variable annuity contract with a Guaranteed Lifelong Withdrawal Benefit (GLWB) rider. These contracts, very popular in North-America and Asia since the new century, are unit linked-type vehicles in which the initial premium paid by the policyholder is used to buy units of a well diversified mutual fund in order to build an investment portfolio which remains in the property of the policyholder. The value of such portfolio (*account value*) obviously depends on the investment performance. Moreover, the cost of additional options and guarantees purchased by the policyholder (*insurance fees*), along with *management fees*, are periodically subtracted from this account. Finally, the account value is also affected by possible withdrawals made by the policyholder, if allowed.

The GLWB rider gives the policyholder the possibility to (periodically) withdraw a guaranteed amount from her account for the rest of her life, and if there are no available funds these payments are covered by the insurer. Any remaining funds on the policyholder's death are paid back to her heirs. In addition, since the policyholder does not lose access to the fund, she usually has the possibility to surrender the contract at any time, by receiving the account value net of some surrender penalties.

In our paper we are particularly concerned with the *initiation option*. In many contracts, in fact, there is not a fixed date from which the policyholder can start withdrawals, and a delayed initiation is encouraged by an increase in the guaranteed withdrawal amount. Our goal is to maximize the contract value with respect to the initiation date. We assume that withdrawals, once initiated, are *static*, i.e., exactly equal to the guaranteed amount.

There are many papers that deal with aspects connected to the valuation of the GLWB rider in a variable annuity, some of which are concerned with the optimization of dynamic withdrawals that start immediately at the valuation date. As far as we know, there are instead very few papers dealing with the initiation option. In particular, [3] use Partial Differential Equations to tackle the optimal initiation problem and analyse initiation decisions based on moneyness, while [4] solve the problem using dynamic programming combined with Fourier analysis to approximate the value function, and assume dynamic withdrawals.

Both papers can hardly be generalized to high dimensional frameworks. For this reason we exploit the Least Squares Monte Carlo approach (LSMC), that has been proposed by [2] and [5] for the valuation of purely financial American claims and has been subsequently extended in order to value the surrender option in life insurance contracts (see, e.g., [1] for the case of variable annuities). We also allow for optimal surrender. The optimization problem can then be framed as a double optimal stopping problem, that we transform into a two-stage problem. The advantage of this approach is that complicate models and contract features can be easily accomodated.

The paper is structured as follows: in the next section we describe the contract structure and introduce our notation. In Section 3 we define the optimization problem that has to be solved in order to find the initial contract value. In Section 4 we

outline the numerical algorithms that allow to solve the problem through (repeated) applications of the LSMC approach, all based on the same set of simulations. Section 5 concludes the paper.

2 The contract structure

We consider a variable annuity contract with a GLWB rider issued at time 0, when the policyholder pays a single premium P that is entirely invested in a well-diversified and non-dividend paying mutual fund. We denote by S_t the market price at time t of each unit of this fund, that drives the return on the investment portfolio built up with the policyholder's payment. The value of such portfolio, i.e., the account value, is denoted by X_t.

The policyholder is allowed to (periodically) withdraw from her account, for the rest of her life, a given percentage (*withdrawal rate*) of the so-called *base amount*,[1] whose value is initially set equal to the single premium, and can decide when to start such withdrawals. The withdrawal rate, denoted by $g(t)$, usually increases with the initiation time t, to compensate for the decreasing withdrawal duration. To encourage late initiation, the withdrawal base is periodically increased, until initiation, at the *roll-up rate*, denoted by β. Moreover, we assume also a *ratchet* feature, according to which the withdrawal base is periodically reset to the account value, if lower, and this occurs before and after initiation. The withdrawal base at time t is denoted by M_t.

The account value evolves according to the investment performance. Moreover, all withdrawals are subtracted from this account, along with the management fees, proportional to the account value (ψ =management fee rate) and the insurance fees, proportional to the base amount (ϕ =insurance fee rate). If the account value is not enough to allow for guaranteed withdrawals, then these are paid out of the insurer's own funds, and fees are no longer charged. If there are still available funds when the policyholder dies, then the account value is entirely paid to the beneficiaries as a death benefit.

Finally, the policyholder is allowed to surrender the contract at any time and receive the account value, net of a surrender penalty. We distinguish between *early surrender*, i.e., before initiation, and *late surrender*, after initiation. We assume that the penalty rate can be different according to the time of surrender and/or to the duration of withdrawals. In particular, we denote the penalty rate by $p^e(t)$, if surrender takes place at time t, before initiation, and by $p^l(h)$ if surrender takes place h years after initiation of withdrawals.

[1] Also referred to as *withdrawal base*.

3 The model

We assume a discrete time model, where all variables are defined on a time grid $\mathbb{T} = \{0, 1, \ldots, N\}$. Typically, N represents the time when the policyholder attains an extremal age, beyond which her survival probability is null.

We denote by τ the time of death of the policyholder, λ the initiation time and π the surrender time. These are *stopping times* taking values in the grid $\mathbb{T}$. In particular, we conventionally set $\lambda = \tau$ and/or $\pi = \tau$ when no action (initiation and/or surrender) is taken. Then we have three possible situations: i) $\lambda < \pi \leq \tau$ (the contract is initiated and remains in force until death, if $\pi = \tau$, or late surrender, if $\pi < \tau$); ii) $\pi < \lambda = \tau$ (early surrender); iii) $\lambda = \pi = \tau$ (no action is taken and the contract remains in force until death).

The evolutions of the account value and the base amount, for any fixed initiation and surrender times λ and π, are given by

$$X_{t+1} = \max\left\{X_t\left(\frac{S_{t+1}}{S_t} - \psi\right) - \left(\phi + g(\lambda)1_{\{\lambda<t\}}\right)M_t,\ 0\right\}1_{\{\pi>t\}} \tag{1}$$

$$M_{t+1} = \max\left\{M_t\left(1+\beta 1_{\{\lambda>t\}}\right),\ X_{t+1}\right\}1_{\{\pi>t\}}, \tag{2}$$

where $X_0 = M_0 = P$ and 1_A denotes the indicator function of A.

We introduce a risk-neutral measure $\mathbb{Q}$, assumed to be chosen by the insurer among the infinitely many equivalent martingale measures characterizing incomplete and arbitrage-free markets. The fair value of the GLWB variable annuity is defined as the largest expected value, under $\mathbb{Q}$, of all contract cash-flows for any feasible initiation and surrender. The time 0 fair contract value, V_0, is then

$$V_0 = \sup_{\lambda,\pi} E^{\mathbb{Q}}\left[F^S_\pi 1_{\{\pi<\lambda=\tau\}}B_{0,\pi} + F^I_{\lambda,\pi}1_{\{\lambda<\pi\leq\tau\}}B_{0,\lambda} + F^D_\tau 1_{\{\lambda=\pi=\tau\}}B_{0,\tau}\right], \tag{3}$$

where $E^{\mathbb{Q}}$ denotes the expectation operator under $\mathbb{Q}$, $B_{u,v} = \exp\{-\int_u^v r_h \mathrm{d}h\}$ the (stochastic) discount factor at the risk-free rate r, $F^S_u = X_u(1-p^e(u))$ the surrender benefit at the early surrender time u, $F^I_{u,v} = g(u)\sum_{t=u+1}^v M_{t-1}B_{u,t} + X_v(1-p^l(v-u)1_{\{v<\tau\}})B_{u,v}$ the present value at the initiation time u of a contract terminated at $v > u$ for late surrender or death, and $F^D_u = X_u$ the death benefit at u for a contract that has never been initiated or surrendered.

We now transform the double optimal stopping problem (3) into a two-stage problem: in the first stage, for each stopping time $\lambda < \tau$, we need to solve the single optimal stopping problems

$$F^{I*}_\lambda = \sup_\eta E^{\mathbb{Q}}[F^I_{\lambda,\eta}|Z_\lambda], \tag{4}$$

where η is a stopping time such that $\lambda < \eta \leq \tau$ and Z_t is the vector of relevant state-variables at time t; then the second-stage problem gives the initial fair contract value as

$$V_0 = \sup_{\lambda,\pi} E^{\mathbb{Q}} \left[F_\pi^S 1_{\{\pi<\lambda=\tau\}} B_{0,\pi} + F_\lambda^{I*} 1_{\{\lambda<\pi\leq\tau\}} B_{0,\lambda} + F_\tau^D 1_{\{\lambda=\pi=\tau\}} B_{0,\tau} \right] \quad (5)$$

over all (λ,π) such that $\pi<\lambda=\tau$, or $\lambda<\pi\leq\tau$,[2] or $\lambda=\pi=\tau$.

4 Valuation algorithms

To solve problems (4) and (5) we need to estimate expectations conditional on future levels of all state-variables. To this end, we exploit the LSMC approach, that combines Monte Carlo simulation with Least-Squares regression. The dynamic backward recursive algorithms that allow to get V_0 are outlined in what follows.

Preliminary.

- STEP 0. Simulate forward H paths of the state vector Z and H times of death τ over the time grid $\mathbb{T}$. Using the subscript h $(=1,2,\ldots,H)$ for the h-th simulated value of each quantity of interest, set $n=\max_h \tau^h$ and, for $j=0,1,\ldots,n-1$, set $A_j=\left\{1\leq h\leq H:\tau^h>j\right\}$.

First-stage problems.

For any fixed $\lambda<n$ execute the following steps:

- STEP 1. For $h=1,2,\ldots,H$ set $\eta^h=\tau^h$.
- STEP 2. For $j=n-1,n-2,\ldots,\lambda+1$ and $h\in A_j$:[3]

  (i) set $P_j^h=g(\lambda)\sum_{k=j+1}^{\eta^h} M_{k-1}^h B_{j,k}^h + X_{\eta^h}^h (1-p^l(\eta^h-\lambda)1_{\{\eta^h<\tau^h\}}) B_{j,\eta^h}$;

  (ii) estimate the continuation value $\tilde{C}_j^h$ by regressing $(P_j^v)_{v\in A_j}$ on the vector of (basis functions of) the simulated state-variables $(Z_j^v)_{v\in A_j}$;

  (iii) if $X_j^h(1-p^l(j-\lambda))>\tilde{C}_j^h$, then set $\eta^h=j$.

- STEP 3. For $h\in A_\lambda$ set $F_\lambda^{I*,h}=F_{\lambda,\eta^h}^{I,h}=g(\lambda)\sum_{k=\lambda+1}^{\eta^h} M_{k-1}^h B_{\lambda,k}^h + X_{\eta^h}^h (1-p^l(\eta^h-\lambda)1_{\{\eta^h<\tau^h\}}) B_{\lambda,\eta^h}$.

Second-stage problem.

- STEP 1. For $h=1,2,\ldots H$ set $\rho^h=\lambda^h=\pi^h=\tau^h$ and $P_{\rho^h}^h=F_{\rho^h}^{D,h}=X_{\rho^h}^h$.
- STEP 2. For $j=n-1,n-2,\ldots,0$ and $h\in A_j$:

  (i) estimate the continuation (i.e., no action) value $\tilde{C}_j^h$ by regressing $(P_{\rho^v}^v B_{j,\rho^v}^v)_{v\in A_j}$ on the vector of (basis functions of) the simulated state-variables $(Z_j^v)_{v\in A_j}$;

  (ii) estimate the initiation value $\tilde{I}_j^{I*,h}$ by regressing $(F_j^{I*,v})_{v\in A_j}$ on the vector of (basis functions of) the simulated state-variables $(Z_j^v)_{v\in A_j}$;

[2] Note that, if the optimal solution of the second-stage problem implies $\lambda<\pi$, then the particular value of π is quite irrelevant.

[3] Clearly STEP 2 is skipped if $\lambda=n-1$.

(iii) calculate the early surrender benefit $F_j^{S,h} = X_j^h(1 - p^e(j))$;[4]

(iv) if $\max\{\tilde{C}_j^h, \tilde{I}_j^{I*,h}, F_j^{S,h}\} = \begin{cases} \tilde{I}_j^{I*,h} & \text{then set } \lambda^h = j \\ F_j^{S,h} & \text{then set } \pi^h = j \end{cases}$, update $\rho^h = \min\{\lambda^h, \pi^h\}$ and set $P_{\rho^h}^h = \max\{\tilde{C}_j^h, \tilde{I}_j^{I*,h}, F_j^{S,h}\}$.

- STEP 3 compute $V_0 = \frac{1}{H}\sum_{h=1}^H P_{\rho^h}^h B_{0,\rho^h}^h$.

If early and/or late surrender are not admitted, to get V_0 one could execute the algorithms after setting $p^e(t) \equiv 1$ and/or $p^l(h) \equiv 1$, but it would be much more efficient to simplify the algorithm. In particular, if late surrender is not admitted, one could completely skip the first-stage problems and replace $(F_j^{I*,h})_{h\in I_j}$ with $(F_{j,\tau^h}^{I,h})_{h\in I_j}$ in STEP 2(ii) of the second-stage problem. The second-stage problem can be clearly simplified if also early surrender is not admitted.

5 Conclusions

In this paper we have analysed the initiation option in a GLWB variable annuity, combined with the surrender option. We have assumed a discrete-time model, and defined the contract value as the solution of a double optimal stopping problem. We have shown that this problem can be transformed into a two-stage one, and solved it through dynamic programming algorithms that involve repeated applications of the LSMC approach, all based on the same set of simulations. This approach, although being time-consuming, allows the model to be completely flexible and not constrained to specific assumptions. The next step is to implement it numerically to capture comparative static properties of the results with respect to contract and model parameters.

References

1. Bacinello, A.R., Millossovich, P., Olivieri, A., Pitacco, E.: Variable annuities: A unifying valuation approach. Insur. Math. Econ. **49**, 285–297 (2011)
2. Carrière, J.F.: Valuation of the early-exercise price for options using simulations and nonparametric regression. Insur. Math. Econ. **1**, 19–30 (1996)
3. Huang, H., Milevsky, M.A., Salisbury, T.S.: Optimal initiation of a GLWB in a variable annuity: No Arbitrage approach. Insur. Math. Econ. **56**, 102–111 (2014)
4. Huang, Y.T., Zeng, P., Kwok, Y.K.: Optimal initiation of Guaranteed Lifelong Withdrawal Benefit with dynamic withdrawals. SIAM J. Financial Math. **8**, 804–840 (2017)
5. Longstaff, F.A., Schwartz, E.S.: Valuing American options by simulation: A simple least-squares approach. Rev. Financ. Stud. **14**, 113–147 (2001)

[4] For $j = 0$ we assume that surrender is not admitted and hence set $p^e(0) = 1$.

Modern design of life annuities in view of longevity and pandemics

Progettazione di benefici di rendita vitalizia, in un contesto di longevità e pandemia

Annamaria Olivieri

Abstract We consider annuity designs in which the benefit amount is contingent on a given longevity/mortality experience. This means, in particular, that the benefit amount is allowed to increase in case of higher mortality than expected, while it can decrease in the opposite case. Guarantees can still be maintained (for example, setting a minimum and a maximum benefit amount), but relaxed in respect of traditional annuity arrangements. This should result in lower premium loadings, thus helping to make the product more popular. In this research we investigate a pricing approach based on periodic fees applied to the policy account value, instead of the usual up-front loading at issue, solution that can make the product structure more flexible.

Abstract *Consideriamo rendite vitalizie il cui beneficio può essere modificato nel corso del contratto, sulla base della mortalità osservata in una popolazione di riferimento. In particolare, il beneficio può aumentare in caso di sovramortalità rispetto a quanto atteso, così come può diminuire nel caso opposto. Garanzie possono essere comunque essere mantenute (ad esempio, fissando un importo minimo e un importo massimo per il beneficio), anche se meno stringenti rispetto a quelle tradizionali. Questa scelta dovrebbe richiedere caricamenti meno onerosi, contribuendo così a migliorare la popolarità del prodotto. In questa ricerca studiamo un sistema di tariffazione che prevede caricamenti periodici, anziché l'usuale caricamento unico, soluzione che può contribuire ad una maggiore flessibilità del prodotto di rendita.*

Key words: Longevity-linked annuities, Aggregate longevity/mortality risk, Longevity guarantee, Periodic longevity fee.

Annamaria Olivieri
Department of Economics and Management, University of Parma, Italy
e-mail: annamaria.olivieri@unipr.it

1 Introduction

The need for individuals to take autonomous decisions regarding their post-retirement income has been largely discussed, especially starting from the late Nineties (of the last century), when major longevity improvements have been reported in many countries. Among the private post-retirement income solutions, traditional life annuities are perhaps the most protective choice for individuals, thanks to the longevity and financial guarantees they embed. However, since such guarantees expose the provider to major risks, they are matched by the rigidity of the benefit and investment structure, and loadings are considered to be too high. This may explain, at least in part, why annuity markets remain underdeveloped.

The current pandemic may suggest that the problem is no longer relevant. On the contrary, longevity remains an issue not to be underestimated, neither by individuals nor by providers. While it is clear that the COVID-19 pandemic is currently causing a mortality shock, it is not so clear what its impact on future mortality will be; among the scenarios that we can figure out for the future, there is also one predicting an increase in the life expectancy of the survivors. Further, social security could be in greater difficulty in the near future, due to the ongoing economic crisis. We must therefore continue to consider situations where a significant part of the post-retirement income will have to be covered with private resources.

In a scenario characterized by longevity but also by mortality shocks (which in the current pandemics are particularly severe at high ages), it is convenient to reconsider the design of longevity guarantees, so to make them more appealing both for individuals and providers. In particular, if the benefit is linked to an appropriate longevity experience, losses and profits originated by longevity/mortality are shared between individuals and providers, resulting first of all in reduced loadings, but also opening up the opportunity to make the product structure more flexible.

A number of Authors have addressed longevity-linked annuity benefits, considering different linking coefficients; see, for example [5, 11, 7, 2, 3, 4, 12]. The problem mainly discussed is the fair valuation of the arrangement, as well as optimality issues for the individual, in an expected utility framework. It must be mentioned that longevity-linking structures are common in self-insured arrangements, such as the best known Group-Self Annuitization pools (see, for example, [10]), without any guarantee provided. A general structure describing longevity-linked post-retirement benefits, proving as particular solutions the alternative linking coefficients described in the literature with some other possible choices, is developed by [8].

This research further develops [8, 9], in particular in respect of the pricing choice. While in [8, 9], as it is traditional for life annuities, the premium loading is defined and charged at policy issue, in this research we introduce periodic premium loadings, through a fee applied to the policy account value, whose level can be updated (though with limitations) to the ongoing experience. Charging periodic fees is in line with the pricing of guarantees in Variable Annuities; see, for example, [1]. [6] explore an idea similar to what we discuss in this research, in the case of pure endowments. A periodic fee, in particular when subject to possible updates, opens the way for greater product flexibility, and can be beneficial both for the individual (as

the loading is deferred in time), and the provider (as, at least to some extent, pricing adjustments based on the evolving longevity/mortality scenario are admitted).

In the Sections that follows, we outline the general structure of a longevity-linked benefit (Sect. 2), the setting for defining the periodic fee (Sect. 3), and the valuation performed in order to check whether pricing through periodic fees joint to longevity-linked annuity benefits can result in more satisfactory benefit profiles both for the individual and the provider.

2 Longevity/mortality-linked annuity benefits: A general structure

We consider a discrete-time annuity immediate in arrears, i.e. with payments at the end of the year. One cohort is addressed, aged x at time 0.

Of the several risks affecting the management of a life annuity, we focus on mortality/longevity risk, while disregarding others. In particular, we assume a deterministic financial setting. Conversely, we adopt a stochastic mortality model.

Benefits can be updated after issue, depending on a given longevity/mortality experience. Let b_0 denote the benefit amount stated at time 0. Following [8, 9], we describe the benefit amount updated at time t with the following general structure:

$$b_t = b_{t-1} \cdot \frac{p_{x+t-1}(\tau')}{\widetilde{p}_{x+t-1}} \cdot \frac{1+a_{x+t}(\tau')}{1+a_{x+t}(\tau'')}\,, \tag{1}$$

where: $p_{x+t-1}(\tau')$ denotes the annual survival probability at age $x+t-1$ provided by the best-estimate mortality assumption at time τ', $0 \le \tau' \le t-1$; $\widetilde{p}_{x+t-1}$ represents the survival probability (or longevity index) observed in a chosen population; $a_{x+t}(\tau)$ denotes the actuarial value at age $x+t$ of a unitary discrete-time annuity in arrears, based on the best-estimate assumption at time τ, with $\tau' \le \tau'' \le t$.

We note that the first ratio in Eq. (1) updates the benefit amount in relation to the realized longevity/mortality in respect of a benchmark; in particular, $p_{x+t-1}(\tau') > \widetilde{p}_{x+t-1}$, i.e. higher mortality than expected, implies an increase of the benefit amount, while $p_{x+t-1}(\tau') < \widetilde{p}_{x+t-1}$, i.e. higher longevity, requires a reduction of the benefit amount. The second ratio involves a comparison between mortality assumptions, where the assumption in the actuarial value at the denominator can be updated in respect of what considered in the numerator. If higher longevity is forecasted, then $a_{x+t}(\tau'') > a_{x+t}(\tau')$ and a reduction of the benefit amount is required; a benefit amount increase follows from $a_{x+t}(\tau'') < a_{x+t}(\tau')$.

We mention that Eq. (1) could be further extended in respect of a participation to the return on investments (see [8]); we disregard this aspect, as we are working in a deterministic financial setting.

Particular solutions for the linking design follows from Eq. (1), after having chosen the benchmark in the first and second ratio, as well as the reference population in which to observe the realized survival probability. It seems reasonable that the

linking is realized by letting only one of the two ratios take a value $\neq 1$; this way, the linking is based only on the survival probability (first ratio) or the actuarial value of the annuity (second ratio). After having investigated premium loadings, the present value of future profits and other quantities, [8, 9] suggest that a benchmark set at time $\tau' = 0$ could represent a satisfactory choice both for the individual and the provider, where the most appropriate linking solution seems to be the one based on the survival probabilities (first ratio of Eq. (1)). Explicit guarantees can be introduced, for example, by setting bounds for the benefit amount. Further, it is also reasonable to accept a maximum age to apply the benefit adjustment (say, age 95).

3 Pricing through periodic fees

In this Section we outline the setting for the assessment of the premium loading required because of the guarantees provided by the provider. We assume that periodic loadings are charged to the policy account value, at the beginning of each policy year, while the benefit is paid at the end of the year. The benefit amount paid at time t is the one updated at time $t-1$. Each individual pays an amount S at time 0.

We denote with A_t the policy account value at time t, whose dynamics during the life of the individual can be described recursevily as follows:

$$A_t = (A_{t-1} \cdot (1 - \pi_{t-1})) \cdot m(t-1,t) - b_{t-1} \cdot 1_{T>t}, \tag{2}$$

where: π_{t-1} is the proportional premium loading (or fee) that is charged to the policy account value at time $t-1$; $m(t-1,t)$ is the accumulation factor during year $(t-1,t)$; T is the random lifetime of the individual; 1_E is the indicator of event E, which takes value 1 if E is true, and 0 otherwise. At time 0, clearly we have $A_0 = S$.

For an individual alive at time t, the present value at time t of future benefits is defined as follows:

$$\text{PVFB}_t = \sum_{h=1}^{\omega-(x+t)} b_{t+h} \cdot {}_h\widetilde{p}_{x+t} \cdot v(t,t+h) \ , \tag{3}$$

where $v(t,t+h)$ is an appropriate discount factor, while ${}_h\widetilde{p}_{x+t}$ is the realized survival probability (i.e., the proportion of survivors) from age $x+t$ to age $x+t+h$, in a properly defined population.

We note that PVFB_t can be interpreted as a function of the fee, as it represents the value of the annuity contract net of premium loadings; thus, $\text{PVFB}_t \equiv \text{PVFB}_t(\pi)$. At time t, the fair fee is such that $\text{PVFB}_t(\pi) = A_t$. We point out that at time 0 such a condition defines the initial benefit amount b_0; at later times, an increase of the fee could require a reduction in the benefit amount, interacting with the linking adjustment. However, any subsequent adjustment of the fee must comply with what is guaranteed in the contract, in particular the lifelong annuity payment, the linking rule, and the minimum benefit amount. This means that not necessarily the provider

will be able to increase the fee, even if this would be required by the assessment in a revised scenario.

4 Investigation and assessments (outline)

Considering a stochastic mortality framework, we first assess at time 0 the required periodic premium loadings, for alternative benefit structures, testing different levels of the guarantee. We further compare the case of a single premium loading at time 0 with that of periodic premium loadings, in order to understand what advantages may come from this alternative structure, both in the individual and the provider perspective. We focus, in particular, on the time-profiles of fees and benefit amounts. In the longevity-linking design, we admit participation both to mortality profits and losses; this way the benefit amount can react both to situations of extra-mortality (as occurs, for example, in a pandemic scenario), and to situations of increasing longevity, which remains the most predictable scenario in the medium-long term. We point out that from the point of view of the individual, it is important to rely on increases of the benefit amount, in the event the initial mortality forecast turns out to have been too conservative.

In the provider's perspective, we address the impact of the longevity-linking and the pricing structure on the business value. Linking the annuity benefit to the longevity/mortality experience impacts on the business value in two opposing directions: possible losses are reduced, but also possible profits. Further, charging periodic fees instead of a single fee implies a deferral of the provider's earnings; on the other hand, provider's earnings should benefit from the possible update of the fee. It is, then, necessary to understand what is overall the trade-off reached by this arrangement. We also assess the loss probability for the provider, comparing alternative linking arrangements, as well as the periodic and the single loading.

References

1. Bacinello, A.R., Millossovich, P., Olivieri, A., Pitacco, E.: Variable annuities: A unifying valuation approach. Insurance: Mathematics and Economics **49(3)**, 285–297 (2011).
2. Bravo, J.M., de Freitas, N.E.M.: Valuation of longevity-linked life annuities. Insurance: Mathematics and Economics **78**, 212–229 (2018).
3. Chen, A., Hieber, P., Klein, J.K.: Tonuity: a novel individual-oriented retirement plan. ASTIN Bulletin **49(1)**, 5–30 (2019).
4. Chen, A., Rach, M.: Options on tontines: An innovative way of combining tontines and annuities. Insurance: Mathematics and Economics **89**, 182–192 (2019)
5. Denuit, M., Haberman, S. Renshaw, A.: Longevity-indexed life annuities. North American Actuarial Journal **15(1)**, 97–111 (2011).
6. Hanbali, H., Denuit, M., Dhaene, J., Truffin, J.: A dynamic equivalence principle for systematic longevity risk management. Insurance: Mathematics and Economics **86**, 158–167 (2019).
7. Milevsky, M.A., Salisbury, T.S.: Optimal retirement income tontines. Insurance: Mathematics and Economics **64**, 91–105 (2015).

8. Olivieri, A., Pitacco, E.: Linking annuity benefits to the longevity experience: Alternative solutions. Annals of Actuarial Science (2020) doi:10.1017/S1748499519000137.
9. Olivieri A., Pitacco, E.: Longevity-Linked Annuities: How to Preserve Value Creation Against Longevity Risk. In: Borda M., Grima, S., Kwiecień, I. (eds) Life Insurance in Europe. Financial and Monetary Policy Studies, **50**. Springer, Cham doi: 10.1007/978-3-030-49655-5_8 (2020)
10. Piggott, J., Valdez, E.A., Detzel, B.: The simple analytics of a pooled annuity fund. The Journal of Risk and Insurance **72(3)**, 497–520 (2005).
11. Richter, A., Weber, F.: Mortality-Indexed annuities. Managing longevity risk via product design. North American Actuarial Journal **15 (2)**, 212–236 (2011).
12. Weinert, JH., Gründl, H.: The modern tontine. Eur. Actuar. J. (2020) doi: 10.1007/s13385-020-00253-y

Risk Management from Finance to Production Planning: An Assembly-to-Order Case Study

La Gestione del Rischio dalla Finanza alla Produzione: Il Caso Assembly-to-Order

Paolo Brandimarte and Edoardo Fadda and Alberto Gennaro

Abstract Production planning, just like finance, is affected by risk factors: demand uncertainty, capacity outages, quality issues, foreign currency risk, etc. The relevance of each risk factor depends on the planning time horizon. Here, we focus on a medium-term problem and only consider demand uncertainty. The classical streams of literature include deterministic decision models and stochastic inventory management models, where average performance is optimized. Due to the shrinking life of products, such a risk-neutral attitude has become more and more questionable. In the paper, we deal with an assembly-to-order problem, modeled as a two-stage stochastic linear programming problem, and consider the introduction of a classical risk measure from finance, conditional value-at-risk.

Abstract *In produzione, come in finanza, giocano un ruolo fattori di rischio quali l'incertezza sulla domanda, variazioni di capacità produttiva, problemi di qualità, rischio valuta, etc. La rilevanza di ciascun fattore dipende dall'orizzonte di pianificazione. Qui trattiamo un problema di medio termine, limitandoci all'incertezza sulla domanda. In molta letteratura classica, o si ignora l'incertezza o la si considera per definire una misura di prestazione basata su una media (come tipico nella gestione delle scorte). In un contesto in cui il ciclo di vita dei prodotti si restringe, un approccio neutrale al rischio può essere discutibile. Nell'articolo, consideriamo un problema assembly-to-order, modellato come problema di programmazione lineare stocastica a due stadi, e l'introduzione di una misura di rischio derivata dalla finanza, il conditional value-at-risk.*

Key words: risk management; production planning; stochastic programming.

Paolo Brandimarte
Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, e-mail: paolo.brandimarte@polito.it

Edoardo Fadda
Politecnico di Torino, e-mail: edoardo.fadda@polito.it

Alberto Gennaro
Politecnico di Torino, e-mail: alberto.gennaro@polito.it

1 The connection between finance and operations management

The traditional literature on operations management focused on long and short term production planning, capacity planning, and inventory management. This range of problems spans different levels in the decision hierarchy, involving different risk factors: demand uncertainty, quality issues, capacity outages, regulation risk, macroeconomic factors, oil price, exchange rate risk, etc. A wide array of optimization models have been proposed over the years. Some, like standard capacitated lot sizing optimization models, plainly disregard any source of uncertainty, possibly adding some buffer like safety stocks [4]. On the contrary, the standard literature on inventory control addresses demand uncertainty, with the aim of optimizing some performance measure linked with a long-term average performance [5]. The increase in computing power has made some more powerful optimization approaches, like stochastic programming with recourse, a usable tool [2]. For instance, in [3] an application of multistage stochastic programming to lot sizing is proposed. Nevertheless, the aim was typically to optimize the expected value of an objective function, which is based on a risk neutrality assumption.

If we step up the decision hierarchy and deal, e.g., with capacity planning problems, the level of uncertainty and the potential economic impact increase considerably, calling for a better way to deal with investment risk. This has motivated the introduction of financial concepts into the domain of operations management. See, e.g., [11] for an illustration of capacity planning and its connection with financial options, or [12] for an introduction to supply chain finance. The proper design of contracts for supply chain management also call for concepts from economics, game theory, and contract design to share risk and align stakeholders' incentives [6].

1.1 Risk management and newsvendor-like models

When we get closer to medium or short term planning, it may be the case that the level and nature of uncertainty makes traditional approaches viable, but this is not necessarily true. An important example involves the wide class of newsvendor-like models. In the literal newsvendor problem, we have to buy or make an amount of items before observing uncertain demand on a limited time window. After the time window (e.g., the day during which we may sell the newspaper) elapses, we have to scrap unsold items or sell them at a markdown price. If the decision has to be repeated several times, or for several independent items, a risk-neutral approach whereby we maximize expected profit, may be warranted. However, in some domains like fashion, the short term losses may motivate a more careful approach. This is more and more relevant, given the pace of innovation and the rapid obsolescence of products. In fact, a considerable stream of literature involves mean–risk models, where risk measures originally developed for financial applications make their way into production models, like mean–variance models [8] or models involving quantile-based risk measures like conditional value-at-risk [7].

1.2 Optimizing conditional value-at-risk

Conditional value-at-risk (CV@R in the following) is a quantile-based risk measure meeting reasonable coherence properties [1]. It is the tail expectation of loss, conditional on loss being larger than a quantile at a given probability level, which is value-at-risk. Its application to simple ordering problems involving a single item is not too complicated. When dealing with more complicated production planning problems, whereby different items interact through shared capacity resources with limited capacity, or the use of common components to assemble end items, we need to develop full-fledged mathematical programming models. Fortunately, the coherence of CV@R implies its convexity, and results originally described in [10] allow the development of relatively simple linear programming models, based on sampling risk factors, to minimize such a risk measure, subject to constraints on expected values.

The essential result is that, in order to minimize CV@R at level $1-\alpha$, we may minimize the auxiliary objective function

$$H_{1-\alpha}(\mathbf{x},\zeta) = \zeta + \frac{1}{\alpha}\int [L(\mathbf{x},\mathbf{z}) - \zeta]^{+} f_Z(\mathbf{z})\, d\mathbf{z}, \tag{1}$$

where: $\mathbf{x}$ is the vector of decision variables; ζ is an auxiliary variable, representing value-at-risk at $1-\alpha$ level; $\mathbf{z}$ is a vector of random risk factors with joint density $f_Z(\mathbf{z})$; L a loss function depending on decisions and realized risk factors. We use the standard notation $[z] = \max\{z, 0\}$. When dealing with a discrete set of scenarios, the above integral boils down to a sum that can be easily dealt with within a linear programming modeling approach.

2 Production planning and risk management in an assembly-to-order environment

In this paper we consider a sort of multiproduct newsvendor problem, where a set of end items is assembled using a set of components (modules). We consider flat bills of materials, comprising only two levels corresponding to end items and components, which are dealt with in radically different way, within an Assembly-to-Order (ATO) manufacturing strategy. An ATO strategy is sensible when, on the one hand, the long lead time to manufacture or procure components makes a pure Make-to-Order approach not feasible and, on the other hand, a pure Make-to-Stock approach, whereby end items are stocked, is ruled out by the the large number of end item configurations that arise by combining even a relatively small number of components.

If final assembly is relatively fast, we may order components under demand uncertainty, while delaying final assembly until customer orders are actually received. Hence, the ATO strategy is a risk mitigation strategy, as it relies on a partial postponement of decisions. Furthermore, the idea can also reduce risk by a pooling

effect. Even though demand for end items is highly variable, demand for common components (i.e., components that are used in multiple end items) may be less variable. Hence, product design by emphasizing component commonality may be helpful in reducing risk [9]. Nevertheless, component commonality may increase cost, and we have to pay attention to other problem features: the tightness of manufacturing capacity, the profit margin from selling end items, the correlation between different end item demands, the demand variability, as well as its skewness and possible multimodality.

The bottom line is that we need a very flexible modeling framework. The ATO strategy naturally fits within the framework of two-stage stochastic programming models with recourse. The first-stage, here-and-now decisions concern the production of components, and the second-stage, wait-and-see decisions concern the final assembly of end items. Furthermore, the framework may deal with risk-neutral optimization of expected profits, as well as the introduction of a risk measure like CV@R.

3 Two-stage stochastic programming models in the ATO environment

In order to state the decision models describing the ATO production planning problem, let us introduce the following sets: the set of components $\mathscr{C} = \{1,\ldots,I\}$, the set of end items $\mathscr{E} = \{1,\ldots,J\}$, the set of production resources (machine groups) $\mathscr{M} = \{1,\ldots,M\}$, and the set $\mathscr{S} = \{1,\ldots,S\}$ of scenarios that we use to discretize the distribution of random demand. We denote by d_j^s the demand for end item $j \in \mathscr{E}$ in scenario $s \in \mathscr{S}$; the probability of scenario s is denoted by π^s. In the following, we will consider Monte Carlo scenario sampling; hence, $\pi^s = 1/S$ and scenarios are equiprobable, but this need not be the case.

Furthermore, let us introduce the following parameters: C_i, cost of component $i \in \mathscr{C}$, and P_j, selling price of end item $j \in \mathscr{E}$; L_m availability (in terms of time) of machine group $m \in \mathscr{M}$; T_{im} processing time for component $i \in \mathscr{C}$ on machine $m \in \mathscr{M}$; G_{ij} number of components of type $i \in \mathscr{C}$ needed to assemble one end item of type $j \in \mathscr{E}$; in manufacturing parlance, these numbers are called *gozinto factors*. Note that the limited capacity of machine groups refers only to the production of components. In an ATO environment, final assembly should not be a bottleneck, so we disregard resource constraints at the assembly level. Finally, we need to introduce two sets of decision variables: the first-stage variables x_i, the amount of component $i \in \mathscr{C}$ that we produce; the second-stage variables y_j^s, the amount of end item $j \in \mathscr{E}$ assembled and sold in scenario s, after observing actual demand. Decision variables may be continuous or integer. The choice depends on demand volume. Since integer linear programs (LP) are more difficult to solve, when demand volume is large enough and rounding effects are negligible, a continuous LP framework is preferred. Then, the resulting stochastic LP model is:

$$\max_{y_j^s, x_i \geq 0} \quad -\sum_{i \in \mathscr{C}} C_i x_i + \sum_{s \in \mathscr{S}} \pi^s \left(\sum_{j \in \mathscr{E}} P_j y_j^s \right) \tag{2}$$

$$\text{s.t.} \quad \sum_{i \in \mathscr{C}} T_{im} x_i \leq L_m \qquad \forall\, m \in \mathscr{M} \tag{3}$$

$$y_j^s \leq d_j^s \qquad j \in \mathscr{E},\ s \in \mathscr{S} \tag{4}$$

$$\sum_{j \in \mathscr{E}} G_{ij} y_j^s \leq x_i \qquad i \in \mathscr{C},\ \ s \in \mathscr{S} \tag{5}$$

The objective function of the problem is the expected net profit, expressed in (2) as expected revenue at the second stage minus cost at the first stage. Constraints (3) limit the machine availability; constraints (4) state that it is not possible to sell more than demand, and constraints (5) precludes assembling items for which we lack the necessary components, thereby linking the two decision stages.

This risk-neutral model may be extended by introducing CV@R. We take advantage of the natural discretization of Eq. (1), where loss is expressed by component cost minus revenue and depends on first-stage variables $\mathbf{x}$, second-stage variables $\mathbf{y}^s$, and demand $\mathbf{d}^s$:

$$H_{1-\alpha}(\mathbf{x}, \mathbf{y}, \mathbf{d}, \zeta) = \zeta + \frac{1}{\alpha} \sum_{s \in \mathscr{S}} \pi^s [L(\mathbf{x}, \mathbf{y}^s, \mathbf{d}^s) - \zeta]^+$$

where $L(\mathbf{x}, \mathbf{y}^s, \mathbf{d}^s) = \sum_{i \in \mathscr{C}} c_i x_i - \sum_{j \in clJ} p_j y_j^s$. Then, the linear model becomes

$$\min_{y_j^s, x_i, z^s \geq 0} \quad \zeta + \frac{1}{\alpha} \sum_{s \in \mathscr{S}} \pi^s z^s \tag{6}$$

$$\text{s.t.} \quad z^s \geq \sum_{i \in \mathscr{C}} C_i x_i - \sum_{j \in \mathscr{E}} P_j y_j^s - \zeta \qquad s \in \mathscr{S} \tag{7}$$

$$\sum_{s \in \mathscr{S}} \pi^s \left[\sum_{i \in \mathscr{C}} C_i x_i - \sum_{j \in \mathscr{E}} P_j y_j^s \right] \geq \beta \tag{8}$$

subject to (3), (4), and (5). Here, we also introduce auxiliary variables z^s to linearize expressions $[z]^+$ and enforce a minimum target expected profit β in Eq. (8).

4 An outline of computational results

Space does not allow to include a detailed computational study. Figure 1 may give a feeling for the different behavior of the risk-neutral and the risk-averse model. The histogram is obtained by fixing the first-stage solution and then drawing a set of out-of-sample scenarios on which we optimize final assembly. The preferred solution is clearly a matter of taste: the tail of low profits is reduced, and we pay this in terms of expected profit. The amount if reduction depends on the probability level and the constraint on minimum expected profit.

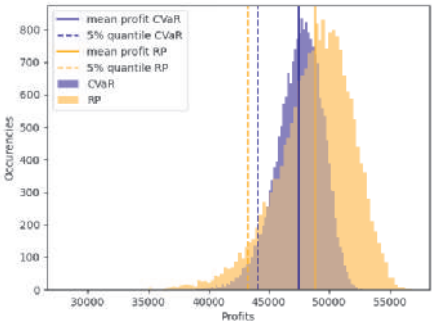


Fig. 1 Out-of-sample profit performance with a risk-neutral and risk-averse ATO planning model.

As a general observation, the risk-averse model is more challenging to solve, especially when we insist on using integer variables, and a larger number of scenarios is needed in order to achieve a stable first-stage solution. The results are also heavily affected by specific problem features. In particular, with a large number of valuable common components, stochastic programming becomes less useful. On the contrary, with skewed demand distributions, and more so with bimodal distributions, taking uncertainty into account becomes necessary to avoid scenarios with considerable loss.

References

1. Artzner, P., Delbaen, F., Eber, J.-M., Heath, D.: Coherent Measures of Risk. Mathematical Finance. **9**, 203–228 (1999)
2. Birge, J.R., Louveaux, F.: Introduction to Stochastic Programming (2nd ed.). Springer, Heidelberg (2011)
3. Brandimarte, P.: Multi-Item Capacitated Lot-Sizing With Demand Uncertainty. International Journal of Production Research. **44**, 2997–3022 (2006)
4. Brandimarte, P., Villa, A.: Advanced Models for Manufacturing Systems Management. CRC Press, Boca Raton (1995)
5. Brandimarte, P., Zotteri, G.: Introduction to Distribution Logistics. Wiley, Hoboken (2007)
6. Cachon, G.: The Allocation of Inventory Risk in a Supply Chain: Push, Pull and Advance-Purchase Discount Contracts. Management Science. **50** 222–238 (2004)
7. Chen, Y., Xu, M., Zhang, Z.G.: A Risk-Averse Newsvendor Model Under the CVaR Criterion Operations research. **57**, 1040–1044 (2009)
8. Choi, T,-M., Li, D., Yan, H.: Mean—Variance Analysis for the Newsvendor Problem. IEEE Transactions on Systems, Man, and Cybernetics. **38**, 1169–1180 (2008)
9. Hillier, M.S.: Using Commonality as Backup Safety Stock. European Journal of Operational Research. **136**, 353–365 (2002)
10. Rockafellar, R., Uryasev, S.: Optimization of Conditional Value-At-Risk. Journal of Risk. **2**, 21–42 (2000)
11. Van Mieghem J.A., Allon, G.: Operations Strategy: Principles and Practice (2nd ed). Dynamic Ideas, Charlestown MA (2015)
12. Zhao, L., Huchzermeier, A.: Supply Chain Finance: Integrating Operations and Finance in Global Supply Chains. Springer, Cham (2018).

Some probability distortion functions in behavioral portfolio selection

Alcune funzioni di distorsione di probabilità nella selezione di portafoglio comportamentale

Diana Barro, Marco Corazza, Martina Nardonthors

Abstract In this paper we propose some portfolio selection models based on Cumulative Prospect Theory. In particular, we consider two different probability distortion functions, respectively advanced by Tversky and Kahneman and by Prelec. The resulting mathematical programming problem turn out to be highly non-linear and non-differentiable. Then, we apply the portfolios selected under the behavioral approach to the European equity market as represented by the STOXX Europe 600 Index and compare their performances.

Abstract *In questo lavoro proponiamo alcuni modelli di selezione di portafoglio basati sulla Teoria del Prospetto nella sua versione cumulativa. In particolare, consideriamo due differenti funzioni di distorsione di probabilità, avanzate rispettivamente da Tversky e Kahneman e da Prelec. Il problema di programmazione matematica che ne deriva risulta fortemente non lineare e non differenziabile. Successivamente, applichiamo i portafogli selezionati secondo l'approccio comportamentale al mercato azionario europeo così come rappresentato dallo STOXX Europe 600 Index e confrontiamo le loro performances.*

Key words: Behavioral finance, Cumulative Prospect Theory, portfolio selection, probability distorsion function.

1 Introduction

In this paper we apply Prospect Theory (PT) in the cumulative version (CPT) of [6] to the portfolio selection problem, similarly to what done in [1], assessing two different probability distortion functions. Previously, [5] propose a behavioral portfolio model under PT.

[2] proposed PT as an alternative theory to Expected Utility (EU) in order to explain actual behaviors. In PT, individuals do not always take their decisions consistently with the maximization of EU; they display risk aversion with respect to gains and risk proneness with respect to losses, and are more sensitive to losses than gains of same magnitude. Investment choices are evaluated through a *value function*, v, in terms of potential gains and losses, instead of final wealth. The objective is the maximization of the prospect value $V = \sum_{i=-m}^{n} \pi_i \cdot v(z_i)$, with *decision weights* π_i, where z_i denotes negative outcomes for $-m \leq i < 0$ and positive outcomes for $0 < i \leq n$, with $z_i \leq z_j$ for $i < j$. Outcomes are interpreted as deviations from a *reference point*.

The value function is typically concave for gains and convex and steeper for losses. A function which is largely used in the literature and also applied in this work is

$$v(z) = \begin{cases} v^+(z) = z^a & z \geq 0 \\ v^-(z) = -\lambda(-z)^b & z < 0 \end{cases}, \tag{1}$$

with positive parameters that control risk attitude, $0 < a \leq 1$ and $0 < b \leq 1$, and loss aversion, $\lambda \geq 1$. Function (1) is continuous, strictly increasing, and has 0 as reference point[1].

Decision weights π_i are biased with respect to objective probabilities: medium and high probabilities tend to be underweighted and low probabilities of extreme outcomes are overweighted. In CPT [6], the prospect value depends also on the rank of the outcomes and the decision weights are differences in transformed countercumulative probabilities of gains and cumulative probabilities of losses.

Hence, risk attitude and loss aversion are modeled through a value function v, whereas a *probability weighting function* (or *probability distortion*)w models probabilistic risk perception via a distortion of probabilities of ranked outcomes. Actual investment decisions of a Prospect Investor (PI) depend on the shapes of these two functions as well as their interaction.

In the present work, we focus on the effects on the portfolio choices and performances of the shaped of two different probability distortion functions, respectively proposed by Tversky and Kahneman (see [6]) and by Prelec (see [4]).

The remainder of this paper is organized as follows. Section 2 synthesizes the main features of the probability weighting function. Section 3 presents the BP selection models. In Section 4, an application to the European equity market is discussed.

2 The probability weighting function

A probability weighting (or probability distortion) function w is a strictly increasing function which maps the probability interval $[0,1]$ into $[0,1]$, with $w(0) = 0$ and

[1] In the application, we use the parameters estimated by Tversky and Kahneman [6]: $\lambda = 2.25$ and $a = b = 0.88$.

$w(1) = 1$. Here we assume continuity of w on $[0,1]$, even thought in the literature discontinuous weighting functions are also considered.

Empirical evidence suggests a typical *inverse-S shape*: the function is initially concave (probabilistic risk seeking or optimism) for probabilities in the interval $(0, p^*)$, and then convex (probabilistic risk aversion or pessimism) in the interval $(p^*, 1)$, for a certain value of p^*. The *curvature* of the weighting function is related to the risk attitude toward probabilities; a linear weighting function describes probabilistic risk neutrality or objective sensitivity towards probabilities, which characterizes EU. Moreover, individuals are more sensitive to changes in the probability of extreme outcomes than mid outcomes: small probabilities of extreme events are overweighted, $w(p) > p$, whereas medium and high probabilities are underweighted, $w(p) < p$.

Different parametric forms for the weighting function with the above mentioned features have been proposed in the literature, and their parameters have been estimated in many studies[2].

[6] (TK) use function of the form

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}}, \tag{2}$$

with $\gamma > 0$ (with some constraint in order to have an increasing function). The parameter γ captures the degree of sensitivity toward changes in probabilities from impossibility ($p = 0$) to certainty ($p = 1$). When $\gamma < 1$, one obtains the typical inverse-S shape form; the lower the parameter, the higher is the curvature of the function.

[4] (PR) suggests a two parameter *compound-invariant* function[3] of the form

$$w(p) = e^{-\delta(-\ln p)^\gamma}, \tag{3}$$

with $\delta \in (0,1)1$ which governs elevation of the weighting function relative to the 45^o line, while $\gamma > 0$ governs curvature and the degree of sensitivity to extreme results relative to medium probability outcomes. When $\gamma < 1$, one obtains the inverse-S shape function. In this model, the parameter δ influences the tendency of over- or underweighting the probabilities.

As an alternative, we adopt the more parsimonious single parameter version of Prelec's function:

$$w(p) = e^{-(-\ln p)^\gamma}. \tag{4}$$

Note that, the unique solution of equation $w(p) = p$ for $p \in (0,1)$ is $p = 1/e$ and elevation of the function does not depend on γ.

In the applications, we use the parameters estimated by [6]: $\gamma^+ = 0.61$ and $\gamma^- = 0.69$, for w^+ and w^-, which denote the weighting function for probabilities of gains and losses, respectively. Fig. 1 compares the TK weighting function (2) and the

[2] See [3] for a review.

[3] In the same paper, two other probability weighting functions are derived: the *conditionally-invariant exponential-power* and the *projection-invariant hyperbolic-logarithm* function.

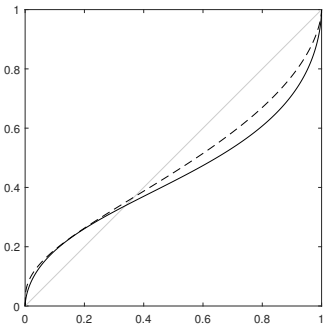


Fig. 1 A comparison between the TK (2) (solid line) and the PR (4) (dashed line) probability weighting functions, with $\gamma = 0.61$. The grey line represents objective probability.

PR one (4) for $\gamma = 0.61$. It is worth noting that, for this choice of the paramter γ, the TK function displays higher curvature and lower elevation; moreover, extreme sensitivity is slightly higher for the PR function.

3 A Behavioral CPT Portfolio Selection

Similarly to what done in [1], we assume that a PI selects the portfolio weights in order to maximize her prospect value subject to the usual budget constraint and short selling restrictions. Let us consider m possible scenarios, with r_{ij} the return of equity j in scenario i, and p_i the probability of each i. In this work we considered equally probable scenarios.

The portfolio returns, measured relative to a fixed reference point r_0, are the results subjectively evaluated by the PI with decision weights computed through the probability weighing functions discussed in the previous section. Formally, the BP selection model is defined as:

$$\begin{aligned} \max_{\mathbf{x}} & \sum_{i=1}^{m} \pi_i \cdot v\left(\sum_{j=1}^{n}(x_j r_{ij} - r_0)\right) \\ \text{s.t.} \quad & \sum_{j=1}^{n} x_j = 1 \\ & x_j \geq 0, \quad j = 1,2,\ldots,n. \end{aligned} \tag{5}$$

where $\mathbf{x} = (x_1,\ldots,x_n)$ is the vector of portfolio weights.

The resulting optimization problem is highly non-linear and non-differentiable so it cannot be solved applying traditional optimization techniques. For these reasons, according to what already done in [1], we adopt a solution approach based on the metaheuristic Particle Swarm Optimization.

4 Case study

To asses the CPT portfolio selection model and especially the role of the here considered probability distorsion functions, we carry out an analysis based the European equity market as represented by the ten sectorial indices in the STOXX Europe 600 Index. We compare in an out-of-sample analysis the performances and the risk profiles of four portfolios. In particular, for each specification of the weighting function (TK and PR), we tried out two different values for the reference point, $r_0 = 0\,\%$ and $r_0 = 2.5\,\%$. The overall testing period goes from July, 2018 to June, 2019. The out-of-sample analysis is carried out using a rolling window procedure with a 1-year in-sample period used to select the optimal portfolios whose compositions are then kept constant for the consecutive 3-month out-of-sample period. This scheme is then repeated as follows: at each step the in-sample and out-of-sample testing period slide 3-month forward until the entire 1-year out-of-sample evaluation period is covered.

In Table 1 we give some statistics related to the out-of-sample returns achieved by the optimal BPs, and in Fig. 2 we provide the relative frequencies of the same returns.

In general, both kinds of BP, i.e. TK and PR, show to perform better than the index (see the row "Mean" of the table), however having slightly better levels of risk than the latter (see the row "Standard deviation" of the table). These findings are highlighted by their Sharpe ratios which range about from 5.4 to 7.3 times that of the index (see the last row of the table).

Then, the optimal BPs are also characterized by levels of skewness and of kurtosis evidently greater than those associated to the index (see the rows "Skewness" and "Kurtosis" of the table). These evidences are the consequence of biasing the objective probabilities of the various portfolios' performances according to CPT. In this regard, notice that the two probability distortion functions considered here, i.e. (2) and (4), behave similarly, with equal r_0 (see again the rows "Skewness" and "Kurtosis" of the table, and Fig. (2)). In particular, when $r_0 = 2.5\,\%$, portfolio's performances in $[0\,\%, 2.5\,\%)$ are assessed as losses, while they are not when $r_0 = 0\,\%$. As PI is more sensitive to losses than gain of same magnitude, this explains why skewness and kurtosis are lower in the case $r_0 = 2.5\,\%$ than in the case $r_0 = 0\,\%$.

Lastly, we highlight that the selected BPs result enough diversified in all the 3-months out-of-sample periods. The determinants of such a finding will be inves-

Table 1 Statistics for the out-of-sample returns of the optimal BPs. (The pedix indicates the value of the reference point.)

	Index	$TK_{0.0}$	$TK_{2.5}$	$PR_{0.0}$	$PR_{2.5}$
Mean	0.0002	0.0013	0.0013	0.0012	0.0010
Standard deviation	0.0176	0.0162	0.0173	0.0168	0.0167
Skewness	−0.5630	−1.1612	−0.7257	−1.1560	−0.9898
Kurtosis	3.0943	5.1846	4.1753	5.6486	4.8756
Sharpe ratio	0.0114	0.0827	0.0726	0.0744	0.0620

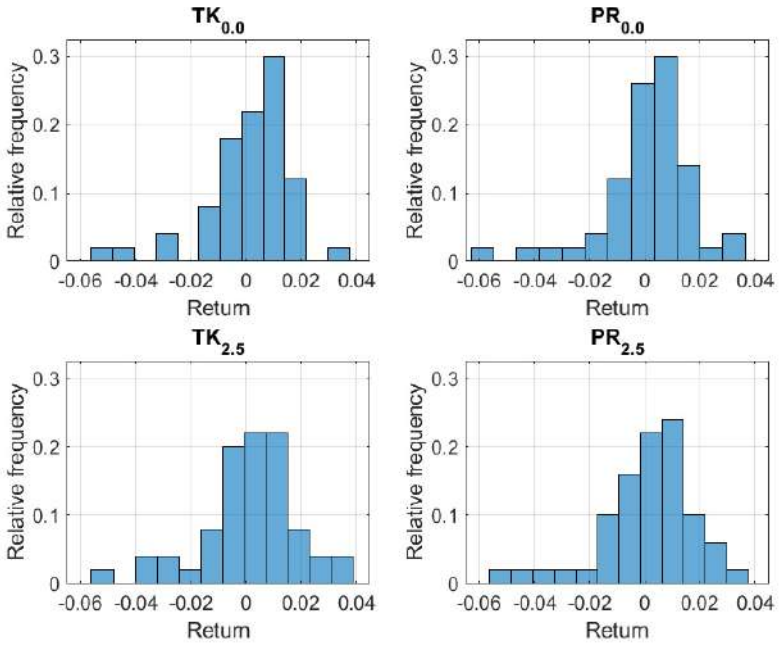


Fig. 2 Relative frequencies of the out-of-sample returns of the optimal BPs.

tigated in future experiments; in particular, we will perform sensitivity analysis on the parameters involved both in the value function and in the probability weighting ones.

References

1. Authors
2. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. Econometrica **47**, 263–291 (1979)
3. Authors
4. Prelec, D.: The probability weighting function. Econometrica **66**, 497–527 (1998)
5. Shefrin, H., Statman, M.: Behavioral portfolio theory. J. Fin. Quant. Analysis **35**, 127–151 (2000)
6. Tversky, A., Kahneman, D.: Advances in prospect theory: cumulative representation of the uncertainty. J. Risk Uncertain. **5**, 297–323 (1992)

3.12 Multiple system estimation

Multiple Systems Estimation in the Presence of Censored Cells

Stima di sistemi multipli in presenza di celle censurate

Ruth King, Oscar Rodriguez de Rivera Ortega and Rachel McCrea

Abstract We consider multiple systems estimation for contingency table data, where the observed cell entries may be censored and only presented to lie within a given interval. We describe how we can analyse such data using log-linear models and obtain estimates of the corresponding total population size taking into account the interval censoring. We compare this approach to some form of bounded approach, by setting the censored cell entries to the lower and upper limits of the interval. For the dataset that we consider we demonstrate that even for relatively small specified intervals, the estimate of the total population size, and selected log-linear model, is sensitive to how these censored cells are dealt with in the analysis.

Abstract *Il nostro lavoro si basa sulla stima di sistemi multipli con dati estratti da tabelle di contingenza dove le osservazioni in ogni cella possono essere censurate e solo presentate all'interno di un dato intervallo. Discutiamo come analizzare questi dati utilizzando modelli log-lineari per ottenere stime della dimensione della popolazione totale che tengono in considerazione l'intervallo di censura. Compariamo il nostro metodo a un tipo di approccio con soglia settando le celle censurate ai valori inferiori e superiori dell'intervallo. Con i dati analizzati in questo lavoro dimostriamo che, anche per intervalli relativamente piccoli, la stima della dimensione della popolazione totale e il modello log-lineare selezionato sono influenzati dal modo in cui le celle censurate sono specificate nell'analisi.*

Key words: Incomplete contingency table; log-linear model; maximum likelihood estimate

---

Ruth King
Thomas Bayes' Chair of Statistics, School of Mathematics, University of Edinburgh e-mail: Ruth.King@ed.ac.uk

Oscar Rodriguez de Rivera Ortega
Postdoctoral Research Associate, School of Mathematics, Statistics and Actuarial Science, University of Kent e-mail: O.Ortega@kent.ac.uk

Rachel McCrea
Professor of Statistics, School of Mathematics, Statistics and Actuarial Science, University of Kent e-mail: R.S.McCrea@kent.ac.uk

1 Introduction

Multiple systems estimation focuses on the estimation of hidden population sizes[1]. Applications range from the number of people who inject drugs in an area[6], civilian casualties within wars[9], webpages on a given topic[4] or modern day slaves[11]. Data are collected on the given population via a series of lists that partially observe individuals in the population. Assuming that individuals are uniquely identifiable, lists can be cross-classified and the corresponding data summarised via an incomplete contingency table, providing the number of unique individuals observed by each distinct combination of lists. The table is incomplete as the number of individuals not observed by any of the lists is not observable. Multiple systems estimation involves fitting statistical model(s) to the observed contingency table data, which in turn permits estimation of the number of individuals not observed by any of the lists.

When dealing with sensitive individual cross-classified data it is important that privacy issues are respected, and deductive disclosure risks are minimised. This issue typically arises when observed cell counts are small. To address this issue, approaches include aggregating cells (for example providing only marginal counts) or perturbing the data[5]. In this paper we consider contingency table data that are interval censored, such that small cell counts are not provided exactly, but instead presented to lie within a given interval. We focus on assessing the impact of censored cells within the statistical analysis, fitting log-linear models to the incomplete contingency table, and subsequent estimation of the total population size. Further, we describe how we can formally account for such censored cells within the analysis, accounting for the additional uncertainty of these cell entries.

2 Log-linear Models

We consider the general case where we have I administrative data lists, labelled $S_1,\ldots,S_I$. Each list has two levels, $j_i = 0,1$, corresponding to whether an individual is observed (=1) or unobserved (=0) by list S_i, for $i = 1,\ldots,I$. The set of possible list combinations is given by $\mathscr{K} = \{0,1\}^I$; and the set of observed combinations by $\mathscr{J} = \{0,1\}^I \backslash \{0,\ldots,0\}$, since we do not observe an individual not observed by any of the lists. The data are presented via an incomplete 2^I contingency table where the cell entries correspond to the number of individuals observed by the given distinct combination of lists. Notationally, we let n_j denote the number of individuals observed by the list combination $j \in \mathscr{J}$; and $n = \{n_j : j \in \mathscr{J}\}$ the observed data. The number of individuals not observed by any list is denoted by $n_{\{0,\ldots,0\}}$; and we let $N = \sum_{j\in\mathscr{J}} n_j + n_{\{0,\ldots,0\}} = \sum_{j\in\mathscr{K}} n_j$ denote the total number of individuals.

We consider log-linear models[3] so that for $j \in \mathscr{K}$,

$$n_j|\lambda_j \overset{ind}{\sim} \text{Poisson}(\lambda_j) \quad \text{such that} \quad \log\lambda_j = X_j\theta,$$

where X_j denotes the associated jth row of the specified design matrix and θ the column vector of parameters. The set of Poisson means is denoted by $\lambda = \{\lambda_j : j \in$

$\mathscr{K}\}$. The design matrix specifies the known relationship between each contingency table cell and log-linear parameters corresponding, to an intercept (θ^0), main-effect terms ($\{\theta^i : i = 1,\ldots,I\}$) and k-way interactions ($\{\theta^i : i \subset \{1,\ldots,I\}, |i| = k\}$), for $k = 2,\ldots,I-1$, such that $|i|$ denotes the number of elements in i. We specify the standard corner-point constraints on the log-linear parameters to define the design matrix. The corresponding likelihood function of the observed data is given by:

$$f(n;\lambda) = \prod_{j \in \mathscr{J}} Poisson(n_j;\lambda_j) \equiv \prod_{j \in \mathscr{J}} \frac{\exp(-\lambda_j)\lambda_j^{n_j}}{n_j!},$$

where Poisson$(n_j;\lambda_j)$ denotes the probability mass function of a Poisson random variable with mean λ_j evaluated at n_j.

Given the incomplete contingency table data we estimate the total population size by (i) fitting a given log-linear model to the observed contingency cell entries n to estimate the parameters θ (and hence λ); and (ii) subsequently estimating the unobserved cell entry, $n_{\{0,\ldots,0\}}$, given the fitted model, which is then combined with the observed data to estimate the total population size. The parameters can be estimated, for example, via maximum likelihood estimation, or using a Bayesian approach. The estimate of the total population size is generally dependent on the given log-linear model fitted to the data, in terms of the interactions present. Thus, a model selection process is usually applied and/or a model-averaging approach applied[8]. We will consider a classical model-fitting approach.

3 Censored cells

The observed contingency table data are often imperfect or only partially available. For example the lists may be "corrupted" permitting the inclusion of individuals who are not members of the target population of interest, leading to cell(s) corresponding to upper bounds rather than observed numbers[10]. Alternatively, the data may be presented in "censored" form where the exact cell entries are not provided but an interval is given which contains the observed value, such as a value ≤ 4 or in the range 1-4[7]. This may arise, for example, for data privacy issues to avoid potential deductive disclosure of given individuals in the data.

We describe a formal statistical approach to analyse such contingency table data with interval censored cells. Let $\mathscr{J}' \subseteq \mathscr{J}$ denote the set of censored contingency table cells; and $\mathscr{J}^* = \mathscr{J} \setminus \mathscr{J}'$ the set of observed (non-censored) cells. For simplicity we assume that the censoring applied to each of the censored cells is the same, but in general this need not be the case. Further, that the censored cells are such that the cell entry lies in the interval $[a,b]$. The corresponding likelihood contribution for a given censored cell, $j \in \mathscr{J}'$ can be obtained by summing over the range of possible values in the interval of the corresponding Poisson probability mass function. Mathematically, we have that for $j \in \mathscr{J}^*$, the corresponding likelihood contribution is given by,

$$f(n_j;\lambda_j) = \sum_{k=a}^{b} Poisson(k;\lambda_j) = \sum_{k=a}^{b} \frac{\exp(-\lambda_j)\lambda_j^k}{k!}.$$

The likelihood of the observed data is thus given by,

$$f(n;\lambda) = \prod_{j\in\mathscr{J}^*} \frac{\exp(-\lambda_j)\lambda_j^{n_j}}{n_j!} + \prod_{j\in\mathscr{J}'} \left[\sum_{k=a}^{b} \frac{\exp(-\lambda_j)\lambda_j^k}{k!}\right].$$

Maximising the likelihood using a numerical optimisation algorithm we obtain the MLEs of the log-linear parameters, $\hat{\theta}$, and hence $\hat{\lambda}_{0,\ldots,0}$. The corresponding MLE for the missing cell is $\hat{n}_{\{0,\ldots,0\}} = \hat{\lambda}_{\{0,\ldots,0\}}$. For the total population size we also require the MLEs for the censored cells. For $j \in \mathscr{J}'$ the MLE is given by $\hat{n}_j = \max[\min(\hat{\lambda}_j, b), a]$. In other words if $\hat{\lambda}_j \in [a,b]$, then $\hat{n}_j = \hat{\lambda}_j$; else if $\hat{\lambda}_j < a$, then $\hat{n}_j = a$; finally if $\hat{\lambda}_j < b$, then $\hat{n}_j = b$. Finally,

$$\hat{N} = \sum_{j\in\mathscr{J}^*} n_j + \sum_{j\in\mathscr{J}'} \hat{n}_j + \hat{n}_{\{0,\ldots,0\}}.$$

To obtain the corresponding 95% confidence interval (CI) for the total population size (and the censored cells) we use a parametric bootstrap.

4 Application

We consider data relating to people who inject drugs in England in 2005-2006[7]. Four lists are used corresponding to:

- S_1 - Probation
- S_2 - Drug intervention programme (DIP) prison assessments
- S_3 - Drug treatment
- S_4 - DIP community assessments.

Data are further cross-classified across 9 geographical regions, gender and age (young = 15-34; old = 35+). However, for potential deductive disclosure, cell entry values in the range 1-4 were not explicitly provided, but indicated by an *. We consider the data for North East England for young females, presented in Table 1.

We consider the set of hierarchical log-linear models and implement a bottom-up search algorithm, using the AIC/BIC criteria for the interval censoring modelling approach. For comparison we consider two further approaches where we set all censored cells to either their minimum value (of 1) or maximum value (of 4). Table 2 summarises the results in terms of MLEs, 95% CI and AIC/BIC for the models with largest support. We observe that the total population estimate is sensitive to the given model fitted, a feature commonly observed within standard multiple systems estimation[2]. As expected, for each individual model fitted, the MLE of the total population size using the interval censoring modelling approach, taking into account the uncertainty of the cell entries, lies within those obtained when simply fixing each

cell to be equal to the minimum or maximum possible value. However, the associated variability of the MLE varies across models. For example, in some cases broadly similar MLEs are obtained, as for model $\{12,13,14,34,134\}$; whereas for other models, such as model $\{12,13,23,14,34\}$ the MLEs are very different. Further, we note that the model deemed optimal differs for the case where all cell values are set to their maximum value, compared to the other two approaches.

Table 1 Contingency table for the number of injecting drug users for the North East region for young females. An * indicates that the cell entry is interval censored and lies within the range 1-4.

S_1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
S_2	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
S_3	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
S_4	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
	74	17	*	584	63	12	9	35	*	*	*	35	17	*	*

Table 2 MLEs and 95% CIs of the total population size for the models identified; and associated ΔAIC (= AIC − min(AIC)) and ΔBIC (= BIC − min(BIC)) statistics for each of the different methods applied to the censored contingency table data. The model is specified via the maximum interaction terms present; and the model deemed optimal is highlighted in bold for each method.

Method	Model	MLE	95% CI	ΔAIC	ΔBIC
Cells = minimum value	$\{12,13,14,34\}$	2107	(1496, 3567)	6.86	6.14
	{12, 13, 23, 14, 34}	**4541**	**(1998, 7126)**	**0**	**0**
	$\{12,13,14,34,134\}$	1648	(1238, 2319)	2.85	2.84
	$\{12,13,23,14,34,134\}$	3179	(1232, 4981)	1.39	2.09
	$\{12,13,23,14,34,123\}$	4206	(1662, 6592)	2.00	2.69
	$\{12,13,23,14,34,123,134\}$	1390	(1071, 1692)	1.81	3.22
Cells = maximum value	$\{12,13,14,34\}$	1642	(1236, 1801)	3.25	1.12
	$\{12,13,23,14,34\}$	1794	(1180, 1814)	4.67	3.25
	$\{12,13,14,34,134\}$	1417	(1071, 1560)	1.87	0.45
	$\{12,13,23,14,34,134\}$	1331	(973, 1376)	3.79	3.07
	$\{12,13,23,14,34,123\}$	1646	(1066, 1679)	6.30	5.59
	{12, 13, 23, 14, 34, 123, 134}	**1010**	**(908, 1084)**	**0**	**0**
Cells = interval	$\{12,13,14,34\}$	1933	(1379, 2636)	2.46	1.75
	{12, 13, 23, 14, 34}	**3060**	**(1583, 3858)**	**0**	**0**
	$\{12,13,14,34,134\}$	1597	(1178, 2273)	0.82	0.82
	$\{12,13,23,14,34,134\}$	2293	(1104, 3456)	1.40	2.11
	$\{12,13,23,14,34,123\}$	2970	(1417, 4073)	1.99	2.70
	$\{12,13,23,14,34,123,134\}$	1113	(924, 1282)	1.32	2.73

5 Discussion

We have described an approach for dealing with general interval censored data within a multiple systems estimation framework, taking into account the uncertainty of the censored contingency table cells. Such data may arise, for example, due to privacy issues, where small cell counts are not published exactly, but instead an interval is given for the cell count. Further we have explored the sensitivity of the population size estimates in relation to how these censored cells are dealt

with in the corresponding associated statistical analysis of data relating to people who inject drugs in North East England. In particular, the estimates are compared with alternative approaches of replacing all censored values with the minimum or maximum possible values to give some indication of potential bounds on the estimates. As expected, the estimates obtained using the approach that accounts for the uncertainty in the censored cells lies between those obtained in the bounded case replacing the censored cells with the lower or upper limits of the interval. Further, different log-linear models, in terms of the interactions present in the model, are identified dependent on the approach taken, which in turn further influences the estimate of the total population size.

In some cases additional information on marginal totals may also be available limiting the set of possible combinations of observed cell entries. In these circumstances the corresponding likelihood is now expressible as a summation over all possible cell combinations, removing the independence across the interval censored cells in terms of the unobserved entries. However, due to the number of possible combinations, the likelihood may quickly become computationally expensive to calculate, even for a relatively moderate number of censored cells and interval widths. Potential avenues in such situations include the use of an EM algorithm, or a Bayesian data augmentation approach. These are the focus of current research.

References

1. Bird, S. M., King, R.: Multiple systems estimation (or capture-recapture estimation) to inform public policy. Annu. Rev. Stat. Appl. **5**, 95-118. (2018)
2. Cruyff, M., Overstall, A., Papathomas, M., McCrea, R.S.: Multiple system estimates for modern slavery: Model assessment and model selection. Crime and Delinq. In press (2021)
3. Fienberg, S.E.: The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika. **59**, 591–603 (1972)
4. Fienberg S., Johnson M., Junker B.: Classical multilevel and Bayesian approaches to population size estimation using multiple lists. J. R. Stat. Soc. A. Stat. Soc. **163**, 383–405 (1999)
5. Fienberg, S.E., Slavkovic, A B: A survey of statistical approaches to preserving confidentiality of contingency table entries. In Privacy-Preserving Data Mining 291–312. Springer, Berlin (2008)
6. Frischer, M., Leyland, A., Cormack, R., Goldberg, D.J., Bloor, M., Green, S.T., Taylor, A., Covell, R., McKeganey, N., Platt, S.: Estimating the population prevalence of injection drug use and infection with human immunodeficiency virus among injection drug users in Glasgow, Scotland. Am. J. Epidem. **138**, 170–181 (1993)
7. King, R., Bird, S.M., Overstall, A., Hay, G., Hutchinson, S.H.: Estimating prevalence of injecting drug users and associated heroin-related death rates in England using regional data and incorporating prior information. J. R. Stat. Soc. Ser. A. Stat. Soc. **77**, 209-236 (2014)
8. King, R., Brooks, S.P.: On the Bayesian analysis of population size. Biometrika. **88**, 317-336 (2001)
9. Lum, K., Price, M., Banks, D.: Applications of multiple systems estimation in human rights research. Am Stat. **67**, 191-200 (2013)
10. Overstall, A., King, R., Bird, S.M., Hutchinson, S.H., Hay, G.: Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. Stat. Med. **33**, 1564-1579 (2014)
11. Silverman, B.W.: Multiple-systems analysis for the quantification of modern slavery: classical and Bayesian approaches (with discussion). J R Stat Soc Ser A Stat Methodol. **183**, 691-736 (2020)

Bayesian population size estimation by repeated identifications of units. A semi-parametric mixture model approach.

Stima Bayesiana della numerosità di una popolazione attraverso identificazione ripetuta delle unità. Un approccio semi-parametrico basato sui modelli mistura

Tiziana Tuoto, Davide Di Cecco and Andrea Tancredi

Abstract The use of mixture models for estimating the size of an elusive population when capture rates vary among individuals has received strong attention from researchers involved in multiple system estimation. In this paper we propose a Bayesian semi-parametric approach by considering a truncated infinite dimensional Poisson mixture model for capture recapture count data. An application in official statistics regarding the estimate of the size of criminal populations is used to illustrate the proposed methodology.

Abstract *I modelli mistura vengono spesso utilizzati per stimare la numerosità di una popolazione elusiva quando i tassi di cattura variano da individuo a individuo. In questo lavoro le frequenze delle catture individuali verranno modellate attraverso un approccio semi-parametrico bayesiano basato su una mistura infinito-dimensionale troncata di distribuzioni di Poisson. Il modello verrà utilizzato per stimare il numero di individui connessi a specifiche attività criminali*

Key words: Criminal populations. Capture-recapture. Dirichlet process mixture. Official statistics.

1 Introduction

The aim of the research is to estimate the size of a hidden criminal population, for instance people working in markets of drug trafficking, prostitution exploitation and

Tiziana Tuoto
ISTAT, e-mail: tuoto@istat.it

Davide Di Cecco
Università di Roma La Sapienza e-mail: davide.dicecco@uniroma1.it

Andrea Tancredi
Università di Roma La Sapienza e-mail: andrea.tancredi@uniroma1.it

smuggling in Italy during a given reference time. The estimate of the size of people involved in these kinds of illegal economic activities, is envisaged at European level: according to European regulations, national accounts aggregates have to include illegal activities covering exhaustively the economic transactions which occur in the economic system.

In this paper, we aim at estimating the size of people involved in smuggling. In Italy, smuggling activities mainly regards cigarettes, and it is related to the importation and exportation of products that are legal in some other countries. Illegal cigarettes arrive in Italy especially from Eastern European countries, China and the United Arab Emirates. It is worthwhile nothing that there are other economic aspects, such as organized crime and corruption of the legal economy by money laundering that came from these activities to facilitate them.

Illegal activities for their nature are difficult to measure as people involved have obvious reasons to hide these activities. For this purpose, in this study we exploit administrative registers coming from the Ministry of Justice, which report alleged crimes for which the judicial authority started a criminal proceeding. Crimes records in the registers of the Public Prosecutor's offices, contain soft identifiers of the denounced subjects, namely date and place of birth and gender, as well as some characteristics of the denounced subjects and the crime acts, like age at the moment of the crime, nationality, the association with other subjects and previous crimes.

On the basis of the soft identifiers, crime authors can be identified and followed in a specific time span. In this way, the administrative source can be considered as a list of potential criminals with the count (i.e. the number of times) that they appear in the Prosecutor's offices registers. In the list we can observe individuals who are charged $1,2,3,\ldots,$ times, however we cannot observe units not caught by the Justice system. Hence, the registers can be considered as incomplete lists of potential criminals, since only denounced crimes and suspected criminals are reported. We want to estimate the hidden part of the population, i.e. the size of it not reported on the registers of the Public Prosecutor's offices.

Notice that the capture recapture data for the problem at hand are usually called repeated counting data. To model individual heterogeneity in these kind of data the use of mixture models has a long tradition, especially in the frequentist literature, see for example [8, 1]. Here we follow a Bayesian approach based on the Dirichlet process mixture model. In the next Section we specify the model and outline the resulting simulation algorithm. In the last Section we briefly illustrate the resulting posterior inference on the size of the smugglers in 2014.

2 The model

We assume that the population of potential criminals in a given year is a closed population of unknown size N. To take account of criminals heterogeneity we assume that the number of times Y that a criminal appears in the Prosecutor's offices registers is a mixture of Poisson distributions. In particular, we assume the *trun-*

cated version of the Dirichlet process mixture, see [6], where the weights of a fixed number Poisson components follow a finite stick breaking process.

Denote as k^* the number of components. Moreover denote as $p_j = Prob(Y = j)$ the probability of a unit being captured j times, and as p_j^i the probability of being captured j times in the i–th component, $p_j^i = \lambda_i^j e^{-\lambda_i}/j!$. Finally let $\pi_1, \ldots, \pi_{k^*}$ be the mixing weights, so that

$$p_j = \sum_{i=1}^{k^*} \pi_i \, p_j^i. \tag{1}$$

Denote as n_j the number of observed units that have been captured j times, and as D the set of all observed counts, $D = \{n_j\}_{j>0}$. We have $\sum_{j>0} n_j = n_{obs}$. We want to estimate the number of uncaptured units n_0, or, equivalently, the total number of units in the population $N = n_{obs} + n_0$.
We set a conjugate Gamma prior for each parameter λ_i, a truncated stick–breaking process with parameter ϕ over the mixture weights, and a Gamma prior over ϕ.

$$\begin{aligned} \lambda_i &\sim Gamma(\alpha_i, \beta_i), \quad i = 1, ..., k^* \\ (\pi_1, ..., \pi_{k^*}) &\sim SB(\phi) \\ \phi &\sim Gamma(\alpha_\phi, \beta_\phi) \end{aligned}$$

A similar modeling approach was considered by [7] in the standard multiple system framework with a fixed number of lists. In particular a Dirichlet process mixture was proposed to model the heterogeneity in the capture histories. See also [4]. Moreover note that a semi-parametric mixture of Poisson distributions driven by the Dirichlet process with the censoring of zero counts was proposed also by [5] for modeling gene expression sequence abundance distributions.

2.1 MCMC algorithm

In this section we detail the Gibbs–based MCMC algorithm to sample from the posterior distribution of N. Let us denote as Θ all the parameters $\{\pi_i\}$ and $\{\lambda_i\}$. Moreover let n_j^i be the (latent) number of units in the i–th component that have been captured j times. Let n^i be the total number of population units (captured or uncaptured) in component i: $n^i = \sum_{j \geq 0} n_j^i$. Then, at iteration t we have the following steps:

1. Sample all parameters $\lambda_i^{(t)}$

$$\lambda_i \sim Gamma\left(\sum_{j \geq 0} j \cdot n_j^i + \alpha_i \, , \, n^i + \beta_i\right) \qquad \text{for } \; i = 1, \ldots, k^*$$

2. In order to sample all mixing weights $\pi_i^{(t)}$, we first sample

$$V_i \sim Beta\left(1+n^i, \phi + \sum_{h=i+1}^{k^*} n^h\right) \qquad i = 1,\ldots,k^*-1$$

then take

$$\pi_i = V_i \prod_{h,h<i} (1-V_h) \qquad \text{for } i = 1,...,k^*$$

where $V_{k^*} = 1$.

3. Sample $\phi^{(t)}$, $\phi \sim Gamma(\alpha_\phi - 1 + k^*, \beta_\phi - \log \pi_{k^*})$
4. Sample $N^{(t)}$ from $P(N \mid \Theta^{(t)}, D)$. Note that

$$\begin{aligned} P(N \mid \Theta, D) = P(N \mid \Theta, n_{obs}) &\propto P(N) P(n_{obs} \mid N, \Theta) \\ &\propto P(N) \binom{N}{n_{obs}} p_0^{N-n_{obs}} (1-p_0)^{n_{obs}}, \end{aligned} \qquad (2)$$

where the probability p_0 of not being captured is calculated according to (1). Then, if we choose the improper prior $P(N) \propto 1/N$, we have

$$N^{(t)} \sim NegBin\left(n_{obs}, 1-p_0^{(t)}\right)$$

5. Sample vector $\mathbf{n}_j^{(t)} = (n_j^1, ..., n_j^{k^*})$ from $P(\mathbf{N}_j \mid \Theta^{(t)}, N^{(t)}, D)$:

$$\mathbf{N}_j \sim Mult\left(n_j, (p_{1|j}, ..., p_{k^*|j})\right) \qquad \text{for } j \geq 0$$

where $n_0 = N^{(t)} - n_{obs}$, and the probabilities $p_{k|j}$ of belonging to the k–th component conditionally on the number of captures j and the current values of Θ are updated as:

$$p_{k|j} = \frac{\pi_k \, p_j^k}{\sum_{i=1}^{k^*} \pi_i \, p_j^i}.$$

3 Results on smuggling data

The distribution of observed counts for smuggling captures in 2014 is reported in Figure 1. We consider a data set with a total number of observed smugglers equal to $n = 3349$. Note also the fat right tail of the capture distribution with a maximum number of captures equal to 27.

Figure 2 (left panel) shows the estimated posterior distribution of the population size obtained with the algorithm described in previous section. In the algorithm we fixed $k^* = 7$; the hyper-parameters of the Gamma prior for the λ_is are set equal to $\alpha_i = 1$ and $\beta_i = 0.05$ for $i = 1,...,k^*$, to provide substantial probability for a large range of y_i values, as common in long tailed distributions; the hyper-parameters of the Gamma prior for ϕ, i.e., the concentration parameter of the Dirichlet process, are fixed as $\alpha_\phi = \beta_\phi = 1$. Finally note that the number of replications of the MCMC

algorithm is 10^6 and the starting values for n_0 and α are respectively equal to 5000 and 1.

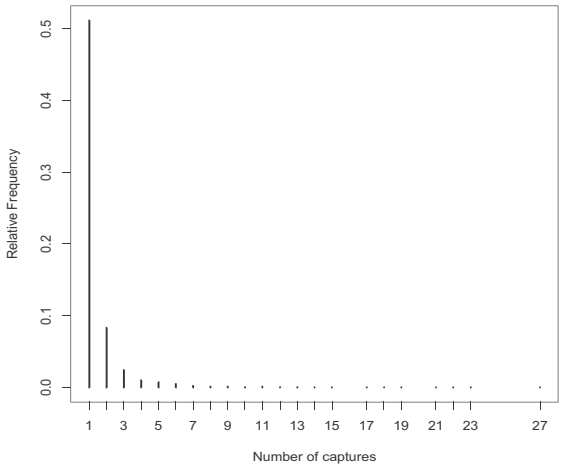


Fig. 1 Relative frequencies of observed counts for smuggling crimes in 2014

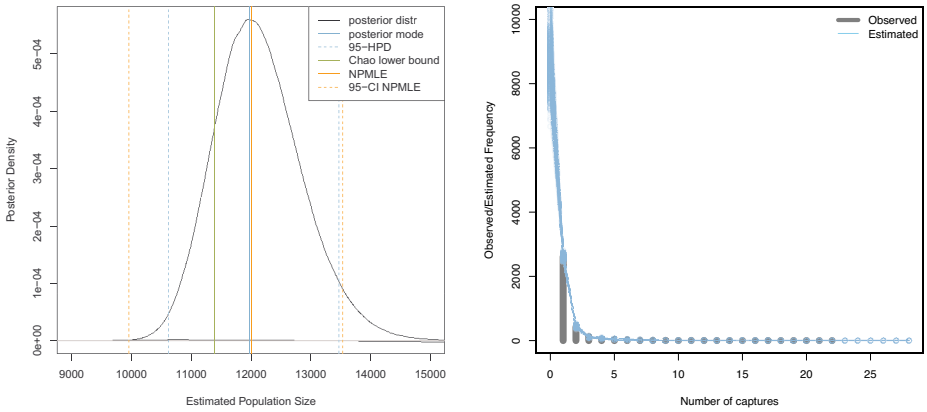


Fig. 2 Posterior distribution for N and different estimates with alternative methods (left panel). Observed and estimated frequencies for smuggling crimes in 2014 (right panel).

In Figure 2 (left panel) we highlighted the posterior mode and the posterior 95% credible interval for N. We also report some frequentist estimates, like the well-known Chao lower bound (see [3] for a review), the unconditional non parametric maximum likelihood estimate proposed by [8], and its 95% confidence interval ob-

tained by the bootstrap procedure proposed in [2]. Figure 2 (right panel) shows the observed and estimated frequency of counts for smuggling crimes in 2014. The estimated frequencies (in light blue) which comprise also the estimate of the unobserved individuals, correspond to the last 1000 simulations of the MCMC algorithm.

References

1. Böhning, D., Dietz, E., Kuhnert, R., Schön, D. (2005). Mixture models for capture-recapture count data. Statistical Methods and Applications **14** 29–43.
2. Böhning, D., Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(4), 721–737.
3. Chao, A. (2014). Capture-Recapture for Human Populations. Wiley StatsRef: Statistics Reference Online, 1-16.
4. Di Cecco, D., Di Zio, M., Liseo, B. (2020). Bayesian latent class models for capture–recapture in the presence of missing data. Biometrical Journal, **62**, 957–969.
5. Guindani M., Sepulveda N., Paulino C.D., Muller P. (2014). A Bayesian semiparametric approach for the differential analysis of sequence counts data. Journal of the Royal Statistical Society, Series C **63** 385–405.
6. Ishwaran, H., James L.F. (2001) Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association. **96** 161–173
7. Manrique-Vallier, D. (2016). Bayesian population size estimation using Dirichlet Process mixtures. Biometrics **72**, 1246–1254.
8. Norris, J.L., Pollock, K.H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. Biometrics **52**, 639–649.

3.13 Network sampling and estimation

Targeted random walk sampling

Campionamento random walk mirato

Li-Chun Zhang

Abstract We develop a novel strategy for estimating finite-order graph parameters based on stationary successive sampling probabilities of the eligible sample motifs under targeted random walk sampling. Illustration is given where the interest is the relative transitivity in a given graph with a core-periphery structure.

Abstract *Viene proposta una nuova strategia nel campionamento random walk mirato per stimare i parametri di un grafo di ordine finito in base alle probabilità di campionamento successive stazionarie. La procedura viene illustrata attraverso un esempio in cui l'interesse è la transitività relativa in un dato grafo caratterizzato da una struttura nucleo-periferia*

Key words: Markov chain, graph sampling, multiplicity, incidence weight

1 Targeted walk sampling for 1st-order graph parameters

Random walk in a graph can be considered as a probabilistic depth-first search algorithm, which can be attractive for real, large and often dynamic graphs, if the walk is fast-moving. We show that it is possible to achieve consistent generalised ratio type estimation of finite-order graph parameters based on the stationary successive sampling probability under targeted random walk at equilibrium.

Let $G = (U, A)$ be a simple undirected graph. Under random walk in G, let $X_t = i$ be the node (or *state*) at step time t. Let d_i be the number of edges incident node i, or its degree. For time $t+1$, one selects one of these edges randomly, $(ij) \in A_{i+}$, which yields $X_{t+1} = j$ as the next state of the random walk. Thus, $\{X_0, X_1, X_2, ...\}$ form a Markov chain, where X_0 is the initial state, with transition probability

$$p_{ij} := \Pr(X_{t+1} = j | X_t = i) = a_{ij}/d_i \, .$$

Li-Chun Zhang
Univ. of Southampton, Statistics Norway, Univ. of Oslo. e-mail: L.Zhang@soton.ac.uk

Let P be the $N \times N$ matrix of *transition probabilities*, with elements p_{ij} and $N = |U|$. Let p_0 be the *row* N-vector of initial node selection probabilities, the probabilities of X_t are given by

$$p_t = p_0 P^t \ .$$

If G consists of a single component, in which case every node can be reached from any other node eventually, then the chain $\{X_t\}$ is irreducible and its stationary probabilities, denoted by $\pi_i = \Pr(X_t = i)$ and $\pi = (\pi_1, ..., \pi_N)$, for $i \in U$, satisfy

$$\pi_j = \sum_{i \in U} \pi_i p_{ij} \qquad \text{or} \qquad \pi = \pi P \ .$$

where $\sum_{j \in U} \pi_j = 1$, and π is the left eigenvector of P with eigenvalue 1. We have

$$\pi_i \propto d_i \ .$$

A walk may be said to be *targeted* if its stationary probability π_i is subject to one's choice. Firstly, random jumps are generally needed in multi-component graphs. The probability of taking a random jump at any given time step affects π. Moreover, Metropolis-Hastings (MH) acceptance mechanism can be applied (Thompson, 2006), by which one can e.g. achieve the $\pi_i \propto d_i + 1$ walk. However, such a *targeted MH walk* with random jumps demands an *observation procedure*, where one needs to observe d_j of all the nodes adjacent to the current $X_t = i$, that is, $\{d_j : j \in U, a_{ij} = 1\}$, before one can make a move.

Avrachenkov et al. (2010) devise a more practical algorithm of *targeted random walk (TRW)* , which requires only d_i at $X_t = i$. Let there be an imaginary node, denoted by $\star \notin U$, which is connected to all the nodes in G, such that a random jump can be accomplished by two successive 'adjacent moves' via $\star$. Given $X_t = i$ at time step t, let the probability of moving to $\star$ be $r_i = r/(d_i + r)$ and, having moved to $\star$, one then takes immediately another random move away from it, to reach $X_{t+1} = j$ with probability $1/N$, for some $j \in U$. Notice that the probability of random jump varies with d_i under this scheme. The transition probability from $X_t = i$ to $X_{t+1} = j$ is then given by

$$p_{ij} = \begin{cases} \frac{1}{d_i + r}\left(1 + \frac{r}{N}\right) & \text{if } a_{ij} = 1 \\ \frac{r}{d_i + r} \cdot \frac{1}{N} & \text{if } a_{ij} = 0 \text{ including } i = j \end{cases} \tag{1}$$

and the resulting stationary probability for $i \in U$ is given by

$$\pi_i \propto d_i + r \ . \tag{2}$$

Given constants $y_U = \{y_i : i \in U\}$ associated with the nodes of the graph, let $\mu = \theta/N$, where $\theta = \sum_{i \in U} y_i$ and $N = |U|$, which is called a 1st-order graph parameter (Zhang and Patone, 2017). Until recently, sampling theory for targeted walk in graphs has only dealt with the estimation of such 1st-order graph parameters.

Let a targeted walk have stationary probabilities $\pi \propto c$, where the values c_i may be unknown for the unobserved nodes. Uniform walk is the special case with $c_i \equiv 1$;

indeed, $\pi_i \equiv 1/N$, if N is known. Random walk in an undirected connected graph is the special case with $c_i = d_i$, where c_i is unknown for any unvisited node. TRW is another special case, with $c_i = d_i + r$ by (2) in undirected graphs.

Targeted walk sampling (TWS) is stationary *draw-by-draw* at equilibrium, where π is the same for each draw whatever t. One can use an extraction of n states,

$$s = \{X_{t_1}, X_{t_2}, ..., X_{t_n}\} \quad \text{with} \quad t_1 < t_2 < \cdots < t_n \ ,$$

which need not to be successive, and a *generalised ratio estimator* of μ given by

$$\hat{\mu} = \Big(\frac{1}{n}\sum_{i\in s}\frac{y_i}{c_i}\Big) \Big/ \Big(\frac{1}{n}\sum_{i\in s}\frac{1}{c_i}\Big) = \sum_{i\in s}\frac{y_i}{c_i} \Big/ \sum_{i\in s}\frac{1}{c_i} \ . \tag{3}$$

This estimator is approximately unbiased for μ given sufficiently large n.

Within-walk auto-correlations do exist among the states in s. One can reduce the correlations by extracting time steps that are far apart from each other, in order to treat s approximately as an IID sample, when it comes to variance estimation. An alternative is to administer multiple walks independently. It is then simple to average the multiple estimators and use the between-walk variance as the basis for variance estimation, regardless the within-walk auto-correlations of each walk.

2 A strategy for finite-order graph parameters

For a strategy to finite-order graph parameters under TWS from graphs generally, we define below the sample graph, the stationary successive sampling probabilities that can be used for inference and the corresponding condition of eligible sample motifs eligible, which are the fundamental elements in any graph sampling situation (Zhang, 2021). Without loss of generality, we shall assume TRW with the incident forward observation procedure (Zhang and Patone, 2017) in the following, noting that the development is valid for any TWS method.

Sample graph The definition of sample graph G_s by Zhang and Patone (2017) needs a tweak of the node set U_s, in order to accommodate the isolated nodes in G, which are not incident to any edges and can only be visited by random jumps. Given s_{ref} and $A_s = A \cap s_{ref}$ by a T-step walk in G, let

$$G_s = (U_s, A_s) \qquad \text{where} \qquad U_s = s \cup \text{Inc}(A_s)$$

and s is the *seed sample*, to which the observation procedure of TWS is applied. Under T-step TRW sampling (T-TRWS), we observe all the edges incident to node i given $X_t = i$ at t, including the last step T. The reference set is given by

$$s_{ref} = s \times U \cup U \times s \ .$$

A subgraph motif $[M]$ such as triangle or K-star is observed if $M \times M \subseteq s_{ref}$, where M is the set of nodes of the motif. Let Ω_s be all such observed motifs in G_s.

Stationary successive sampling probability (S3P) Let $M = \{X_{t_1}, ..., X_{t_q}\}$ be a set of states given in the order by which the states are sampled, where $t_1 < \cdots < t_q$ and $q = |M|$. At equilibrium, we have

$$\pi_M = \Pr(X_{t_1}, ..., X_{t_q}) = \pi_{X_{t_1}} \prod_{i=1}^{q-1} p(X_{t_i}, X_{t_{i+1}})$$

where $\pi_{X_{t_1}}$ is the stationary probability of state X_{t_1}, and $p(X_{t_i}, X_{t_{i+1}})$ is the transition probability from X_{t_i} to $X_{t_{i+1}}$ over exactly $t_{i+1} - t_i$ time steps.

The *stationary sampling probability* π_M includes π_i as the 1st-order special case with $M = \{i\}$ and $|M| \equiv 1$. It is known if M consists of a set of *successive* states, apart from an unknown proportionality constant in π_{X_t}. We shall refer to π_M as the *stationary successive sampling probability (S3P)* if $t_q - t_1 = |M|$.

Other S3P π_M can be calculated, for any $M \subseteq s$, even though M is not a successive set of states of the *actual* walk, because the sub-matrix of P corresponding to $s \times s$ is known, even if the full matrix P is not. For instance, given the actual sequence $(X_t, X_{t+1}, X_{t+2}) = (1,2,3)$, we can also calculate the transition probability $p_{32}p_{21}$ had $(X_t, X_{t+1}, X_{t+2}) = (3,2,1)$ been the actual walk.

Given T-step walk with seed sample s, let the collection of such node sets be

$$\mathscr{C}_s = \{M : M \subseteq s\}$$

We shall refer to $\mathscr{C}_s$ as the *generating (sets of) states* of a T-step walk. The subset of generating states that are parts of the actual walk are given by

$$\mathscr{C}_w = \{\{X_t, ..., X_{t+q}\} : 0 \le t \le t+q \le T\}$$

Eligible sample motifs Let the sample motif κ in Ω_s be observed from the *actual sampling sequence of states (AS3)* $s_\kappa = (X_t, ..., X_{t+q})$, for some t and $q = |s_\kappa| - 1$. An *equivalent sampling sequence of states (ES3)* of s_κ, denoted by $\tilde{s}_\kappa \sim s_\kappa$, is any possible sequence of states of the same length, $|\tilde{s}_\kappa| = |s_\kappa|$, such that the motif κ is observed given $(X_t, X_{t+1}, ..., X_{t+q}) = \tilde{s}_\kappa$ but not based on any subsequence of $\tilde{s}_\kappa$.

Lemma 1 *Under TWS at equilibrium, a motif $\kappa \in \Omega_s$ observed from AS3 s_κ is eligible for estimation, iff all its ES3s belong to the generating states $\mathscr{C}_s$.*

Under TWS at equilibrium, a motif κ whose AS3 is of order greater than 1 can be sampled *sequence-by-sequence*, for which its ES3s constitute the multiplicity (of sampling). The motif is eligible for estimation if it satisfies the condition of Lemma 1, as the S3P of any sequence $\tilde{s}_\kappa$ is known up to a proportionality constant if $\tilde{s}_\kappa \in \mathscr{C}_s$. Thus, incidence weighting estimation (Patone and Zhang, 2020; Zhang and Oguz-Alper, 2020; Zhang, 2021) of finite-order graph parameters based on TWS sequence-by-sequence generalises estimation of 1st-order parameters based on TWS draw-by-draw.

3 Illustrations

Let $G = (U, A)$ be an undirected simple graph with 100 nodes, $N = |U| = 100$. Let $y = 1$ be the value associated with the first 20 nodes $i = 1, ..., 20$, or the cases; and let $y = 0$ be the value for the rest 80 nodes, $i = 21, ..., 100$, or the noncases.

Let the edges be generated randomly, with different probabilities for a given pair of nodes: (i) if both have $y = 1$, (ii) if one of them has $y = 1$ and the other $y = 0$, and (iii) if both have $y = 0$. In the resulting graph, there are altogether 299 edges, $|A| = 299$; the cases have an average degree 13.5, and the noncases have an average degree 4.1. The population graph G exhibits thus a core-periphery structure.

This valued graph will be held fixed for the illustrations below.

Estimation of case prevalence Let $\mu = \sum_{i \in U} y_i / N$ be the population case prevalence. Let $s = \{X_0, ..., X_T\}$ be the states obtained by T-TRWS, where X_0 is drawn with $p_{0,i} = 1/N$. Apply (3) to s yields $\hat{\mu}$. The burn-in stage is between 8-16 steps here. Using all the states is instructive for appreciating the convergence of $\hat{\mu}$.

Table 1 Estimation of case prevalence $\mu = 0.2$ under T-TRWS, 1000 simulations.

	$r = 1$			$r = 0.1$		
T	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	ψ	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	ψ
50	0.200	0.081	0.346	0.204	0.091	0.321
100	0.199	0.059	0.538	0.205	0.068	0.501
500	0.200	0.027	0.938	0.201	0.031	0.893
1000	0.199	0.019	0.987	0.201	0.022	0.959

Table 1 gives the results for $T = 50, 100, 500, 1000$, $r = 1$ or 0.1, each based on $B = 1000$ simulations of the T-TRWS. The consistency of $\hat{\mu}$ is already evident at $T = 50$, even without removing the initial burn-in states. The last column shows the average of the *traverse* ψ, which is the ratio between the number of distinct nodes visited by the walk and $N = |U|$, indicating how extensively the walk has travelled through the population graph. How quickly TRW traverses the graph is affected by the isolated nodes that can only be visited by random jumps, the probabilities of which are slightly reduced from $r = 1$ to $r = 0.1$, as is the convergence of $\hat{\mu}$, which can be seen by comparing the corresponding SD($\hat{\mu}$).

Estimation of a 3rd-order graph parameter Let $\mu = \theta / \theta'$, where θ is the total number of triangles among cases where all the nodes have $y = 1$, and θ' is the total number of other triangles involving at least one noncase. The larger the value of μ, the higher is the transitivity among cases compared to the overall transitivity in the graph. We have $\mu = 4.667$ in this population graph.

Table 2 gives the results for $T = 100, 500, 1000$, $r = 0.1$ or 6, where X_0 is selected with $p_{0,i} = \pi_i$, to avoid any details of handling the burn-in states. TRW sampling uses $r = 0.1$ for the results in the left part of Table 2, whereas $r = 6$ for the right part of the table. Since the average degree is about 6 in the population graph here,

Table 2 Estimation of $\mu = \theta/\theta' = 4.667$ under T-TRWS, 1000 simulations.

	$r = 0.1$			$r = 6$		
T	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	ψ	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	ψ
100	6.119 (0.142)	4.498	0.498	6.362 (0.160)	5.069	0.606
500	4.737 (0.028)	0.893	0.893	4.805 (0.034)	1.075	0.983
1000	4.669 (0.019)	0.593	0.958	4.704 (0.022)	0.702	0.999

setting $r = 6$ in (1) makes a random jump on average at least as probable as an adjacent move at each time step. This raises the traverse of the walk, e.g. TRW of length $T = 1000$ can be expected to cover almost the whole population graph.

Here, the convergence of $\hat{\mu}$ is not greatly affected by r, which seems to be at least almost the case given $T = 1000$, where each value in the parentheses in Table 2 is the estimated simulation error of Mean($\hat{\mu}$). Clearly, the walk needs to be longer for estimating this 3rd-order graph parameter, compared to that for the 1st-order parameter population case prevalence. This is not surprising because not every two successive states (X_t, X_{t+1}) correspond to an adjacent move, nor does one necessarily observe any triangle based on every adjacent move. In contrast, every state X_t contributes to the estimation of a 1st-order graph parameter.

References

1. Avrachenkov, K., Ribeiro, B. and Towsley, D. (2010). Improving Random Walk Estimation Accuracy with Uniform Restarts. *Research report*, RR-7394, INRIA. inria-00520350
2. Thompson, S.K. (2006). Targeted random walk designs. *Survey Methodology*, 32, 11–24.
3. Patone, M. and Zhang, L.-C. (2020) Incidence weighting estimation under bipartite incidence graph sampling. `arXiv:2004.04257v1`
4. Zhang, L.-C. (2021). Graph sampling: An introduction. *The Survey Statistician*, 83:27–37. `http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2021_January_N83_04.pdf`
5. Zhang, L.-C. and Oguz-Alper, M. (2020). Bipartite incidence graph sampling. `https://arxiv.org/abs/2003.09467`
6. Zhang, L.-C. and Patone, M. (2017). Graph sampling. *Metron*, 75:277.

Estimation of poverty measures in Respondent-driven sampling

Stima delle misure di povertà nel campionamento guidato dai rispondenti

María del Mar Rueda, Ismael Sánchez-Borrego and Héctor Mullo

Abstract Poverty measures are important socio-economic indicators to assess the economic situation of a region or a state. Respondent-driven sampling (RDS) is an advanced snowball-type sampling method used to survey hard-to-reach populations. We consider the problem of estimating poverty measures for RDS data by proposing estimators of the distribution function and in particular, estimators of poverty measures. The performance of the proposed estimators is illustrated with a RDS survey of ethnic minorities in Ecuador.

Abstract *Le misure di povertà sono importanti indicatori socioeconomici per valutare la situazione economica di una regione o di uno stato. Il campionamento guidato dai rispondenti (RDS) è un metodo avanzato di campionamento a palla di neve ulizzato per esaminare le popolazioni difficili da raggiungere. Consideriamo il problema della stima delle misure di povertà per i dati RDS proponendo stimatori della funzione di distribuzione e in particolare stimatori delle misure di povertà. La performance degli stimatori proposti è illustrata attraverso un'indagine RDS sulle minoranze etniche in Ecuador.*

Key words: poverty measures, RDS, distribution function.

María del Mar Rueda
Department of Statistics and Operations Research, University of Granada, 18071 Granada, Spain, e-mail: mrueda@ugr.es

Ismael Sánchez-Borrego
Department of Statistics and Operations Research, University of Granada, 18071 Granada, Spain, e-mail: ismasb@ugr.es

Héctor Mullo
Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo (ESPOCH), 060155 Riobamba, Ecuador, e-mail: hmullo@espoch.edu.ec

1 Introduction

Economic indicators can assess economic characteristics of countries and regions for analysis and prediction purposes. Poverty is one the most relevant indicators of social wellfare and it has important implications in a person's subjective well-being and overall satisfaction. The study of economic indicators such as poverty measures is becoming increasingly relevant to society and policy makers. Some well-known poverty measures are the headcount index, the poverty line and the Gini coefficient, among others. They are usually estimated from survey data, as information at the population level is typically not available.

If we consider income as a variable of interest, estimating the distribution function is essential to estimate important poverty measures, especially those based on quantiles. Estimators of the distribution function have been studied extensively in the overall probabilistic context ([8, 10]). Nevertheless, collecting accurate information on groups that represent only a small fraction of the population can become challenging, as their members are often stigmatized and typically difficult to reach. This generally results into a lack of a well-defined sampling frame for drawing a random sample [6]. Therefore, using standard sampling for surveying such a group is unfeasible in practice.

Respondent driven sampling (RDS) is a network-based method for sampling hard-to-reach, stigmatized and/or elusive populations. It was first introduced by [6] and was developed afterwards by [11] and [13]. Some examples includes injection drug users, LGBTI communities, HIV at risk persons, commercial sex workers, migrants and homeless.

RDS does not require an ordinary sampling frame and it enables privacy protection, which has been proven useful for those groups relunctant to be acknowledged because of social prejudice or stigma. It also reduces the costs compared to classical sampling, as the very same respondents recruit other participants.

We propose estimators of the distribution function and quantiles and particularly, estimators of poverty measures. We illustrate the performance of these estimators with a RDS survey conducted in Riobamba (Ecuador) ([9]) to study the living conditions of Indigenous, Montubios and Afro-Ecuadorian young people.

We propose estimators of the distribution function and quantiles in Section 2. An application to a RDS real data survey is carried out to illustrate their performance in Section 3. Finally, Section 4 presents concluding remarks.

2 Estimation of poverty measures

We consider the target population consists of N people (nodes) with labels $1,...,N$. We assume the target population is connected by a network of mutual relations with $N \times N$ adjacency matrix $\mathbf{Z}$. This means that $z_{kj} = z_{jk} = 1$ if k and j are connected and 0 otherwise. We define the nodal degree of a the person k, $\delta_k = \sum_j z_{kj}$, as the number of network ties or alters of node k.

The RDS selection process starts with a set of initial members of the target population called seeds, that represent the wave 0 of the sample. These respondents are given recruitment coupons (typically three), so that they recruit the next wave of participants, among their known contacts within the hidden group, usually with incentives [6]. When these respondents return their coupons, they recruit the next wave of participants. This process is repeated until the desired sample size n, is attained [11].

Let y_k be the value of the variable of interest for the respondent k in the sample s. We propose an estimator of the distribution function $F_y(t)$ as

$$\widehat{F}_y(t) = \frac{1}{N}\sum_{k\in s}\delta_k^{-1}\Delta(t-y_k), \tag{1}$$

where

$$\Delta(t-y_k) = \begin{cases} 0 \text{ if } t < y_k \\ 1 \text{ if } t \geq y_k \end{cases} \tag{2}$$

with δ_k the degree reported by respondent k. This estimator is similar to the Hájek estimator of the distribution function and the degree plays a similar role in a RDS setting to the role played by the first-order probability in a probabilistic survey sampling context.

Then, the quantile $Q_y(\alpha)$ can be estimated as

$$\widehat{Q}_y(\alpha) = inf\{t : \widehat{F}_y(t) \geq \alpha\} = \widehat{F}_y^{-1}(\alpha) \tag{3}$$

Once we have estimators of the distribution function and the quantiles, we can now estimate poverty measures, such as the Gini coefficient, the poverty risk *HCI* and the interquantile and interdecile ratios.

The Gini coefficient is a measure of inequality of a distribution. Eurostat [5] defined the Gini coefficient as the relationship of cumulative shares of the population arranged according to the level of income, to the cumulative share of the income received by them. The Gini coefficient [3] is estimated by

$$\widehat{G}_y = \frac{\sum_{k\in s}\delta_k^{-1}(2\widehat{F}_y(y_k)-1)y_k}{\sum_{k\in s}\delta_k^{-1}y_k}. \tag{4}$$

The poverty risk *HCI* [7] is the proportion of individuals with a disposable income below the at risk-of-poverty threshold, which is set at 60 % of the national median equivalised disposable income. It is estimated as

$$\widehat{HCI} = \frac{1}{N}\sum_{k\in s}\delta_k^{-1}I(y_k < 0.6\widehat{Q}_y(0.5)). \tag{5}$$

The interquartile and the interdecile ratios are measures of spread of a distribution. The first one is estimated as

$$\widehat{IQR} = \frac{\widehat{Q}_y(0.75)}{\widehat{Q}_y(0.25)}, \tag{6}$$

and the interdecile ratio is estimated as

$$\widehat{IDR} = \frac{\widehat{Q}_y(0.90)}{\widehat{Q}_y(0.10)}. \tag{7}$$

3 Application to a real survey

In this section, the proposed estimators are applied to a RDS survey on ethnic minorities conducted in the Canton of Riobamba, Ecuador [9]. The survey intended to study the living conditions and socioeconomic issues of young Indigenous, Montubios and Afro-Ecuadorians. These ethnic groups have been studied in the some national surveys (CPV, ECV, and ENEMDU), but people in these social groups find it difficult to self-identify. Moreover, there are evidence of exclusion and underrepresentation ([4, 2, 1, 12]) and therefore, this group lacks a reliable sampling frame. Nevertheless, as the RDS methodology reduces privacy concerns and they form a well-connected social network, RDS is a convenient method for sampling this population [6].

A total of 814 people were recruited in six waves and questioned on their social and economic background and living conditions using a dual system of incentives to motivate recruitment. The reported income of the household is the variable of interest and we consider the estimators of the distribution function and quantiles to compute poverty measures of this group of young ethnic minorities in Ecuador.

Table 1 Estimated poverty measures for the RDS survey on ethnic minorities in Ecuador

$\widehat{G}_y$	$\widehat{HCI}$	$\widehat{IQR}$	$\widehat{IDR}$
0.4241	0.2647	3.0722	12.7551

A Gini coefficient of 0 expresses equality, where all the incomes in a group are the same. On the other hand, a Gini coefficient of 1 expresses maximal income inequality. Here we have an intermediate distribution of income among young ethnic minorities in Ecuador. The value of $\widehat{HCI}$ shows that 26.47 % of the participants receive an income below the at-risk-of-poverty threshold.

The closer the $\widehat{IQR}$ value is to 1, the smaller the spread in the central 50% of the distribution. Here it is 3.072, which represents a gap in incomes larger than three times in the central part of the distribution. The $\widehat{IDR}$ value is 12.7551, showing important differences along the distribution of incomes.

4 Discussion

The literature on survey sampling is mainly focused on the estimation of linear parameters. Nevertheless, estimation of the distribution function is crucial when we consider income as a variable of interest, or when it comes to studying important socio-economic indicators such as poverty measures. We have proposed estimators of poverty measures for RDS data in a non probability survey sampling setting. To the best of our knowledge, no such an approach has been considered so far. RDS has been proven useful for sampling hard-to-reach populations, as it can collect information on people who are reluctant to self-identify as part of a hidden population, such as ethnic minorities. We have used the RDS survey on ethnic minorities as an illustration of the application of the proposed estimators. Future lines of research should be focused on extending the class of estimators of poverty measures, as well as studying their practical properties in simulation studies.

References

1. Araki, Hidekazu: Movimientos étnicos y Multiculturalismo en el Ecuador: Pueblos Indígenas, Afrodescendientes y Montubios. (2012) Master's Dissertation, University of Kanagawa, Kanagawa, Japan (2012)
2. Chisaguano, S.: La población indígena del Ecuador (Análisis de Estadísticas Socio-Demográficas) (2006) Available: https://www.acnur.org/fileadmin/Documentos/Publicaciones/2009/7015.pdf (accessed on 10 August 2020).
3. Handcock, M.S., Morris, M.: Relative Distribution Methods in the Social Sciences; Springer Science & Business Media, Berlin, Germany (2006)
4. Larrea, C., Torres, F., López, N., Rueda, M.: Pueblos Indígenas, Desarrollo Humano y Discriminación en el Ecuador. Abya Yala, Quito, Ecuador (2007)
5. Eurostat: Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC). Doc. LC-ILC/39/09/EN-rev.1, Unit F-3: Living conditions and social protection, Directorate F: Social and information society statistics. Eurostat, Luxembourg (2009)
6. Heckathorn, D.: Respondent-driven sampling: A new approach to the study of hidden populations. Soc. Probl. **44**, 174–199 (1997)
7. Martínez, S, Illescas, M., Martínez, H., Arcos, A.: Calibration estimator for Head Count Index. Int. J. Comput. Math. **97**, 51–62 (2020)
8. Martínez, S. Rueda, M., Illescas, M.: The optimization problem of quantile and poverty measures estimation based on calibration. J. Comput. Appl. Math. **113054** (2020)
9. Mullo, H.S., Sánchez-Borrego, I., Pasadas, S.: Respondent-driven sampling for surveying ethnic minority in Ecuador. Sustainability **12**, 9102 (2020)
10. Rueda, M., Martínez, S., Illescas, M.: Treating nonresponse in the estimation of the distribution function Math Comput. Simulat. **186**, 3 (2020)
11. Salganik, M., Heckathorn, D.: Sampling and estimation in hidden populations using respondent-driven sampling. Sociol. Methodol. **34**, 193–240 (2004)
12. Uquillas, J., Carrasco, T., Rees, M.: Exclusión Social y Estrategias de vida de los Indígenas Urbanos en Perú, México y Ecuador (2003) Available online: http://repositorio.minedu.gob.pe/handle/123456789/524 (accessed on 10 August 2020).
13. Volz, E., Heckathorn, D.: Probability based estimation theory for respondent driven sampling. J. Off. Stat. **14**, 79–97 (2008)

Sampling Networked Data for Semi-Supervised Learning Algorithms

Campionamento di dati su rete per l'apprendimento semi-supervisionato

Roberto Benedetti[a], Simone Di Zio[b], Lara Fontanella[c], Francesco Pantalone[d], Piersimoni Federica[e]

Abstract In this paper, focusing on classification for relational data represented through graphs, we address the problem of sampling on networks and propose a probabilistic design that produces well-spread samples over the networked data. We assess the sampling impact on the classification obtained through label propagation, a standard simple semi-supervised learning method. The proposed sampling design's performance is evaluated on a real-world network representing follower/friend relationships on Twitter

Abstract *In questo articolo, concentrandoci sulla classificazione per dati relazionali rappresentati attraverso grafi, affrontiamo il problema del campionamento su rete e proponiamo un disegno probabilistico che produca campioni ben distribuiti sul network. L'impatto del campionamento è valutato rispetto ai risultati della classificazione ottenuta attraverso la label propagation, un metodo di apprendimento semi-supervisionato. La performance del disegno di campionamento proposto viene valutata su una rete che rappresenta le relazioni di follower/friends su Twitter*

Key words: Network sampling, Semi-supervised learning, Label propagation

1 Introduction

Network sampling is of interest to researchers from many distinct fields due to the range of complex datasets that can be represented as graphs. In general, the accu-

(a) Department of Economic Studies, "G. d'Annunzio" University, Viale Pindaro 42, Pescara, Italy. e-mail: roberto.benedetti@unich.it. (b) Department of Legal and Social Sciences, "G. d'Annunzio" University, Viale Pindaro 42, Pescara, Italy. e-mail: simone.dizio@unich.it (c) Department of Legal and Social Sciences, "G. d'Annunzio" University, Viale Pindaro 42, Pescara, Italy. e-mail: lara.fontanella@unich.it (d) Department of Economics, University of Perugia, Via Alessandro Pascoli 20, Perugia, Italy. e-mail: francesco.pantalone@studenti.unipg.it (e) Istat, Directorate for Methodology and Statistical Process Design, Via Cesare Balbo 16, Roma, Italy. e-mail: piersimo@istat.it

racy of sampling from large networks has been assessed in terms of robustness of standard topological properties like degree distribution, diameter, centrality measures, or clustering [Leskovec and Faloutsos, 2006, Wagner et al., 2017, Ruggeri and De Bacco, 2020]. Focusing on classification tasks, Ahmed et al. [2012] and Espìn-Noboa et al. [2018] show how the accuracy of collective attribute inference in networks depends on the strategy used to create the initial set of labelled nodes. In particular, Espìn-Noboa et al. [2018] examine how network sampling strategy and sample size affect the accuracy of classification considering different synthetic networks where nodes have a binary attribute and a tunable level of homophily, i.e. different tendency of similar nodes to be linked to each other [McPherson et al., 2001]. A comprehensive analysis of classes and objectives of network sampling is presented by Ahmed et al. [2013]. The first taxonomy of sampling methods depends on the sampling units, and the algorithms can be categorised as node, edge, and topology-based sampling. As for the goals, it is possible to distinguish between estimation of network parameters (e.g. average degree of nodes), selection of a subgraphs whose representativeness is evaluated with respect to a set of topological properties, and estimation of node/edge attributes. In this paper, we consider the estimation of nodes' attributes and, in particular, we focus on a classification task for relational data. In the typical classification setting, independent and identically distributed training instances are assumed, and the goal is to predict independent test instances drawn from the same distribution. This independence assumption is violated in relational data, that encode dependencies among data instances. Relational data are typically represented through graphs, where the nodes are instances, and the edges between them imply dependence. This graph representation facilitates inferences on the pre-defined categories by exploiting the relationships in the network. In this context, researches have primarily focused on developing algorithms to sample small(er) sub-graphs, which are used to learn models, evaluate and compare algorithms' performance, and study complex network processes. A variety of techniques have been introduced for collectively classifying the nodes in a network [see Li and Pi, 2020, and references therein]. The usual approach is to use a set of previously labelled instances to learn a model, which can be used to classify new instances. This supervised learning process requires an extensive collection of labelled examples which are often expensive and difficult to obtain. The semi-supervised approach [see van Engelen and Hoos, 2020, for a comprehensive review], learning from both labelled and unlabelled instances, addresses the classification problem reducing the effort required to label training data. Semi-supervised learning methods are effective for classification in sparsely labelled networks, but, given that the classes of unlabelled nodes are inferred from a small number of seed nodes, it is relevant to assess the impact of seeds' choice on inference error.

In this paper, we address the problem of node sampling and propose a probabilistic design that produces samples that are well-spread over the networked data. We assess the sampling impact on the classification obtained through label propagation [Zhu and Ghahramani, 2002, Zhu et al., 2003], a standard simple semi-supervised learning method, which tries to set the label probabilities of nodes so that connected

nodes have similar probabilities. In the application, we consider a network of Twitter users connected by follower/friend relationships.

2 A well-spread sampling design for networked data

In this work, we propose a sampling strategy aiming to obtain well-spread samples over the graph structure, such that the number of selected nodes is close to what is expected on average in every part of the network. These types of sampling designs avoid the selection of neighbouring instances. Different spatially balanced sampling designs have proposed for spatial and geo-referenced data, and an exhaustive review can be found in Benedetti et al. [2017a,b].
Using a balanced sample on the graph, our goal is to partition the nodes into nodes to be labelled (*seeds*) and nodes to be predicted only considering network information. Let the network be represented by a simple undirected graph $G = (V,E)$, where V denotes the set of N nodes and E is the set of edges. The relational structure is summarised by the $N \times N$ adjacency matrix $\mathbf{A}$. For unweighted graph, $A_{ij} = 1$ if the node i connects to the node j, and $A_{ij} = 0$ otherwise. The adjacency matrix can be exploited to derive a distance matrix D over the graph, which is the base for our sampling procedure and is considered a sufficient synthesis of the relational information. The distance between any two nodes can be defined as the shortest path's length, and this geodesic distance can be obtained computing powers of the adjacency matrix. The problem to select well-spread samples is to define a design with a selection probability of each sample s proportional to some synthetic index $M(D_s)$ of the within sample distance matrix D_s. In this paper we use the product of the within sample distances (PWD), $M_0(D_s) = \prod_{i \in s} \prod_{j \in s} d_{ij}$, where d_{ij} is the distance between nodes i and j belonging to the sample s. We adapted the PWD iterative algorithm [Benedetti and Piersimoni, 2017], to the network sampling context, where the elements of s are nodes of the network. The PWD algorithm starts at iteration $t = 0$, with an initial sample $s^{(0)}$, obtained through node random sampling with constant inclusion probabilities and fixed size n. In a generic iteration t the elements of $s^{(t)}$ are updated according to the following steps:

1. select at random one node (V_i) included in the sample in the previous iteration, and another node (V_j) not included, and denote with $s_e^{(t)}$ the sample where the nodes V_i and V_j exchange their status;
2. randomly decide whether or not to update the sample $s^{(t+1)} = s_e^{(t)}$ with probability:
$$p = min\left[1, \left(\frac{M_0(D_{s_e^{(t)}})}{M_0(D_{s^{(t)}})}\right)^{\beta}\right]$$
otherwise retain the previous configuration $s^{(t+1)} = s^{(t)}$;
3. repeat steps (1) and (2) $q \times N$ times, where q is the maximum number of iterations each consisting of N attempts.

The acceptance rate within the N attempts in any of the q iterations is a natural candidate as a stopping criterion. In particular we stop the algorithm when within an iteration no attempt was accepted. A detailed discussion of the properties of this sampling scheme in a spatial context is provided in Benedetti and Piersimoni [2017].

3 Sampling design performance

The PWD sampling design's performance was assessed considering a network representing the follower/friend relationships between $N = 902$ Twitter's users, manually labelled according to their pro-Brexit or anti-Brexit stance. The labelling procedure was based on users' account profile and the textual content of the tweets containing the term "Brexit" that they posted between 1 January 2019 and 1 February 2020. A detailed description of the data collection can be found in del Gobbo et al. [2020]. The whole network is represented in Figure 1.

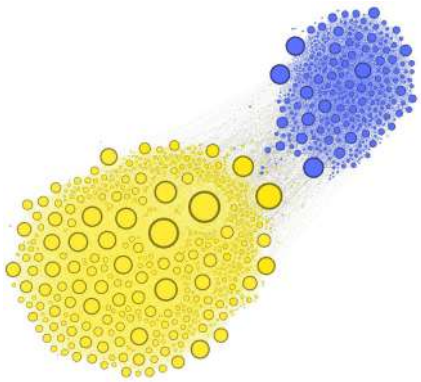


Fig. 1 Brexit stance network: blue nodes represent pro-Brexit users, yellow nodes represent anti-Brexit ones. *Node size represents user' degree*

We considered three different values for the PWD algorithm's tuning parameter, namely $\beta = \{1,5,10\}$, indicated as a suffix after the design's acronym in Table 1. Furthermore, we standardized the distance matrix to unit row and column products. As pointed out in Benedetti and Piersimoni [2017], this expedient allows to obtain a set of probabilities of inclusion π_i, $i = 1,\ldots,N$, approximately constant and equal to n/N. Finally, the sample size was fixed to $n = \{10,20,30,40,50\}$. We assess the sampling impact on the classification obtained through label propagation. To evaluate how well the not-sampled nodes' labels are recovered, we use the accuracy index derived from the two-class confusion matrix. We used simple random node sampling (RNS) as a benchmark for comparison. Besides, as possible alternatives to the suggested design, we considered the spatially correlated Poisson sampling (SCPS) [Grafström, 2012] and the local pivotal method (LPM2) [Grafström et al.,

2012]. To be comparable with the suggested design, for all these alternative designs, we set $\pi_i = n/N$ for all the nodes in the networks. For each design, we run 10000 replications and the mean of the accuracy measures across the replications, along with their variance, are provided in Table 1. As it can be noted, the PWD sampling design provides samples which yield to a higher classification accuracy, and the gain is larger for smaller sample sizes. It is worth noting how the accuracy increases with the increase of the tuning parameter β which controls the spreading of the sample on the network.

	n=10		n=20		n=30		n=40		n=50	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
RNS	0.65	0.03	0.76	0.03	0.84	0.02	0.90	0.01	0.94	0.01
LPM2	0.67	0.02	0.78	0.03	0.86	0.02	0.92	0.01	0.96	0.00
SCPS	0.68	0.02	0.78	0.02	0.87	0.02	0.93	0.01	0.96	0.00
PWD1	0.68	0.03	0.79	0.03	0.89	0.02	0.94	0.01	0.97	0.00
PWD5	0.81	0.03	0.94	0.01	0.96	0.00	0.97	0.00	0.98	0.00
PWD10	0.84	0.02	0.93	0.01	0.96	0.00	0.97	0.00	0.98	0.00

Table 1 Means and corresponding variances of the accuracy index for the different sampling designs and different sample sizes.

4 Conclusion

In this work, we present a probabilistic sampling procedure for networked data that, based on the within-sample distance, helps spread the sampled nodes on the graph. Moreover, the choice of the tuning parameter β allows to adjusts the design for the amount of spread required. The well-spread samples enable us to achieve more accurate classification results in a semi-supervised learning framework.
The proposed sampling procedure can be exploited to select the training set for a supervised task where node features are taken into account, in addition to the relational information enclosed in the graph. In this setting, the sampling design's probabilistic inherent nature can also be proved useful in splitting a sub-graph into training set and validation set, for cross-validation purposes. Finally, scalability of the sampling procedure to large networks can be achieved by using the algorithm proposed by Piersimoni and Benedetti [2017] which allows a fast selection of well-spread samples.

References

Ahmed, K. N., Neville J., and Kompella R. R.: Network Sampling Designs for Relational Classification. *ICWSM*, 2012.

Ahmed, K. N., Neville J., and Kompella R. R.: Network Sampling: From Static to Streaming Graphs. *ACM Trans. Knowl. Discov. Data*, 8(2), 2013. doi: 10.1145/2601438.

Benedetti R. and Piersimoni F.: A spatially balanced design with probability function proportional to the within sample distance. *Biom. J.*, 59(5):1067–1084, 2017. doi: 10.1002/bimj.201600194.

Benedetti R., Piersimoni F., and Postiglione P.: Alternative and complementary approaches to spatially balanced samples. *Metron*, 75(3):249–264, 2017a. doi: 10.1007/s40300-017-0123-1.

Benedetti R., Piersimoni F., and Postiglione P.: Spatially Balanced Sampling: A Review and A Reappraisal. *Int. Stat. Rev.*, 85(3):439–454, 2017b. doi: 10.1111/insr.12216. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12216.

del Gobbo E., Fontanella S., Sarra A., and Fontanella L.: Emerging Topics in Brexit Debate on Twitter Around the Deadlines. *Soc. Indic. Res.*, 2020. doi: 10.1007/s11205-020-02442-4.

Espìn-Noboa L., Wagner C., Karimi F., and Lerman K.: Towards Quantifying Sampling Bias in Network Inference. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1277–1285, 2018. doi: 10.1145/3184558.3191567.

Grafström A.: Spatially correlated Poisson sampling. *J. Stat. Plan. Inference*, 142(1):139–147, 2012. doi: 10.1016/j.jspi.2011.07.003.

Grafström A., Lundström N.L.P. , and Schelin L.: Spatially Balanced Sampling through the Pivotal Method. *Biometrics*, 68(2):514–520, 2012. doi: 10.1111/j.1541-0420.2011.01699.x.

Leskovec J. and Faloutsos C.: Sampling from Large Graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 631–636, New York, NY, USA, 2006. Association for Computing Machinery. doi: 10.1145/1150402.1150479.

Li B. and Pi D.: Network representation learning: a systematic literature review. *Neural. Comput. Appl.*, 32(21):16647–16679, 2020.

McPherson M., Birds of a Feather: Homophily in Social Networks L., and Cook J. M.: Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.*, 27(1):415–444, 2001. doi: 10.1146/annurev.soc.27.1.415.

Piersimoni F. and Benedetti R.: Fast Selection of Spatially Balanced Samples. *arXiv: Methodology*, 2017. https://arxiv.org/abs/1710.09116

Ruggeri N. and De Bacco C.: Sampling on networks: estimating spectral centrality measures and their impact in evaluating other relevant network measures. *Appl. Netw. Sci.*, 5(1), 2020. doi: 10.1007/s41109-020-00324-9.

van Engelen J.E. and Hoos H.H.: A survey on semi-supervised learning. *Mach. Learn.*, 109(1), 2020. doi: 10.1007/s10994-019-05855-6.

Wagner C., Singer P., Karimi F., Pfeffer J.n , and Strohmaier M.: Sampling from Social Networks with Attributes. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1181–1190, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. doi: 10.1145/3038912.3052665.

Zhu X. and Ghahramani Z.: Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107, Carnegie Mellon University, 2002.

Zhu X., Ghahramani Z., and Lafferty J.: Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 912–919. AAAI Press, 2003.

A sequential adaptive sampling scheme for rare populations with a network structure

Uno schema di campionamento adattivo sequenziale per popolazioni rare con una struttura di rete

Emilia Rocco

Abstract The design of an efficient sample for surveying rare-trait populations still continues to be a challenging task. This contribution discusses a sampling strategy, which, integrating an adaptive component into a sequential selection, aims to intensify the detection of positive cases both by exploiting the links among population units and by spreading the sample over the whole population. Obviously, if a separate frame for a rare population exists the sample may be selected using standard methods as well as if adequate auxiliary information are available. The suggested approach refers to situations in which neither of these two conditions happen. Many factors related to the structure more or less clustered of the population and to the characteristics of the sampling design may be decisive in such a situation.

Abstract *La definizione di un disegno di campionamento efficiente per lo studio di popolazioni con tratti rari continua a essere un compito impegnativo. In questo contributo viene discusso un disegno di campionamento che, integrando una componente adattativa in uno schema sequenziale, mira ad intensificare l'individuazione di casi positivi sia sfruttando i legami tra le unità della popolazione sia distribuendo il campione sull'intera popolazione. Ovviamente, se sono disponibili una lista separata delle unità rare, o adeguate informazioni ausiliarie si può utilizzare un disegno tradizionale. L'approccio suggerito riguarda casi in cui nessuna di queste due condizioni si verifica. In essi, molti fattori legati alla struttura più o meno raggruppata della popolazione e a scelte relative al disegno campionario sono rilevanti.*

Key words: adaptive web sampling, graph setting, link tracing, Markov chain Monte Carlo methods, natural groups, sequential selection

Emilia Rocco
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Viale G. Morgagni 59, 50134 Firenze (Italy), e-mail: emilia.rocco@unifi.it

1 Introduction

Often, rare populations, such as endangered species, drug users and individuals infected by rare genetic or contagious diseases, tend to be highly clumped with clumps widely separated. For surveying these populations, when neither separate frames of adequate coverage exist, nor information correlated to the rare trait are available in advance, adaptive link-tracing designs provide the only practical way to obtain a large enough sample. The basic idea behind these designs is to start with an initial sample of units and then to follow links among units for adding more members of the rare population to the sample. The decision whether to follow a link from a specific units may depend on the value of the variable of interest or on values of any other variable observed during the survey. In some of these designs , the link-tracing procedure goes on as long as relevant values of the variable of interest continue to be found, i.e. as long as the encountered clusters or natural aggregations of units are completely sampled. This happens in ordinary adaptive cluster sampling (Thompson 1990), in some types of network sampling (Birnbaum and Sirken 1965) and in some snowball designs (Frank and Snijders 1994), just to name a few examples. It is evident that in these cases the final sample size is unknown and can be quite variable depending on the patchiness of the population. This is a limit for the practical use of these sampling designs, above all in case of social phenomena with population units connected in networks of huge size. Our interest is in an adaptive sampling design with the following features:

1. the final sample size is fixed in advance;
2. the initial sample is a probabilistic one;
3. the selection procedure cares for intensifying both the detection of positive cases and the spread of the sample over the whole population;
4. the selection procedure may exploit auxiliary information, when available, but it is usable in spite of them.

We consider an extension of the adaptive web sampling introduced in Thompson (2006). In adaptive web sampling firstly, an initial sample is selected by means of a conventional sampling design and then the remaining units - until the desired sample size is reached - are sequentially selected in order to concentrate most of the remaining sampling effort in the areas that appear to be of the greatest interest based on previously observed survey values. Rocco (2016) has examined the use of adaptive web sampling for surveying finite populations that are distributed over space, showing the advantages of selecting the initial sample through a spatially balanced sampling, i.e. through a design that produces an initial well spread sample. When a population has not a spatial distribution but auxiliary variables are available, we can use a similar approach for selecting an initial sample that is well spread in the auxiliary space, i.e. a sample that is balanced, or approximately balanced, on the auxiliary variables. On the other hand, when we have only a list of labels for selecting a sample, we can only try to use the information observed during the survey for both to concentrate a large portion of the sampling effort in the areas that

appear to be of the greatest interest and to spread the other sampling units as much as possible over the whole population.

2 Sampling scheme and estimation

Let us consider a population U of N units labelled $i = 1, ..., N$ and a study variable y associated to each unit. The variable y may be a dichotomous one with $y_i = 1$ if unit i is a case (a member of the rare population) and $y_i = 0$ otherwise, or may be a quantitative variable that designates a characteristic/trait of the rare units. In addition, we assume that the value w_{ij} of an indicator variable is associated to each pair of units i and j with $w_{ij} = 1$ if there is a link from unit i to unit j, and $w_{ij} = 0$ otherwise. The link variable, like the study variable y, is observed only through the sample.
In order to select a sample of fixed size n that is as spread as possible among the whole population and at the same time includes as many as possible relevant units, we suggest the following sequential scheme in $K \leq n$ steps. In the first step, an initial sample s_0, of $n_0 \leq n$ units is selected by simple random sampling. More generally, the initial sample may be selected with any conventional sampling design, but designs different from the simple random one assume some knowledge of the population distribution. For each unit in s_0 the value y_i and all the link values out from i are observed. If the unit i is a case, i.e. $y_i = 1$ or, more in general, y_i satisfies a condition of interest specified in advance, then the unit i together with any associated variable of interest is enclosed in a subset of the sample so far selected. This subset is called *active set*. In particular, all the links out from i are considered variables of interest associated to unit i and enclosed in the active set. If the unit i is not a case, the positive link values out from i going to other units not already in the sample are used to partition the set of the total units not already in the sample in two subsets: all the units having a positive link with at least one selected negative case ($y_i = 0$) and, all the other units. Both the active set and the partition in two subsets of units not already in the sample are then updated after each of the subsequent selection steps. At each step after the initial sample, the selection is made unit by unit. At the kth $(k = 1, ..., n - n_0)$ step, one unit, i.e. a sample s_k of size 1, is selected from a mixture distribution, so that with high probability p, one of the links in the current active set is selected at random and followed to bring a new unit into the sample, and with low probability $(1 - p)$, a new unit is randomly selected from one of the two partitions of units not already in the sample. It is selected from the set of units not linked to the selected negative case with probability $(1 - p)q$, with q close to 1, and from the other set with probability $(1 - p)(1 - q)$. Therefore, the probability that unit i is selected in the kth step is defined as:

$$q_{ki} = p\frac{w_{\alpha_k i}}{w_{\alpha_{k+}}} + (1-p)\left(q\frac{1}{N - n_{s_{ck}} - N_{\bar{l}}} - (1-q)\frac{1}{N_{\bar{l}}}\right) \tag{1}$$

where $s_{ck} = \bigcup_{i=0}^{k-1} s_i$ and α_k ($\alpha_k \subset s_{ck}$) denote the current sample and the current active set at step k, $N_{\bar{l}}$ is the number of not sampled units that have at least a link with the negative cases in the current sample, $n_{s_{ck}}$ is the number of units in the current sample, $w_{\alpha_{k+}} = \sum_{i \in \alpha_k, j \in \bar{s}_{ck}} w_{ij}$ is the total number of links out, from the active set to units not in the current sample and $w_{\alpha_k i} = \sum_{j \in \alpha_k} w_{ij}$ is the number of the links out, from the active set to unit i.
Moreover, at each step, if there are no links out from the current active set to any unsampled unit, the next unit is randomly selected from the collection of unsampled units. In these cases, then:

$$q_{ki} = \left(q \frac{1}{N - n_{s_{ck}} - N_{\bar{l}}} - (1-q) \frac{1}{N_{\bar{l}}} \right) \tag{2}$$

If there are no links from the current active set and no unsampled units that have at least a link with the negative cases in the current sample, then:

$$q_{ki} = \frac{1}{N - n_{s_{ck}}} \tag{3}$$

If there are links from the current active set but there are no unsampled units that have at least a link with the negative cases in the current sample, then:

$$q_{ki} = p \frac{w_{\alpha_k i}}{w_{\alpha_{k+}}} + (1-p) \frac{1}{N - n_{s_{ck}}} \tag{4}$$

The sampling continues step by step until the desired sample size n is reached and the ordered sample $\mathbf{s} = \{s_0, s_1, ..., s_{n-n_0}\}$ of fixed size n is obtained.
The probabilities p and q may themselves depend on the current sample, active set or partition of the units not earlier selected.

2.1 Estimation

For adaptive web sampling Thompson (2006) suggests more possible estimators of the population mean that can be used also in our case. Among these, the simplest and most accurate one is obtained by finding, via the Rao-Blackwell approach, the conditional expectation of sample mean of the initial sample, $\bar{y}_0(\mathbf{s})$, given the reduced set of data $d_r = \{(i, y_i, w_{i+}, w_{ij}), i \in \mathbf{s}, j \in \mathbf{s}\}$. Its expression is:

$$\hat{\mu} = \sum_{\mathbf{S}: r(\mathbf{S}) = s} \bar{y}_0(\mathbf{s}) p(\mathbf{s}|d_r) \tag{5}$$

For the expression of its variance we refer to Thompson (2006) and Rocco (2016).
Computation of the estimator $\hat{\mu}$ and of its variance estimator requires the enumeration of all the reorderings of the sample units. For each reordering the probability of that reordering needs to be computed along with the value of the mean estimator

and of its variance estimator. Enumerative calculus is prohibitive even for relatively small values of n. The Markov chain resampling procedure based on the Markov chain accept/reject procedure of Hasting (1970), suggested in Thompson (2006) for making inference computationally feasible, can be extended to our case. Therefore, by denoting with n_r the number of selected resampled permutations and with $\bar{y}_{0j}$ the sample mean of the initial sample for the permutation j, the resampling estimator used to replace $\hat{\mu}$ is:

$$\tilde{\mu} = \frac{1}{n_r} \sum_{j=0}^{n_r-1} \bar{y}_{0j} \tag{6}$$

3 A simulation study and final consideration

In this section the properties of the suggested approach were investigated through sampling simulations carried out from four simulated populations. Figure 1 shows the network structure of these populations for which the parameter of interest is the proportion of units with a rare trait (black nodes in the figure). In all populations, of 200 units, the relevant units tend to be aggregated: they are included in a single network in the populations A and D and in two networks in the populations B and C; have an average of links between them equal to 5 in the populations A, C and D and to 3.6 in the population B; each of them is linked only to 1 or 2 non-relevant units in the populations A, C and D and to zero non-relevant units in the population B. Moreover, for each population a general natural clustering structure is assumed even for the non-relevant units. For each population a Monte Carlo experiment has

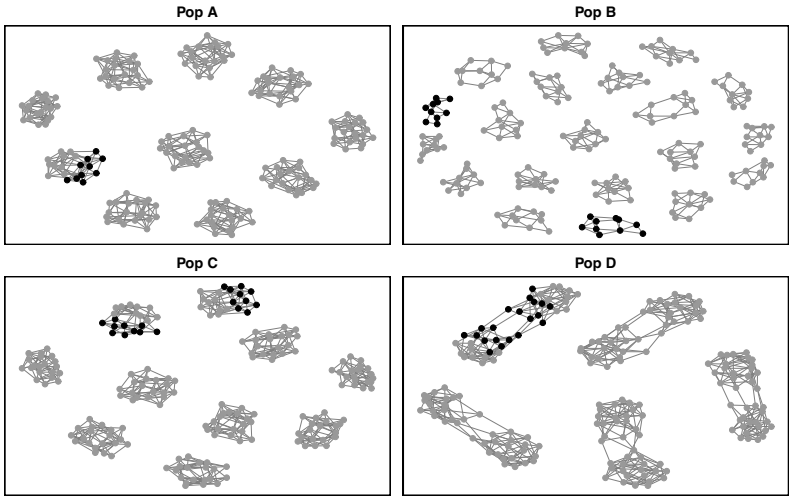


Fig. 1 Four different simulated population having a clustering structure and a rare trait. The black nodes represent the units with the rare trait.

been carried on. The number of simulation runs for each experiment is fixed at 500 and the number of Markov chain resamplings used in the estimation procedure is set to 10,000. The condition that activates the adaptive selection is $y_i = 1$, the size of initial sample is $n_0 = 15$ and the final sample size is $n = 20$. The design used for the initial sample s_0 is the simple random one and the value of the mixing probabilities p and q are chosen as follows: q is constant equal to 0.9; while p varies depending on the current sample: if the number of relevant units in it is less or equal to a fixed threshold (15% of n) it is equal to 0.9 otherwise it is equal to 0.01. This because a high value of p increases the probability of observing more relevant units, but when these are rare in the the population, the inclusion in the sample of a too large number of them can produce very large estimation with certain samples having small conditional selection probabilities.

Table 1 Reative efficiency of $\tilde{\mu}$

Populations	A	B	C	D
$eff(\tilde{\mu})$	0.87	0.95	0.87	0.96

Table 1 shows for each experiment the efficiency of $\tilde{\mu}$ relative to that of the mean of a simple random sample of equal size. The basic idea behind the suggested approach is that units close to observed negative cases are less likely to be positive cases as well as those close to positive cases are more likely to be positive too. For all the considered populations, that simulate such a situation, the suggested approach produces a moderate gain in efficiency respect to the simple random sampling. Many factors affect the effectiveness of the proposed strategy. Some are characteristics of the population like the portion of positive cases, the number of units to which each unit is linked (while in the case of units in space and spatial proximity relationship the number of contiguous units is limited, for other types of relationships, such as social, the number of links can be very high) and how many of those linked to a positive case are in turn positive. Other factors depend on the sampling design like the final sample size, the size and the type of initial sample and its consequent ability to capture positive cases in order to activate the subsequent adaptive selection. Prospective research endeavors will consider an in depth investigation of all them.

References

1. Birnbaum, Z.W., Sirken M.G.: Design of sample surveys to estimate the prevalence of rare Diseases: Three unbiased estimates. Vital and Health Statistics, Series 2, Vol. 6, Washington DC: Government Printing Office (1965)
2. Frank, O., Snijders, T.: Estimating the size of hidden populations using snowball sampling. J Off Stat **10**, 53–67 (1994)
3. Hastings, W.K.: Monte-Carlo sampling methods using MarKov chains and their applications. Biometrika. **57**, 97-109 (1970)
4. Rocco, E.: Spatially-balanced adaptive web sampling. Environ Ecol Stat **23**, 219–231 (2016)
5. Thompson, S.K.: Adaptive cluster sampling. J Am Stat Assoc **85**, 1050–1059 (1990)
6. Thompson, S.K.: Adaptive web sampling. Biometrics. **62**, 1224–1234 (2006)

3.14 New perspectives on multidimensional child poverty

Estimating uncertainty for child poverty indicators: The Case of Mediterranean Countries

La stima dell'incertezza negli indicatori di povertà infantile: il caso dei paesi Mediterranei

Benedetti Ilaria, Crescenzi Federico, De Santis Riccardo

Abstract Over the last few years, there has been increased interest in compiling poverty indicators for children, as well as in providing uncertainty measures associated with point estimates. In this paper, we provide child point and bootstrapped relative standard error estimates for the At-risk-of-poverty Rate and Gini coefficient for Mediterranean countries. Using the 2018 EU-SILC survey, our results show that for these categories, poverty tends to be higher when compared to the national estimates for most of the analysed countries.

Abstract *Negli ultimi anni, c'è stato un crescente interesse nella compilazione di indicatori di povertà infantile e giovanile, nonché nel fornire misure di incertezza associate a stime puntuali. In questo lavoro, forniamo stime puntuali e degli errori standard del tasso di rischio di povertà e indicatore di disuguaglianza di Gini infantile per i paesi mediterranei. A questo scopo, abbiamo adottato il metodo di replicazione Bootstrap grazie alle sue proprietà convenienti. Utilizzando l'indagine EU-SILC 2018, i nostri risultati rivelano che le stime puntuali e gli errori standard per la povertà infantile sono più elevati rispetto alle stime nazionali per la maggior parte dei paesi analizzati.*

Key words: Child poverty indicators, Uncertainty estimates, Mediterranean countries

Benedetti Ilaria
Department of Economics, Engineering, Society and Business organizzation, University of Tuscia. e-mail: i.benedetti@unitus.it

Crescenzi Federico
Department of Statistics, Computer Science, Applications, "G.Parenti", University of Florence. e-mail: federico.crescenzi@unifi.it

De Santis Riccardo
Department of Statistical Sciences, University of Padova. e-mail: riccardo.desantis.1@phd.unipd.it

1 Introduction

In the context of poverty and social exclusion indicators, measuring child poverty is a key topic in social science research, due to its importance for national governments and international organizations. The first of the Sustainable Development Goals (SDGs) has brought out the need to ensure successful outcomes for today's children by building the foundations of our societies' future well-being [2]. The persistence of child poverty at rather high levels compared to national poverty rates explains why reducing child poverty is now high on the social policy agenda of many OECD countries [3].

Despite the fact that several initiatives have been carried out for measuring and monitoring children's poverty over time and European countries, to the authors' knowledge the issues of uncertainty measurements have not yet been fully explored. Given the key role played by poverty indicators in designing and monitoring social progress in the EU, it is essential that the indicators used for measuring poverty are of sufficient high quality, especially in terms of their accuracy and reliability.

During the last years, several statistical authorities and organisations have started investing in identifying ways to measure and communicate data uncertainty. From a methodological point of view the formulae for calculating standard errors also depend on the statistics to be computed and the sampling design included in the survey adopted by each country [5]. A first contribution of this paper provide updated figures regarding child poverty of the population in the Mediterranean countries. Among the income-poverty measures, we selected the at-risk-of-poverty rate (AROP) while among the income-inequality indicators, we selected the Gini coefficient. In order to provide standard error estimations we provide an empirical application using the Bootstrap replication method.

The rest of this paper is organized as follows: Section 2 focuses on the child economic and inequality situation in the Mediterranean countries, in addition it addresses the issue of measuring uncertainty for poverty indicators. Section 3 discussed the Bootstrap approach for variance estimation, while Section 4 illustrates the main characteristics of the EU-SILC data and the main results obtained for the Mediterranean countries. Section 5 reports conclusions and suggestions for further research.

2 Child poverty in Mediterranean countries and the issue of uncertainty measurement

Around 23.4% of European children live in income poverty, 8.5% live in severe material deprivation and 9.3% in workless households. Child poverty is a problem for all Member States though prevalence and intensity is highest in some of the Central Eastern European, Baltic and Mediterranean states. Moreover, children and young people were some of the main victims of the 2008 financial crisis. In particular,

some Mediterranean countries such as Italy and Greece suffered from the economic crisis more than other European countries [1].

In the Mediterranean countries, in 2015 more than a third of children were at risk of poverty or social exclusion. The highest rate was observed in Greece (37.8%), Spain (34.4%) and Italy (33.5%). Moreover, in approximately half the EU member states, the at-risk-of-poverty or social exclusion rate grew from 2010 to 2015, with the highest increases recorded in Greece (from 28.7% in 2010 to 37.8% in 2015), Cyprus (+7.1 percent points) and Italy (+4.0 percent points). The persistence of child poverty at rather high levels compared to national poverty rates and its rebound with the economic crisis explains why reducing child poverty is now high on the social policy agenda of many OECD countries [3]. Several factors could affect child poverty and inequality [13]. Since children's circumstances almost always depend on their parents' and family backgrounds, a lack of education can be a major risk factor for child poverty or social exclusion. Lower educational levels can often mean that parents have less disposable income from wages or salaries. Moreover, children's likelihood of being AROP is also determined by their parents' country of birth. Household composition is a further factor influencing the probability to be at risk of poverty or social exclusion. The study of children's well-being is characterized by a plurality of approaches and measures [1], [6]. Although a wide stream of literature addressed the multidimensional aspect of child poverty, relative monetary measures of poverty are crucial for evaluating children's well-being over time and represent the main indicator to measure child poverty.

This paper contributes to this stream of literature by providing a detailed picture of the current economic and inequality situation in the Mediterranean countries. We used data collected on a regular basis through the EU-SILC survey. To this aim, we focus on the EU-SILC Laeken indicators which comprise both income-poverty and income-inequality measures. In this paper we selected one income-poverty measures, the AROP, which belong to the class of the Foster-Greer-Thorbecke (FGT) measures, and one income-inequality measures: the Gini coefficient. AROP is computed by Eurostat as the share of people with an equivalised disposable income below the at-risk-of-poverty threshold (ARPT), which is set at 60% of the national median equivalised disposable income after social transfers. While, the Gini coefficient measures the extent (0 to 100) to which the distribution of income deviates from a perfectly equal distribution. Given the key role played by poverty indicators in designing and monitoring social progress in the EU, it is paramount to produce and communicate to the public measures of the associated inherent and unavoidable uncertainty of point estimates. Indeed, measuring uncertainty around point estimates is a complex and challenging task, which may involve the use of sophisticated statistical methods as well as the adoption of econometric techniques and subjective judgement [4]. Regarding the issue of uncertainty measurement, numerous variance estimation approaches have been developed for measuring uncertainty of poverty indicators, such as linearization and re-sampling methods. Focusing on re-sampling methods, bootstrap tests based on the FGT poverty measure perform very well as soon as sample sizes are large enough for there to be more than around 10 observations below the poverty line [8].

3 Bootstrap replication method for estimating uncertainty uncertainty

To obtain a variance estimate for the nonlinear statistics considered in this paper we followed an approach based on the bootstrap. In particular, we make use of the so-called naive bootstrap approach to estimate confidence intervals as implemented in [9]. Let $X = (X_1; ...; X_n)'$ denote a survey sample of n observations. The algorithm is implemented as follows:

1. Draw R independent samples $X_1^*, ..., X_R^*$ from X, where each one contains n observations drawn with replacement;
2. Compute the bootstrap replicate estimates $\hat{\lambda}_r^* = \hat{\lambda}(X_r^*)$ for each X_r^* $r = 1, ..., R$ where $\hat{\lambda}$ denotes an estimator of the poverty indicator of interest.
3. Estimate the variance $V\left(\hat{\lambda}\right)$ by using the variance of the R bootstrap replicate estimates: $\hat{V}\left(\hat{\lambda}\right) = (R-1)^{-1} \sum_{r=1}^{R} (\hat{\lambda}(X^r) - R^{-1} \sum_{j=1}^{R} \hat{\lambda}(X^j))^2$
4. The confidence interval at confidence level $(1-\alpha)$ is then calculated as:

$$\left[2\hat{\lambda} - \hat{\lambda}^*_{((R+1)(1-\frac{\alpha}{2}))}, 2\hat{\lambda} - \hat{\lambda}^*_{((R+1)(\frac{\alpha}{2}))}\right]$$

where $\hat{\lambda}^*_{(1)} \leq \hat{\lambda}^*_{(2)} \leq ... \leq \hat{\lambda}^*_{(R)}$. In the case of sampling designs that involve different strata, the observations are re-sampled independently within each stratum.

4 Data and results

We use cross-sectional data from the EU-SILC survey collecting timely and comparable cross-sectional and longitudinal microdata on income, poverty, social exclusion and living conditions.

In this paper, we have selected the following Mediterranean countries: Italy (IT),France (FR), Malta (MT), Spain (ES), Portugal (PT), Cyprus (CY), Greece (EL), Croatia (HR) according to their sampling design by using cross-sectional data for years 2018, corresponding to the income year 2017. Point estimates and relative standard error estimates are reported in Table 1.

It is evident that the most significant discrepancies existing between the group of children (0-15) and youngsters (16-24) are to be found in Greece and Cyprus. On the contrary, very little differences exist in these two strata for Italy and Portugal. On average, we are able to obtain satisfactory estimates of variability for each country - either for the Gini index and at-risk-of-poverty - suggesting a good level of accuracy for the point estimates.

Focusing on (relative) standard error estimates at the national level, it is essential to note that our results show a satisfactory level of reliability, since the estimated relative standard errors are lower than 5%, as emphasized in [12]. Indeed, even

if precision thresholds are generally survey specific and depend on the required reliability and resource-related political decision, specifying the degree of precision is an important step when planning a sample survey.

Table 1 Arop and Gini estimates for mediterranean countries

		Gini		Arop	
Country	stratum	Est.	RSE	Est.	RSE
PT	child	32.230	2.29%	18.189	5.94%
	young	33.717	2.52%	18.174	4.34%
	national	33.577	1.23%	17.506	2.72%
MT	child	25.646	2.69%	10.419	11.46%
	young	27.982	3.82%	11.159	10.95%
	national	27.468	1.16%	11.807	3.76%
IT	child	33.511	1.63%	24.064	2.95%
	young	35.086	1.95%	24.299	2.98%
	national	33.336	0.97%	19.906	1.63%
FR	child	28.585	6.12%	18.107	6.35%
	young	27.105	3.24%	16.472	5.57%
	national	28.706	3.02%	13.091	3.23%
HR	child	31.305	3.08%	20.923	7.70%
	young	29.476	2.19%	22.267	6.71%
	national	30.800	1.36%	21.894	3.03%
ES	child	32.647	2.13%	22.377	4.85%
	young	34.048	1.77%	24.833	3.71%
	national	32.843	1.22%	20.576	3.22%
EL	child	32.749	3.67%	20.546	4.12%
	young	32.900	1.96%	25.579	2.54%
	national	31.249	1.62%	16.897	2.44%
CY	child	31.089	4.23%	20.866	19.75%
	young	28.317	4.29%	16.635	10.34%
	national	30.052	1.90%	17.683	5.58%

5 Conclusion

In this paper we computed measures of uncertainty in children's AROP and Gini indicators based on Mediterranean European countries. Information about the sampling variability of point estimates is essential when comparing poverty rates in different geographical areas or socio-economic groups. The bootstrap method is implemented in order to obtain relative standard error estimates for the AROP and Gini indicators. The computation of standard errors for the main official poverty

measures is a complex task due to the characteristics of these indicators, which are often expressed as non-linear statistics.

The bootstrap turned out to be an easy and effective approach to compare countries that adopt different sampling designs. In fact, there are not particular constraints on the design such as a minimum number of units inside PSUs. The results emerged from this study suggest the existence of relevant differences among mediterranean countries in terms of child AROP and Gini indicators. However, there exist significant differences in the percentage of at-risk-of-poverty children and youngsters when compared to the national values. An integrated and child rights-based approach should be a priority for the EU approach on child poverty. Thus, it is necessary to monitor the effectiveness of the already implemented policies and possibly to propose *ad-hoc* policies to combat and eradicate poverty and exclusion halving education.

References

1. D'Agostino, A. and Gagliardi, F. and Giusti, C. and Potsi, A.: Investigating the impact of the economic crisis on children's wellbeing in four European countries. Social science research. **84**, 102322 (2019)
2. Brazier, C.: Building the Future: Children and the Sustainable Development Goals in Rich Countries. UNICEF. (2017)
3. Cantillon, B., Chzhen, Y. Handa, S. and Nolan, B.: Children of austerity: Impact of the great recession on child poverty in rich countries. Oxford University Press. (2017)
4. Manski, C. F.: Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern. Journal of Economic Literature.**53**, 631–53 (2015)
5. Betti, G. Gagliardi, F., Lemmi, A. and Verma, V.: Subnational indicators of poverty and deprivation in Europe: methodology and applications. Cambridge Journal of Regions, Economy and Society. Oxford University Press. **5**, 129–147 (2012)
6. Betti, G. and Caruana, E. and Gusman, S. and Neri, L.: Economic poverty and inequality at regional level in Malta: focus on the situation of children. Economy of Region. **11**, 114–122 (2015)
7. Palmitesta, P., Provasi, C. and Spera, C.: Confidence interval estimation for inequality indices of the Gini family. Computational Economics, Springer. **16**, 137–147 (2000)
8. Davidson, R. and Flachaire, E.: Asymptotic and bootstrap inference for inequality and poverty measures. Journal of Econometrics, Elsevier. **141**, 141–166 (2007)
9. Alfons, A. and Templ, M.: Estimation of social exclusion indicators from complex surveys: The R package laeken. KU Leuven, Faculty of Business and Economics Working Paper (2007)
10. Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap. CRC press. (1994)
11. Betti, G. Gagliardi, F. and Verma, V.: Simplified Jackknife variance estimates for fuzzy measures of multidimensional poverty. International Statistical Review, **86(1)**, 68–86 (2008)
12. Ardilly, P.: Les techniques de sondage, Editions Technip. (2016)
13. D'Agostino, A., Grilli, G. and Regoli, A.: The determinants of subjective well-being of young adults in Europe. Applied Research in Quality of Life, Springer, **14(1)**, 85–112 (2019)

Child poverty and government social spending in the European Union during the economic crisis

Povertà infantile e spesa pubblica nell'Unione Europea durante la crisi economica

Angeles Sánchez and María Navarro[1]

Abstract: Fighting child poverty is desirable to foster the sustainability of the social well-being for coming generations and, in addition, to ensure equality of opportunities for all children. Using panel data methodology, this work analyses the association between child poverty and government social spending in the 28 Member States of the European Union during the last economic crisis (2008-2014). We confirm that the government social spending as a whole, as well as the government spending on education and health negatively correlated with child poverty during the economic crisis, while government social protection spending did not it. That is to say, within the context of economic crisis in which European Union displayed policies to ensure sustainability of public finance, reductions in social spending and increases in child poverty could be associated.

Abstract: *La lotta alla povertà infantile è importante per promuovere la sostenibilità del benessere sociale per le generazioni future e, anche, per garantire l'uguaglianza di opportunità tra i bambini. Utilizzando dati panel, questo lavoro analizza l'associazione tra povertà infantile, spesa sociale pubblica nei 28 Stati membri dell'Unione Europea durante l'ultima crisi economica (2008-2014). I risultati confermano che la spesa pubblica nel suo insiemecosì come la spesa pubblica per istruzione e salute sono correlate negativamente* con *la povertà infantile durante la crisi economica i, mentre la spesa pubblica per la protezione sociale no. Vale a dire, nel contesto della crisi economica in cui l'Unione Europea ha mostrato politiche per garantire la sostenibilità della finanza pubblica, potrebbero essere associate riduzioni della spesa sociale e aumenti della povertà infantile.*

Key words: Multidimensional child poverty, social spending, tax structure, public policies, welfare state, panel data

[1] Angeles Sánchez, Department of Applied Economics, University of Granada (Spain); sancheza@ugr.es
María Navarro, Department of Economics and Business, University of Almeria (Spain); marianh@ual.es

Introduction

Over the last few years, there has been a growing interest on child poverty and social exclusion for the majority of the governments in developed countries, especially since the last economic crisis.

Two different reasons do the study and prevention of child poverty matter. Firstly, children were the most affected group by the poverty during the last economic crisis in the European Union (EU). In 2014, 27.5% of the population under the age of 16 in the EU-28 was considered at-risk-of poverty or social exclusion (AROPE) compared to 23.7% of the entire population (Eurostat, Statistics on Income and living conditions). Secondly, the social and economic future of a country depends on its capacity to fight child poverty and social exclusion, since these problems represent a threat to future generations in terms of both economic development and social stability (Diris et al., 2017; Esping-Andersen et al., 2002; Frazer and Marlier, 2017). Consequently, fighting child poverty is desirable to achieve equality of opportunities for all children, as well as fostering improvements of living standards and prosperity (Brazier et al., 2017; OECD, 2019).

In this work, panel data methodology is used to analyze the relationship between government social expenditure and child poverty in the 28 Member States of the EU, for the period of the economic crisis between 2008 and 2014. Particularly, our main aims are to check under a framework of economic crisis (1) the relationship between government social expenditure and child poverty, and (2) whether all the categories of social expenditure correlated in the same way.

Child poverty evolution

Figure 1 reports poverty rates through AROPE for people under 16 years old, i.e. child poverty, for the years 2008 and 2014, in the Member States and EU-28 as a whole. As a firts glance, it is worth noting that the child poverty rates are greater in 2014 than in 2008 in almost the countries. We can also observe that there are disparities in child poverty between countries. In some of them, such as, for instance, Austria, Belgium, Germany, Denmark, France or The Netherlands, child poverty rates are below the EU-28 in both years, whereas Romania, Bulgary, Hungary, Greece, Spain, Italy or the United Kingdom report levels above the EU-28 as a whole.

Moreover, countries were affected differently by the economic crisis. Changes in child poverty rates from 2008 to 2014 were modest for the majority of countries, except for Greece, Spain, Italy, Hungary and Malta, where child poverty has considerably increased between both years. Thus, these countries have gotten worse in terms of child poverty from the beginning of the economic crisis, whereas only Croatia or Poland have gotten better.

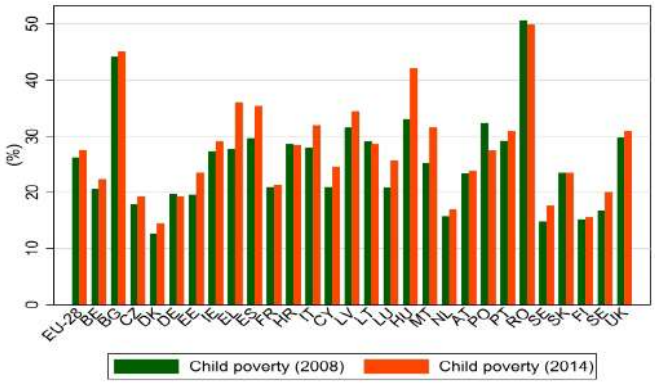


Figure 1. Child poverty across European Union in 2008 and 2018. The values for Croatia in 2008 are referred to 2010 (first available year with information), and for EU-28 are referred to the rate of EU-27. Adapted from Eurostat, Income and living conditions.

Data and Variables

We used highly balanced macro panel data with information from Eurostat on child poverty, government social spending, tax structure, and economic and socio-demographic variables of EU-28 Member States over the period 2008-2014.

Child Poverty is the dependent variable in our analysis. Particularly, we consider the AROPE indicator for children under 16 years old. This represents the percentage of the population younger than 16 that share in at least one of the following three conditions: poverty risk, severe material deprivation and/or low work intensity.

The explanatory variables are classified into two groups: fiscal policy and socio-economic factors. In the first group, we include *Government Education Spending, Government Health Spending* and *Government Social Protection Spending* and the variable *Government Social Spending*, which is the sum of the previous three variables. All of them are measured in constant 2015 Euros per billion inhabitants. We also include a dummy variable to account for the tax structure of the Member State (*Tax Structure*), which measures the relationship between direct and indirect taxes, in order to account for how the expenditures are financed (Diris et al., 2017; Sánchez and Pérez-Corral, 2018). This takes the value 1 if the ratio of the country is greater than the value of the ratio for the EU-28 as a whole, and 0 otherwise.

Within the second group of explanatory variables, *Economic Growth Rate* allows analysing the influence of economic context on child poverty. This variable is introduced in the models with a delay of two years. The variable *Early Leavers* reflects the percentage of population aged 18 to 24 that has dropped out of education and training, and *Age Childbirth* measures the average age of women at birth of first child.

Methodology

In order to study the relationchip between child poverty and fiscal policy and economic growth, we define the following model:

$$P_{it} = \omega_0 + \beta' G_{it} + \lambda' T_i + \phi' E_{it-2} + \eta' X_{it} + \alpha_i + u_{it} \qquad (1)$$

i denotes the country and *t* the year. *P* it is the level of child poverty in the country *i* in the year *t*; G_{it} represents the variables of government social expenditure; T_i is the variable dummy, to have a more progressive tax structure than the EU-28 as a whole; E_{it-2} is the two years lagged *Economic Growth Rate;* X_{it} is a set of socio-economic variables in which *Early Leavers* and *Age Childbirth* are included; α_i is the individual effect of each of the countries or unobservable heterogeneity invariant in time; and u_{it} is the error term.

Based on literature (Baltagi, 2014; Hsiao, 2014), and after doing several test to check the nature of the data (see Table 1), we estimate several fixed effects models according to equation (1) with two estimators: Feasible Generalized Least Squares (FGLS) and Panel Corrected Standard Errors (PCSE). To account for time fixed effects in our analysis, we also include a set of year dummies.

Table 1: Results of the tests to choose the estimation method and to analyse the error term

Restrictive F test	F(27,106) = 52.20 (p < 0.001)
Breush and Pagan Lagrangian Multiplier	Chi2(1) = 195.02 (p < 0.001)
Hausman test	Chi2(6) = 38.27 (p < 0.001)
Modified Wald test	Chi2 = 261.88 (p < 0.001)
Wooldridge test for panel data	F(1,27) = 10.38 (p = 0.0033)
Pesaran'test of cross sectional independence	0.766 (p = 0.4434)
Year dummies	F(4,102) = 4.59 (p = 0.0019)

Note. The tests have been carried out taking as a baseline a specification without the dummy of Tax Structure.

Results

Table 2 presents the descriptive statistics of the variables used in the study. Table 3 reports the results of the models estimated using FGLS (Models 1 and 3) and PCSE (Models 2 and 4). Models 1 and 2 are estimated for the total government social spending and Models 3 and 4 include the different concepts of social spending, that is, spending on education, health and social protection.

As expected, a negative association between child poverty and government social spending as a whole is found. Thus, we can confirn that this kind of spending can buffer child poverty. Regarding the different concepts of social spending, both a higher spending on education and health would help to combaint child poverty, whereas a possitive association between child poverty and spending on social protection is found.

Table 2: Descriptive statistics of child poverty in 28 Members States, 2008-2014

Variable	Mean	SD	Min	Max	Median
Child poverty	27.12	9.3	12.7	52.8	26.25
Government Social Spending	7.93	5.72	1.03	25.49	5.91
Government Education Spending	1.42	1.01	0.19	5.02	1.12
Government Health Spending	1.73	1.18	0.21	4.53	1.33
Government Social Protection Spending	4.78	3.61	0.58	16.42	3.54
Economic Growth	0.27	3.75	-14.8	9.3	0.9
Early Leavers	11.62	6.02	2.8	34.9	10.9
Age Chilbirth	29.84	1.14	26.5	31.8	30

Note. N = 140 observations. SD = standard deviation. Adapted from Eurostat: Income and Living Conditions, Government and Finance Statistics, Annual National Accounts, Labour Force Survey and Fertility.

Table 3: Regression analysis: child poverty and government social spending in the European Union-28, 2008-2014

	Model 1	Model 2	Model 3	Model 4
Government Social Spending	-0.387***	-0.384**		
	(0.119)	(0.169)		
Government Education Spending			-4.738***	-4.505**
			(1.423)	(1.822)
Government Health Spending			-4.998***	-4.529***
			(1.358)	(1.700)
Government Social Protection Spending			2.242***	1.884**
			(0.582)	(0.775)
Tax Structure	-1.894*	0.028	0.675	1898
	(1.149)	(1.528)	(1.302)	(1.630)
Economic Growth(t-2)	-0.254***	-0.282***	-0.249***	-0.258***
	(0.070)	(0.092)	(0.072)	(0.089)
Early Leavers	0.584***	0.550***	0.579***	0.532***
	(0.105)	(0.117)	(0.100)	(0.108)
Age Childbirth	-3.794***	-3.783***	-3.821***	-3.648***
	(0.683)	(0.809)	(0.622)	(0.732)
Observations (countries)	140 (28)	140 (28)	140 (28)	140 (28)
Wald's test (Chi2)	179.25***	102.42***	216.95***	144.33***
R-squared		0.766		0.780
Rho		0.747		0.735

Note. Estimators used: Feasible Generalized Least Squares (Models 1 and 3) and Panel Corrected Standard Errors (Models 2 and 4). Standard errors in parentheses. A constant term and time dummies are included in all the models. * $p < 0.1$, ** $p < 0.05$, ***$p < 0.01$.

Concerning control variables, a negative association between child poverty and economic growth is found. Moreover, dropping out of school is possitively correlated with child poverty, while the higher age of the mother at birth of the first child, the lower child poverty is.

Conclusions

Taking into account that during the years of economic crisis (especially the years of the stability of public finances 2011, 2012 and 2013), both spending on education

and health have registered negative annual variation rates in the whole of the EU-28, the results confirm that these reductions in social spending have contributed to the increase in child poverty. For the government social spending as a whole, the same results are observed. These findings are in line with most papers (Benedetti and Betti, 2020; Cantillon et al., 2017; Thévenon et al., 2018). On the contrary, spending on social protection has registered a better performance (with smaller decreases), especially since part of its programs are automatic stabilizers of the economy (i.e. unemployment benefits). In line with other studies (Diris et al., 2017; Vliet and Wang, 2015), the results of the estimated models indicate that its relationship with child poverty is positive.

These findings should be studied in greater depth, distinguishing between groups of countries with different institutional characteristics and analyzing the aspects of social programmes that can explain the effectiveness in the fight against child poverty.

References

Baltagi, B.H.: Econometric Analysis of Panel Data (reprint, 5th ed.). Wiley, London, UK (2014)

Brazier, C.: Building the future: children and the sustainable development goals in rich countries. UNICEF Office of Research - Innocenti, Florence, Italy (2017)

Diris, R., Vandenbroucke, F., Verbist, G.: The impact of pensions, transfers and taxes on child poverty in Europe: the role of size, pro-poorness and child orientation. Socioecon Rev, 15(4), 745–775 (2017) https://doi.org/10.1093/ser/mww045

Esping-Andersen, G., Gallie, D., Hemerijck, A., Myles, J.: Why we need a new welfare state. Oxford University Press, USA (2002)

Frazer, H., Marlier, E.: Progress across Europe in the Implementation of the 2013 EU Recommendation on Investing in Children: Breaking the Cycle of Disadvantage. A study of national policies, European social policy network, Brussels: European Commission (2017)

Hsiao, C.: Analysis of panel data. Cambridge University Press, Cambridge, UK (2014)

OECD: Can Social Protection Be an Engine for Inclusive Growth? (2019) https://doi.org/10.1787/9d95b5d0-en (accessed 2.14.21).

Sánchez, Á., Pérez-Corral, A.L.: Government social expenditure and income inequalities in the European Union. Hacienda Pública Española 227(4), 133–156 (2018)

Vliet, O.V., Wang, C.: Social Investment and Poverty Reduction: A Comparative Analysis across Fifteen European Countries. Journal of Social Policy 44(3), 611–638 (2015) https://doi.org/10.1017/S0047279415000070

The Children's Worlds Study: New perspectives on children's deprivation research

Lo studio Children's Worlds: nuove prospettive per la ricerca sulla deprivazione dei bambini

Caterina Giusti and Antoanneta Potsi

Abstract This contribution investigates European children's deprivation from a child's perspective in four diverse and interwoven dimensions: deprivation in the school environment (1), in the familial and social relationships (2), in material goods (3) and in basic needs (4). The data analysed are from the Children's Worlds Study dataset 2016-19. To preserve the richness of the data available from this study, a Fuzzy-set methodology originally proposed for the study of multidimensional poverty is used.

Abstract *Questo contributo propone lo studio della deprivazione dei bambini europei dalla prospettiva dei bambini, utilizzando quattro dimensioni diverse ed interconnesse: deprivazione nell'ambiente scolastico (1), nelle relazioni familiari e sociali (2), nei beni materiali (3) e nei bisogni primari (4). I dati analizzati provengono dal database del Children's Worlds Study 2016-19. Per preservare la ricchezza dei dati disponibili, in questo paper si utilizza la metodologia fuzzy-set originariamente proposta per lo studio della povertà multidimensionale.*

Key words: children, deprivation, fuzzy-set methodology.

[1] Caterina Giusti, University of Pisa, caterina.giusti@unipi.it
Antoanneta Potsi, Bielefeld University, anneta.potsi@uni-bielefeld.de

1 Introduction

The United Nations Convention on the Rights of the Children (1989) shifted the policy interest to childhood and raised concerns about children's quality of life, the spectrum of their well-being and their behaviour and attitudes within their homes, schools and communities. Moreover, social and cultural perspectives on childhood have liberated research and policy from over reliance on normative developmental accounts (Woodhead, 2006). As a result, the international concern for the welfare of young children is increasing. International charities such as UNICEF are investing in welfare programmes. Children's welfare is now a universal and multidisciplinary concern, for which educators (as front service providers for children) are expected to take responsibility.

Children's Worlds (www.isciweb.org) is the first global study of childhood from a child's perspective. The study aims to collect solid and representative data on children's lives and daily activities, their time use and, in particular, their own perceptions and evaluations of their well-being. The data are used to improve children's well-being by creating awareness among children, parents and communities to the everyday lives of children, their environment, their relationships with others, their beliefs and satisfaction. By studying children's worlds in as many countries as possible, it is aimed to influence opinions of leaders, decision makers, professionals and the general public. The study gains insights into children's living conditions and deprivation and generates knowledge relevant for creating and enabling the conditions in which children can live flourishing lives. The purpose of this paper is to broaden the discussion on the crucial topic of children's deprivation, adding some interesting points to the current literature.

From a methodological point of view, we present an approach based on the fuzzy methodology introduced by Cheli and Lemmi (1995) and then updated by Betti et al. (2006). This methodology, developed for the study of poverty on a multidimensional perspective, is able to preserve the richness of the data available from the Children's Worlds study.

Differently from previous studies based on the same methodology and similar scope (Potsi et al., 2016; D'Agostino et al., 2020), the present work is the first based on data having a child's perspective. In this sense, the study presents new insides on the study of children deprivation.

2 Deprivation in childhood

Although childhood is characterized by a wide diversity across cultural frames, space and time (Facer et al. 2012, p. 172), is an important life stage with a value in itself. Qvortrup (1994) stressed the tendency to regard children as "human becomings'' rather than ''human beings'' where the ultimate goal and end-point of individual development is adulthood.

This paper focuses on children as active social actors and as subjects with capabilities that have crucial role in society (Comim et al. 2011).
The United Nations Convention on the Rights of the Child (UNCRC) has advanced the debate on childhood and altered the view on children from being merely recipients of freedoms and services or beneficiaries of protective measures, to being subjects with rights and participants in the actions impacting on them. The fundamental difference between present discussions about children's rights and those of previous years lies partly in a different picture of the child as deserving personal rights rather than simply protectionist rights (Sünker and Swiderek 2007). Empirical work in the field of early childhood (Danby and Baker 1998) has shown that children are competent social agents and have an active social world that is located beyond the audible and visual scrutiny. However, children are not seen naively as actors without any limits to their agency, but as actors with limited and unequal access to action (Bühler-Niederberger and König, 2011).
Deprivation can be described as the lack or denial of something(s) considered to be a necessity within the variety of social determinants such as income, gender, caste, class, ethnicity and race, as well as broader issues related to the political and economic governance and demographic realities of a region (Marmot, 2004)
Assuming that deprivation is multidimensional, and that multiple deprivation can be conceptualised as the combination of individual dimensions or domains of deprivation. Specifically, we consider the importance of constructing a child-focused deprivation measurement.

3 Methodology

The data of this study are derived from the Children's Worlds Study. Based on the nature of the available data, it was decided to conduct a quantitative analysis of the living conditions of children, using adequate methodological tools, able to preserve the richness of the data and to improve new measures of deprivation. The methodological approach used in this paper (henceforth Integrated Fuzzy and Relative—IFR) was born on the assumption that poverty is a multidimensional phenomenon and a vague predicate that manifests itself in different shades and degrees (fuzzy concept) rather than an attribute that is simply present or absent for individuals in the population, as the traditional poverty approach assumes.
The fuzzy set vision of poverty is particularly adequate for studying children's living conditions and deprivation mainly for two reasons. Firstly, it includes a non fixed value of poverty risk and deprivation, through the introduction of a membership functions (m.f.), i.e. a quantitative specification of degrees of poverty or deprivation depending on the other individuals or households included in the analysis. A membership function's value of 0 is always associated with the lowest risk of poverty or deprivation, whereas a value of 1 is associated with the highest risk.

Secondly, the multidimensional framework of the IFR approach updated by Lemmi et al. (2010) works up on several non-monetary indicators, assumed to be themanifest representation of a restricted number of underlying domains of deprivation, besides a monetary indicator based on the equivalent disposable income. The multidimensional analysis of poverty seems to be one reasonably grounded way to combine the CA and secondary quantitative data, because it includes monetary and non-monetary dimensions going beyond the traditional approach based only on the economic or financial situation.

4 Findings

In the first step of the analysis, missing data were imputed using a multivariate approach by chained equations (van Buuren and Groothuis-Oudshoorn, 2011). The percentage of missing values was very low, below the 10%, for most of the variables. Using Explorative Factor Analysis (EFA), four dimensions of children deprivation were identified. Then Confirmatory Factor Analysis was used to confirm whether EFA results provide a good fit to the data in the 17 countries. Table 1 reports the indicators identifying each of the four dimensions.

Table 1: Indicators affecting children deprivation dimensions.

Indicators	***Dimension***
My teachers care about me	SCHOOL ENVIRONMENT
If I have a problem at school my teachers will help me	
If I have a problem at school other children will help me	
My teachers listen to me and take what I say into account	
There are people in my family who care about me	SOCIAL RELATIONSHIPS
If I have a problem, people in my family will help me	
We have a good time together in my family	
I feel safe at home	
My parents/carers listen to me and take what I say into account	
My parents and I make decisions about my life together	
I have enough friends	
If I have a problem, I have a friend who will support me	
How many bathrooms are in your home?	MATERIAL GOODS
Does your family own a car, van or truck?	
How many computers do your family own?	
Whether has: Access to the Internet	
Whether has: Equipment/things for sports and hobbies	
Do you have enough food to eat each day?	BAISC NEEDS
Whether has: Clothes in good condition to go to school	
Whether has: Enough money for school trips and activities	
Whether has: Two pairs of shoes in good condition	
Whether has: Equipment/things you need for school	

Table 2 reports the Fuzzy measures by Country and dimension. As we can see, the Fuzzy measures of the fourth dimension, Basic Needs, are all rather small: this general result highlight that European children are not deprived in essential needs such as having enough food, good clothes, shoes, and enough money for school activities and equipment. For the third dimension, Material Goods, the results are overall higher, although still below 0.1 for many of the Countries.

Table 2: Fuzzy results by Country and dimension

Country	*SCHOOL ENVIRONMENT*	*SOCIAL RELATIONSHIPS*	*MATERIAL GOODS*	*BASIC NEEDS*
Albania	0.178	0.186	0.254	0.031
Belgium	0.363	0.293	0.076	0.075
Croatia	0.334	0.207	0.103	0.020
England	0.316	0.253	0.083	0.033
Finland	0.372	0.258	0.092	0.038
France	0.353	0.274	0.083	0.021
Germany	0.442	0.314	0.091	0.027
Greece	0.344	0.215	0.130	0.019
Hungary	0.374	0.191	0.097	0.012
Italy	0.382	0.305	0.150	0.013
Malta	0.258	0.261	0.115	0.047
Norway	0.290	0.219	0.046	0.025
Poland	0.314	0.185	0.106	0.019
Russia	0.442	0.369	0.181	0.057
Spain	0.279	0.206	0.098	0.027
Switzerland	0.279	0.223	0.082	0.017
Wales	0.326	0.268	0.097	0.031

Children are more affected by deprivation in material goods such as family car, computer, number of bathrooms, internet access and equipment for hobbies in Albania, Russia, Italy, Greece and Malta.

The fuzzy results highlight that, although overall not deprived in material goods or basic needs, European children are instead deprived to a greater extent in other dimensions related to the school environment and to the social and familial relationships.

Specifically, from the first two columns of Table 2 we can see that the Fuzzy measures obtained for the first and the second dimensions – School Environment and Social Relationship – are overall higher with respect to those of the third and fourth dimension. The first dimension, School Environment, represents the deprivation that children may experience at school in terms of lack of support from the teachers or other children.

References

1. Betti, G., Cheli, B., Lemmi, A., Verma, V.: Multidimensional and longitudinal poverty: an integrated fuzzy approach. In: Lemmi, A., Betti, G. (eds.) Fuzzy Set Approach to Multidimensional Poverty Measurement, pp. 111–137. Springer, New York (2006)
2. Bühler-Niederberger, D., König, A.: Childhood as a resource and laboratory for the self-project. Childhood 18(2), 180-195 (2011)
3. Cheli, B., Lemmi, A.: A totally fuzzy and relative approach to the multidimensional analysis of poverty. Econ. Notes 24(1), 115–134 (1995)
4. Chiappero-Martinetti, E.: Capability approach and fuzzy set theory: description, aggregation and inference. In: Lemmi, A., Betti, G. (eds.) Fuzzy Set Approach to Multidimensional Poverty Measurement. Springer, New York (2006)
5. Comim, F., Ballet, J., Biggeri, M., Iervese, V.: Introduction-theoretical foundations and the book's roadmap. In: Biggeri, M., B, J., Comim, Flavio (eds.) Children and the Capability Approach Studies in Childhood and Youth. Palgrave Macmillan, Basingstoke (2011)
6. D'Agostino A., Gagliardi F., Giusti C., Potsi A.: Investigating the impact of the economic crisis on children's wellbeing in four European countries. Social Science Research, 84, 102322 (2019)
7. Danby, S., Baker, C.: What is the problem? Restoring social order in the preschool classroom. In: Hutchby, I., Moran-Ellis, J. (eds.) Children and Social Competence: Arenas of Action, pp. 157–186. Falmer Press, London (1998)
8. Facer, K., Holmes, R., Lee, N.: Editorial. Glob. Stud. Child. 2(3), 170–175 (2012)
9. Lemmi, A., Verma, V., Betti, G., Neri, L., Gagliardi, F., Tarditi, G., Ferretti, C., Kordos, J., Panek, T., Szukiełojć-Bienkunska, A., Szulc, A., Zieba, A.: Multidimensional and fuzzy indicators developments. Project Small Area Methods for Poverty and Living Conditions Estimates. No. EU-FP7-SSH- 2007–2011 (2010)
10. Marmot, M.: Social causes of inequalities in health. In: Anand, S., Peter, F., Sen A. (eds.) Public Health, Ethics and Justice, 1st ed. Oxford University Press, Oxford, pp. 37-63.
11. Potsi, A., D'Agostino, A., Giusti, C., Porciani, L. (2016). Childhood and capability deprivation in Italy: a multidimensional and fuzzy set approach. Quality and Quantity, 50(6), 2571–2590.
12. Qvortrup, J.: Childhood matters: an introduction. In: Qvortrup, J., Bardy, M., Sgritta, G., Wintersberger, H. (eds.) Childhood Matters. Social Theory, Practice and Politics, pp. 1–24. Avebury, Aldershot (1994)
13. Sunker, H., Swiderek, T.: Politics of childhood, democracy and communal life: conditions of political socialization and education. Policy Futur. Educ. 5(3), 303–314 (2007)
14. van Buuren, S. Groothuis-Oudshoorn, K.: mice: Multivariate Imputations by Chained Equations in R. Journal of Statistical Software, 45(3) (2011)
15. Woodhead, M.: Changing perspectives on early childhood: theory, research and policy. Int. J. Equity Innov. Early Child. 4(2), 1–43 (2006)

The impact of different definition of "households with children" on deprivation measures: the case of Italy

L'impatto di differenti definizioni di "famiglie con figli": il caso dell'Italia

Laura Neri and Francesca Gagliardi

Abstract This paper analyses multidimensional fuzzy monetary and non-monetary deprivation in households with children, using two different definitions: households with children less than 14 and the EU definition of households with dependent children. Eight dimensions of non-monetary deprivation are found using 34 items from EU-SILC 2016 survey. A focus on Italy and Italian macro-region is presented.
Abstract *Questo studio analizza la deprivazione monetaria e non monetaria delle famiglie con figli secondo due diverse definizioni: famiglie con figli minori di 14 anni e famiglie con figli dipendenti, in linea con la recente definizione EU. A partire da 34 item rilevati con l'indagine EU-SILC 2016, sono state definite otto dimensioni. Un particolare focus è stato dedicato all'Italia.*

Key words: household with children, fuzzy sets, non-monetary poverty.

1 Introduction

Children are especially vulnerable to poverty and deprivation. One issue of particular concern, as poverty experienced by children can compromise their outcomes in future adult life (Del Boca, 2010).

Laura Neri, Dipartimento di Economia Politica e Statistica, Università di Siena; laura.neri@unisi.it

Francesca Gagliardi, Dipartimento di Economia Politica e Statistica, Università di Siena; gagliardi10@unisi.it

In 2018, one out of four children (aged 0-18) were at risk of poverty or social exclusion in the EU. However, as reported in Eurostat (2020), child poverty levels vary significantly between Member States. In RO, BG, EL, and IT, one out of three children were at risk of poverty or social exclusion, in 2018; whilst in DK, NL, CZ and SI it was of one out of six children. Most EU Member States reported that the at-risk-of-poverty rate was highest for single persons with dependent children. As regard to IT there are peculiarity to notice: IT (with ES and EL) reported the highest at risk of poverty or social exclusion rate (nearly 20%) in EU(27) for households composed by two adults with one dependent, whilst nearly 40% of households composed by two adults with three or more dependent children are at risk of poverty (just RO and BG report higher figures). It seems that the burden of dependent children is heavier in Italy with respect to other EU countries. A consideration that could help in understanding this issue, refers to an Italian cultural model: the average age at which they leave home is much higher than in many other European countries, indeed, they depend on their parents for a long time. So, we want to analyse more in depth this point considering two different definition of households with children: households with at least one child in 0-14 years and households with at least one dependent child.
The rest of the paper is organized as follows. Section 2 presents the data used for the analysis and delineates the research methodology. Finally, Section 3 presents the study's findings and articulates the conclusions.

2 Data and methodology

This paper uses the 2016 wave of the EU Statistics on Income and Living Conditions (EU-SILC) survey. It provides multidimensional microdata data on income, poverty, social exclusion and living conditions in the EU. *Ad-hoc* modules are developed each year to complement the permanently collected variables with supplementary ones highlighting unexplored aspects of social inclusion. The 2016 *ad-hoc* module includes variables on "Access to Services" exploring measures in terms of access to childcare, homecare, training, education, and healthcare. Accesses to education and healthcare services are important and closely linked to living conditions for all the household members: education has an important impact on the income of individuals as well as their knowledge and culture, a better access to health care can improve life expectancy as well as well-being. Access to childcare too, has an important impact on the household income: the lack of access to childcare affects the work-family balance of both women and especially make lower active female participation in the labour market. Moreover, childcare service improves the life chances of all children, especially disadvantaged children. It stimulates children's learning and gives them the chance to mix with others from different backgrounds.
The target variables involved in the analysis relate to different types of units. Information on social exclusion, housing condition and material deprivation is

collected mainly at household level, while labour, education and health information are collected at individual level for everyone aged 16 and over. Data on single income components are collected mostly at individual level and then, aggregated at household level to construct the household income. The income variable considered in the current analysis are the equivalised household income (HX090), the total disposable household income (HY020) and the total disposable household income before social transfers other than old-age and survivor's benefits (HY022). The variable from the ad-hoc module chosen for the analysis are those related to the affordability of the service, specifically: affordability of formal education, affordability of healthcare services, affordability of childcare services.

Our analysis starts from the cross-section sample of households included in the 2016 wave of the EU-SILC and, specifically, we are interested in two sets of households: households with at least one child in 0-14 years and households with at least one dependent child. A dependent child is any person aged below 18 as well as aged 18 to 24 years and living with at least one parent and economically inactive. Using this criterion, the sets of households analysed consist, respectively of 42,817 and 52,871 households, which are about 23% and 28% respectively of whole sample.

The countries involved in the analysis are AT, BE, BG, CH, CY, CZ, DK, EE, EL, ES, FR, HU, IE, IS, IT, LU, LV, NO, PL, PT, RS, SE, SK[1].

The adopted methodology is based on the cross-sectional fuzzy multidimensional measures of monetary and non-monetary deprivation that treats poverty as a matter of degree (Zadeh, 1965). This definition has several advantages as underlined by Verma et al. (2017): first, non-monetary poverty is subject to forced non-access to various facilities or possessions that determine basic living conditions and an individual might have access only to some of them; second, but not less important, the fuzzy approach provides more robust indicators (Betti et al., 2018), so it is particularly indicated for studying subpopulations or small domains, as in our case for households with children. In treating monetary and non-monetary poverty with fuzzy approach, the fundamental point is the choice of the membership function that quantify the propensity of each unit to poverty. We chose the membership function defined by Betti and Verma (2008), through which two indicators are defined: Fuzzy Monetary (FM) for monetary poverty and Fuzzy Supplementary (FS) for non-monetary poverty. The step-by-step procedure for measuring the FS proposed in Betti et al. (2015) has been applied.

In this study, 34 items have been identified to investigate non-monetary deprivation within household with children less than 14 or with dependent children. After their transformation into the range [0,1], the exploratory factor analysis enabled us to identify eight hidden dimensions of multidimensional non-monetary poverty; the dimensions with their related items are:

1. Basic lifestyle: Meals with meat, fish or chicken; Household adequately warm; Holiday away from home; Ability to make ends meet.
2. Consumer durables: Car; PC; Telephone; Washing Machine; TV.

[1] Some countries were removed from the analysis due to high number of missing values in the considered variables or variables not collected at all, or because of small sample sizes in households with children.

3. Housing amenities: Bath or Shower; Indoor flushing toilet; Leaking roof and damp; Rooms too dark; Overcrowd house.
4. Financial situation: Inability to cope with unexpected expenses; Arrears on mortgage or rent payments; Arrears on utility bills; Arrears on hire purchase instalments; Financial burden of total housing cost.
5. Environment: Crime, violence, vandalism; Pollution; Noise.
6. Work and education: Early school leavers; Low education; Worklessness; Duration of unemployment.
7. Services' Affordability: Affordability of childcare services; Affordability of formal education; Affordability of health care services.
8. Health related: General health; Chronic illness; Mobility restriction; Unmet need for medical exam; Unmet need for dental exam.

Most of the 34 items have been already used in literature on multidimensional non-monetary poverty and their goodness have been assessed (see Betti et al., 2015). Anyway, we decided to add a new dimension on services' affordability, using two items from EU-SILC ad-hoc module 2016: one item (overcrowd house) is included in the dimension Housing amenities, the other (financial burden of total housing cost) is included in the dimension Financial situation. The eight dimensions have been tested by a confirmatory factor analysis that confirmed for both samples their goodness of fit. Main goodness of fit indices, very similar for both samples, highlight the goodness of the chosen items and dimensions. Then, weights, FS for each of the eight dimensions and overall are computed. Concerning the fuzzy monetary poverty, it has been implemented using three different incomes, namely, household equivalised income (HX090), household disposable income (HY020) and household disposable income before social transfers (HY022).

3 Results

Results of the described analysis are used to compare the deprivation status for households with children less than 14 and for households with at least one dependent child: the figures are reported as ratio (Table 1). Each ratio shows the relative magnitude of the non-monetary and monetary deprivation, so a ratio close to one means that the deprivation level is similar the two sets of households; a ratio greater than one states that the level of deprivation is higher in households with 0-14 children than in households with 0-24 dependent children and a ratio lower than one states that the level of deprivation is higher in households with 0-14 children than in households with 0-24 dependent children. From Table 1, it can be observed that ratios are generally greater than one, meaning that the level of deprivation is higher in households with 0-14 children than in households with 0-24 dependent children. Only three countries, namely EL, IE and IT present ratios generally lower than one. For all countries, the non-monetary dimension related to Consumer Durable goods is greater than one, meaning that the deprivation level is higher in households with 0-14 children than in households with dependent children.

Table 1: Fuzzy non-monetary and monetary deprivation: children less than 14 to households with dependent children ratios.

COUNTRY	*FS=FM=HCR*	*FS1*	*FS2*	*FS3*	*FS4*	*FS5*	*FS6*	*FS7*	*FS8*	*FM HX090*	*FM HY020*	*FM HY022*
AT	1.00	1.03	1.13	1.02	1.02	1.03	0.99	1.00	1.04	1.00	1.15	1.03
BE	1.14	1.13	1.33	1.14	1.13	1.12	1.10	1.14	1.11	1.14	1.19	1.13
BG	1.08	1.07	1.09	1.07	1.06	1.03	1.06	1.07	1.03	1.08	1.05	1.04
CH	1.14	1.15	1.19	1.13	1.11	1.12	1.12	1.14	1.12	1.14	1.08	1.09
CY	1.04	1.11	1.34	1.06	1.00	1.02	1.03	1.09	1.05	1.04	1.07	1.10
CZ	1.26	1.28	1.37	1.21	1.22	1.20	1.21	1.26	1.22	1.26	1.41	1.21
DK	1.17	1.20	1.49	1.13	1.20	1.17	1.17	1.17		1.17	1.66	1.35
EE	1.13	1.14	1.21	1.09	1.11	1.08	1.10	1.10	1.10	1.13	1.14	1.07
EL	0.97	0.99	1.03	1.01	1.00	0.96	0.96	1.02	0.97	0.97	0.96	1.01
ES	1.02	1.04	1.09	1.03	1.03	1.02	1.00	1.04	1.02	1.02	1.07	1.00
FR	1.09	1.08	1.23	1.09	1.09	1.07	1.06	1.08	1.08	1.09	1.28	1.16
HU	1.03	1.04	1.05	1.03	1.02	1.05	1.03	1.04	1.04	1.03	1.13	1.12
IE	0.93	0.93	1.25	0.99	0.99	1.02	0.93	0.95	0.95	0.93	1.16	1.17
IS	0.98	1.02	2.07	1.00	1.01	1.04	0.97	1.00		0.98	1.39	1.23
IT	0.95	0.99	1.15	0.97	0.95	0.96	0.92	1.00	0.94	0.95	0.97	0.95
LU	0.97	1.00	1.35	0.99	1.01	0.96	0.95	0.98	1.01	0.97	0.96	0.99
LV	1.11	1.10	1.09	1.06	1.10	1.10	1.09	1.08	1.07	1.11	1.09	1.13
NO	1.71	1.75	2.86	1.56	1.67	1.48	1.59	1.66		1.71	2.56	1.28
PL	1.00	1.02	1.05	1.00	1.03	1.03	1.02	1.00	1.01	1.00	1.02	1.04
PT	1.00	1.01	1.08	1.00	1.00	1.00	0.98	1.01	0.97	1.00	1.10	1.07
RO	1.07	1.09	1.09	1.09	1.03	1.07	1.04	1.06	1.05	1.07	1.06	1.07
RS	1.04	1.07	1.07	1.03	1.03	1.04	1.02	1.05	1.02	1.04	1.04	1.04
SE	1.06	1.08	1.19	1.06	1.09	1.12	1.03	1.06		1.06	1.25	1.05
SK	1.19	1.18	1.18	1.13	1.16	1.11	1.15	1.20	1.19	1.19	1.17	1.11

As regard to the monetary deprivation, the figures are similar to the non-monetary one: the level of deprivation is more severe in households with 0-14 children than in households with dependent children. Again, few countries, namely EL, IE, IS and LU show a ratio lower than one as regard to the measure of deprivation related to the household equivalised income (HX090), just Italy shows a ratio lower than one for all three monetary variables. According to our figures, among the countries considered, Italy is the only one presenting all the ratios (out of the one related to Consumer Durable dimension) lower the one, meaning that Italy is the only country presenting a deprivation level more severe in households with dependent children than in households with 0-14 children. It is true that in Italy young people are more likely to live at home with their parents accepting being without a job or being in education than in many other EU countries. Despite this, the issue brings together concerns and certainly deserves to be investigated. The analysis of the ratios has been disaggregated by geographical area. From Table 2, two peculiar patterns are evident: a) as regard to dimension on Work and Education, ratios are lower than one

for all the macro-regions; b) all the ratios are lower than one for all the dimensions (out of the dimension Consumer Durable) for the Centre and South of Italy.

A serious concern rises observing pattern (a): with respect to Work and Education deprivation, in Italy, regardless of the macro regions in which they live, households with dependent children are more deprived than households with 0-14 children. The higher level of deprivation of the households with dependent children may be induced by the so-called NEETs (Not Engaged in Education, Employment or Training), because Italy is the EU country with the highest percentage of NEETs either in 20-24 or in 20–34 year-olds (Rosina, 2020). As regard to pattern (b), the most evident peculiarity regards the South, for which, the added deprivation of the households with 0-24 dependent children is evident for Financial situation (FS4), Services' Affordability (FS7), Health related (FS8) that could be connected to the uneven development in the county and the consequent significant gap between the South and the other Italian macro regions. Results highlight a sort of cycle: high deprivation as regard to Services' Affordability and Health related, have an important negative impact on the household income and well-being. It becomes harder and harder to interrupt the intergenerational reproduction of poverty.

Table 2: Households with children less than 14 to households with dependent children non monetary deprivation ratios, by macro-region

	FS	*FS1*	*FS2*	*FS3*	*FS4*	*FS5*	*FS6*	*FS7*	*FS8*
North West	0.99	1.02	1.28	0.98	1.02	0.96	0.91	1.05	0.99
North Est	1.05	1.06	1.43	1.04	1.03	0.96	0.91	1.09	1.05
Centre	0.95	0.96	1.19	0.92	0.91	0.96	0.93	0.96	0.97
South	0.90	0.97	1.21	0.98	0.90	0.94	0.93	0.87	0.82
Islands	1.00	1.02	1.11	1.00	0.91	1.04	0.99	1.07	0.94

References

1. Betti G., Gagliardi F., Lemmi A., Verma V. : Comparative measures of multidimensional deprivation in the European Union, Empirical Economics, 49(3), pp. 1071–1100 (2015)
2. Betti G., Gagliardi F., Verma V.: Simplified Jackknife Variance Estimates for Fuzzy Measures of Multidimensional Poverty, International Statistical Review, 86(1), pp. 68-86 (2018)
3. Betti G., Verma V.: Fuzzy measures of the incidence of relative poverty and deprivation: A multi-dimensional perspective, Statistical Methods and Applications, 17, pp. 225–250 (2008)
4. Del Boca, D.: Child Poverty and Child Well-Being in the European Union: Policy Overview and Policy Impact Analysis - A Case Study: Italy. In: Child poverty and child well-being in the European Union, Vol IV, TARKI-Applica, Budapest/Brussels (2010)
5. Eurostat: Statistics Explained https://ec.europa.eu/eurostat/statisticsexplained (2020)
6. Rosina A., I NEET in Italia Dati, esperienze, indicazioni per efficaci politiche di attivazione, StartNet - Network transizione scuola-lavoro. ISBN 978-88-945226-0-0 (2020)
7. Zadeh L.A.: Fuzzy sets, Information and Control, 8(3), 338–353 (1965)
8. Verma V., Lemmi A., Betti G., Gagliardi F., Piacentini M.: How precise are poverty measures estimated at the regional level?, Regional Science and Urban Economics, 66, pp. 175–184 (2017)

3.15 Perspectives in social network analysis applications

A comparison of student mobility flows in Eramus and Erasmus+ among countries

Un'analisi comparativa dei flussi di mobilità studentesca tra nazioni nei programmi Erasmus

Kristijan Breznik, Giancarlo Ragozini and Marialuisa Restaino

Abstract The internationalization of higher education has become a priority for the university system. It is crucial for institutions to analyse student mobility flows across countries to identify factors pulling and pushing students in a foreign country. In line with related works, the present contribution aims at comparing the characteristics of the student mobility trajectories involved in both Erasmus and Erasmus+ programmes by means of a network analysis approach. Starting from the European Union Open Data Portal, information are extracted and used to define network data structures for the two periods. The role and position covered by each country over time is explored by defining centrality scores and by assessing the sensibility of results in presence of data normalization procedures.

Abstract *L'internazionalizzazione è diventata un obiettivo strategico da perseguire per il sistema universitario. È quindi fondamentale analizzare i flussi di mobilità studentesca tra i paesi, identificando i fattori che spingono gli studenti ad andare in un'università straniera per completare il proprio percorso di studi. Il presente contributo ha come obiettivo lo studio delle caratteristiche delle traiettorie di mobilità studentesca nell'ambito dei programmi Erasmus e Erasmus+ a partire da un approccio di analisi di rete. Le informazioni estratte dal Portale Open Data dell'Unione Europea hanno consentito di definire delle strutture di dati di rete per entrambi i periodi. Il ruolo svolto da ciascun paese nel tempo nella rete è analizzato attraverso misure di centralità e valutando la sensibilità dei risultati rispetto a procedure di normalizzazione dei dati di rete.*

Key words: student mobility, weighted network, normalization, hubs and authorities

---

Kristijan Breznik
International School for Social and Business Studies, Celje, Slovenia e-mail: kristijan.breznik@mfdps.si

Giancarlo Ragozini
Department of Political Science, University of Naples Federico II, Napoli, Italy e-mail: giragoz@unina.it

Marialuisa Restaino
Department of Economics and Statistics, University of Salerno, Fisciano (SA), Italy e-mail: mlrestaino@unisa.it

1 Introduction

Higher education institutions are progressively drawing attention to the internationalization during the last decades, in order to increase international collaborations and cooperation, and to enhance the quality of their research and teaching activities. In particular, they are encouraging students and/or academic staff to participate in international mobility exchanges. Thus, it becomes essential to study mobility flows across different European and non-European countries, and identify the factors pulling and pushing students and academic staff in a foreign country thanks to which it will be possible to implement policies devoted to increasing of the degree of internalization.

In line with related papers [1, 2, 6, 7], the present contribution focuses on the student mobility for studies and aims at analyzing the characteristics of the student mobility trajectories involved in both Erasmus and Erasmus+ programmes by a network analysis approach. Starting from this theoretical and analytical perspective, the main purpose is to discover the role played by each country by computing centrality measures [8], revealing the presence of hubs (i.e. good exporting countries) and authorities (i.e. good importing countries) [4, 9, 10] during the two programmes.

Thanks to the European Union Open Data Portal (EU ODP), a statistical overview of Erasmus and Erasmus+ student mobility for studies from 2007–08 to 2013–14 and from 2014–15 to 2018–19 is obtained.[1] Temporal network data structures, i.e. weighted and directed one-mode network data structures, are defined. In these networks, actors are countries (*vertices*) and relation is defined as student mobility exchange between them (represented by *links/arcs*) with weights proportional to the number of students involved in the exchange's flow between countries. In addition, as the countries are very different in size (and then, in *in-degree* and *out-degree* scores), the arcs' weights are normalized by considering the procedure described in [11]. The main descriptive findings show the presence of some relevant changes between the two periods, exploiting the role played by each country.

The remaining of the contribution is organized as follows. Section 2 reports the network data definition and the methodological approaches for exploring student mobility data in the two Erasmus programmes. In Section 3, the main descriptive results are shown along with suggestions for future lines of research.

2 Erasmus student network data definition

The data used in this study are gathered from the European Union Open Data Portal. Network data structures are applied in order to describe changes in the international student flows among countries in Erasmus and Erasmus+ programmes. For Erasmus+ only Key Action 1 (KA1), and in particular the line KA103, is considered. Then, although mobilities in Erasmus+ programme were under KA1, the idea of student mobility and positive experiences is transferred from the Erasmus programme. Table 1 shows the trend of student mobility for studies (SMS) and for placement (SMP) in both programmes. It is particularly evident the increasing trend

[1] For details see https://data.europa.eu/euodp/en/data/publisher/eac.

for both SMS and SMP. In 2018-2019 the exchange flow is related to the period October – December 2018.

Table 1 Distribution of students mobility in Erasmus and ErasmusPlus programmes.

Year	Total number of exchanges	# of exchanges SMS	SMP	Year	Total number of exchanges	# of exchanges SMS	SMP
Erasmus Programme				Erasmus+ Programme			
2007–2008	182,697	162,694	20,003	2014–2015	269,025	199,551	69,474
2008–2009	198,523	168,193	30,330	2015–2016	287,902	210,021	77,881
2009–2010	213,266	177,705	35,561	2016–2017	297,239	212,837	84,402
2010–2011	231,408	190,495	40,913	2017–2018	310,269	219,295	90,974
2011–2012	252,827	204,744	48,083	2018–2019[1]	186,219	155,937	30,282
2012–2013	268,143	212,522	55,621				
2013–2014	272,497	212,208	60,289				

[1] Only first 3 months in 2018 – 2019.

2.1 Network definition and data treatment

In order to analyze and compare the global structure of the international relationships established through the student mobility flow over the two periods, we model such a mobility through a proper network structure, and we use the network methodological perspective as instrument for capturing the patterns of students' mobility.

The network established by the Erasmus programme at a time t is a directed weighted one-mode network that can be described as a graph $\mathscr{G}_t$, where countries are vertices ($\mathscr{V}_t$), and (directed) edges ($\mathscr{E}_t$) are given by the presence of students moving by one country towards another in year t. The number of students involved in this exchange represents the weight ($\mathscr{W}_t$) of each edge. For the sake of simplicity, we assume that the set of countries is constant along the considered period, i.e., $\mathscr{V}_t \equiv \mathscr{V}_{t'} \equiv \mathscr{V} \quad \forall t \neq t'$. There are 31 countries that are part of this study. We removed countries that were not participating the Erasmus programme in at least one of both periods (Switzerland has been suspended as a participant in the Erasmus programme and North Macedonia joined later; therefore it was impossible to compare the participation of any of them in both periods). Thus, the weighted directed graphs are $\{\mathscr{G}_t(\mathscr{V},\mathscr{E}_t,\mathscr{W}_t)\}_{(t=1,\ldots,T)}$, where $\mathscr{V} = (v_1, v_2, \ldots, v_g)$ is the set of g countries, $\mathscr{E}_t \subseteq \mathscr{V} \times \mathscr{V}$ is the set of edges at time t, $\mathscr{W}_t$ is the set of weights at time t, $w : \mathscr{E}_t \to \Re$, and $w[(v_i, v_j)_t] = w_{ijt}$ is the number of students moving from a country v_i towards another country v_j (with $i \neq j$) at time t. It is possible to consider the corresponding adjacency matrices $\mathbf{A}_t$ with elements $a_{ijt} = 0$ if $(v_i, v_j)_t \notin \mathscr{E}_t$, and $a_{ijt} = w_{ijt}$ otherwise.

The derived network structures are quite dense, well linked and unbalanced as the weights are strongly affected by the different country population (and consequently student population) sizes (ranging from 38,700 inhabitants of Liechtenstein to 83,166,700 of Germany). These features make the analysis of such networks not easy to interpret and call for specific normalization procedures to disentangle such complex structure. Among the different proposals available in literature, we opt for that proposed in [11] and already used by [4, 5] to analyze the Erasmus networks. All elements in adjacency matrix are normalized by dividing them by the root square of the product of the row and column total marginals, i.e., given $a_{ijt} \neq 0$, the nor-

malized version $\tilde{a}_{ijt} = \frac{a_{ijt}}{\sqrt{a_{i\cdot t} a_{\cdot jt}}}$, where $a_{i\cdot t}$ and $a_{\cdot jt}$ are the number of outgoing students from the i-th country and the number of incoming students in the j-th country, respectively. Note that this kind of normalization resembles the independence hypothesis of the χ^2 test or the system of weights given by a random graph.

In case of weighted directed network, the hub and authority scores [8] are centrality measures able to identify the most attractive and active countries in the network. By definition, a good authority is a country that is pointed to by many good countries, that is *good importer*, whereas a good hub is one that points to many good authorities, that is *good exporter*.

The Kleinberg's algorithm solution converges to the dominant eigenvectors of the cross-product of the adjacency matrix and its transposed one. In particular, the authority scores $auth_{it}$ are determined by the values of the dominant eigenvector of the authority matrix $\mathbf{A}_t'\mathbf{A}_t$, and the hub scores hub_{it} are given by the entries of the dominant eigenvector the hub matrix $\mathbf{A}_t\mathbf{A}_t'$. Finally, by considering the normalized adjacency matrices $\tilde{\mathbf{A}}_\mathbf{t}$, $\tilde{auth}_{it}$ and $\tilde{hub}_{it}$ are the corresponding authority and hub scores.

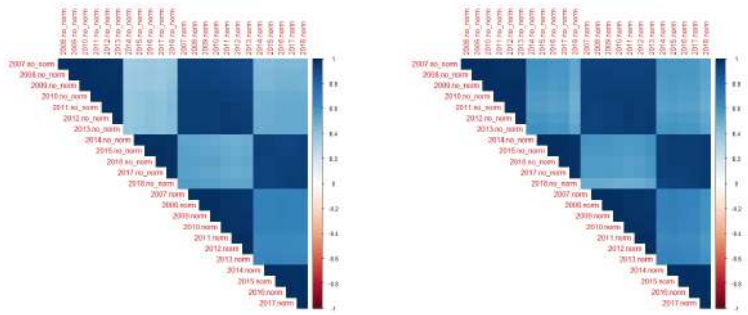
3 First findings and concluding remarks

In a first step, we evaluate the correlations among the authority (and hub) scores along the time, i.e. $corr(\underline{auth}_t, \underline{auth}_{t'})$, for both original adjacency matrices and the normalized adjacency matrices. Looking at the heatmaps in Figure 1, it appears that in both cases the higher correlations are inside both periods, i.e. Erasmus and Erasmus+ programmes. It is worth to note that also the correlations among the normalized and not normalized values are very high, denoting a substantial stability of the hub and authority measures with respect to the this kind of normalization. On the contrary, the inter-programme correlations are generally less than 0.5. These first results denote that there was a deep change in the network structure passing to the Erasmus+ programme. This means that some countries that in the first edition of the Erasmus programme were not central, have gained centrality in the mobility network, and vice versa. These changes could derive by some socio-political events that shaped differently the country reputation, such as the terroristic attack, the economic and/or political crisis, etc..

In order to gain insights in such a change, we report the ranking of the authority and hub scores referred to 2010 and to 2016. In Figure 2, we notice that the first five authorities in 2010 were the five largest European countries (Spain, France, Germany, United Kingdom and Italy). In particular, Spain, France, Germany and Italy were also the first four hubs which reveals their double central role in the network. With small changes these four countries are central along all the years of the first Erasmus programme. After six years, in the Erasmus+ programme, Finland and Greece gained centrality as both authorities and hubs, while France and Spain lost their leading roles. Italy and Germany, but also United Kingdom, Poland, Portugal and The Netherlands, remain almost stable as authorities of the network. These changes are confirmed on Dumbell Charts in Figure 3 that show, for each country, the variations between the authority (hub) scores averaged over the first Eras-

mus programme (red dots) and the Erasmus+ programme (blue dots). Spain and France had the largest negative variations, while Finland and Greece experimented the largest positive variations. Also Poland, Czech Republic and Croatia, with different positions in the networks, had large positive variations, gaining centrality in the network.

All these results confirm that the new rules of the Erasmus+ programme and some socio-political changes over the last ten years have changed the macro mobility patterns of university students. Further analysis should be performed taken into account shifts that happened in the countries analyzed. In addition, research at university level by considering other types of mobilities allowed by the Erasmus programme can explain these changes.



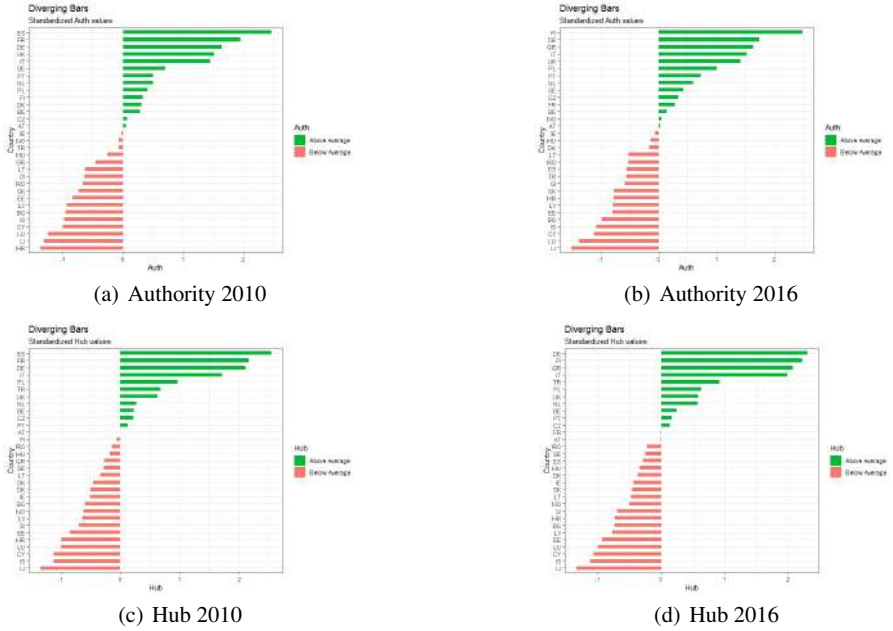
(a) (b)

Fig. 1 Correlation of Authority and Hub scores measured at country level (normalized and not normalized) between the two Erasmus programmes.

Acknowledgements This work has been partially funded by the Slovenian Research Agency within the Programme P5-0049 and by the research grant "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide" MIUR-PRIN 2017HBTK5P-CUP B78D19000180001.

References

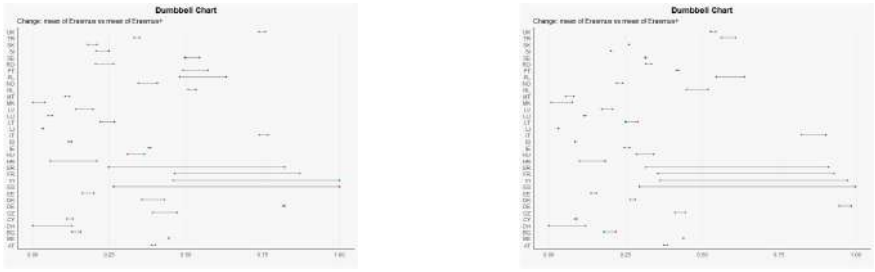
1. Amendola, A., Restaino M.: An evaluation study on students international mobility experience. Qual. and Quant., **51**(2), 525–544 (2017)
2. Barnett, G.A., Ke Jiang, M.L., Park, H.W.: The flow of international students from a macro perspective: a network analysis. Compare: J. Comp. Int. Edu., **46**(4), 533–559 (2016)
3. Benzi, M., Estrada, E., Klymko, C.: Ranking hubs and authorities using matrix functions. Linear Algebra Appl, **438**(5), 2447–2474 (2013)
4. Breznik, K., Ragozini, G.: Exploring the italian erasmus agreements by a network analysis perspective. 2015 IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min. (ASONAM), 837–838 (2015)
5. Breznik, K., Skrbinjek, V.: Erasmus student mobility flows. Eur. J. of Edu., **55**(1), 105–117 (2020)
6. Derszi, A., Derszy, N., Kaptalan, E., Neda, Z.: Topology of the Erasmus student mobility network. Phys. A: Stat. Mech. and its Applic., **390**(13), 2601–2610 (2011)
7. De Benedictis, L., Leoni, S.: Gender bias in the Erasmus network of universities. App. Netw. Scie., **5**(1), 1–25 (2020)



(a) Authority 2010 (b) Authority 2016

(c) Hub 2010 (d) Hub 2016

Fig. 2 Ranking of countries according to Hub and Authority standardized scores. Bar color: green above average; pink below average.



(a) (b)

Fig. 3 Ranking of countries according to the average student between the two Erasmus programmes.

8. Kleinberg, J. M.: Authoritative Sources in a Hyperlinked Environment. J. of the ACM, **46**(5), 604–632 (1999)
9. Kondakci, Y., Bedenlier, S., Zawacki-Richter, O.: Social network analysis of international student mobility: uncovering the rise of regional hubs. High. Educ., **75**(3), 517–535 (2018)
10. Restaino, M., Vitale, M.P., Primerano, I.: Analysing international student mobility flows in higher education: A comparative study on European Countries. Soc. Indic. Res., **149**(3), 947–965 (2020)
11. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, (1983)

Network-based approach for the analysis of LexisNexis news database

Analisi del database LexisNexis mediante un approccio basato sulle reti

Carla Galluccio and Alessandra Petrucci

Abstract The recent COVID-19 sanitary emergency has contributed to the increase of public attention towards several issues of social interest, such as educational inequality, primarily due to the introduction of the government's protective measures. In this vein, we aimed to study the journal narratives on educational inequality in Italy from September 2019 to May 2020, namely before and during the first wave of the COVID-19 pandemic (started in Italy at the end of February 2020). To do this, we carried out a study on four of the most important Italian journals included in the LexisNexis news database. In particular, we exploited network analysis and text mining methods to extract information from this unstructured textual data so as to examine and infer the structure of the semantic relationships regarding this topic.
Abstract *La pandemia da Covid-19 ha comportato l'incremento dell'attenzione da parte dell'opinione pubblica verso numerose tematiche di interesse sociale, come il tema della disuguaglianza nell'accesso all'istruzione, determinato principalmente dall'introduzione delle misure protettive da parte dal governo. A tal proposito, questo lavoro si propone di studiare la narrazione della disuguaglianza educativa trasmessa attraverso i giornali in Italia da settembre 2019 a maggio 2020, ovvero nel periodo precedente e durante la prima ondata della pandemia da COVID-19 (iniziata in Italia alla fine di febbrario 2020). Per fare ciò, abbiamo condotto uno studio su quattro dei principali giornali italiani contenuti nel database LexisNexis, utilizzando metodi di text mining e network analysis per estrarre informazioni ed esaminare la struttura delle relazioni semantiche presenti nelle notizie pubblicate su questo argomento.*

Key words: multilayer network analysis, text mining, educational inequality, LexisNexis news database, COVID-19

Carla Galluccio
University of Florence, Viale Morgagni 59 Firenze (FI), e-mail: carla.galluccio@unifi.it

Alessandra Petrucci
University of Florence, Viale Morgagni 59 Firenze (FI), e-mail: alessandra.petrucci@unifi.it

1 Introduction

Educational inequality can be defined as objective and systematic intergroup disparities in terms of academic achievement opportunities. These disparities may relate to various educational and social factors, such as resources, treatment, access, and/or results [3]. During the last decades, the literature on educational inequality in Italy has been focused mainly on the socioeconomic determinants of inequality, gender, ethnic minorities, and geographical differences (for instance, between Northern and Southern Italy or between suburbs and city centres [4]).

In this vein, the recent COVID-19 pandemic has represented a breaking point, bringing to light several issues of social interest, such as educational inequality. Indeed, school closures and distance learning have revealed significant disparities, especially for children of families with low socioeconomic status or sociocultural barriers [9]. In this regard, news media data can be employed to inquire into the transmission and perception of events with respect to which the attention and sensitivity of public opinion have increased due to the sanitary emergency.

In order to study the journal narratives on educational inequality in Italy over the last year, we carried out a study on four of the most important Italian journals, namely the Corriere della Sera, il Resto del Carlino, Il Giorno, and La Nazione, included in the LexisNexis news database. LexisNexis is an online platform that collects European and worldwide newspaper articles regarding different fields. More specifically, we focused on news about educational inequality published from September 2019 to May 2020, namely before and during the first wave of the COVID-19 pandemic (started in Italy at the end of February 2020). To this end, we exploited network analysis and text mining methods to extract information from this unstructured textual data so as to examine and infer the structure of semantic relationships regarding educational inequality. In particular, we extracted news about educational inequality from the Italian journals filtering news headlines and bodies through selected keywords; then, we pre-processed news body text by means of text mining methods. Afterwards, we obtained a multiplex network, a particular case of multilayer network where each layer describes a different type of links between the same set of nodes [7]. Herein, the multiplex data structure consists of different journals representing the layers, the nodes are the words, and the links are given by the semantic relationship between the words.

2 Statistical methods

Network analysis and text mining methods were carried out in order to extract information from Italian news included in the LexisNexis news database about educational inequality. In particular, the analysis is made up of two steps.

The first step of the analysis regarded the extraction of news of interest and the selection of words. To do this, we firstly filtered news by means of selected keywords on both news headlines and bodies. More specifically, we considered articles with

one or more keywords in the headline or articles that included at the same time two or more words of interest in their body.[1]

Afterwards, we pre-processed news body text by means of text mining methods [1]. Firstly, we removed numbers and non-alphanumeric characters; then, we removed proper nouns identified using the NER algorithm and conversed the text to the lower case. After removing the Italian "stop words", we divided the text into tokens and lemmatised it. Finally, the analysis was restricted to words common all over the journals, which were used to create co-occurrence matrices within a window of two concepts for statistical analysis [6].

Regarding the second step of our analysis, many complex systems can be represented as multilayer networks [8]. In this case, the word co-occurrence matrices obtained from different journals and transformed into networks can represent a complex system of multiple documents written about the same topic. So, in order to extract information about the topic of interest and similarities between documents, we exploited a multiplex network analysis in which each journal represents a layer, nodes are the words, and edges are given by the semantic relationship between the words [11]. So, let $\mathcal{M}$ a multiplex network consisting of a sequence of graphs $\{G_k\}_{k=1,\ldots,K} = \{(V,E_{kk})\}_{k=1,\ldots,K}$, with $V = (v_1,\ldots,v_n)$ the set of n nodes of each network, and $E_{kk} \subseteq V \times V$ the set of edges [5]. From each network G_k we can define the adjacency matrices of the K layers $\mathbf{A}_k = (a_{ijk})$ with $a_{ijk} = 1$ if $(v_i, v_j) \in E_{kk}$, and $a_{ijk} = 0$ otherwise, so that each layer is an unweighted network.

The degree of a node i on a layer k can be expressed as $d_{ik} = \sum_j a_{ijk}$, with $0 \leq d_{ik} \leq n-1 \; \forall i, \forall k$. Besides, in according to [2], we defined the aggregated topological adjacency matrix $\mathbf{A} = \{a_{ij}\}$, where $a_{ij} = 1$ if $\exists\, k : a_{ijk} = 1$, and 0 otherwise. The degree of node i on $\mathbf{A}$ can be computed as:

$$d_i = \sum_j a_{ij} \tag{1}$$

However, the aggregated topological adjacency matrix disregards the possible existence of multi-ties between nodes in different layers. In this case, it is possible to define the aggregated overlapping adjacency matrix $\mathbf{O} = \{o_{ij}\}$, where $o_{ij} = \sum_k a_{ijk}$ is the edge overlap of the edge between node i and node j. The total number of connections of node i can be defined as overlapping degree of node i:

$$o_i = \sum_j o_{ij} = \sum_k d_{ik} \tag{2}$$

with $o_i \geq d_i$.

When the connections among nodes are weighted, the multiplex network can be fully described by the vector of its weighted adjacency matrices, with the generic $\mathbf{W}_k = (w_{ijk})$. In analogy with the unweighted case, it is possible to define both the aggregated topological adjacency matrix $\mathbf{A}^w = \{a_{ij}\}$, where $a_{ij} = 1$ if $\exists\, k : w_{ijk} >$

[1] For example, we considered articles in which there were the "disuguaglianza educativa" or "disuguaglianza nell'accesso all'istruzione" keywords in the headline or articles in whose body there were words such as "disuguaglianza" and "educativa" at the same time.

0, and 0 otherwise, and the aggregated overlapping adjacency matrix $\mathbf{O}^w = \{o^w_{ij}\}$. In this case, the degree of node i on the aggregated topological network can be expressed as $s_i = \sum_j w_{ij}$, whereas the weighted overlapping degree of node i can be computed as $o^w_i = \sum_j o^w_{ij} = \sum_k s_{ik}$ [2].

Besides, we computed the Kendall rank correlation coefficient to quantify the correlation between the aggregated degree sequences and the degree of the nodes at each layer [2].

Finally, the participation of node i to the different layers can be computed as:

$$P_i = \frac{K}{1-K}\left[1 - \sum_{k=1}^{K}\left(\frac{d_{ik}}{o_i}\right)^2\right] \tag{3}$$

where P_i is the participation coefficient, which takes value in $[1,0]$. P_i measures if the links of node i are uniformly distributed among the K layers. In the case of node i presents the same number of links in all layers, P_i is equal to 1. Conversely $P_i = 0$ if all the links of node i are concentrated in one layer [2].

3 Results and discussion

The total number of articles about educational inequality published before and during the first wave of the COVID-19 pandemic is 34 in il Resto del Carlino (on a total of 244287 articles), 54 in the Corriere della Sera (on a total of 76044 articles), 15 in Il Giorno (on a total of 96050 articles), and 45 in La Nazione (on a total of 246023 articles). After the text pre-processed stage, il Resto del Carlino presented a total of 5850 words, the Corriere della Sera 21124, Il Giorno 2768, and La Nazione 8682. Among all the words, we found 839 common words used to create four co-occurrence matrices, one for each journal, within a window of two concepts.

These co-occurrence matrices were used to create four undirected weighted networks, from which a multiplex network was defined. Hence, the resulted multiplex network presented as many layers as journals analysed and consisted of a fixed set of nodes, namely the 839 common words. The edges represented the semantic relationship between words, weighted based on the frequency of their connection in a given layer.

Afterwards, we focused on basic node properties, in particular the degree of nodes at each layer and across layers. For this reason, we computed the aggregated topological degree s_i, the aggregated overlapping degree o^w_i, and the degree of the nodes in each layer (see Table 1). Looking at the degree of nodes in different layers and in the aggregated matrices, we noticed that the words with the highest degree were “disuguaglianza”, “scuola”, and “scolastico”, but also “sociale”, “potere”, and “politico”, namely words that outline the role of the government in addressing educational inequality. This finding is consistent with findings discussed in [10], where it is supposed that distance learning does not create disparities but turns the spotlight

on them. Hence, inequalities need to be addressed both in the case of face-to-face and distance learning equally.

Table 1 Top 10 words with highest degree at each layer (il Resto del Carlino s_{iCa}, the Corriere della Sera s_{iCo}, Il Giorno s_{iG}, and La Nazione s_{iN}) and across layers (s_i and o_i^w).

s_{iCa}		s_{iCo}		s_{iG}		s_{iN}		s_i		o_i^w	
lavorare	178	politico	502	scuola	104	lavorare	410	politico	380	lavorare	958
sociale	176	potere	502	economico	84	scuola	354	scuola	378	scuola	946
disuguaglianza	172	sociale	386	lavorare	84	disuguaglianza	212	potere	377	sociale	830
scuola	170	disuguaglianza	340	disuguaglianza	79	sociale	212	lavorare	376	disuguaglianza	803
economico	124	volere	322	famiglia	74	economico	188	sociale	338	potere	782
famiglia	96	scuola	318	sociale	56	potere	170	disuguaglianza	330	politico	770
potere	90	dovere	304	aumentare	44	educativo	164	dovere	283	economico	624
parlare	82	lavorare	286	politico	40	politico	162	economico	275	dovere	536
bambino	80	venire	246	studente	40	famiglia	150	famiglia	264	volere	454
scolastico	78	pensare	231	dovere	36	dovere	144	venire	241	famiglia	446

Besides, in order to quantify the correlation between all the degree sequences, we computed the Kendall rank correlation coefficient. Results reported in Table 2 showed that the highest correlation is between s_i and o_i^w, whereas there is a moderate correlation between other degree sequences. However, due to the moderate correlation observed between degree sequences, we decided to compute the participation coefficient P_i, in order to quantify the richness of the connectivity patterns across layers. The 5 words with the highest P_i are: “adozione”, “afflitto”, “agricolo”, “altamente”, and “analogo”, with $P_i = 0.94$. Regarding this result, we speculate that the journal narratives might change in relation to their political orientation and audience. Indeed, we observed that the most significant differences were between Il Corriere della Sera, leftist and national journal, and the other three journals, rightist and local. In particular, the former is more focused on the political aspects of educational inequality, whereas the attention of the latter was on its social and regional aspects. The participation coefficient results could support this finding since the words with the highest degree and those with the highest P_i are not the same, showing that journals used the most significant words about educational inequality in a different way.

Table 2 Kendall rank correlation coefficient between all the degree sequences.

	s_i	o_i^w	s_{iCa}	s_{iCo}	s_{iG}	s_{iN}
s_i	1					
o_i^w	0.91	1				
s_{iCa}	0.64	0.68	1			
s_{iCo}	0.80	0.80	0.54	1		
s_{iG}	0.48	0.53	0.50	0.40	1	
s_{iN}	0.67	0.71	0.59	0.52	0.51	1

4 Conclusion

The present study aimed to investigate the Italian journal narratives about educational inequality before and during the first wave of the COVID-19 pandemic, exploiting multiplex network analysis and text mining methods.

Undoubtedly, the COVID-19 pandemic and the government's protective measures have increased the attention of public opinion towards educational inequality. Indeed, the primary question is if the introduction of distance learning has amplified the disparities, particularly for students of families with low socioeconomic status.

However, our results showed that there are not particularly differences in the journal narratives on educational inequality throughout the period considered, underlying the need to address educational inequality both in face-to-face and distance learning equally.

In conclusion, we believe that the approach we proposed could represent a useful tool to extract novel and non-trivial information about the journal narratives regarding topics of interest. Indeed, newspapers articles that deal with the same topic transformed into informative word-pairs networks can be seen as a multilayer network that allows us to compare the texts.

In the future, we plan to use factorial methods in order to analyse and visually explore the obtained multilayer networks.

References

1. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: A brief survey of text mining: Classification, clustering and extraction techniques. arXiv prepr. arXiv:1707.02919. (2017)
2. Battiston, F., Nicosia, V., Latora, V.: The new challenges of multiplex networks: Measures and models. Eur Phys J Spec Top. **226**, 401-416 (2017)
3. Crahay, M., Dutrévis, M.: Educational Inequality. In: Michalos, A.C. (ed.) Encyclopedia of Quality of Life and Well-Being Research, pp. 1830-1836. Springer Netherlands, Dordrecht (2014)
4. Gentili, A., Pignataro, G.: Disuguaglianze e istruzione in Italia. Dalla scuola primaria all'università. Carocci editore, Roma (2021)
5. Giordano, G., Ragozini, G., Vitale, M. P.: Analyzing multiplex networks using factorial methods. Soc. Netw. **59**, 154-170 (2019)
6. Jiang, K., Barnett, G.A., Taylor, L.D.: Dynamics of culture frames in international news coverage: A semantic network analysis. Int. J. Commun. **10**, 3710–3736 (2016)
7. Kanawati, R.: Multiplex Network Mining: A Brief Survey. IEEE Intell. Inform. Bull. **16**, 24-27 (2015)
8. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. J. Complex Netw. **2**, 203-271 (2014)
9. Poletti, M.:Hey teachers! Do not leave them kids alone! Envisioning schools during and after the coronavirus (COVID-19) pandemic. Trends Neurosci Educ. (2020)
10. Roncaglia, G.: L'età della frammentazione: cultura del libro e scuola digitale. Gius. Laterza & Figli Spa, Roma-Bari (2020)
11. Sebestyén, V., Domokos, E., Abonyi, J.: Multilayer network based comparative document analysis (MUNCoDA). Methodsx. **7**, (2020)

A multiplex network approach to study Italian Students' Mobility

Un approccio multiplex per lo studio delle reti di mobilità studentesca in Italia

Ilaria Primerano, Francesco Santelli and Cristian Usala

Abstract The aim of this contribution is to explore the main patterns of Italian students' mobility network in the transition from bachelor's to master's degree by relying upon administrative data regarding students' careers. To explore these flows, we define a multiplex network structure, where the disciplinary groups represent the layers, each university is a node, and the weighted and directed links are defined as the number of students moving between nodes. This empirical strategy allows us to highlight the presence of the universities that play a fundamental role within each field, and to verify if the same central structure is observed in different layers.

Abstract *L'obiettivo del presente contributo è quello di esplorare i flussi di mobilità degli studenti nel passaggio dalla laurea triennale alla magistrale considerando i dati amministrativi sulla carriera degli studenti. Sulla base dei flussi di mobilità osservati, le singole reti di mobilità studentesca sono state organizzate in una struttura di tipo multiplex, dove ogni layer consiste in uno specifico gruppo disciplinare, ogni università è un nodo, e i legami (direzionati e pesati) sono dati dal numero di studenti in mobilità tra i nodi. In questo modo è stato possibile evidenziare il ruolo delle università nei diversi gruppi di studio e confrontare i differenti gruppi disciplinari.*

Key words: University mobility, Multiplex networks, Migration flows

Ilaria Primerano
Department of Political and Social Studies, University of Salerno, Italy e-mail: iprimerano@unisa.it

Francesco Santelli
Department of Political Sciences, University of Naples Federico II, Italy e-mail: francesco.santelli@unina.it

Cristian Usala
Department of Political and Social Sciences, University of Cagliari, Italy e-mail: cristian.usala@unica.it

1 Introduction

Understanding the Italian internal migration patterns of students in the transition from bachelor's to master's degree programmes is a key factor in the analysis of students' mobility pathways. The Italian Student mobility has been studied by following different approaches of analysis, depending both on the the kind of data and the statistical methods used. In the last decade, in order to detect the determinants of student mobility, the literature has been mainly focused on first level mobility (i.e. from diploma to bachelor), whose main route has been traced from the Southern to Central and Northern regions of Italy [1] [9]. Indeed, only few recent contributions have dealt with second level mobility (i.e. from bachelor to master), by performing longitudinal analysis [11] and by focusing mainly on Southern Italian students' flows [2]. In this framework, other studies have used micro-data to explore the main patterns of the incoming and outgoing flows of Italian studies by means of Network Analysis, taking into account also some aggregate data referred to macro-areas, such as provinces, regions and marco-regions [7] [12], and the specific field of study chosen by students [6].

In this scenario, we addresses second level mobility by considering students with a bachelor degree enrolling in a master's programme, with an emphasis on the choice of the field of study. We focus on the flows of students who change university for the master's degree, regardless of the distance from their residence. We propose a network perspective that considers the university where students achieved the bachelor's degree as the origin and the university where students enrolled for the master's degree as the destination. Moreover, we also consider those students who choose to complete their university career enrolling at a telematic university since we do not account for the geographical aspect of mobility. Even if previous studies have already specified the network structure of the student mobility flows, to the best of our knowledge, no studies have been conducted to explore the second level mobility taking into account, at the same time, several aspects: including telematic universities, considering the field of study, and neglecting the geographical settings. In this framework, multiplex network approaches [14] represent a valid opportunity to consider the role played by the field of study in students' choice process in second level mobility. Specifically, each disciplinary field is a layer of the network, the Italian universities are the nodes, and the flows of students between them are the links.

The paper is structured as follows. The micro-data we use are described in section 2, the methodological approach is addressed in section 3, main results and concluding remarks are presented in section 4.

2 Data description

Our analysis relies upon the micro-data extracted from the database MOBYSU.IT [10] which includes information on students university careers provided by the Ital-

ian National Student Archive (NSA).[1] We consider all the Italian students who started their university career in a bachelor's programme between a.y. 2011-12 and a.y. 2015-16, and have enrolled in a master's degree between a.y. 2014-15 and a.y. 2018-19. More specifically, starting from the population of 1,171,006 Italian bachelor students, we keep the information regarding the 621,075 (53%) students that have graduated in the time frame considered. Secondly, we retain in our data only those students that have enrolled in a master's programme after graduation. Following this strategy, our data consists of 367,725 students belonging to 92 universities (of which 11 are exclusively on-line universities, usually called telematic). We classify the degree programmes provided by universities into 10 disciplinary groups according to the ISCED-F 2013 classification [17]. Table 1 presents, for each disciplinary group, the number of students observed in bachelor's degree along with information on students in mobility and their field choices. The two most chosen groups are 'Engineering' and 'Social sciences' while the least chosen is 'ICTs'. Concerning the second level mobility, we observe that at least 14% of students in each group has changed university after graduation with a maximum of 38.2% in 'Health'. Moreover, the tendency to change field of study seems to vary depending on the field considered. For example, 94.26% of 'Engineering' graduates decided to stay in the same field after graduation, whereas more than one-half (62.98%) of students from 'Services' decided to enroll in a different field.

Table 1 Distribution of students in mobility according to disciplinary groups

Code	ISCED - F 2013	Graduated students	Students in mobility	Same field	Different field
L1	Agriculture, forestry, fisheries and veterinary	10,959	2,631 (24.0%)	65.79%	34.21%
L2	Arts and humanities	63,756	19,440 (30.4%)	72.38%	27.62%
L3	Business, administration and law	59,949	15,849 (26.4%)	78.06%	21.94%
L4	Education	12,728	3,245 (25.4%)	83.39%	16.61%
L5	Engineering, manufacturing and construction	82,753	12,319 (14.8%)	94.26%	5.74%
L6	Health and welfare	11,258	4,309 (38.2%)	61.06%	38.94%
L7	Information and Communication Technologies (ICTs)	4,339	816 (18.8%)	80.64%	19.36%
L8	Natural sciences, mathematics and statistics	41,195	10,945 (26.5%)	93.15%	6.85%
L9	Services	12,484	3,979 (31.8%)	37.02%	62.98%
L10	Social sciences, journalism and information	68,304	23,713 (34.7%)	75.41%	24.59%

3 Multiplex data definition and analysis

We describe students mobility flows as a Multiplex network where a common set of nodes is connected through multiple types of relationships represented by different layers [14] [8]. Formally, a multiplex network $\mathscr{M}$ is a set of K graphs $G(V,E_k)$, with $k = 1,\ldots,K$ where V is the set of common nodes and E_k is the set of both intra-layer

[1] Data drawn from the Italian 'Anagrafe Nazionale della Formazione Superiore' has been processed according to the research project 'From high school to the job market: analysis of the university careers and the university North-South mobility' carried out by the University of Palermo (head of the research program), the Italian 'Ministero Università e Ricerca', and INVALSI.

edges (E_{kk}) and inter-layers edges (E_{kh}). For each layer k the corresponding adjacency matrix is $A_k = (a_{ijk})$, with $a_{ijk} = 1$ if $(v_i, v_j) \in E_k$, and $a_{ijk} = 0$ otherwise. In our case, the layers of the network are the disciplinary groups. Each of these layers holds a set of common nodes represented by Italian universities that are linked according to students flows. Intra-layer connections are observed if the edges links pairs of universities in the same field, while inter-layer connections are observed if the edges links pairs of universities in different disciplinary groups. Since Italian institutions are very different in size (and then in both out and in degree), we normalize the observed matrix applying the Multidimensional Iterative Proportional Fitting [3], already used for applications on migration flows [16]. This procedure accounts for both the attractiveness effect (columns marginal) and the loss of students (rows marginal) of universities by allowing a value ranging from 0 to 1 for each observed edge. Since the normalized network is very dense, we dichotomize the obtained results by setting a cut-off threshold at the median value of the non-zero entries in order to remove the less important links. On the resulting network, we propose a layer-by-layer comparison [5] aiming to *i)* identify the core universities and compare the results obtained for each layer *ii)* highlight the disciplinary fields that determine the largest mobility flows by applying layer similarity measures. In particular, we compute the Pearson Degree Similarity coefficient [4] and the Jaccard Layer Correlation Coefficient. The Pearson coefficient allows us to quantify the similarity between nodes' degree across layers to assess universities' centrality across different disciplinary fields. The Jaccard Layer Correlation Coefficient allows to measure the overlap between pairs of layers. This coefficient takes values between 0 and 1, with 0 indicating no overlap and 1 perfect overlap between the layers. It highlights the presence of edges among the same actors on different layers, i.e. the presence of edges linking the same group of universities on different disciplinary fields. This allow us to determine the turnover that takes place between the layers.

4 Main results and concluding remarks

Our dataset consists of 10 layers (disciplinary groups) and a set of 92 nodes (Italian universities) linked by 5,689 total intra-layer directed links of students flows. To compare the results obtained for each layer, and to highlight their main structural characteristics some network indexes, such as the density and the clustering coefficient, have been computed. Results show that the density is always pretty low, with maximum values of 0.13 for both 'Health' and 'Social Sciences' and a minimum value of 0.06 for 'Agriculture'. Global clustering coefficient [5] is homogeneous across layers, ranging from 0.31 to 0.46. We compute the in-degree and the out-degree centrality indexes for both the flattened network, when all the layers are considered jointly, and the single-layer ones. These measures are computed to identify, in both cases, the most attractive universities (i.e. high in-degree centrality) and those losing students (i.e. high out-degree centrality). As regards the results obtained for the flattened network, the top five universities attracting students are

Pisa, Bologna, Florence, Milan, and Turin, while the top five exporters are Parma, Catania, Pisa, Florence and Modena. Considering the difference among in-degree and out-degree in the flattened network, the pattern shows in the last positions many universities of the South (e.g. Magna Grecia and Kore). It is remarkable that in the top 10 position we find several telematic univeristies (UNITEL, Niccolò Cusano, Pegaso, and UniNettuno), together with universities located in metropolitan areas (e.g. Turin, Milan, and Rome). As regards single-layer networks, the in-degree centrality results show that the top five universities differ among the 10 layers considered. This element indicates that each field has its own structure, with few universities present in more than one layer (e.g. Pisa ranks in the top positions in five different layers). A similar heterogeneity in layers' structures also appears from the out-degree results, but involving different groups of universities. The results indicate that many of the outgoing flows originate from Southern Italy to Central and Northern universities. For example, considering the two most chosen groups, we notice that the top five universities reached by incoming flows in the 'Social Sciences' field are IULM, Carlo Bo, Perugia, Pisa and Siena, while the top five exporters are Bologna, Brescia, Modena, Parma and Pavia. As concerns the exchanging flows in the 'Engineering' field, the top five importing universities are La Sapienza, Bologna, Milan, Pisa, and Turin, while those losing students are Bicocca, Padua, Salerno, Trento and Udine. Another dynamic shown by the results is related to intra-region and intra-city mobility (main cities in Italy have more than one university). Overall, the pattern shows that it is almost impossible for a university, even if big, historical and prestigious, to be in a leading position in every different field. Indeed, it is more likely to observe some peculiar 'departments of excellence' that are able to consistently attract students, making that university in that specific layer a central node. Interestingly, the results show that telematic universities are central in different layers and appears mainly as importers rather than exporters.

Results are also confirmed by the multiplex networks layer similarity measures that allow us to compare layers properties. The Pearson coefficients computed on the student mobility multiplex network results show positive values for each pair of layers. In particular, the higher values involves the 'ICTs' disciplinary field whose nodes degree distributions are strongly correlated with those in the 'Agriculture' (0.74), 'Business', and 'Social Sciences' (0.60). Other positive correlations concerns the 'Services' disciplinary field with 'Arts' (0.63) and 'Social sciences' (0.63). The results show very low values of the Jaccard coefficients (close to 0) indicating that students flows among universities have not defined a common and stable network of universities for all the disciplinary fields here considered and that the relations linking them define groups of universities which are typical of each specific field. This is consistent with the previous findings; actors at the center of the network vary across layers.

In conclusion, in this contribution we have introduced the study of second level student mobility into the framework of multiplex network analysis. Several aspects of the statistical analysis to be used when dealing with students mobility data organized into multilayer structures have yet to be explored, such as Community detection methods [15]. Future lines of research will also concern the possibility of

applying these procedures to deepen the analysis on our student mobility multiplex network, without considering the telematic universities, and including also the inter-layer connections.

Acknowledgements This contribution has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide.", n. 2017HBTK5P - CUP B78D19000180001.

References

1. Attanasio M., Enea M. & Priulla, A.: Quali atenei scelgono i diplomati del Mezzogiorno d'Italia?. Neodemos, ISSN: 2421-3209 (2019)
2. Attanasio M., Enea M., Albano A.: Dalla triennale alla magistrale: continua la "fuga dei cervelli" dal Mezzogiorno d'Italia. Neodemos, ISSN: 2421-3209(2019)
3. Barthélemy, J., & Suesse, T. . mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. J. Stat. Softw, **86**(2) (2018)
4. Berlingerio, M., Coscia, M., Giannotti, F.: Finding and characterizing communities in multidimensional networks. In: 2011 International Conference on Advances in Social Networks Analysis and Mining. IEEE. pp. 490–494 (2011)
5. Bródka, P., Chmiel, A., Magnani, M., & Ragozini, G.: Quantifying layer similarity in multiplex networks: a systematic study. R. Soc. Open Sci., **5**, 171747 (2018)
6. Columbu, S., Porcu, M., Primerano, I., Sulis, I., and Vitale, M.P.: Geography of italian student mobility: A network analysis approach. Socio-Econ. Plan. Sci. **73**, 100918 (2021)
7. Columbu, S. and Primerano, I. (2020). A multilevel analysis of university attractiveness in the network flows from bachelor to master's degree. In A. Pollice, N. Salvati, and F. Schirippa Spagnolo, editors, Book of short Papers SIS 2020, pages 480–485.
8. Dickison, M. E., Magnani, M., & Rossi, L.: Multilayer social networks. Cambridge University Press (2016)
9. D'Agostino A., Ghellini G. & Longobardi S.: Exploring determinants and trend of STEM students internal mobility. Some evidence from Italy. Electron. J. App. Stat. Anal., **12**(4), 826–845 (2019)
10. Database MOBYSU.IT [Mobilità degli Studi Universitari in Italia], research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenza, Cattolica and Napoli Federico II, Scientific Coordinator Massimo Attanasio (UNIPA), Data Source ANS-MUR/CINECA.
11. Enea, M.: From South to North? Mobility of southern italian students at the transition from the first to the second level university degree. In C. Perna, M. Pratesi, and A. Ruiz-Gazen, editors, Studies in Theoretical and Applied Statistics. SIS 2016., pages 239–249. Springer Proceedings in Mathematics and Statistics, vol 227. Springer, Cham. (2018)
12. Genova, V. G., Tumminello, M., Enea, M., Aiello, F., and Attanasio, M.: Student mobility in higher education: Sicilian outflow network and chain migrations. Electron. J. App. Stat. Anal., **12**(4), 774–800 (2019)
13. Giordano, G., Ragozini, G., & Vitale, M. P. (2019). Analyzing multiplex networks using factorial methods. Soc.Netw. **59**, 154-170.
14. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A.: . Multilayer networks. Journal of complex networks, **2**, 203–271 (2014)
15. Magnani, M., Hanteer, O., Interdonato, R., Rossi, L., Tagarelli, A.: Community Detection in Multiplex Networks. arXiv preprint arXiv:1910.07646. (2019)
16. Slater, P. B.: Multiscale network reduction methodologies: Bistochastic and disparity filtering of human migration flows between 3,000+ us counties. arXiv preprint arXiv:0907.2393 (2009)
17. UNESCO Institute for Statistics. ISCED Fields of Education and International Standard Classification of Education 2011, Montrèal (2014)

Ego-centered Support Networks: a Cross-national European Comparison

Reti egocentrate di supporto sociale: confronto tra Paesi Europei

Emanuela Furfaro, Elvira Pelle, Giulia Rivellini and Susanna Zaccarin

Abstract This contribution aims at comparing patterns of social support -provided and received- among the elderly population in European countries. Adopting an ego-centered network perspective, by means of multivariate techniques for categorical data, we intend to map the ego-support network structures of the elderly, as well as ego-network functional content of provided and received support in the different countries. Individual and country socio-demographic characteristics will be also considered in interpreting results.

Abstract *Questo contributo intende confrontare i profili di supporto sociale -fornito e ricevuto- della popolazione anziana nei paesi europei, secondo la prospettiva delle reti egocentrate. Mediante l'utilizzo di tecniche multivariate per dati categoriali, si propone una mappatura delle reti di supporto, considerando per ciascun Paese la struttura di tali reti-egocentrate e il tipo di supporto scambiato. Le caratteristiche socio-demografiche individuali e dei singoli Paesi verrano considerate nell'interpretazione dei risultati.*

Key words: Social support, ego-centered network, SHARE Wave 7 data, MCA

Emanuela Furfaro
Department of Statistics, University of California, Davis; Department of Statistical Science, Università Cattolica del Sacro Cuore, e-mail: emanuela.furfaro@unicatt.it

Elvira Pelle
Department of Communication and Economics, University of Modena and Reggio Emilia, e-mail: elvira.pelle@unimore.it

Giulia Rivellini
Department of Statistical Science, Università Cattolica del Sacro Cuore, e-mail: giulia.rivellini@unicatt.it

Susanna Zaccarin
Department of Economics, Business, Mathematics and Statistics, University of Trieste, e-mail: susanna.zaccarin@deams.units.it

1 Introduction

Population aging has been a predominant phenomenon in twentieth century Europe and it is going to most likely intensify throughout the current century [12]. According to the latest data published by the European Institute of Statistics (Eurostat), the percentage of people aged ≥ 65 years in European Union -27 countries- represents the 20.3% of the total population in 2019, with an increase of more than 16 percentage points compared with 10 years earlier. This is due to consistently low birth rates, coupled with higher life expectancy in all of the EU Member States [10].

In order for societies to face the severe aging of population, it is crucial to ensure a healthy and active aging. Remaining active means to prevent mental and physical decline, to sustain health and well-being, and to enhance quality of life as people age [15]. In this context, active aging has been defined as the propensity to be engaged in activities for oneself or for others in later life. This definition also gives relevance to social support, that is considered a specific characterization of social participation. In particular, social support is defined as a set of helpful functions performed for an individual by significant others, for instance by family, friends, relatives and neighbours [1]. Received support, regardless the type (emotional, informational and instrumental), has been largely studied, highlighting the positive influence of social support on various health outcomes and well-being. However, minor attention has been devoted to support provided to others by the elderly. Indeed, providing support, especially to members outside of the household, instead of receiving it, can be considered a sign of an active lifestyle and participation in social life.

The network perspective in describing social support is widely suggested [8]. In particular, it is often investigated through *ego-centered support networks*. These are composed by the focal person (ego), and the persons or institutions - usually referred as "alters" - to which ego is related by some support tie of interest.

Drawing from the considerations mentioned above, this contribution aims to analyse patterns of social support among elderly population in European countries adopting an ego-centered perspective. We use Wave 7 data release [5] from the largest European survey on elderly, the "Survey of Health, Aging and Retirement in Europe" (SHARE, [4]). By means of multidimensional analysis techniques, we synthesize the support provided and received, providing cross-national comparisons in order to highlight specificities of the European countries in ego-support networks and in the type of support exchanged between ego and alters. Since ego-centered data on social support in SHARE are built from information on alter categories providing (receiving) support and type of support exchanged, Multiple Correspondence Analysis (MCA) [11] is an appropriate method to analyse the different ego-centered network characteristics, and to compare countries' patterns.

2 Data description and related literature

Covering 28 European countries and Israel, SHARE is the data source for most European studies on aging [4]. It is a multidisciplinary and cross-national panel database of micro data that includes information on health, socio-economic status and on social and family networks. It is a large project comprising about 140,000 individuals aged 50 or older. The first wave was carried out in 2004 with only 11 European countries, while with the last wave (Wave 7) completed in 2017 a full coverage of the EU was achieved by including 8 new countries (Finland, Lithuania, Latvia, Slovakia, Romania, Bulgaria, Malta and Cyprus). Moreover, in June 2020, a sub-sample of SHARE's panel respondents was interviewed via a Computer Assisted Telephone Interview (CATI) to collect data targeted to the COVID-19 living situation of people who are 50 years and older.

Given its richness on social networks, SHARE data have been largely used to study the social participation domain in the active ageing framework with different approaches.

More specifically, from SHARE data [3] identified clusters of elderly people with similar patterns of social participation, considering also the type of activities and their frequency. [6] investigated the level of values and personal orientations, as well as relational networks among Italian active young elders, underlining the existence of a relationship between the magnitude of the social network and the propensity to exchange with other generations.

A recent study based on the fourth wave of the SHARE analysed the effect of structural social capital on the health (measured through self-perceived health) of individuals aged 60 and above living in European countries [2]. In accordance with previous studies, results underlined that self-perceived health generally worsens with age. However, the physical health of an individual is only one of several factors influencing the perception of their own health: social capital in the form of networking, volunteering, and attending clubs appeared to be preventive against poor self-perceived health, stressing the beneficial effect of support networks in elderly people. These results were in line with previous SHARE-based evidence: not only socialising is highly beneficial for one's health, but the effect intensifies with increasing frequency and heterogeneity of social contacts.

While these studies underlined the association between social participation and health, or analyzed specific types of interaction (such as the intergenerational) the elderly entertain with others (family or non family members), we focus on a cross-national comparison of social support network structures.

3 Ego-centered support network definition in SHARE data

We consider the Wave 7 [5] release of the SHARE data, carried out in 2017. In this wave, information on received/provided support are contained in two specific modules. The "social support" module (SP) contains information about help the

respondents might receive from or give to family or non family members. Four types of received or provided support are investigated: personal care, practical household help and help with paperwork. Information on the intensity of the support -expressed as the frequency of each type of help- are also collected as well as specific questions are devoted to deepen children's care.

The second module on "financial transfers" (FT) collects information on any financial transfers and payments given or received from others.

Both modules are part of the "regular panel" questionnaire administered to a subsample of 13,959 respondents (among which 11,390 are aged 65+) living in Austria, Germany, Sweden, Spain, Italy, France, Denmark, Greece, Switzerland, Belgium, Czech Republic and Poland (the countries involved in SP and FT modules).

The set of support information allows to build ego-centered support networks [14]. The ego-centered network is composed by a focal person, ego, and a set of alters, i.e. people to whom ego is related through given or received support. The relation of interest (the different types of given or received support) existing between ego and alters is considered as "a tie" between them. In SHARE Wave 7, respondents are allowed to indicate up to three alters –family members from outside the household, friends, neighbours and others– who helped them or they have helped in the last twelve months prior to the interview. For each alter, the role relation with ego is selected from a list of 28 role relations. In particular, the first twenty (20) categories are devoted to family and kinship roles (i.e. partner/spouse, mother, father, brother, sister, child and step-child, etc.), while the four (4) next entries are devoted, respectively, to friends, (ex-)colleagues/co-workers, neighbour and ex-spouse/partner. The last entries comprise religious (minister, priest, or other clergy) and "professional" (therapist or other professional helper; housekeeper/home health care provider) roles, plus a residual category if the alter role does not fit no one in the proposed list (see Figure 1).

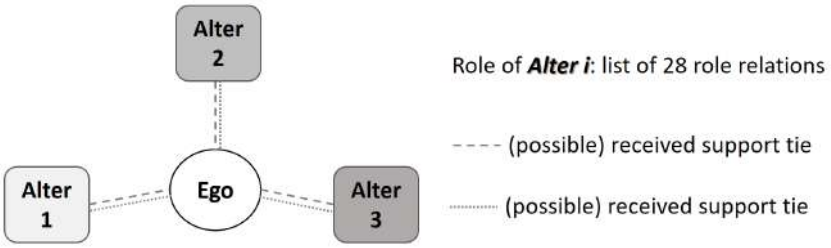


Fig. 1 Example of ego-centered support network from SHARE data.

4 Preliminary results

Table 1 provides the frequency distribution of the elders that have provided/received support outside their household. Some significant country-level differences can be noted: Southern European countries less often provide and receive support outside their household, possibly suggesting a lower propensity to seek support outside their family circle.

Table 1 Percentage of respondents who provided or received support to/from outside their household in the 12 months prior to the interview.

Country	Provided support (%)	Received Support (%)
Austria	27.6	38.6
Germany	32.9	31.8
Sweden	32.9	23.5
Spain	7.3	18.7
Italy	11.2	17.1
France	32.1	30.1
Denmark	47.5	38.8
Greece	8.6	22.7
Belgium	26.2	26.5
Czech Republic	28.0	44.0
Poland	11.3	17.5

We build the ego-centered networks of all respondent units and also consider ego's structural characteristics (gender and age), the alters' role, and the type of support. Gender and birth generation of alters can be inferred in most cases, allowing to deepen mechanisms driven by intergenerational support and/or by peer homophily support, that is the preference to be related to alters in the same birth generation.

Following [13], by using MCA [11] we map the relationship among egos, the network structure, and the type of provided/received support. MCA is suitable to deal with categorical information expressing social support, and it allows to detect and visualize underlying structures in the data. Among others, we expect to highlight patterns of inter-generational support [7].

As a final result, ego's and alters' characteristics, and type of support, will be represented as points in a bi-dimensional Euclidean space which will allow to highlight patterns of social support networks across European countries.

In interpreting results, we refer to socio-demographic and welfare system [9] characteristics of the analyzed countries. For instance - with respect to Northern European countries - Southern European countries are characterized by a "familistic" regime where people rely more on support from their family and personal social network rather than on public welfare state. We expect the detected social support patterns to reflect on such differences across countries.

References

1. Amati, V., Meggiolaro, S., Rivellini, G., Zaccarin, S.: Relational Resources of Individuals Living in Couple: Evidence from an Italian Survey. Social Indicators Research, 134(2), 547-590 (2017).
2. Arezzo M. F., Giudici C. The Effect of Social Capital on Health Among European Older Adults: An Instrumental Variable Approach. Social Indicators Research 134(1), 153–166 (2018).
3. Bordone, V., B. Arpino.: Active Ageing Typologies: A Latent Class Analysis of the Older Europeans: Data. In: Zaidi, A. and S. Harper et al. Building Evidence for Active Ageing Policies, 295–300. United Kingdom: Palgrave Macmillan DOI: 10.1007/978-981-10-6017-5. (2018).
4. Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, S. Zuber. Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). International Journal of Epidemiology. DOI: 10.1093/ije/dyt088 (2013).
5. Börsch-Supan, A. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w7.700 (2019).
6. Bramanti, D., Meda, S. G., Rossi, G.: Intergenerational Exchanges and Social Networks of Italian Active Elders: A Quantitative Analysis. The International Journal of Aging and Society, 6(4), 27–45 (2016).
7. Craveiro, D., Delerue Matos, A., Martinez-Pecino, R., Silva, S. G., Schouten, M. J. C.: Intergenerational support: the role of gender and social networks. Active ageing and solidarity between generations in Europe, 359-368 (2013).
8. Dykstra, P. A.: Aging and social support. In G. Ritzer (Ed.), Wiley-Blackwell Encyclopedia of Sociology, 2nd edition (2016).
9. Dykstra, P. A.: Cross-national Differences in Intergenerational Family Relations: The Influence of Public Policy Arrangements. Innovation in Aging, 2(1), 1-8 (2018).
10. Eurostat: Population structure and ageing - Statistics Explained. Available online at https://ec.europa.eu/eurostat/statistics-explained/
11. Greenacre, M., Blasius, J. (Eds.): Multiple correspondence analysis and related methods. CRC press (2006).
12. Grundy, E. M., Murphy, M.: Population ageing in Europe. Oxford Textbook of Geriatric Medicine, 11–17 (2017).
13. Lumino, R., Ragozini, G., Vitale, M. P.: Investigating social support patterns of single mothers from a social network perspective. International Review of Social Research 6.4, 182–194 (2016).
14. Perry, B. L., Pescosolido, B. A., Borgatti, S. P.: Egocentric network analysis: Foundations, methods, and models. Cambridge University Press (2018).
15. WHO: Active aging. Geneva: A Policy Framework. World Health Organization (2002).
16. Zaidi, A., Stanton, D.: Active ageing index 2014: analytical report. Report produced at the Centre for Research on ageing, University of Southampton, under contract with UNECE (Geneva), co-funded by European Commission, Brussels (2015).

3.16 Statistical analysis of energy data

Machine learning models for electricity price forecasting

Modelli machine learning per la previsione dei prezzi elettrici

Silvia Golia, Luigi Grossi, Matteo Pelagatti

Abstract In this paper machine learning models are estimated to predict electricity prices. As it is well known, these models are extremely flexible, can be used to include exogenous variables and allow to account for possible non-linear behavior of observed time series. Random forests (RF) and Support Vector Machines (SVM) are considered and their performances are compared with those of linear AutoRegressive (AR) models, with and without LASSO penalization. The application to Italian electricity spot prices (day-ahead market) with the inclusion of exogenous variables like forecast demand and wind generation and intra-day prices, has revealed that the prediction performance of the simple AR model is mostly better than the machine learning models. Only the SVM model seems to be a good competitor of the AR model, but even when its loss function is lower, the performance gain is hardly statistically significant.

Abstract *In questo lavoro vengono analizzate le performance predittive di alcuni modelli machine learning con riferimento ai prezzi del mercato osservati sul mercato elettrico all'ingrosso. I modelli machine learning si sono spesso rivelati efficaci nella previsione dei valori futuri delle serie storiche per la loro flessibilità che li rende capaci di includere regressori esogeni e di catturare l'eventuale non-linearità delle serie. Random Forests (RF) a Support Vector Machines (SVM) sono confrontati, dal punto di vista previsivo, con modelli AutoRegressivi lineari. L'applicazione di tali modelli ai prezzi elettrici ha rivelato che spesso i modelli AR sono preferibili rispetto ai modelli machine learning. Solo i SVM sono in grado di*

Silvia Golia
Department of Economics and Management, University of Brescia, Brescia, Italy, e-mail: silvia.golia@unibs.it

Luigi Grossi
Department of Statistical Sciences, University of Padova, Padova, Italy e-mail: luigi.grossi@unipd.it

Matteo Pelagatti
Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy e-mail: matteo.pelagatti@unimib.it

reggere il confronto con i modelli AR, ma anche quando le loro funzioni di perdita sono inferiori, raramente tale vantaggio può essere considerato statisticamente significativo.

Key words: Electricity spot prices, Forecasting, Intra-day electricity prices, Random Forests, Support Vector Machines.

1 Introduction

Electricity price forecasting is one of the most important topic in the analysis of wholesale energy markets. Many market operators are interested in future values of electricity prices: regulators, generators, traders and final users. Many models have been applied in the literature and a plethora of exogenous regressors has been considered in order to explain the dynamics of prices and thus improve the ability to predict future values. Very recently, machine learning models have received special attention. However, it is not clear whether the application of very sophisticated black-box models is really motivated by clear best forecasting performance with respect to very simple and easy to interpret linear models. The present paper try to fill this gap by comparing the performance of two machine learning models, Random Forests (RF) and Support Vector Machines (SVM), with that of linear AutoRegressive (AR) models with and without LASSO penalization. The application to the Italian electricity market is of great interest for two main reasons. First, the Italian market is one of the most transparent electricity market in the world. Second, the zonal structure allows researchers and practitioners to explore the main pros and cons of markets integration, which is a hot topic in view of the European energy markets integration pursued by the European Union. Another original contribution of the present paper is represented by the inclusion of intra-day prices among the set of regressors which, to the best of our knowledge, has not been explored yet.

2 Methods

This section contains a brief recall to the theory underlying the models used in this paper to the prediction of the electricity prices in Italy.

The first model, which can be interpreted as the benchmark model, is the AutoRegressive model with exogenous variables (ARX). The general formulation of an ARX(p) with k exogenous variables is the following:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{k} \Lambda^i(B) v_t^i + a_t \tag{1}$$

where a_t is a white noise with zero mean and $\Lambda^i(B) = \lambda_0^i + \lambda_1^i B + \cdots + \lambda_{r_i}^i B^{r_i}$, with B the backshift operator. This is a general formulation which allows the exogenous variables to be delayed. Nevertheless, in the present work only one time span for each exogenous variable is used. Equation (1) implies a linear model in which the regressors are the set of delayed y_t and the exogenous variables. They can be arranged in a $(T-p+1)\times(p+k)$ matrix, whose columns are the p delayed y_t and the k exogenous variables, so that, for the generic time span t, with $t = p+1, p+2, \ldots, T$ (T is the length of the time series), the record of the regressors is $\boldsymbol{x}_t = \{y_{t-1}, y_{t-2}, \ldots, y_{t-p}, z_{1t}, z_{2t}, \ldots, z_{kt}\}$.

A first modification of the ARX model is denoted as ARX-Lasso model and it implements a selection of explanatory variables through the LASSO-regularized linear least-squares regression [6]. The coefficient estimates are obtained by minimizing:

$$\frac{1}{2}\sum_{t=p+1}^{T}\left(y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta} + \beta_0\right)^2 + \lambda ||\boldsymbol{\beta}||_1.$$

The second modification of the ARX model is called ARX-Lasso Int. and it takes into account not only the regressors but also the interactions between each couple of them and then it performs a selection of explanatory variables through the LASSO-regularized linear least-squares regression.

Support Vector Machines (SVM; [6], [7]) are machine learning models born for classification and then adapted to regression. The regularized loss function for SVM is

$$\sum_{t=p+1}^{T} V_\varepsilon(y_t - f(\boldsymbol{x}_t)) + \frac{\lambda}{2}||\boldsymbol{\beta}||_2^2,$$

with loss

$$V_\varepsilon(r) = \begin{cases} 0 & \text{if } |r| < \varepsilon \\ |r| - \varepsilon & \text{otherwise.} \end{cases}$$

When $f(\cdot)$ is linear, it can be shown that the regressors $\boldsymbol{x}_i$ enter the solution of the minimization only through the inner products $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \boldsymbol{x}_i^\top \boldsymbol{x}_j$ and this allow the use of the so-called *kernel trick* for expanding the class of functions that approximate $f(\cdot)$. Indeed, we can approximate the unknown expectation function using a basis expansion:

$$f(\boldsymbol{x}) \approx \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{x}) + \beta_0,$$

where $h_m(\cdot)$ are basis functions. Since only inner products are relevant for the solution, one does not need to explicitly compute this expansion, but one can simply substitute the inner products with the kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{M} h_m(\boldsymbol{x}_i) h_m(\boldsymbol{x}_j)$. In our application we used the radial-basis function: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2\right)$.

The last model considered is the Random Forests (RF; [3]). RF belong to the family of ensemble learning models, with the decision tree as base learner. Decision trees were also applied in the time series context, as shown in [1] and [8]. RF is a modification of the bagging algorithm [2] by means of introducing a random

selection of input features which causes a collection of de-correlated trees. The parameters involved in a RF specification are the number of trees forming the forest (N_T), the number of covariates that are randomly selected at each split step ($mtry$) and the minimum number of observations presented into a leaf node (min_u). When the target variable is continuous, a typical choice for $mtry$ is $\lfloor r/3 \rfloor$, where r is the number of the considered explanatory variables, whereas 5 is the choice for min_u. When using RF, the well interpretable structure of the tree is lost, but it is possible to retrieve some information regarding the explicative role played by the regressors via variable importance measurement [4].

3 Results

The available time series concern the hourly electricity prices from the Italian Power Exchange (IPEX) market for the six zones of the Italian market: North (NOR), Center-North (CNOR), Center-South (CSOU), South (SOU), Sicily (SIC), Sardinia (SAR). The covered period goes from January 1st, 2015 to 31st August 2019. The data related to the first four years are used to train all the models described in Sect. 2, whereas the 8 months of year 2019 were destined for out-of-sample forecasting. The data have an hourly frequency; therefore, each day consists of 24 load periods with 00:00-01:00am defined as period 1. In this study, following a widespread practice in literature, each hourly time series is modeled separately, thereby eliminating the problem of modeling intra-daily periodicity. Moreover, the models were estimated separately for each zone of the Italian market.

The exogenous regressors considered [9] are: the day of the year, the day of the week (categorical variable with 7 classes), a calendar dummy for the festivities, the intraday market prices (MI1, MI2, MI3, MI4) at time $t-1$, the day-ahead predicted demand of electricity (source: Italian Electricity Market Manager, GME) and the day-ahead predicted wind generation (source, TERNA SpA). The day of the year as a discrete variable with values from 1 to 365 was used only for the RF, whereas it was replaced by the first 16 harmonics with base period 365 for all the other models.

The tuning of the hyper-parameters involved in the analyzed methods was done as follows. For the ARX-Lasso and ARX-Lasso Int. the hyper-parameter for L_1 penalization, λ, was fixed zone by zone and hour by hour by 10-fold cross-validation. For SVM ε and γ, were determined by 10-fold cross-validation whereas for Rf, $mtry$ and min_u were set equal to the typical choices, and $N_T = 10000$.

One-day ahead predictions were computed for a rolling window of the 8 months of 2019 and the evaluation of the forecasting performance of each estimator was carried out by means of the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Moreover, the significance of prediction differences was evaluated with the one-tailed Diebold and Mariano test [5], with the null hypothesis declared as "prediction performance of model A is equal or lower than model B".

Table 1 reports the mean value and the standard deviation of RMSE over peak (14 hours) and off peak (10 hours) hours. It is of interest to distinguish between

these two types of hours because companies adopt different strategies when demand is high or low. All the models perform better during the off peak hours, since when demand is low the supply curve is rather flat and the elasticity of the price rather low (i.e., unexpected changes in demand have small effect on the equilibrium price). Moreover, the predictive performances seems equivalent for the North and the Center-North of Italy, whereas they become worse moving from the North to the South and islands. Electricity prices for Sicily appear quite difficult to accurately predict due to the frequent and scarcely predictable switch between two regimes: as a separate market with a higher price and as part of the Italian market with a common (lower) price. Looking at the average values of the RMSE, in the peak hours SVM seems to outperform the other models in the North whereas the ARX seems to show better predictive capacities in the remaining areas. In the off peak hours ARX seems to outperform the other models in all the areas except for Sardinia where RF seems to predict slightly better the spot prices.

Table 1 *Mean value of RMSE over peak and off peak hours. In parenthesis the standard deviation*

	Area	ARX	ARX-Lasso	ARX-Lasso Int.	RF	SVM
Peak hours	NOR	6.795 (0.707)	7.282 (0.886)	7.220 (0.859)	7.080 (0.738)	6.775 (0.747)
	CNOR	6.971 (0.613)	7.353 (0.765)	7.311 (0.757)	7.344 (0.671)	7.043 (0.679)
	CSOU	7.481 (0.903)	7.805 (1.172)	7.790 (1.156)	7.720 (0.936)	7.590 (0.998)
	SOU	9.909 (2.771)	10.332 (2.885)	10.254 (2.773)	10.134 (2.664)	9.932 (2.728)
	SIC	21.311 (2.372)	22.108 (2.680)	22.108 (2.607)	21.675 (2.349)	21.662 (2.354)
	SAR	8.705 (1.519)	8.953 (1.772)	8.905 (1.746)	8.830 (1.768)	8.693 (1.656)
Off peak hours	NOR	5.216 (0.608)	5.459 (0.722)	5.382 (0.696)	5.399 (0.771)	5.293 (0.672)
	CNOR	5.951 (0.507)	6.100 (0.508)	6.115 (0.510)	6.117 (0.513)	6.061 (0.511)
	CSOU	7.360 (1.177)	7.414 (1.326)	7.481 (1.362)	7.374 (1.257)	7.403 (1.303)
	SOU	7.793 (1.327)	7.907 (1.437)	7.981 (1.446)	7.846 (1.389)	7.872 (1.387)
	SIC	14.769 (4.350)	15.096 (5.030)	15.010 (4.774)	14.752 (4.277)	14.806 (4.310)
	SAR	7.620 (0.939)	7.649 (1.019)	7.751 (1.021)	7.520 (1.022)	7.601 (1.070)

In order to inspect more deeply the predictive performances of the models, the performances on the single hours were considered. Table 2 reports the number of hours for which a given model has the lowest RMSE and in parentheses the number of times the RMSE is significantly lower according to the Diebold and Mariano test.

The results show that for the majority of the hours the ARX has a lowest RMSE in all the areas, with the exception of North and peak hours, where the SVM outperforms for most of the hours, and Sardinia and off peak hours, where RF outperforms for most of the hours. Nevertheless, in very few cases there is a significant Diebold and Mariano test.

These results have shown that the size of the prediction errors is very similar among all the models and higher in Sicily than in the other zones, especially during the peak hours. Moreover, looking at all zones, ARX (with and without LASSO and interactions) seems to perform better than nonlinear models.

Table 2 *Number of times a model is the best predictor according to RMSE. In parenthesis the number of significant Diebold and Mariano tests*

	Model	NOR	CNOR	CSOU	SOU	SIC	SAR
	ARX	5 (0)	11 (0)	12 (1)	8 (1)	12 (1)	9 (0)
	ARX-Lasso	0	1 (0)	2 (2)	1 (1)	0	0
Peak	ARX-Lasso Int.	0	1 (1)	0	0	0	1 (1)
hours	RF	0	0	0	1 (0)	2 (0)	3 (1)
	SVM	9 (0)	1 (0)	0	4 (0)	0	1 (0)
	ARX	7 (3)	8 (5)	5 (0)	6 (1)	1 (0)	3 (0)
	ARX-Lasso	0	0	2 (1)	1 (0)	1 (0)	0
Off peak	ARX-Lasso Int.	0	1 (0)	0	0	1 (0)	1 (0)
hours	RF	3 (3)	0	2 (1)	3 (0)	3 (0)	5 (0)
	SVM	0	1 (0)	1 (0)	0	4 (0)	1 (0)

4 Conclusions

The conclusions achieved in the paper can be summarized as follows. First, intra-day prices have revealed to be effective in the prediction of spot prices. Second, AR models seem to perform better than the other models in most of the zones and hours. SVM seem to be slightly better than AR in case of peak hours at least in the North zone. This result can be explained by the highest volatility of peak-hour prices which is better captured by more complex non-linear models.

References

1. Ahmed, N.K., Atiya, A.F., El Gayar, N., El-Shishiny, H.: An Empirical Comparison of Machine Learning Models for Time Series Forecasting. Econometric Reviews *29*(5-6), 594–621 (2010)
2. Breiman, L.: Bagging predictors. Machine Learning *24*(2), 123–140 (1996)
3. Breiman, L.: Random Forests. Machine Learning *45*(1), 5–32 (2001)
4. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman and Hall, London (1984)
5. Diebold, F.X., Mariano, R.S.: Comparing Predictive Accuracy. Journal of Business & Economic Statistics *13*(3), 253–263 (1995)
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference and Prediction (VII Edition). Springer, New York (2009)
7. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing *14*, 199–222 (2004)
8. Tsay, R.S, Chen, R.: Nonlinear Time Series Analysis. Wiley (2019)
9. Weron, R.: Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting *30*(4), 1030–1081 (2014)

The impact of hydroelectric storage in the Italian power market

L'impatto del pompaggio da idroelettrico nel mercato elettrico Italiano

Filippo Beltrami

Abstract The literature highlights ambiguity in the effect of storage from hydroelectric power production over the levels of carbon emissions. This paper examines the external benefit related to charge and discharge operations of hydroelectric storage power plants, applied to the case of the Northern area of the Italian wholesale electricity market. The OLS estimations based on data for year 2018 indicate that storage generation reduces carbon emissions in aggregate terms, being the estimated storage marginal emission factor (MEF) equal to 0.13 tCO_2/MWh. This finding is largely explained by the value of the MEF during off-peak hours (0.17 tCO_2/MWh), thus showing effectiveness of storage in the displacement of the carbon-intensive baseload generation acting on the margin during night-hours. However, the calculation of the MEF for peak-demand hours indicates that storage generation, individually taken, is not able to affect the structure of marginal generation in the considered area. Finally, the use of a simulation approach indicates that pumped hydroelectric storage (PHS) contributed to reduce carbon emissions into the atmosphere by 471 ktCO_2. The obtained result is consistent with the typical coefficient of round-trip efficiency of PHS documented in the literature, which amounts to 74%.

Abstract *La letteratura riporta risultati discordanti riguardo agli effetti ambientali derivanti dall'accumulazione di energia da parte degli impianti idroelettrici sui livelli di emissione di CO_2 nell'atmosfera. Il presente articolo analizza il beneficio esterno netto risultante dalle operazioni di accumulazione e rilascio di energia, tipico degli impianti di produzione idroelettrica di media-larga taglia (e, in particolare, degli impianti di pompaggio), con un focus sulla zona Nord del mercato elettrico Italiano. Le stime OLS basate sui dati relativi all'anno 2018 indicano che la produzione di energia da fonte idroelettrica è legata ad una diminuzione media delle emissioni di anidride carbonica pari a 0.13 tCO_2/MWh (MEF, fattore marginale di emissione). Tale valore è influenzato in misura maggiore dal MEF stimato per*

Filippo Beltrami
Department of Economics, University of Verona, Via Cantarane, 24, Verona (Italy).
E-mail: filippo.beltrami@univr.it

le ore fuori-picco (pari a 0.17 tCO_2/MWh), evidenziando un effetto significativo di sostituzione della produzione idroelettrica a discapito degli impianti termoelettrici ad alta intensità di carbonio, che tipicamente operano durante le ore notturne (baseload generation) in assenza di produzione da fonti rinnovabili. Al contrario, il calcolo del MEF per le ore di picco della domanda indica che la produzione di energia da fonte idroelettrica, considerata individualmente, non è in grado di influenzare in maniera significativa la struttura dell'ordine di merito degli impianti per l'area analizzata. Infine, l'utilizzo di un algoritmo di simulazione ha permesso di computare il risparmio netto di emissioni di CO_2 legato alle operazioni di accumulazione e rilascio degli impianti di pompaggio (PHS), pari ad un totale di 471 $ktCO_2$ evitate. Il risultato ottenuto è in linea con il coefficiente di efficienza media degli impianti di pompaggio tipicamente riportato in letteratura, pari al 74%.

Key words: CO_2 emissions; electricity markets; renewable energy sources; pumped hydroelectric storage

1 Introduction and data

The present article contributes to the literature on the role of storage within electricity markets by applying an OLS model with fixed effects to calculate the environmental benefit connected to hydrostorage production for the case of the Italian day-ahead (DA) power market. On top of this, this study introduces a novel simulation methodology that builds counterfactual scenarios to compute the net balance of carbon emissions connected to charge and discharge operations of *pumped hydroelectric storage* (PHS) power plants.
The Italian wholesale power market has been explored by many authors (Clò et al., 2015; Graf et al., 2020) for its high level of data transparency, spatial heterogeneity across market zones and for the increasing penetration of *Renewable Energy Sources* (RES),[1] which progressively substitute conventional polluting sources for electricity generation and become relevant to the determination of the structure of the marginal power generation mix. For the whole study, hourly data for year 2018 are explored for the zone North of the Italian market, given the large diffusion of modulable hydroelectric power plants in the area and their relevant contribution to the power generation mix.[2]

[1] From now onwards, RES are referred as renewable energy sources including bioenergies (biomass, biogas and waste), hydroelectric, photovoltaic, geothermic and eolic. This group can be further divided into programmable (such as hydropower) and non-programmable (such as wind and solar) renewable energy sources.

[2] The main source of data is the list of public offers (in Italian *Offerte Pubbliche*) provided by the GME - *Gestore del Mercato Elettrico*, the independent Market Operator (MO) - which allows to collect the full list of electricity supply and demand bids submitted on the day-ahead power market (*Mercato del Giorno Prima*, MGP), the marketplace where producers and buyers exchange power in Italy. Crucially, the preliminary technical codification of power plants and, in particular, of hydroelectric units (further distinguished in run-of-river, large hydro reservoirs and pumped hydroelectric storage) is at the origin of this article.

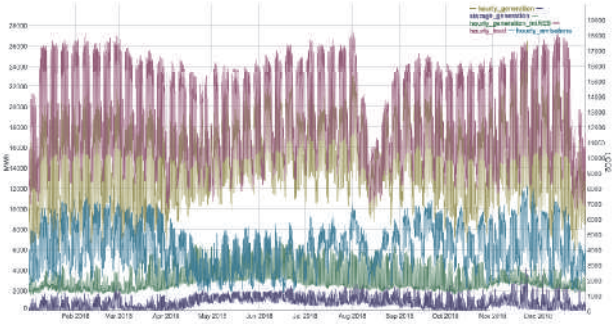


Fig. 1 *Time series data for the zone North in 2018. Note that hourly carbon emissions (light blue) are measured on the secondary vertical axis. Source: own elaboration from GME data.*

2 Research methodology

2.1 Econometric model

To assess the average environmental benefit related to the levels of power generated from storage facilities, the identification strategy suggested by Hawkes (2010) with regard to the calculation of the marginal emission factor (MEF) from electricity production is adopted. The main relationship object of investigation is formalised in Equation (1), which is then replicated for off-peak and peak sub-samples, to discriminate between relevant periods in the analysis (Carson and Novan, 2013).[3]

$$E_{h,z} = \beta_0 + \beta_1 S_{h,z} + \beta_2 intRES_{h,z} + \beta_3 netload_{h,z} + \gamma_4 D_{h,z} + \varepsilon_{h,z} \quad (1)$$

In detail, $E_{h,z}$ is the amount of hourly carbon emissions from thermoelectric power plants, $S_{h,z}$ measures the amount of hourly generation from storage power plants (large reservoirs and PHS units), $intRES_{h,z}$ is the hourly generation from intermittent RES (large and small-scale wind and solar); $netload_{h,z}$ is the net hourly demand of electricity from the grid, i.e. the total load at the net of the demand coming from PHS units;[4] lastly, $D_{h,z}$ is a set of dummies that control for low-scale seasonality, i.e. hourly and day-of-week fixed effects. The estimated coefficient β_1 represents the marginal CO_2 effect arising from a 1 MWh increase of hydrostorage generation.[5]

[3] This methodology for the estimation of MEFs is defined as "US fixed-effects" approach (Beltrami et al., 2020).

[4] As stated by Clò et al. (2015), the consumption of electricity is mainly price insensitive and inelastic due to the cyclical request of power from the grid. Thus, the inclusion of this component assures exogeneity in the model.

[5] In terms of characterisation of the marginal effect, the estimated coefficient β_1 is depurated from the increased emissions produced by the consumption of electricity of PHS units for their accumulation of energy, which tipically takes place during off-peak hours.

2.2 Simulating the effect of PHS

As argued by Staffell (2017), "pumped hydro storage does not create CO_2 emissions by itself; however, the electricity it stores is not carbon-free, and it only redelivers this electricity with 73.6% round-trip efficiency [...]". Moreover, "with round-trip losses, these plants deliver electricity with a carbon intensity 31±9% above the system average [...]. Here these emissions are accounted for when the electricity was first generated, and are attributed to the technologies which produce that electricity". McKenna et al. (2017) indicate that "the overall storage emissions factor is dependent on the marginal emissions factors during charging and discharging, and the storage round-trip efficiency".[6]
In this paper, the identification of the amount of net carbon emissions due to charge and discharge operations of PHS power plants is modelled in Equation (2).

$$\Delta E_{PHS} = \Delta E^S - \Delta E^P \tag{2}$$

ΔE_{PHS} accounts for two components. The first component is the quantity ΔE^P, which represents the extra CO_2 produced due to charging operations of PHS power plants. The request of electricity from PHS plants typically takes place during off-peak hours and needs to be fulfilled by other power generators, thus increasing potential carbon leakages. To simulate the effect of the demand from PHS on the levels of emissions, it is assumed that the demand curve is shifted under the scenario $S = PHS$ of omitting the demand bids from PHS plants in the market. This results in the creation of a new type of counterfactual equilibrium, which is characterised by lower quantities, lower price and less emissions. Thus, the comparison between the levels of CO_2 emissions under the status-quo and the counterfactual scenarios, i.e. $\Delta E^P = E^* - E^{scenario,D}$, enables to compute the amount of extra CO_2 produced due to charging operations of PHS power plants.
The second component is the quantity ΔE^S, which represents the amount of CO_2 saved due to discharging operations of PHS. In this second case, where $\Delta E^S = E^{scenario,S} - E^*$, the reasoning is opposite. In particular, E^* is the amount of emissions under the status-quo scenario whereas $E^{scenario,S}$ is the amount of emissions that would happen by simulating the removal of supply bids from PHS on the actual supply curve. In this situation, the shift of the supply curve would result in higher emissions, lower equilibrium quantities and higher equilibrium price. Hence, ΔE^S computes the CO_2 avoided from PHS generation for the crowding-out of polluting fossil fuel generators.
Eventually, Equation (2) identifies the net environmental benefit from charge and discharge operations of PHS power plants from the outcomes of the negotiations within the day-ahead wholesale power market.

[6] This is formalised by McKenna et al. (2017) with the following equation: $\varepsilon_{storage} = \varepsilon_{charge} - \varepsilon_{discharge}\eta_{storage}$ where the letter ε stands for emission factor and η for the assumed round-trip efficiency.

3 Results

3.1 The MEF from storage generation

Table 1 reports the resulting estimated storage marginal emission factor (MEF) in Column (4), based on the specification of Equation (1).
The calculations show that generation from storage is found to have a decreasing effect on carbon emissions. When fixed effects are included, the storage marginal emission factor is equal to 0.13 $tCO2/MWh$. This evidence is opposite to the one reported by Carson and Novan (2013), who documented that storage generation has an adverse impact on carbon emissions in the Texas electricity market, reporting that each MWh of electricity stored causes an average increase of 0.19 tons of CO_2 into the environment. Arguably, the result shown in Table 1 suggests that there is a sufficiently high level of integration between RES and hydroelectric storage in the Northern area of Italy, which indicates an effective displacement of baseload carbon-intensive generation.

Table 1 *OLS results - Full sample 2018. Physical market zone: North.*

	Dependent variable:			
	hourly_emissions			
	(1)	(2)	(3)	(4)
storage_generation	0.764***	0.925***	-0.195***	-0.138***
	(0.026)	(0.027)	(0.016)	(0.020)
hourly_generation_intRES		-0.230***	-0.488***	-0.587***
		(0.014)	(0.008)	(0.012)
net_load			0.331***	0.293***
			(0.002)	(0.003)
Constant	3,557.156***	4,144.441***	-347.439***	391.771***
	(25.805)	(44.716)	(37.944)	(64.834)
Observations	8,760	8,760	8,760	8,760
Fixed effects	No	No	No	Yes
R^2	0.091	0.117	0.755	0.767
Adjusted R^2	0.091	0.117	0.755	0.766
Residual Std. Error	1,401.222 (df = 8758)	1,381.331 (df = 8757)	727.656 (df = 8756)	711.025 (df = 8727)
F Statistic	881.139*** (df = 1; 8758)	580.865*** (df = 2; 8757)	8,995.911*** (df = 3; 8756)	897.136*** (df = 32; 8727)

Note: hourly and day-of-week fixed effects. *p<0.1; **p<0.05; ***p<0.01

3.2 Net carbon balance of PHS

Table 2 reports the resulting net carbon balance from the impact of PHS charge and discharge operations in the market zone North in 2018.

The total amount of saved carbon emissions from generation of PHS power plants for North in 2018 results in 631 $ktCO_2$. This figure represents only 5.74% of the total yearly amount of saved CO_2 from hydroelectric power plants in the North, as

Table 2 *Reduced CO_2 emissions from PHS generation, extra CO_2 emissions from charge of PHS, net carbon balance for North in 2018. Values in tCO_2.*

	ΔE^S	ΔE^P	ΔE_{PHS}
North	631,044.9	159,273.1	471,771.8

calculated by Beltrami et al. (2021). The total CO_2 emitted by thermoelectric power plants to serve the demand of power coming from PHS power plants during charging times was equal to nearly 159 $ktCO_2$. Hence, the net balance of carbon emissions from PHS operations was positive and equal to 471,771.8 tCO_2.
This result is consistent with the argument by Staffell (2017), who specified 73.6% as the coefficient of round-trip efficiency of PHS power plants. Reassuringly, the empirical round-trip efficiency ratio derived for PHS plants for North, based on the results of Table 2, results in 74.8%. This value is in line with the coefficient found in the literature, thus confirming the value of round-trip efficiency for a typical modern PHS facility.

References

Beltrami, F., Burlinson, A., Grossi, L., Giulietti, M., Rowley, P., and Wilson, G. (2020). Where did the time (series) go? Estimation of marginal emission factors with autoregressive components. *Energy Economics*, 104905.

Beltrami, F., Fontini, F., and Grossi, L. (2021). The value of carbon emission reduction induced by renewable energy sources in the Italian power market. *Working Papers 04/2021, University of Verona, Department of Economics*.

Carson, R. T. and Novan, K. (2013). The private and social economics of bulk electricity storage. *Journal of Environmental Economics and Management*, 66(3):404–423.

Clò, S., Cataldi, A., and Zoppoli, P. (2015). The merit-order effect in the Italian power market: The impact of solar and wind generation on national wholesale electricity prices. *Energy Policy*, 77:79–88.

Graf, C., Quaglia, F., and Wolak, F. (2020). Simplified Electricity Market Models with Significant Intermittent Renewable Capacity: Evidence from Italy. *National Bureau of Economic Research*, NBER Working Paper No. 27262.

Hawkes, A. D. (2010). Estimating marginal CO2 emissions rates for national electricity systems. *Energy Policy*, 38(10):5977–5987.

McKenna, E., Barton, J., and Thomson, M. (2017). Short-run impact of electricity storage on CO 2 emissions in power systems with high penetrations of wind power: A case-study of Ireland. *Journal of Power and Energy*, 231(6):590–603.

Staffell, I. (2017). Measuring the progress and impacts of decarbonising British electricity. *Energy Policy*, 102:463–475.

Jumps and cojumps in electricity price forecasting

Peru Muniain[1], Aitor Ciarreta[2] and Ainhoa Zarraga[3]

Abstract This paper analyzes the potential for including jumps and cojumps in electricity price forecasting models. The study is carried out on the German-Austrian day-ahead electricity market. Two price series are considered: The original price series and the Probability Integral Transformation with normal distribution (N-PIT) price series. First, the ARX model is estimated, then the residuals of the estimated ARX model are used to detect jumps and cojumps. Next, the ARX models with jumps and cojumps are estimated. To prevent over fitting, we estimate the models using elastic net. Finally, the forecasting performances of the models are compared. Results show that overall, the ARX model using the transformed price series is the best-performing model.

Key words: Forecasting, Jumps, Cojumps, ARX model, Elastic net, N-PIT transformation

[1]University of the Basque Country
Department of Applied Mathematics, School of Engineering, e-mail: peru.muniain@ehu.eus

[2]University of the Basque Country
Department of Economic Analysis, Business School, e-mail: aitor.ciarreta@ehu.eus

[3]University of the Basque Country
Department of Quantitative Economics, Business School, e-mail: ainhoa.zarraga@ehu.eus

1 Introduction

The modeling of price spikes plays a crucial role in electricity price modeling and forecasting. Many research papers have tackled the modeling and forecasting of volatility in electricity prices. Examples include [Chan et al., 2008] and [Ciarreta et al., 2017]. In both cases, volatility is modeled and predicted by including jumps in the models. Both these papers conclude that jumps add relevant information to the modeling and forecasting of volatility. [Ciarreta et al., 2020] consider jumps detected using different tests and prove that volatility forecasting performance is better when the jump test proposed by [Lee and Mykland, 2007] (LM) is used than when other jump tests are used. Moreover, [Dumitru and Urga, 2012] conclude that the LM jump test performs quite well using financial data.

Recent literature has also proven that including cojumps as well as jumps adds relevant information to time series modeling. Cojumps are jumps that occur at the same time in different time series. One of the pioneering papers in cojump analysis is [Gilder et al., 2014], where the predictive power of the models is improved by including cojumps as regressors.

Electricity prices may be transformed to improve forecasting performance.For instance, [Uniejewski et al., 2018] compare electricity price forecasting performance for several variance stabilizing transformations in different electricity markets. They conclude that the probability integral transformation with a normal distribution (N-PIT) is the best transformation overall in terms of forecasting accuracy.

In the literature, many approaches have been applied to model electricity prices. Recent research focuses on autoregressive models with exogenous variables: ARX models. In [Ziel and Weron, 2018], univariate and multivariate models are compared. They conclude that there is a slight gain in forecasting performance when multivariate models are considered. Computation time is also lower for multivariate models. However, ARX models are usually over-parameterized, which makes them difficult to estimate using OLS. Estimation methods with a shrinkage property have therefore been applied in the literature. See [Tibshirani, 1996] and [Zou and Hastie, 2005] for for lasso and elastic net estimation methods, respectively. [Uniejewski et al., 2016] apply different estimation methods with a variable selection property in electricity price forecasting, and conclude that the best performing method is the elastic net.

Finally, the most common criteria used in the literature to assess the forecasting performance of different models are the root mean squared error (RMSE) and the mean absolute error (MAE) . Examples can be found in [Uniejewski et al., 2018] and [Uniejewski et al., 2016].

The research seeks to forecast German-Austrian day-ahead electricity prices as accurately as possible. The day-ahead market is the market with the highest liquidity in Germany and Austria, so many agents participate in it. The participating agents need signposts to decide what bidding strategy will maximize their profits optimally. Those signposts are the forecasts made for the following days' prices. This article analyzes whether jumps and cojumps add useful information to electricity price modeling and forecasting.

In the analysis, two price series are considered: (i) the original price series; and (ii) the transformed price series applying the N-PIT. Both time series are separated into 24 time series, each corresponding to one hour of the day. Thus, there are two multivariate time series, the original and the transformed price series, and each has 24 individual price series. We analyze the forecasting performance of both the original and the N-PIT transformed price series when jump and cojump information is added. The benchmark model (ARX) estimated is an ARX-type model similar to the so-called fARX (full ARX) model by [Ziel and Weron, 2018]. The residuals of the ARX model are analyzed to detect jumps and cojumps.

Jumps are detected as proposed by [Lee and Mykland, 2007]. We use the LM jump test because the test needs to be applicable for daily data, and of the most popular jump tests applied in the literature only the LM works with daily observations. See for example [Ciarreta et al., 2020] and [Dumitru and Urga, 2012]. Cojumps are constructed following [Gilder et al., 2014]. Cojumps are considered as jumps that occur on the same day at different hours which takes into account the correlation between jumps at different hours.

After jumps and cojumps are detected, ARX-type models with jump and cojump variables (the so-called ARX-J and ARX-J-CJ models) are estimated. All three models (ARX, ARX-J and ARX-J-CJ) are estimated by applying the elastic net estimation method and using both the original and the N-PIT transformed prices. Then forecasting is carried out, with the following seven days being predicted for all 24 hours of the day. Finally, MAE and RMSE criteria are used to assess the forecasting performance.

2 Methodology

The modeling of the prices follows several steps. In the first step the price series is transformed using the N-PIT transformation proposed by [Uniejewski et al., 2018]. Then we separate original and N-PIT transformed price series, one for each hour, so as to reduce forecast errors. The ARX model specified below is estimated using elastic net estimation. The next step is to detect the jumps in the residuals of the ARX model; to that end the jump test proposed by [Lee and Mykland, 2007] is applied. After all the jumps in each of the 24 time series have been detected, cojumps are detected as per [Gilder et al., 2014]. Finally, the two models proposed in Subsection 2.1 are estimated, i.e. the ARX-J and ARX-J-CJ.

After the models are estimated, forecasting is carried out. MAE and RMSE criteria assess the accuracy of the predicted values.

2.1 Models

Three different ARX-type models are estimated. The first is the ARX model, based on the fARX model proposed by [Ziel and Weron, 2018],

$$p_{d,h} = \underbrace{\beta}_{\text{Constant}} + \underbrace{\sum_{h=1}^{24}\sum_{i=1}^{8}\beta_{i,h}p_{d-i,h}}_{\text{Autoregressive effects}} + \underbrace{\sum_{i=1}^{8}\beta_{min,i}p_{d-i,min} + \sum_{i=1}^{8}\beta_{max,i}p_{d-i,max}}_{\text{Non-linear effects}}$$
$$+ \underbrace{\sum_{j=1}^{6}\left[\left(\gamma_{0,j} + \gamma_{1,j}p_{d-1,h} + \gamma_{24,j}p_{d-1,24}\right)\mathrm{W}_d^j\right]}_{\text{Day-of-the-week effects}} + \varepsilon_{d,h} \tag{1}$$

where $p_{d,min}$ and $p_{d,max}$ are the minimum and maximum prices throughout the 24 hours on day d, $p_{d-1,24}$ is the price on day d-1 and at hour 24, W_d^j is a dummy variable for day j of the week, and $\varepsilon_{d,h}$ is the error term with mean 0 by construction. The second term accounts for up to eighth order autoregressive and cross-period effects (effects of each hour from up to 8 days ago). The third term accounts for non-linear effects. The fourth term accounts for seasonality. In the ARX model there are 227 parameters in total to estimate for each hour.

In the second case, jumps are included in model (1), resulting in the ARX-J model. The jumps are detected in the residuals of the ARX model. Because the sign of the jumps might be different depending on the hour of the day, this model considers both positive and negative jumps. The ARX-J model is expected to capture the behavior of prices more accurately at the tails of the distribution.

The last model proposed is the ARX model with jumps and cojumps. The ARX-J-CJ model accounts for correlation between jumps by considering cojumps, which are jumps that occur on the same day across different hours. The ARX-J-CJ model not only accounts for correlation by cross-period effects; it also takes into account correlation in the tails, through the cojump variable.

2.2 Forecast

A window of D observations is considered for estimating the models and the prices for the following H days are predicted for each hour. The window is then moved one day forward and the estimation and forecasting procedure is repeated. In total, there are N different windows of equal size.

The RMSE and MAE criteria are used to assess forecasting performance over the N rolling windows and the H horizons. By construction, the MAE criterion is optimal for median forecasts while the RMSE is optimal for mean forecasts.

3 Data description

The data used in this paper are day-ahead prices from the German-Austrian electricity market. The data run from 1^{st} January 2014 to 30^{th} September 2018. In total, there are 1727 days. 30^{th} September 2018 was the last day on which the German-Austrian day ahead market operated.

The data consists of 24 different time series, one for each hour of the day. The length of each estimation window is $D = 730$, the initial rolling window starts on 1^{st} January 2014 and ends on 31^{st} December 2015, $H = 7$ horizons are predicted in each window, and $N = 997$ different rolling windows are considered. Note that the first 730 observations of the sample are used to forecast the first price.

Fig. 1 Transformed price series and histograms (hours 4 and 19)

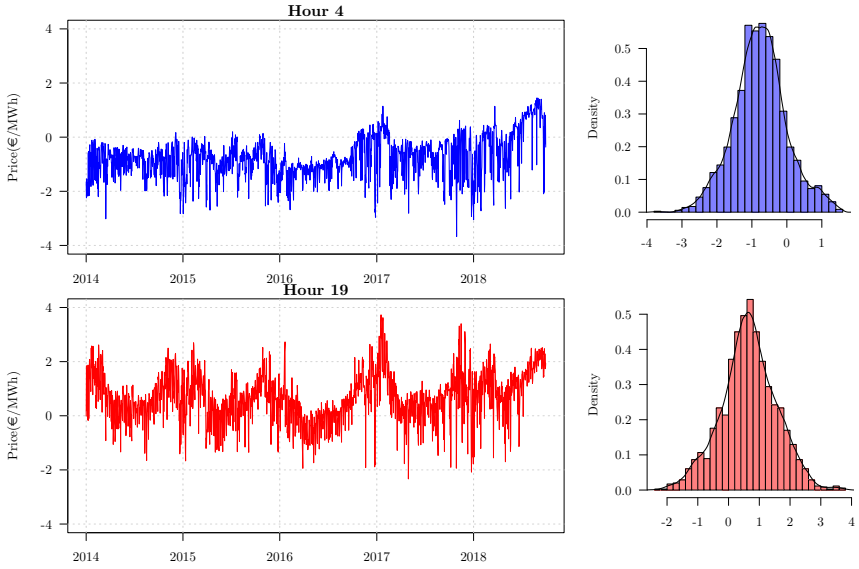
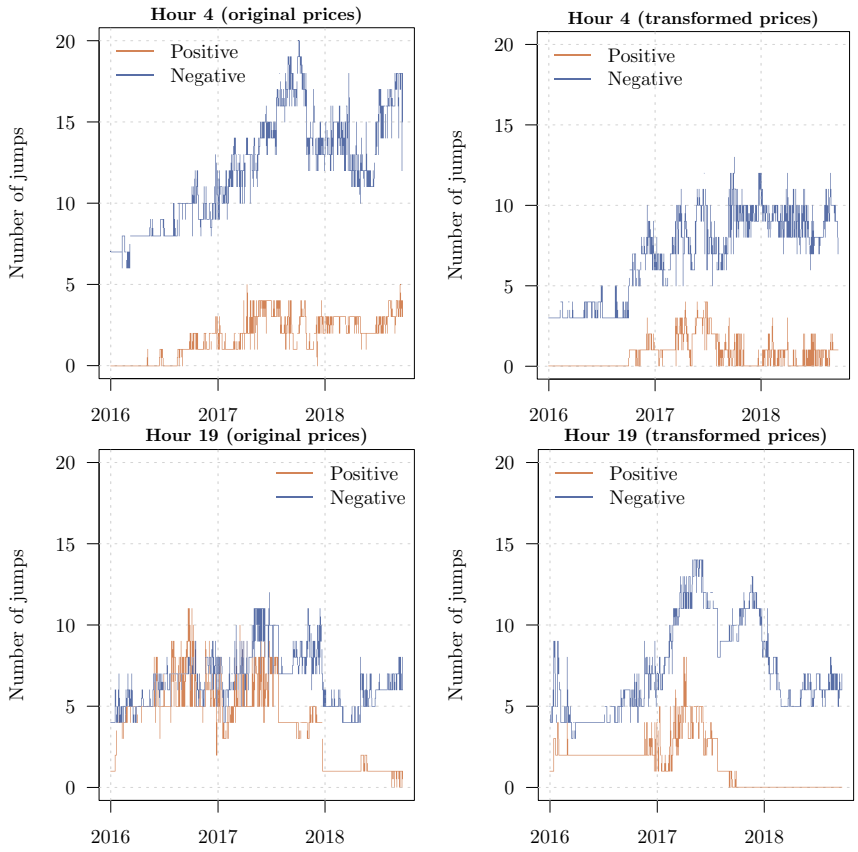


Figure 1 shows the price series and its histogram after applying the N-PIT transformation. This transformation converts the original data into normally distributed data, and after forecasting the inverse transformation is applied to the predicted values. The histograms show that the distributions of both series are closer to the normal distribution. After the transformation, all 24 time series follow a close-to-normal distribution. In these transformed time series, the tails are thinner, so the LM test should detect only a few jumps.

According to the LM test, there are significant jumps in both original and transformed prices in most of the 24 hours. Figure 2 shows the total number of positive and negative jumps detected in the residuals of the ARX model in each rolling window for hours 4 and 19, and both the original and transformed prices. In hour 4 there

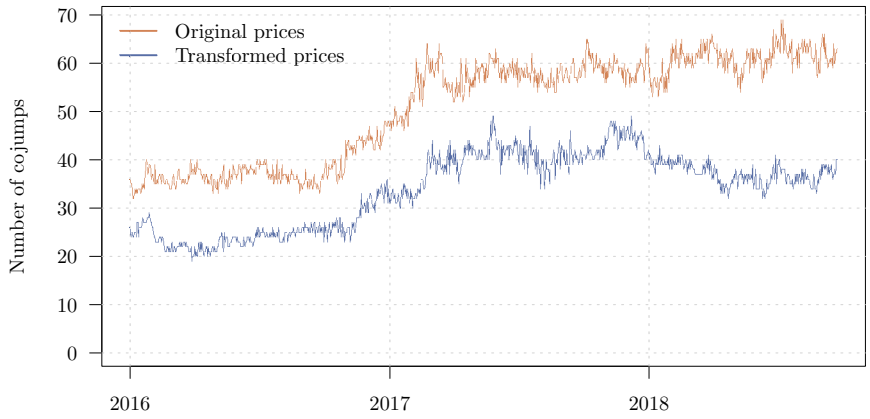
Fig. 2 Number of jumps detected in each rolling window



are fewer jumps in the transformed price series than in the original price series and more negative jumps than positive jumps. By contrast, in hour 19, the number of jumps is similar for both price series but the difference between the number of positive and negative jumps is less clear in most of the rolling windows for the original prices.

Figure 3 shows the number of cojumps detected in the residuals of the ARX model in each rolling window for both the original and transformed prices. As expected, the number of cojumps detected is lower in the N-PIT transformed price series than in the original time series. Observe that in the latter the number of cojumps detected increases to almost 70 at the end of the period in which the market was in place. This seems to be particularly so in early 2017 and late 2018.

Fig. 3 Number of cojumps detected in each rolling window



4 Empirical results

Regarding the estimation results in the elastic net estimation jump and cojump variables are often included in the ARX-J-CJ model which suggests that they might add relevant information to the models. Hence, the ARX-J-CJ model is expected to outperform the ARX-J and ARX models in price forecasts.

The forecasting performance of the different models is measured using MAE and RMSE criteria. As expected, the performance of each model varies depending on the hour of the day and the horizon.

Tables 1 and 2 show the forecasting performance in each horizon (H1 to H7) according to the RMSE criterion, taking the mean of off-peak hours and peak hours, respectively, as explained in Section 2.2. In general, the ARX model for N-PIT transformed prices provides the most accurate forecasts for off-peak hours, while jumps and cojumps should be included in the models when original prices are used. For peak hours, models that use the original prices and include jumps or jumps and cojumps are the best in terms of forecasting performance. Furthermore, the N-PIT transformation can be seen not to perform well with the ARX-J and ARX-J-CJ models. The poor performance of these models with the transformed prices may be due to the inverse transformation effect, where the size of the jumps might not be adequately accounted for when the inverse of the transformation is applied.

Overall, it can be observed that the errors according to the RMSE criterion are larger for peak hours than for off-peak hours. This may be because the volatility of the time series in peak hours is greater than in off-peak hours, which makes it more difficult to forecast prices accurately (see, for example, Figures 1 and 2)[1].

[1] Results using MAE criterion mostly choose the ARX model with N-PIT transformed prices. The table is available on request.

Table 1 RMSE off-peak hours

Model	Transf.	H1	H2	H3	H4	H5	H6	H7
ARX	Original	9.231	12.561	13.592	13.767	13.528	13.721	14.338
	N-PIT	8.993	12.208	12.97	13.331	13.648	13.883	13.716
ARX-J	Original	9.271	12.406	13.361	13.788	13.743	14.161	14.248
	N-PIT	11.148	13.324	13.943	13.958	14.303	13.733	14.075
ARX-J-CJ	Original	9.26	12.628	13.225	13.75	13.973	14.121	14.039
	N-PIT	10.096	12.62	13.447	13.552	13.786	13.814	14.125

Table 2 RMSE peak hours

Model	Transf.	H1	H2	H3	H4	H5	H6	H7
ARX	Original	13.374	15.967	16.97	16.947	17.029	17.138	17.469
	N-PIT	13.544	16.294	16.397	17.124	17.343	18.119	17.726
ARX-J	Original	13.581	15.465	16.585	17.229	16.808	17.923	17.793
	N-PIT	16.126	18.072	18.187	18.428	19.18	18.2	18.948
ARX-J-CJ	Original	13.345	16.087	16.447	17.398	17.132	17.634	17.534
	N-PIT	15.457	17.124	17.957	18.134	18.128	18.449	18.321

5 Conclusions

Forecasting results differ for peak and off-peak hours. On the one hand, the forecasting performance of the ARX-J and ARX-J-CJ models using the original price series is better than that of the ARX model for peak hours. This emphasizes the importance of including jumps and cojumps in forecasting models. Indeed, the ARX-J-CJ using the original price series is the best model for the first horizon predicted. The forecasting performance of the ARX-J and ARX-J-CJ models using the original price series becomes weaker when the forecasting horizon increases. The poor forecasting performance of these models with longer horizons indicates that they do not adequately capture the jump and cojump effects.

On the other hand, in off-peak hours, the relevance of jumps seems to be lower as the ARX model using the transformed prices outperforms the other models. The ARX-J and ARX-J-CJ models using transformed price series perform quite poorly in all cases. As may happen with any transformation, at the tails of the transformed price distribution the difference between predicted and observed values, even if the two are close together, may be large when the inverse of the transformation is applied.

At off-peak hours we recommend using the ARX model with the transformed prices, while at peak hours we suggest applying the ARX-J-CJ model with the original price series. These forecasts are also of interest to participants in the futures market. Electricity markets around the world are encouraging market agents to participate in futures markets, and the decision is taken after profitability analyses. For instance, Phelix futures are traded on the EPEX market. In such futures it is possible to trade base and peak prices, which are the mean of all hours of the day and the mean of peak hours, respectively. Hence, day-ahead price forecasting helps partici-

pants to optimize their bidding strategies for the following days and decide whether to participate in the futures market or not.

Finally, due to the differences in price formation for peak and off-peak hours, research needs different models to be estimated for each hour of the day in order to reduce forecast errors significantly. These differences emerge because the different technologies are marginal at different hours of the day.

Acknowledgements

Financial support from Dpto. de Educación, Universidades e Investigación del Gobierno Vasco under research grant IT1336-19 is acknowledged. The authors are grateful for valuable comments from participants in the Workshop on Forecasting in Electricity Markets held in Bilbao in 2019. The authors also thank Susan Orbe, Luiggi Grossi and Rafał Weron for helpful comments.

References

[Chan et al., 2008] Chan, K. F., Gray, P., and van Campen, B. (2008). A new approach to characterizing and forecasting electricity price volatility. *International Journal of Forecasting*, 24(4):728–743.

[Ciarreta et al., 2017] Ciarreta, A., Muniain, P., and Zarraga, A. (2017). Modeling and forecasting realized volatility in German–Austrian continuous intraday electricity prices. *Journal of Forecasting*, 36(6):680–690.

[Ciarreta et al., 2020] Ciarreta, A., Muniain, P., and Zarraga, A. (2020). Realized volatility and jump testing in the japanese electricity spot market. *Empirical Economics*, 58(3):1143–1166.

[Dumitru and Urga, 2012] Dumitru, A.-M. and Urga, G. (2012). Identifying jumps in financial assets: a comparison between nonparametric jump tests. *Journal of Business & Economic Statistics*, 30(2):242–255.

[Gilder et al., 2014] Gilder, D., Shackleton, M. B., and Taylor, S. J. (2014). Cojumps in stock prices: Empirical evidence. *Journal of Banking & Finance*, 40:443–459.

[Lee and Mykland, 2007] Lee, S. S. and Mykland, P. A. (2007). Jumps in financial markets: A new nonparametric test and jump dynamics. *The Review of Financial Studies*, 21(6):2535–2563.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Uniejewski et al., 2016] Uniejewski, B., Nowotarski, J., and Weron, R. (2016). Automated variable selection and shrinkage for day-ahead electricity price forecasting. *Energies*, 9(8):621.

[Uniejewski et al., 2018] Uniejewski, B., Weron, R., and Ziel, F. (2018). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, 33(2):2219–2229.

[Ziel and Weron, 2018] Ziel, F. and Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, 70:396–420.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

3.17 Statistical methods and models for the analysis of sports data

Football analytics: a Higher-Order PLS-SEM approach to evaluate players' performance

Analisi statistica nel calcio: un approccio Higher-Order PLS-SEM per valutare la performance dei calciatori

Mattia Cefis and Maurizio Carpita

Abstract Nowdays, data science is applied in several area of our life, and also many applications in sports fields are increasing. In this context, we are focusing on football (e.g. soccer); thanks to this work we have the aim to give a new approach in the evaluation of football players' performance given from the EA Sports experts and available on Kaggle in the KES dataset. For this purpose, we adopt a Higher-Order PLS-SEM approach to the *sofifa* KPIs (e.g. Key Performance Indicators) in order to compute a composite indicator and compare it with the well-known overall index from EA Sports. The final goal is to suggest a new performance index for helping coaches and scouting staff of professional teams to take strategic decisions, in order to evaluate impartially players' performance.

Abstract *Oggi la data science è applicata in diversi contesti della nostra vita e anche in ambito sportivo le sue applicazioni sono in crescita. Nel nostro contesto ci siamo focalizzati sul calcio e con questo lavoro proponiamo un approccio innovativo all'analisi della performance dei calciatori partendo da quella già offerta dagli esperti di EA Sports e disponibile sulla piattaforma Kaggle grazie al KES dataset. A tale scopo adottiamo un approccio Higher-Ordered PLS-SEM agli indici di performance di sofifa per calcolare un nuovo indice composito, confrontandolo con quello di EA Sports. L'obiettivo finale è quello di proporre un nuovo indice di performance per aiutare allenatori e l'area scouting di una società calcistica a prendere decisioni strategiche e a valutare oggettivamente i calciatori.*

Key words: football performance indicators, PLS-SEM, composite indicators.

Mattia Cefis
University of Brescia, Department of Economics and Management, e-mail: mattia.cefis@unibs.it

Maurizio Carpita
University of Brescia, Department of Economics and Management
e-mail: maurizio.carpita@unibs.it

1 Introduction

The latest developments in sports research, especially in football, are driven from a sort of a new "data-culture" approach. Players' performance evaluation is becoming a strategic key for football coaches and for the management of a football team. We know that players' performance on the soccer field has been extensively measured and described by soccer experts: in literature, very important are the detailed classification by the experts from Electronic Arts (EA)[1]. In their opinion, players' performance can be thought as a multidimensional construct made up of 6 performance composite indicators (e.g. *defending*), each of which consists of several, more specific skills (e.g. *marking*, *standing tackle* and *sliding tackle* as elements for the *defending* dimension), which combined form an *overall* index that sums up the performance; here the main problem is that experts' opinions are not statistically supported [2, 3].

In this paper, our goal is to propose the use of the Higher Order PLS-SEM approach, starting from the data a relevant Data Science platform (e.g. Kaggle) in order to build a new composite index and to compare it with the well-known *overall* index from EA Sports experts, in order to give a significant statistics support to the experts' opinion.

2 Literature overview and data employed

In order to give an overview about literature, we can say that there are two main approaches in football analytics: an explorative method oriented on analysis and classification of the KPIs (e.g., Key Performance Indicators) with the aim to evaluate players' performance [2, 3] and another one oriented in the prediction of football match results [4]. Furthermore, in order to evaluate the single player's performance there exist different methods: for example Pappalardo [8] adopted a SVM observing match outcome, Schultze and Wellbrock [10] created a rating performance index thanks to a plus-minus metric, Carpita [1] adopted an unsupervised method to classify different area of performance. We will focalize our attention on this last issue (e.g. evaluation of single player's performance), in fact our goal is to explore players' performance variables (e.g. KPIs), in order to evaluate some different strategic skills of each one; it can be useful for understanding any key choice of coaches, as well as to guide player transfer decisions, transfer fees and contract negotiations or to improve future predictive modelling.

In the European framework, the Kaggle European Soccer (KES) database is the biggest open one devoted to the soccer leagues of European countries: it contains data about 10000 players and 21000 matches of the championship leagues of 10 countries and 7 seasons from 2009/2010 to 2015/2016. It is composed mainly by two big tables:

[1] Link to the website: https://www.easports.com/

- The Match table contains the date, the positions (X and Y coordinates) on the pitch for the 22 players of the two teams and the final result of each match.
- The Player Attributes table contains other 29 variables (e.g. KPIs), with periodic player's performance on a 0–100 scale with respect to different abilities.

For our work we are interest just in the Player table and in particular we will take into account just midfielder's players from Italian Serie A 2015/2016, with stats relying the beginning of the season, in order to have a toy dataset of 106 players and 29 KPIs for each one. As said in the introduction, for what concerns attributes' description, experts of Electronic Arts (EA/*sofifa*) Sports are considered the main authority: players' performance is defined as a multidimensional entity made up of 6 latent traits (e.g. *attacking*, *skill*, *movement*, *power*, *mentality*, *defending*), but they are not statistically supported [2, 3]. Our goal is to apply to these KPIs a Higher-Order PLS-SEM model, in order to create a new synthetic composite indicator and compare it with the *overall* index of EA Sports experts.

2.1 The proposed Higher-Order PLS-SEM approach

PLS-SEM [11], also called PLS Path-Model, is a very interesting tool that offers us a valid alternative to the well-known covariance-based model [6]. Its goal is to measure causality relation between concepts (e.g. latent variables, the 6 *sofifa* latent traits in our case), starting from some manifest variables (e.g. MVs, in our case the *sofifa* KPIs), thanks to an explorative approach: the explained variance of the endogenous latent variables (e.g. LVs, variables that we see as a sort of outcome, the performance in our case) is maximized by estimating partial model relationships in an iterative sequence of ordinary least squares regression [7]. Another essential point of PLS-SEM is that does not require any preliminary assumptions for the data, so it's called a soft-modelling technique. PLS-SEM estimates simultaneously two model:

- Measurement (or outer) model $\Rightarrow$ links MVs (e.g. KPIs in our case) to their LVs (e.g. the 6 *sofifa* dimensions). Each block of MVs $\mathbf{X}_g$, $g = 1, ..., G$ (with $G = 6$) must contain at least one MV and this relation can be treated in two ways: reflective (where the MVs are the effects of their own LV) and formative (where the MVs are the causes of their own LV). In our work we will assume a formative structure for the outer model where each LV ξ_g is considered to be formed by its KPIs following a multiple regression:

$$\xi_g = \mathbf{X}_g \mathbf{w}_g + \delta_g \tag{1}$$

and

$$E[\delta_g | \mathbf{X}_g] = \mathbf{0} \tag{2}$$

where $\mathbf{w}_g$ is the vector of the outer regression weights and δ_g is the vector of error terms. So, the vector of the outer weights for the g-th LV is estimated by

least squares:

$$\mathbf{w}_g = (\mathbf{X}_g^T \mathbf{X}_g)^{-1} \mathbf{X}_g^T \xi_g \tag{3}$$

- Structural (or inner) model $\Rightarrow$ thanks to this model LVs are divided into two groups: exogenous and endogenous. The first one does not have any predecessor in the path diagram, the rest are endogenous. For the *j*-th endogenous variable in the model, the linear equation of its own structural model is:

$$\xi_j = \beta_0 + \sum_{r=1}^{R} \beta_{rj} \xi_r + \zeta_j \tag{4}$$

  where R is the number of exogenous LVs that affect the endogenous one and β_{rj} is so called path coefficient, a sort of linkage between the *r*-th exogenous LV and the *j*-th endogenous LV and ζ_j is the error term.

Moreover, for our work we will assume a PLS-SEM with Higher-Order Constructs, also known as Hierarchical Models [9]. In this framework we can include LVs that represent an "higher-order" of abstraction. In fact, for our purpose, we will assume players' performance as extra-latent construct of higher (second) order. Since this LV is virtual, and so without any apparent MVs, literature suggested us an interesting technique in order to modelling this framework: a two-step or patch approach [9]. In the first step of this approach, we can compute thanks to PCA (e.g. Principal Component Analysis) the scores of the lower-order LVs (e.g. the first principal component -I PC- of each one), while in the second one we can apply the classical PLS-SEM using the computed scores as MVs for the endogenous (e.g. the performance) LV. In our work, we will build two different frameworks, following the experts' suggestion[2] , in order to replicate the EA Sports *overall*:

- In the first framework, with the classical *sofifa* LVs classification (6 groups of LVs), we assume a conceptual structure behind the performance [9] with the presence of 3 endogenous LVs: *attacking*, *defending*, and the player's performance (e.g. PLS Path in Fig. 1). Note that for the performance (the only II order construct), we used the I PCs of *movement*, *defending* and *attacking* as MVs.
- In the second framework we take in consideration the EA FIFA cards ability classification (a little bit different classification of the same 29 MVs into others 6 LVs); here we assume just one endogenous Higher-Order LV (e.g. performance) influenced directly from the others 5 exogenous (Fig. 1).

For the work we used the R package *plspm* [9] and bootstrap for the validation of the models. In the next section we will share our results and a brief discussion.

[2] For details see the website: https://www.fifauteam.com/fifa-19-attributes-guide/

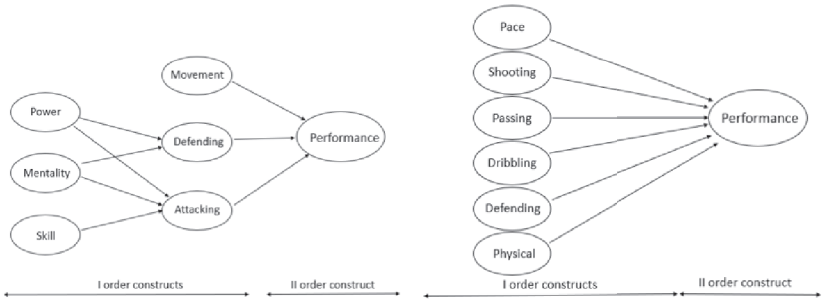


Fig. 1 PLS Path *sofifa* model vs FIFA cards model

3 Results and discussion

Preliminary results are showed in Fig. 2, where we can see an example of loadings (in a formative way) for the *defending* LV in both *sofifa* and FIFA cards model. We can see immediately the differences in these two classifications (3 KPIs for the first and 5 for the second): loadings are not exactly the same but are very high in both the cases.

In Table 1 instead we can see a comparison regards some assessments index between our two models: the unidimensionality holds in both frameworks, while the goodness of fit index (e.g. GoF) is good (e.g. > 0.7, [9]) and reveals that the second framework is a bit better than the first one. Then we computed the *rho* index (e.g. the correlation) between our Higher-Ordered PLS-SEM performance index and the true *overall* index computed from EA experts. It shows us a very high concordance ($rho > 0.9$) between our index and the EA index, in both frameworks.

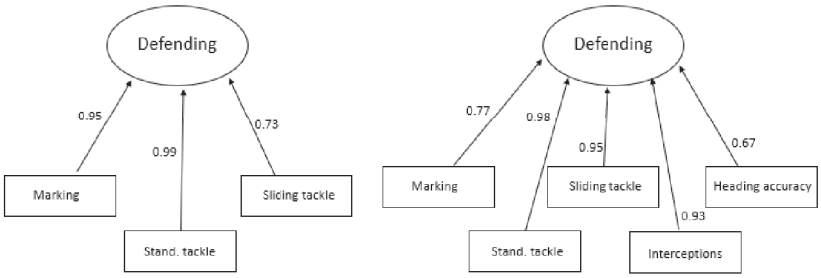


Fig. 2 Loadings comparison for *defending* LV between *sofifa* vs FIFA cards model

In Table 2 we can see the output of the bootstrap validation (with 200 samples) and how both models have a significative R^2 index for their own endogenous LVs. Interesting to note how in the second model all MVs and LVs are significative for their respectively outer and inner model.

Table 1 The two models goodness index comparison

Model	Unidim. LVs	GoF	Corr. with the EA *overall* index
sofifa	OK	0.71	0.94
FIFA cards	OK	0.82	0.93

Table 2 The two models validation comparison

Model	Non-sign. MVs	Non-sign. LVs	CI 95% for R^2 of the endogenous LVs
sofifa	1	1	$A.: [0.89; 0.95]$ $D.: [0.67; 0.83]$ $P.: [0.95; 0.98]$
FIFA cards	0	0	$P.: [0.98; 0.99]$

In summary, we have seen how both models are good and so we reapplied them across data of others European leagues and players' roles, discovering some little differences between the path coeffiecients: because of this, for future research it could be interesting, as in-depth analysis, to focus on the problem of observed and unobserved heterogeneity for players' performance (e.g. roles, leagues, teams...), maybe thanks the REBUS-PLS algorithm [5].

References

1. Carpita, M., Ciavolino, E., Pasca, P.: Players' role-based performance composite indicators of soccer teams: A statistical perspective. Social Indicators Research (2020): 1-16.
2. Carpita, M., Golia, S.: Discovering associations between players' performance indicators and matches' results in the European Soccer Leagues. Journal of Applied Statistics (2020): 1-16.
3. Carpita, M., Ciavolino, E., Pasca., P.: Exploring and modelling team performances of the Kaggle European Soccer database. Statistical Modelling 19.1 (2019): 74-101.
4. Carpita, M., et al.: Discovering the drivers of football match outcomes with data mining. Quality Technology & Quantitative Management 12.4 (2015): 561-577.
5. Esposito Vinzi, V., et al.: REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling. Applied Stochastic Models in Business and Industry 24.5 (2008): 439-458.
6. Jöreskog, K.G.: Structural analysis of covariance and correlation matrices. Psychometrika 43.4 (1978): 443-477.
7. Monecke, A., Leisch, F.: semPLS: structural equation modeling using partial least squares. Journal of Statistical Software, 48 (3),(2012) 1-32.
8. Pappalardo, L., et al.: PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Transactions on Intelligent Systems and Technology (TIST) 10.5 (2019): 1-27.
9. Sanchez, G.:. PLS path modeling with R. Berkeley: Trowchez Editions 383 (2013)
10. Schultze, S.R., Wellbrock, C.M.: A weighted plus/minus metric for individual soccer player performance. Journal of Sports Analytics 4.2 (2018): 121-131.
11. Wold, H.: Encyclopedia of statistical sciences. Partial least squares. Wiley, New York, (1985): 581-591.

Bayesian regularized regression of football tracking data through structured factor models

Regressione bayesiana di dati di tracking nel calcio con regolarizzazione via modelli fattoriali strutturali

Lorenzo Schiavon and Antonio Canale

Abstract The recent spread of football tracking data motivates the development of statistical tools able to extract and summarize valuable knowledge from the large amount of information available. Factor analysis is routinely used in statistics to reduce dimensionality and when it is applied to a set of regressors it induces regularization that can improve the out-of-sample prediction performances of the linear model. In this article, we propose to use a structured infinite factor model on a set of tracking performance indicators used as covariates of a model for dangerousness of football actions. Such factor model is able to induce a flexible penalty structure on the linear regression model which can be, on the other hand, easily interpreted, providing useful insights in terms of football strategy.

Abstract *La recente diffusione di dati di tracciamento posizionale nel calcio incentiva lo sviluppo di metodologie statistiche in grado di estrarre conoscenza dalla grande quantità di informazione disponibile. L'analisi fattoriale è comunemente utilizzata in statistica per ridurre la dimensionalità dei dati, migliorando le performance predittive di un modello lineare se applicata alla matrice di regressori. In questo articolo proponiamo di applicare un modello fattoriale strutturale ad un set di indicatori di performance usati come covariate di un modello per la pericolosità di azioni di calcio. Tale modello fattoriale è in grado di indurre una struttura di penalità che sia flessibile e permetta, allo stesso tempo, una facile interpretazione al fine di ottenere informazioni strategiche per lo sviluppo di azioni nel calcio.*

Key words: Dimensionality reduction; Factor analysis; Group penalty; Increasing shrinkage; Infinite factorization; Tracking football data.

1 Introduction

Data and statistics in football association (football hereafter) are commonly based on ball related events, due to the manual system of data collection. Recently, advances in computer vision techniques make it possible to automatically track every

Lorenzo Schiavon · Antonio Canale
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy. Correspondence: Lorenzo Schiavon - e-mail: lorenzo.schiavon@phd.unipd.it

player on the pitch at discrete but frequent time points. It should be emphasised that despite these methods were introduced around ten years ago [4], it is just in the last year that they started to be routinely used, making available huge quantities of data, which require suitable statistical methods to extract valuable information. For instance, each action can now be described by a large number of Key Performance Indicators (KPIs) based on the exact positions of the the players of both teams during the entire action. When the interest is focused on a single aspect of the football action, the proliferation of KPIs represent an undoubted advantage, since it is likely that a suitable KPI addresses such specific aspect exists. On the other hand, coaching teams are more often interested in detecting general traits of the actions and to evaluate the impact of strategic decisions on the match outcome. Recognizing and isolating such traits could be challenging.

Therefore, we propose to apply a Bayesian factor model to summarize the large amount of information contained in the KPIs in a lower dimensional set of meaningful aggregated indicators. Factor models are of particular interest to shed light on the underlying covariance structure among the KPIs, but also they can be particularly suitable when we want to predict a variable of interested through the KPIs, by performing a variable selection approach replacing the original very many predictors with the low-dimensional latent factors [9]. In the football context we may be interested in modelling the dangerousness y_i of the action i through a regression on a set of p KPIs x_i. Then, we reduce covariates dimensionality by considering the Gaussian linear factor model for the $n \times p$ covariate matrix x

$$x_i = \Lambda \eta_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \Sigma), \qquad i = 1, \dots, n, \tag{1}$$

with Λ a $p \times k$ loadings matrix, η_i a k dimensional factor, and ε_i a p-dimensional error term with covariance matrix $\Sigma = I_p(\sigma_1^2, \dots, \sigma_p^2)^\top$. The usual linear regression model for the n-variate response vector y on the latent covariates is

$$y_i = b_0 + \eta_i^\top b + \nu_i, \qquad \nu_i \sim N(0, \sigma_\nu^2), \qquad i = 1, \dots, n, \tag{2}$$

where b_0 is the intercept coefficient and b is the k-variate coefficient vector, with $k \ll p$. Let δ denote a p-variate coefficient vector such that $b = \Lambda \delta$, the model above can be re-written as the usal p-variate regression

$$y_i = b_0 + x_i^\top \delta + \nu_i^*, \qquad \nu_i^* = \nu_i - \varepsilon_i \delta \qquad i = 1, \dots, n.$$

Notice that model (2) is equivalent to model (3) where regularization on δ is applied through k linear constraints determined by the columns of Λ. A sparsity pattern on Λ implies that the linear constraints act only on subsets, possibly overlapped, of elements of δ. This can be seen as a group penalty, which is a quite common approach in Bayesian literature, with the grouped Lasso [10] providing a notable example. Nonetheless, inducing the group penalty through a carefully specified sparse factorization of the covariate matrix could present several benefits. Firstly, the k latent factors can be interpreted as latent covariates that summarize the information cointained in the observed covariates. In our case, this means the construction of a new

set of k more informative KPIs which can be directly analysed by the coaching teams. Secondly, the constraints definition is flexible, specially if we rely on the recent literature about over-fitted factorization models. Such approaches, introduced by [2], assume an increasing shrinkage prior on infinite columns of the loadings matrix Λ, allowing to adaptively truncate the model and choose the number of latent components—i.e. the number k of linear constraints on δ–as well as the weigths of such components without imposing any fixed structure. In this literature, the novel class of generalized infinite factorization priors proposed by [8] is designed to include additional information about the columns of x to help the identification of sparsity pattern on Λ, implying a coherent group penalty on δ. For example, we could consider how each KPI is constructed to induce the k linear constraints on groups of coefficients of similar KPIs. Motivated by the recent advances both in data collection technology and in Bayesian factor analysis, in this article we apply a structured increasing shrinkage factorization to model the covariate matrix constituted by tracking KPIs and then to perform regularized regression of an action dangerous index.

The article is organized as follows: Sect. 2 presents the model; in Sect. 3 we apply the methodology to the football dataset. Finally, in Sect. 4 a discussion of the results and the of the method limitations is provided.

2 The structured increasing shrinkage prior for factor models

Consider the notation introduced in the previous section. Let 1_s and 0_s denote the s-variate all-ones vector and null vector, respectively. The model of the data is

$$[x,y] \sim N(1_n[0_p^\top, b_0] + \eta[\Lambda^\top, b], \Sigma^*), \qquad \Sigma^* = I_{p+1}(\sigma_1^2, \ldots \sigma_p^2, \sigma_v^2)^\top. \tag{3}$$

Following common practice in the Bayesian factor analysis literature [2, 3, 8], we avoid imposing identifiability constraints on Λ and induce a class of scale-mixture of Gaussian shrinkage priors [5] for the loadings, specifying

$$\lambda_{jh} \mid \theta_{jh} \sim N(0, \gamma_h \phi_{jh}), \qquad h = 1, \ldots, \infty \tag{4}$$

where the column-specific γ_h and the local ϕ_{jh} scales are all independent *a priori*. In particular we specify the structured increasing shrinkage (SIS) prior proposed in [8], which induces the increasing shrinkage behaviour through the prior on γ_h

$$\gamma_h = \vartheta_h \rho_h, \qquad \vartheta_h^{-1} \sim \mathrm{Ga}(a_\theta, b_\theta), \quad a_\theta > 1, \qquad \rho_h = \mathrm{Ber}\,(1-\pi_h)\,,$$

where $\mathrm{Ber}(\pi)$ denote the Bernoulli distribution with mean π and $\mathrm{Ga}(a,b)$ denote the gamma distribution with mean a/b and variance a/b^2. The parameter $\pi_h = \mathrm{pr}(\gamma_h = 0)$ follows a stick-breaking construction similarly to [3],

$$\pi_h = \sum_{l=1}^{h} w_l, \qquad w_l = v_l \prod_{m=1}^{l-1} (1-v_m), \qquad v_m \sim \mathrm{Be}(1,\alpha),$$

with $\text{Be}(a,b)$ indicating the beta distribution with mean $a/(a+b)$, such that $\pi_{h+1} > \pi_h$ is guaranteed for any $h = 1,\ldots,\infty$ and $\lim_{h\to\infty}\pi_h = 1$ almost surely.

Differently from most of the existing literature on shrinkage priors, SIS defines a non-exchangeable structure that includes *meta* covariates z informing the sparsity structure of Λ. Meta covariates provide information to distinguish the p different covariates. As a simple example regarding the football application discussed above, meta covariates could inform on which team—i.e. attacking or defending team—each KPI refers to or if the KPI is based on spatial or temporal measurements. Letting z denote a $p \times q$ matrix of such meta covariates, each local scale ($j = 1,\ldots,p$; $h = 1,\ldots,\infty$) is specified as

$$\phi_{jh} \mid \beta_h \sim \text{Ber}\{\text{logit}^{-1}(z_j^\top \beta_h)\, 2e\log(p)/p\}, \qquad \beta_h \sim N_q(0, \sigma_\beta^2 I_q),$$

where $\text{logit}^{-1}(u) = e^u/(1+e^u)$. Equation (5) plays a key role in promoting posterior samples of the loadings matrix characterized by recognizable sparsity pattern associated to the meta covariates, helping the interpretation of the latent factors. The intuition behind is based on expecting that meta covariates inform on KPIs similarities that can be expressed in terms of high or low loadings on the same latent factors. Note that such structure is allowed while not imposed.

The prior specification is completed assuming usual conjugate priors in regression models, namely $\sigma_v^{-2} \sim \text{Ga}(a_v, b_v)$, $\sigma_j^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$ ($j = 1,\ldots,p$), and $b_h \sim N(0,\sigma_b^2)$ ($h = 0,1,\ldots,\infty$).

Posterior inference is conducted via Markov chain Monte Carlo sampling. Consistently with infinite factorization literature, we use an adaptive Gibbs algorithm, which attempts to infer the number of latent factors k while it runs. Non-identifiability of the latent structure creates problems in interpretation of the results from Markov chain Monte Carlo samples and therefore we summarize $[\Lambda^\top, b]$ and β through $[\Lambda^\top, b]^{(t^*)}$ and $\beta^{(t^*)}$ sampled at the iteration t^*, characterized by the highest marginal posterior density function. Computational details follow [8].

3 Application and results

We consider a dataset provided by MathAndSport[1] composed by $n = 125$ independent actions of three matches of a professional European league. The covariate matrix, opportunely standardized, include $p = 21$ KPIs such as the length and width of the teams during the action, the distance run, the number of players involved and other spatial, temporal and tactical metrics. The response variable is the dangerous index computed by MathAndSport as a weighted estimate of the maximum probability to score during the action, assuming values in $(0,1)$. Meta covariates include two categorical variables on KPIs characteristics. The first one indicates if each KPI is referred to the attacking, to the defefending team or to both, while the second meta covariate classifies the KPIs according to the type of measure they consider: spatial measurements, physical performances, or possession choices. This leads to

[1] MathAndSport s.r.l. is a sport analytics company based in Milan: www.mathandsport.com

$q = 5$. Consistently with [8], we fix the hyperparameters $a_\sigma = 1$, $b_\sigma = 0.3$, $\sigma^2_\beta = 1$, and $a_\theta = b_\theta = 2$. We set $a_\nu = 1$, $b_\nu = 2$, $\sigma^2_b = 1$ and $\alpha = 4$. Then, we run the algorithm for 15000 iterations after a burnin of 10000 iterations and we thin the Markov Chain, discarding all but every third sampled parameters. We verified satisfying convergence and low autocorrelation in the sampled parameters. To evaluate the model in terms of predictions, we consider y as missing values in a subset of $n_v = 25$ randomly sampled actions. Their root mean squared error (RMSE) with respect to the the predictive posterior mean is 0.1637.

In addition to the possible advantages in terms of prediction provided by regularized regression, the proposed structured prior helps in interpreting the relations among the large set of KPIs and the response variable. The estimate of k^a

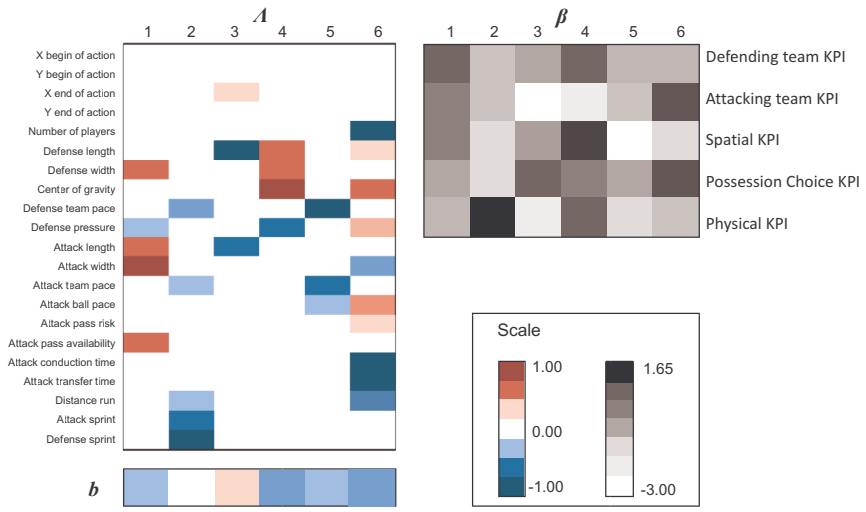


Fig. 1 Posterior summaries $[\Lambda^\top, b]^{(t^*)}$ and $\beta^{(t^*)}$ of the SIS regression model, where the rows of the left matrix refer to the 21 KPIs considered, while the rows of the right matrix refer to the five meta covariates. Light colored cells of $\beta^{(t^*)}$ induce shrinkage on corresponding cells of $\Lambda^{(t^*)}$.

strongly suggests six main factors, whose impact can be illustrated by the estimates of $[\Lambda^\top, b]^\top$ and meta covariate coefficients β reported in Fig. 1. The loadings matrix is quite sparse, indicating that each latent factor impacts a small group of KPIs. Lower elements of $\beta^{(t^*)}$, represented with light cells on the right panel, induce higher shrinkage on the group of KPIs described by the corresponding meta covariate. The KPIs influenced by the first factor are fairly homogeneous, all related to the amount of spaces between defenders. The more spaces there are during the action, the more is likely that the action has not developed dangerous situations, when usually defenders collapse and attackers have few possible choice to catch the oppurtunity. The high level of $\beta^{(t^*)}_{52}$ suggests that KPIs which measure physical performance tend to have loadings different from zero for the second factor and an

overall very low influce on the dangerousness of the action, suggesting that increasing physical capacity of the players can impact the probability to score only when they produce differences and advantages in strategic and technical aspects. Focusing on the most important factors in terms of explaining dangerousness, we note that high levels of both the fourth and the sixth factors decrease the probability of scoring, even if they represent distinct aspects of the action, as it is clearly visible from the fourth and sixth columns of the meta covariates coefficient matrix. The KPIs influenced by the fourth factor are mostly related to the defenders attitude during the action: as expected, we observe low, narrow, and high pressure defense when the ball is in dangerous areas and close to the goal. The last factor describes the attacking strategy providing the most interesting insights. Long actions that involves a lot of players well distributed along the width of the attacking pitch are generally more dangerous than other actions. Surprisingly, the loadings $\lambda_{14\cdot 6}^{(t^*)}$ and $\lambda_{15\cdot 6}^{(t^*)}$ indicate that risky and fast passes are generally not worth, since they are not rewarded in terms of scoring probability. Switching the signs of $[\Lambda^\top, b]^\top$ columns does not change considerations above.

4 Discussion

The paper shows the potentiality of the generalized infinite factorization framework also in the regression context to induce flexible and interpretable coefficient regularization. Although an extended comparison with alternative regularized regression models on a bigger dataset is needed, the football insights obtained display that structured incresing shrinkage prior can help in construct meangiful relations between tangible variables even in a new and complex context as that provided by football tracking data.

Acknowledgements The authors are grateful to MathAndSport s.r.l. for providing the data.

References

1. Bhattacharya, A., Dunson, D.B.: Sparse Bayesian infinite factor models. Biometrika. **98**, 291–306 (2011)
2. Legramanti, S., Durante, D., Dunson, D. B.: Bayesian cumulative shrinkage for infinite factorizations. Biometrika. **107**, 745–752 (2020)
3. Liu, J., Tong, X., Li, W., Wang, T., Zhang, Y., Wang, H.: Automatic player detection, labeling and tracking in broadcast soccer video. Pattern Recognit. Lett. **30**, 103–113 (2009)
4. Polson, N.G., Scott, J.G.: Shrink globally, act locally: Bayesian sparsity and regularization. Bayesian Statistics. **9**, 1–16 (2010)
5. Roberts, G. O., Rosenthal, J. S.: Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. J. Appl. Probab. **44**, 458–475 (2007)
6. Schiavon, L., Canale, A., Dunson, D.B.: Generalized infinite factorization models. arXiv preprint arXiv:. (2021)
7. West, M.: Bayesian factor regression models in the large p, small n paradigm. Bayesian stat. **7**, 733–742 (2003)
8. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B. Stat. Methodol. **68**, 49–67. (2006)

A dynamic matrix-variate model for clustering time series with multiple sources of variation

Un modello dinamico matrix-variate per il clustering di serie storiche con molteplici fonti di variazione

Mattia Stival

Abstract In this paper we present a dynamic matrix-variate model for clustering online several multivariate time series. The specific matrix state-space structure of the model allows for the inclusion of multiple sources of variation, such as those due to dynamics typical of time series, or the presence of time-dependent regressors. The use of matrix-variate normal distributions for the error terms and the introduction of an unknown selection matrix among the model matrices for clustering ensure flexibility, interpretability and usability in many contexts, while keeping the number of parameters small. We outline the estimation method and analyze, as an example, the training activities of one athlete using biometric data collected with a sport-watch.

Abstract *In questo lavoro presentiamo un modello dinamico matrix-variate per il clustering in tempo reale di numerose serie storiche multivariate. La struttura state-space matriciale del modello permette di includere molteplici fonti di variazione, come ad esempio quelle dovute a dinamiche temporali tipiche delle serie storiche, oppure alla presenza di regressori tempo dipendenti. L'utilizzo di distribuzioni normali matrici-variate per i termini di errore e l'introduzione di una matrice di selezione ignota per il clustering tra le matrici del modello, garantiscono flessibilità, interpretabilità ed utilizzabilità in numerosi contesti, mantenendo ridotto il numero di parametri. Deliniamo il metodo di stima e analizziamo, come esempio, le attività di allenamento di un atleta tramite i dati biometrici raccolti con uno sport-watch.*

Key words: matrix time series, sport analytics, time series clustering

Mattia Stival
Department of Statistical Sciences, University of Padova, Via C. Battisti, 241 - 35121 Padova e-mail: mattia.stival@phd.unipd.it

1 Introduction

Thanks to new technologies, the possibility of collecting a huge amount of data is offering new stimuli for the development of statistics. Generally, the need for easy-to-use and interpretable tools clashes with data characterized by multiple types of complexity: on the one hand, data are increasingly numerous, posing challenges to the scalability of the tools, on the other hand, they present a complex structure, due both to the nature of the data itself, and to the fact that the data collected often come from non-homogeneous sources. In this context we insert our work, which proposes a dynamic matrix model for clustering multivariate time series, as a new alternative to the models reviewed in [8], allowing in a parsimonious way for the need of flexibility and interpretability of many recent applied contexts. In the time series literature, there is growing interest in analyzing time series beyond the typical vector form of the observations, and examples of models that consider a time dependence between matrices are numerous in different contexts, such us engineering, medicine, finance and econometrics (see, e.g [4, 19, 18, 2, 3]). The particular state space matrix form that we will give to the model, places it among the models for the analysis of matrix-variate time series, as a general case that can include most of the cited contributions. We emphasize the fact that time series of matrices can emerge in two different situations. In the first case, the observation at time t is a matrix by nature of the data, as in the case of images, while in the second case the observation at time t should be thought as a mathematical expedient to deal compactly and parsimoniously with elements whose nature is not that, such as the case of the vectors of observations of different economic indexes for different nations, which are stacked in a matrix for convenience.

To provide an example of our model, we propose the analysis of an athlete's continuous running activities, using geophysical and bio-metrical data collected by his sport-watch during a training period, as a new development of the model proposed by [16] and [17]. The use of GPS-enabled tracking devices and heart rate monitors is common in several disciplines, and the reason for such interest relies on the primary need to improve the knowledge and individualize the design of training activities and exercise programs in order to maximize the improvements, and avoid over-training, which may lead to impaired health, and typically under-performance [1]. In this context, data are collected as a sequence of N activities, where each activity is represented by a high frequency multivariate time series collecting P different response variables, such as speed, heart rate, cadence, and R covariates, such as GPS position and altitude. The most valuable statistical contribution in this field has been provided by [7], for an advanced retrospective statistical analysis of these kind of data using the R statistical software. The main novelty of our contribution relies in the fact that the model is designed for real-time analysis of time series of this type. In fact, we think that having real-time knowledge of how an athlete is performing is critically important, with the idea that training in a sub-optimal condition can greatly affect his or her health status.

2 The model

Let Y_t be the $P \times N$ matrix of observations storing, in the n–th column, the P–dimensional vector of observed variables for the n–th time series at time t, for $n = 1, \ldots, N$, and $t = 1, \ldots, T$. We assume that Y_t can be described by the following dynamic model

$$Y_t = \sum_{m=1}^{M} Z_{m,t} A_{m,t} S_{m,t}^{\top} + \Upsilon_t, \quad \Upsilon_t \sim MN_{P,N}(0, \Sigma^R \otimes \Sigma^C), \tag{1}$$

$$A_{m,t+1} = T_{m,t} A_{m,t} U_{m,t}^{\top} + \Xi_{m,t}, \quad \Xi_{m,t} \sim MN_{Q_m,G_m}(0, \Psi_m^R \otimes \Psi_m^C), \tag{2}$$

for $A_{m,1} \sim MN_{Q,G}(\hat{A}_{m,1|0}, P_{m,1|0}^R \otimes P_{m,1|0}^C)$, $m = 1, \ldots, M$, $t = 1, \ldots, T$.

In the above specification, the measurement equation in Eq. 1 links the matrix of observations Y_t to the latent states $A_{m,t}$ ($m = 1, \ldots, M$), which follow a matrix autoregressive process of order 1, described by the state equation in Eq. 2. Both the equations involve a left and right multiplication of the latent states by left structural matrices, $Z_{m,t}(P \times Q_m)$ and $T_{m,t}(Q_m \times Q_m)$ and right structural matrices, $S_{m,t}(N \times G_m)$ and $U_{m,t}(G_m \times G_m)$ for all $t = 1, \ldots, T$, and $m = 1, \ldots, M$, which may depend on the R covariates. Moreover, the error terms Υ_t and disturbance terms $\Xi_{m,t}$ follow independent and serially uncorrelated matrix-variate normal distributions, with covariance matrices that can be decomposed by a Kronecker product (see, e.g. [10]), a practical choice that allows us to reduce drastically the number of parameters involved in the model, preserving an interpretable row-column dependence structure (see,e.g [13]).

Let

$$y_t = vec(Y_t), \quad \upsilon_t = vec(\Upsilon_t),$$

$$\alpha_t = (vec(A_{1,t})^{\top}, \ldots, vec(A_{M,t})^{\top})^{\top}, \quad \xi_t = (vec(\Xi_{1,t})^{\top}, \ldots, vec(\Xi_{M,t})^{\top})^{\top},$$

$$Z_t^{(vec)} = [S_{1,t} \otimes Z_{1,t} | \ldots | S_{M,t} \otimes Z_{M,t}] \quad T_t^{vec} = blkdiag(U_{1,t} \otimes T_{1,t}, \ldots, U_{M,t} \otimes T_{M,t}),$$

$$\Sigma = (\Sigma^C \otimes \Sigma^R), \quad \Psi = blkdiag(\Psi_1^C \otimes \Psi_1^R, \ldots, \Psi_M^C \otimes \Psi_M^R).$$

The model can be represented in a vector state space form

$$y_t = Z_t^{(vec)} \alpha_t + \upsilon_t, \quad \upsilon_t \sim MVN_{PN}(0, \Sigma) \tag{3}$$

$$\alpha_t = T_t^{(vec)} \alpha_t + \xi_t \quad \xi_t \sim MVN_{Q(G_1+\ldots+G_M)}(0, \Psi) \tag{4}$$

for $\alpha_1 \sim MVN_{Q(G_1+\ldots+G_M)}(\hat{\alpha}_{1|0}, P_{1|0})$. If $Z_t^{(vec)}$, $T_t^{(vec)}$, Σ, and Ψ are fixed, we recognize a linear and Gaussian state space model (see, e.g [6]).

Under this setting, we assume $S_{m,t} = X_{m,t} S$, for all $m = 1, \ldots, M$. Here, $X_{m,t} = diag(x_{m,1,t} \ldots, x_{m,N,t})$ is an $N \times N$ diagonal matrix of covariates for the N time series, and S is an $N \times G$ unknown selection matrix with n-th row that takes value $I(S_n = g) = 1$ in its g-th column if the n-th time series belongs to the g-th group, with the role of linking the group-specific dynamics of the states to the covariates of

the N time series. We adopt a fully conjugate Bayesian approach. We note that $\Sigma^R \otimes \Sigma^C = c\Sigma^R \otimes \Sigma^C/c$ for any $c > 0$, leading to a potential identifiability issue of the model. We deal with this issue by adopting the prior proposed by [14] for a similar problem with the multinomial probit model, setting to 1 the 1-1 elements of all all the row-covariance matrices (those with R in superscript). A Multinomial-Dirichlet prior is used for the rows of S.

2.1 Posterior simulation

Although a Gibbs sampler is available, we propose to simulate from the posterior distribution via a Metropolis-Hasting within Gibbs procedure, which is possible since it is straightforward to obtain the full conditional distributions of the parameters. The algorithm iterates the following steps: 1) simulation of the states; 2) simulation of covariance matrices; 3) update of cluster allocations and weigths; 4) update of structural matrix parameters. We briefly discuss how we tackle step 1 and 3, highlighting some interesting feature of the model. Given S and the model parameters, state simulation can be obtained without loss of information with the Simulation Smoothing algorithm by [6], after transforming the measurement equation on a reduced form, thanks to the reduction by transformation technique by [12]. Large computational savings are possible if $rank(Z_t^{(vec)}) \ll NP$, as in standard cases where with a number of groups lower than the number of time series, and a moderate number of covariates. Differently from standard mixture models, in which allocation are drawn one at a time, independently of each others, we propose to draw in one-shot the selection matrix S with a Metropolis-Hasting step, by combining the moves that do not change the number of groups of the Allocation Sampler by [15], and a random walk proposal on the space of selection matrices. The reason of this choice resides into two facts: firstly, both Ψ_m^C and Σ^C may be not diagonal matrices, leading to states of different group which are apriori dependent, and time series that, conditional to the states, are still dependent; secondly, the latent states can be easily integrated out by the use of a Kalman filter routine, leading to chain with better mixing properties. Updating the entire matrix S one row at a time would require the use of $N(G-1)$ Kalman filter routines, an unfeasible step for large N and long time series (large T). The evaluation of the proposed matrix requires instead the use of just 2 Kalman filter routines.

3 Conclusion

In the application we cluster $N = 90$ continuous running activities into $G = 3$ groups, for which we use the variable Heart Rate (bpm), Speed (m/s), and Cadence (spm) as response variables measured over time, and variation in Altitude (m/s) as time dependent explanatory variable. Activities belonging to the same group share both the

trends which describes the global dynamics of the performance, and the time varying coefficients which measure the effect of one meter variation in altitude during the activities, described by 3 stationary around the mean (estimated) AR(1) processes. The results are compatible with those presented in [17], with the effects of variation in Altitude slightly shrunk toward zero. In Figure 1 we present the results in terms of the dissimilarity matrix $D = \frac{1}{B}\sum_{b=1}^{B} S^{(b)}S^{(b)^\top}$, where $S^{(b)}$ denotes the b-th draw selection matrix S, after a pre-selected burn-in period. An intense yellow cell indicates activities that under different MCMC draws are allocated in the same cluster. As the $N = 90$ activities are sequentially ordered, the presence of yellow squares determined by contiguous activities highlights the presence of temporal dependence (a posteriori) between cluster allocations, a relevant aspect that suggests that temporally close activities are more likely to be allocated in the same cluster.

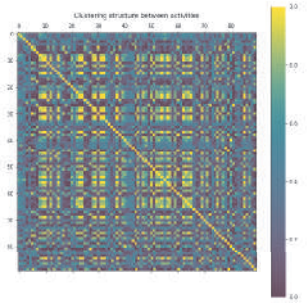


Fig. 1 Clustering structure between activities expresses in terms of the dissimilarity matrix $D = \frac{1}{B}\sum_{b=1}^{B} S^{(b)}S^{(b)^\top}$, where $S^{(b)}$ denotes the b-th draw selection matrix S, after pre-selected burn-in period. The $N = 90$ activities are sequentially ordered.

The proposed model involves the use of an unknown selection matrix that determines the clustering of the time series and whose dimensions are given by the number of time series N, and the number of groups G. If G is unknown, its selection can be done in different ways. On the one side, one can use information criteria, such as DIC and some of its variants. On the other hand, it is possible to put an a priori distribution on the unknown number of groups G, i.e. on the number of columns of S. The use of reversible jump techniques are extremely hard, as well as practical tries with the Allocation sampler by [15] shows the difficulties of the moves that change the number of group. To avoid these difficulties, one may combine the moves that not change the number of group, which seems to work well also in high dimensional setting, with the Telescoping Sampling by [9]. From the modelling perspective, we may also consider a different selection matrices S_m for each of the additive terms in the measurement equation. Finally, as the interest resides on the real time use on these data, we may consider for further developments the development of an online learning methodology, such as those related to online EM and online variational ap-

proximation algorithms.

Funding: This research was supported by funding from the University of Padova Research Grant 2019-2020, under grant agreement BIRD203991.

References

1. Cardinale, M. and Varley, M.C. (2017). Wearable training-monitoring technology: applications, challenges, and opportunities. *International Journal of Sports Physiology and Performance*, 12(s2), 55 – 62.
2. Chen, E. Y., Tsay, R. S., and Chen, R. (2020a). Constrained factor models for high-dimensional matrix-variate time series. *J. Amer. Statist. Assoc.*, 115(530):775–793.
3. Chen, R., Xiao, H., and Yang, D. (2020b). Autoregressive models for matrix-valued time series. *Journal of Econometrics*.
4. Choukroun, D., Weiss, H., Bar-Itzhack, I. Y., and Oshman, Y. (2006). Kalman filtering for matrix estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 42(1):147–159.
5. Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–615.
6. Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, second edition.
7. Frick, H. and Kosmidis, I. (2017) trackeR: Infrastructure for Running and Cycling Data from GPS-Enabled Tracking Devices in R. *Journal of Statistical Software*, 82(7), 1 – 29.
8. Frühwirth-Schnatter, Sylvia. (2011). Panel data analysis: a survey on model-based clustering of time series. *Advances in Data Analysis and Classification*, 5(4):251–80
9. Frühwirth-Schnatter, S., Malsiner-Walli, G., Grün, B. (2020) Generalized mixtures of finite mixtures and telescoping sampling. *arXiv*, 2005.09918
10. Gupta, A. K. and Nagar, D. K. (2000). *Matrix variate distributions*, volume 104 of *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*. Chapman & Hall/CRC, Boca Raton, FL.
11. Huang, L., Bai, J., Ivanescu, A., Harris, T., Maurer, M., Green, P., and Zipunnikov, V. (2019). Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations. *J. Amer. Statist. Assoc.*, 114(526):553–564.
12. Jungbacker, B. and Koopman, S. J. (2008). Likelihood-based analysis for dynamic factor models. Technical report, Tinbergen Institute Discussion Paper.
13. Lovison, G. (2006). A matrix-valued Bernoulli distribution. *J. Multivariate Anal.*, 97(7):1573–1585.
14. McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173 – 193.
15. Nobile, A. and Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Stat. Comput.*, 17(2):147–162.
16. Stival, M. and Bernardi, M. (2019) Dynamic Bayesian clustering of running activities. *In: Smart Statistics for Smart Applications.*, Milano:Pearson, ISBN: 9788891915108.
17. Stival, M. and Bernardi, M. (2020) Dynamic Bayesian clustering of sport activities. *In: Proceedings of the 35 th International Workshop on Statistical Modelling.*, CIP. Biblioteca Universitaria. ISBN: 9788413192673.
18. Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of econometrics*, 208(1):231–248.
19. Wang, H. and West, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika*, 96(4):821–834.

Evaluating football players' performances using on-the-ball data

Valutazione della performance dei giocatori di calcio utilizzando on-the-ball data

David Dandolo

Abstract We introduce a model-based approach to estimate the probability that a particular action performed by a football player is leading to scoring a goal in the immediate next actions. After the construction of an appropriate model, and given the probabilities for the event of interest, we build an overall index summarizing the offensive impact of each player, leading to a completely data-driven approach to performances evaluation.

Abstract *In questo paper introduciamo un modello per stimare la probabilità che una particolare azione condotta da un giocatore di calcio abbia come esito un goal in una azione immediatamente vicina. Dopo aver costruito un modello appropriato e calcolate le probabilità degli eventi di interesse, costruiamo un indice in grado di sintetizzare l'impatto offensivo di ciascun giocatore.*

Key words: Sport analytics, evaluation of players performances, Bayesian inference.

1 Introduction

Data analysis is becoming a crucial aspect in a lot of contexts, including the sports analysis. The bookmakers use predictive models to calculate better odds, but also managers and coaches make extensively use of statistical methods and softwares to gain information about the performance of the teams. The evolution of the computing capacity gives now the possibility to study a lot of more data to study new approaches, in particular in the soccer context, regarding the players evaluation. The actual state of art includes use of descriptive measurements (number of shots, percentage of shots on goal, cross or pass completed, and others). In addition, others studies are based on the Expected Goals measure [2], and just a few more recent works try to includes in their analysis more information about what happen in the pitch, for example [3]. An interesting approach is that of [4], which evaluate the

Table 1 Description of the available dataset.

#	Variable name	Variable type	Variable description
Identifiers of team, players and match			
1	Id	numerical	
2	playerId	numerical	identifier of the player performing the action
3	teamId	numerical	team in which the player performing the action plays
4	matchId	numerical	identifier of the match
5	matchPeriod	categorical	match period: in our data it indicates the first or second half more generally it can also indicate any extra time or penalty kicks
Event attributes			
6	eventSec	numerical	time in seconds elapsed since the start of the game period
7	eventId	numerical	identifier of the event
8	eventName	textual	textual description of the event
9	subEventId	numerical	identifier of the sub-event
10	subEventName	textual	textual description of the sub-event
11-18	tagId1,...,tagId8	textual	8 variables indicating the identifier of the Tag
19-27	tagsLabel1,...,tagsLabel8	textual	8 variables indicating additional description of the event
Coordinates			
28	x1	textual	coordinates of the point where the action begins, i.e. the starting point
29	y1	textual	of the ball, indicated as a percentage distance from the left corner of the team door of the player performing the action
30	x2	textual	coordinates of the point where the action ends, i.e. the point of arrival
31	y2	textual	of the ball, indicated as a percentage distance from the left corner of the door of the team of the player performing the action

impact of an action performed by a player as the difference of the probability, estimated by SVM, to realise a goal before and after that action. Instead, in this paper we introduce a model-based approach having the scope of estimating the probability that a particular action performed by a player is leading to scoring a goal in the immediate next actions. After the construction of an appropriate model, knowing the probabilities of interest, we build an index summarizing the offensive impact of each player, leading to a completely data-driven approach to performances evaluation.

2 The data

The data used in this paper are provided by WyScout, the so called on the ball data. On the ball data are collected in such a way that a ball-event is recorded every time one of the football players make a play on the ball. The dataset at our disposal has 1520 matches of the top Italian Soccer League (Serie A), concerning the soccer seasons 2014-15, 2015-16, 2016-17, 2017-18, with more than 2 millions of events, in which are engaged 1111 football players of 27 different teams. The dataset has been organised so that, every match is described by n-tuples composed by 27 variables as described in Table 1.

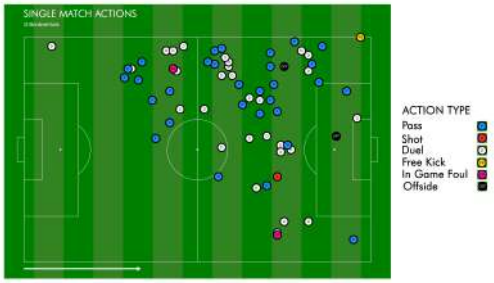


Fig. 1 Actions performed by Giacomo Bonaventura in the football match Empoli versus Milan on September 23, 2014. The white arrow in the left-bottom corner of the figure indicates the attack direction for the analyzed player.

3 Building the design matrix

The dataset provided by WyScout is not directly exploitable "as it is" for the purpose of building a statistical model aiming to evaluate the football players performances. Therefore, an extensive pre-processing of the data for finding and eliminating errors have been done. First, data incongruences such as ID of Event, subEvent or Tag (see Table 1) mismatched with the textual description or events positions not in the correct ranges or again events associated to players who could not have made them have been removed from the dataset. Then, the events regarding the goalkeepers and referee interventions, which are not suitable for our purpose, have been removed form the dataset.

To exploit the information concerning Event, subEvent and Tag, we construct a $n \times p$ design matrix, in which every column represents one of the possible values of the variables listed in Table 1, with value equal to 1 if that particular event is happened, and 0 otherwise. Again, for the purpose of our analysis, the information regarding the event "Foul" could be summarized in two macro tags: "In game Foul", which contains all the fouls that interrupt the game and "Out game Foul", that otherwise contains all the fouls that happened when the game was already stopped. To make the events "Others" on the ball less general, we left value 1 only to the subEvent related cell (when the subEvent information was present) and removed it from the major event, thereby becoming a container for the event for which we do not have enough information.

Another relevant aspect concerns the different roles assigned to the football players. Indeed, in a football team there are different roles for the players, and it would be unfair to evaluate the performance of all the players playing different roles with the same rule, and compare players having offensive roles with those having defensive roles. Therefore, we opt for performing a clustering of the players' starting event positions (x1,y1) using the k-means algorithm. The results provided in Figure 2 show that using 7 clusters we have some interpretable centroid coordinates. Concerning again the players' starting event positions, it is worth noting that, while the event position coordinates are useful for visual representation of player actions in a match (see Figure 1), they are not so informative from the perspective of building a statistical model for assessing the players performances. Indeed, to get a represen-

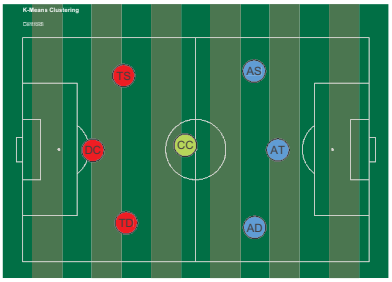


Fig. 2 Centroids identified by the k-means algorithm to assign the player' role. "ST" stands for "striker", "LW" and "RW" stand for "left" and "right" wing, "MD" stands for "midfilder", "LB"and "RB" stands for "left" or "right" back, and "CD" stands for "central defender".

tation of the position accounting for the distance from the opponent's goal line and angle, we propose to transform the position into polar coordinates with reference to the goal-line, see Figure 3. All the information about the event position is now included in the variables angle (α) and distance from the goal-line d.

The last step of our data-building procedure is to aggregate the variables in order to match those present in the original dataset: playerId, teamId, matchId, MatchPeriod, eventSec. Also, for simplicity we replace the numerical values for the playerId and teamId with their names.

Because of in soccer the aim is to score more goal then the opponents, inspiring by [4], we decided to create a dichotomous response variable, with value 1 when the action t of a team h is followed in max k actions by a goal in favour of team h, and 0 otherwise. In this way, we reward all the actions that participates in the network of events leading to the achievement of the goal. Therefore, the vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ where y_i can be interpreted as the realization of a Binomial random variable $Y_i \sim \mathsf{Bin}(1, \pi_i)$. The choice of the value for k is not immediate. Indeed, a value too high would in fact reward even trivial or unimportant actions, temporally distant from the moment when the scoring opportunity was actually created, while a low value would only reward those actions who finalize the goal. Another aspect to be considered is that players in the opponents team could play a relevant role in the current action, but this aspect does not necessarily mean that the events happened till that moment are no longer relevant. As final remark we observe that in an entire Serie A championship, there are about 1000 goals while the average of events on the ball per game is 1600. This leads to an unbalanced dataset, with a lot of 0 and very few 1. Therefore a bigger value of k could help to avoid the consequences of zero-inflation. Keeping in mind all previous considerations we opt for setting $k = 15$. The response variable y_t can be therefore interpreted as follows: "within the next k events, the team that has the possession of the ball in the action t will score a goal".

4 Measuring players' performances

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be a random sample of n binary observations for a given player and let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\intercal$ be the set of p covariates associated to each

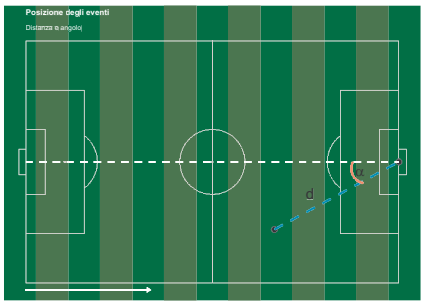


Fig. 3 Representation of the coordinates of the event.

sample unit $i = 1, 2, \ldots, n$, we consider the following Binary regression model:

$$y_i \sim \mathsf{Ber}(1, \psi_i) \tag{1}$$

$$\psi_i = \mathrm{F_{Lo}}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \tag{2}$$

$$\eta_i = \alpha + \mathbf{x}_i^{\mathsf{T}} \beta, \qquad i = 1, 2, \ldots, n, \tag{3}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathsf{T}}$ is the $p \times 1$-vector of the unknown regression parameters, $\alpha \in \mathbb{R}$ is the constant, $\eta_i = \log \frac{\psi_i}{1+\psi_i}$ is the log-odds ratio and $\mathrm{F_{Lo}}(\cdot)$ denotes the logistic-link function. The Bayesian inferential procedure requires the specification of the prior distribution for the unknown vector of parameters $(\alpha, \beta^{\mathsf{T}})^{\mathsf{T}}$. In principle non informative priors can be specified for the vector of regression parameters, i.e., $\pi(\alpha, \beta) \propto 1$. Alternatively, the usual Gaussian prior can be specified for regression parameters, i.e., $\pi(\alpha, \beta) = \pi(\alpha)\pi(\beta)$ with $\pi(\alpha) \sim \mathsf{N}(\mu_\alpha, \sigma_\alpha^2)$ and $\pi(\beta) \sim \mathsf{N}\left(\mu_\beta, \Sigma_\beta\right)$. The Bayesian inference is performed via the Gibbs sampling algorithm leveraging the data augmentation approach of [1], that relies on the Pólya-Gamma representation of the logistic-link function.

Once we get the estimate of $\widehat{\beta}_r \; \forall r$, we calculate the probability of scoring a goal. That probability measures the relevance of that particular event for scoring a goal. However, to get the players ranking, it is important to build an index that summarizes the offensive impact of player i. The simplest method consists to extract, for each subject i, the sub-matrix X_i consisting of all the actions he performed, and calculate the probability $\widehat{\pi}_{i,t}$ for each action t, with $t = 1, \ldots, T_i$, where T_i denotes the total number of actions performed by the i-th player. Also, to account for the temporal evolution of the performances we leverage an exponential weighted moving average (EWMA) approach to update the score after the end of each match. During the g-th match, the i-th player performs $T_{i,g}$ actions, and the rating is estimated as:

$$z_{i,1} = \frac{1}{T_{i,1}} \sum_{t=1}^{T_{i,1}} \widehat{\pi}_{i,1,t}, \quad z_{i,g} = (1-\lambda) z_{i,g-1} + \lambda \left(\frac{1}{T_{i,g}} \sum_{t=1}^{T_{i,g}} \widehat{\pi}_{i,g,t} \right), \tag{4}$$

for $g = 2, \ldots, G_i$, where G_i indicates the total number of games in which the i-th player has played, and λ denotes the tuning parameter that regulates the importance of past player's performances. To include the information about the number of goals scored by a player, we again exploited an EMWA approach:

$$\tilde{z}_{i,1} = \left(\frac{1}{T_{i,1}} \sum_{t=1}^{T_{i,1}} \widehat{\pi}_{i,1,t} \right) (1-\gamma) + \gamma \varsigma_{i,1} \tag{5}$$

$$\tilde{z}_{i,g} = \left[(1-\lambda)\, \tilde{z}_{i,g-1} + \lambda \left(\frac{1}{T_{i,g}} \sum_{t=1}^{T_{i,g}} \widehat{\pi}_{i,g,t} \right) \right] (1-\gamma) + \gamma \varsigma_{i,g} \tag{6}$$

$$\varsigma_{i,g} = \frac{\#\mathsf{GOAL}_{i,g}}{\#\mathsf{GOAL}_g}, \tag{7}$$

for $g = 2, \ldots, G_i$, where $\varsigma_{i,g}$ is the number of goals scored by the player i in the match g, normalized by the total of goals scored in the match. Therefore, for every player the performance value obtained by the EWMA model could be used as rating.

5 Conclusions

Instead, in this paper we propose a model-based approach for estimating the probability that a particular action performed by a player is leading to scoring a goal in the immediate next actions. After the construction of an appropriate model, knowing the probabilities of interest, we build an index summarizing the offensive impact of each player, leading to a completely data-driven approach to performances evaluation.

Funding: This research was supported by funding from the University of Padova Research Grant 2019?2020, under grant agreement BIRD203991.

References

1. Polson, N. G. and Scott, J. G. and Windle, J.: Bayesian inference for logistic models using Pólya-Gamma latent variables. Journal of the American Statistical Association. **108**, 1339–1349 (2013)
2. Eggels, H. H.: Expected goals in soccer:explaining match results using predictive analytics. Working paper, (2016)
3. McHale, Ian G and Scarf, Philip A and Folker, David E.: On the development of a soccer player performance rating system for the English Premier League. Interfaces. **42**, 339–351 (2012)
4. Decroos, Tom and Bransen, Lotte and Van Haaren, Jan and Davis, Jesse.: Actions speak louder than goals: Valuing player actions in soccer. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. **108**, 1851–1861 (2019)

3.18 The social and demographic consequences of international migration in Western societies

Employment and job satisfaction of immigrants: the case of Campania (Italy)

La soddisfazione lavorativa degli immigrati: il caso della Campania

Alessio Buonomo, Stefania Capecchi, Francesca Di Iorio, and Salvatore Strozza

Abstract
Using the two-stage Heckman procedure, a preliminary analysis is proposed on the job satisfaction of immigrants in Campania (Italy). Data come from a sample survey carried out in 2013 based on the "centre sampling technique". Citizenship and gender are prominent features, as well as the type and characteristics of work activity; over-education turns out to be of little relevance to satisfaction. The levels of social and cultural integration are significant factors.
Abstract *Utilizzando la procedura di Heckman a due stadi viene proposta una prima analisi sulla soddisfazione lavorativa degli immigrati in Campania sulla base dei dati di un'ampia indagine campionaria realizzata nel 2013 con il metodo dei centri. Cittadinanza e genere sono importanti, così come tipo e caratteristiche dell'attività lavorativa; essere sovra-istruiti è invece poco rilevante sulla soddisfazione. I livelli di integrazione sociale e culturale sono fattori significativi.*

Key words: Immigrants, Job satisfaction, Ordered Probit with selection, Campania

[1] Alessio Buonomo, University of Naples Federico II; email: alessio.buonomo@unina.it
Stefania Capecchi, University of Naples Federico II; email: stefania.capecchi@unina.it
Francesca Di Iorio, University of Naples Federico II; email: francesca.diiorio@unina.it
Salvatore Strozza, University of Naples Federico II; email: salvatore.strozza@unina.it

1 Introduction

Due to the growth of the immigrant labour force in many countries, several studies have focused on immigrant workers' overall life satisfaction and their well-being at work, only recently highlighting which influential factors do exert some impacts on satisfaction. Self-reported satisfaction at work is frequently investigated as a proxy of global individual well-being and the relationship between job satisfaction and worker characteristics has been heavily researched over the years in various domains, such as Sociology, Economics and Community Management, often uncovering positive links between workers' satisfaction and their productivity levels (for an extensive review, see: Wang and Jing 2018). A lot of job and non-job-related aspects impact on immigrants' job satisfaction, as shown in empirical studies (Covington-Ward, 2017, among others).
The aim of this contribution is to evaluate, through an ordered Probit model the relevance of different determinants on the probability of job satisfaction of a sample of adult immigrants by means of the dataset from the survey titled "*Caratteristiche e condizioni di vita degli immigrati in Campania*" (Characteristics and living condition of immigrants in Campania), carried out in the period May-October 2013.

2 Data and method

The survey involved 3,815 adult immigrants originating from least developed countries and Central and Eastern Europe. Besides the usual demographic and social characteristics, variables of interest are migratory histories, living conditions and level of integration of adult immigrants present in the five provinces of Campania (de Filippo and Strozza 2015). The survey, conducted in 79 municipalities, allows to collect information on both legally and irregularly resident immigrants in the region, since the sample design follows the centre sampling technique (Blangiardo 2004; Baio et al. 2011). In order to explore different response behaviours with respect to immigrants' job satisfaction by country of citizenship, weighted data are considered to reproduce the main characteristics of the universe.
A two-step Heckman procedure (e.g., Winship and Mare 1992) is used in this contribution. In particular, the selection equation is a Probit model on the probability of being employed (78.6% of the sample). In the second step, the categorical variable of interest is immigrants' job satisfaction, as expressed over a 4-point Likert scale where 1=not at all satisfied and 4=very satisfied ("not at all satisfied" respondents are

9.1%; “not very satisfied” are 30.8%, “satisfied” are 46.0%, “very satisfied” are 14.2%). Then, an ordered Probit model is estimated.
Possible explanatory factors are: gender (50.8% male among those employed); resident status (71.8% legally residents); marital status (single 35.2%; married 47.9%; other 16.9%); having children (59.6% have children); education (5.6% no education, 10.3% primary, 29.3% lower secondary school, 38.7% higher secondary school, and 16.2% university degree or more); citizenship (the first 10 citizenships as observed in Campania, and a comprehensive “other” category); type of occupation (general workers are 24.3%, caregivers are 21.5%, hourly housekeepers are 7.3%, housekeepers are 5.9%, specialized workers are 5.4%, shop assistants are 18.0%, cleaners are 7.2%, food services are 6.9%, other occupations are 3.6%); age (39); age squared; years since arrival (8.4); years since arrival squared; following Blangiardo and co-authors (2013), we included two different indicators of integration: the social dimension (related to the active participation in social and public life), and the cultural one (with respect to use of the Italian language, healthcare services, eating habits, etc.).
In the second step, the restriction exclusion will be applied on marital status, having children, education, age, and years since arrival. Furthermore, we included a dummy variable for overeducation (not overeducated respondents are 48.8%), employment legal status (not legal workers are 38.9%) and working hours per week (less than 30 h are 16.7%; 30-39 h are 18.8%; 40-49 h are 29.9%; 50 h and more are 34.6%).

3 Main results and discussion

The model is estimated using the ordered Probit with selection in Stata14. We focus on restricted labour force sample (3,256 employed and unemployed individuals where 434 are the unemployed). Estimated coefficients of the selection equation are reported in the following Table 1. Behavioural equation estimates are reported in Table 2.

Table 1: Selection equation, Probit estimated coefficients: Pr(Employed=1)

	Coef.	pval		Coef.	pval
Gender (ref. male=0)	-0.171	0.055	Years since arrival	0.010	0.511
Residence status (ref. legal=1)	0.152	0.028	(Years since arrival)^2	-0.001	0.112
Marital status (ref. no married=0)			Education (ref. None=0)		
Married	-0.120	0.200	Primary	0.130	0.404
Other	-0.234	0.031	Lower Secondary	0.208	0.134
Having children (ref. Not have children=0)	0.197	0.022	Higher Secondary	0.193	0.167
Citizenship (ref. Ukraine)			University	0.159	0.308
Romania	-0.117	0.235	Occupation (ref. general worker=0)		
Morocco	-0.163	0.189	Caregiver	0.046	0.681
Sri Lanka	-0.239	0.124	Hourly housekeeper	0.367	0.012
Senegal	0.214	0.405	Housekeeper	0.190	0.192
Poland	0.182	0.221	Specialized worker	0.131	0.334
Albania	-0.458	0.004	Shop assistant	0.240	0.023
Bangladesh	-0.247	0.275	Cleaner	0.054	0.700

Russia	0.203	0.390	Food service	0.130	0.310
China	0.817	0.003	Office worker, other	-0.152	0.355
Others	-0.412	0.000	Cultural Indicator	0.487	0.001
Age	0.016	0.310	Social Indicator	-0.316	0.025
Age^2	0.000	0.392	Constant	0.580	0.105

Specification Wald test results, at the bottom of Tab. 2, indicate a good model fit, whereas the Likelihood Ratio test does not allow to reject the null hypothesis that the errors for behavioural and selection equations are uncorrelated. This circumstance suggests a mild or not significant bias selection effect. Such a result suggests that further research may be developed without the implementation of a selection equation.
In Tab. 2 collects the estimated parameters of the behavioural equation. As it is evident, Polish, Russians and Romanians do not present a significantly different behaviour with respect to Ukrainians (reference category), whereas experiencing legal job conditions, number of hours worked per week, seem to be relevant, as expected. The interpretation of the estimates in the ordered Probit model is complex, since neither the sign nor the magnitude of the coefficients provides information about the partial effects of a given explanatory variable. Subsequently, the evaluation of the change of the estimated probabilities due to one of the explanatory variables depends on all estimated parameters and data.

Table 2. Behavioural equation, ordered Probit estimated coefficients

	Coef.	**Std.Err.**	**z**	**pval**
Gender (ref. male=0)	-0.145	0.066	-2.20	0.028
Residence status (ref. legal=1)	-0.205	0.054	-3.79	0.000
Employment legal status (ref. Legal=0)	-0.747	0.050	-15.09	0.000
Citizenship (ref. Ukraine)				
Romania	0.051	0.067	0.76	0.449
Morocco	-0.240	0.091	-2.63	0.009
Sri Lanka	0.718	0.113	6.38	0.000
Senegal	-0.475	0.152	-3.13	0.002
Poland	-0.039	0.094	-0.42	0.675
Albania	-0.338	0.134	-2.53	0.011
Bangladesh	-0.583	0.171	-3.41	0.001
Russia	-0.030	0.147	-0.20	0.839
China	0.823	0.128	6.44	0.000
Others	-0.207	0.075	-2.77	0.006
Overeducation (ref.. no=0)	-0.113	0.046	-2.43	0.015
Occupation (ref. general worker=0)				
Caregiver	0.210	0.084	2.49	0.013
Hourly housekeeper	-0.151	0.106	-1.42	0.154
Housekeeper	0.092	0.108	0.85	0.395
Specialized worker	0.625	0.104	6.02	0.000
Shop assistant	0.195	0.079	2.47	0.014
Cleaner	0.071	0.108	0.65	0.514
Food service	0.373	0.095	3.94	0.000
Office worker, other	0.888	0.143	6.19	0.000
Working hours by week (ref.<30 =0)				
30-39	0.253	0.072	3.49	0.000
40-49	0.497	0.068	7.35	0.000
>50	0.418	0.067	6.23	0.000
Cultural Indicator	0.679	0.107	6.36	0.000

Social Indicator	0.302	0.106	2.86	0.004
Cut point 1	-1.652	0.125	-13.22	0.000
Cut point 2	-0.377	0.121	-3.11	0.002
Cut point3	1.187	0.122	9.70	0.000
athrho	0.005	0.241	0.02	0.983
rho	0.005	0.241		

LR test of equation residuals independence (rho = 0): chi2(1) = $5e^{-4}$ pval = 0.9830
Wald chi2(27) = 701.95 pval = 0.0000

In panel A of Figure 1 the predicted probabilities to be "not at all satisfied" are depicted, whereas in panel B of Figure 1 refers to "very satisfied" category, considering a respondent profile for a "man, not legally employed, with 40-49 weekly working hours and legally resident", while continuous variables are set at their mean. The circumstance of being an illegal worker clearly affects the probability of job satisfaction, for all the citizenships, although the highest probability of expressing the lowest satisfaction is that of respondents from Senegal and Bangladesh.

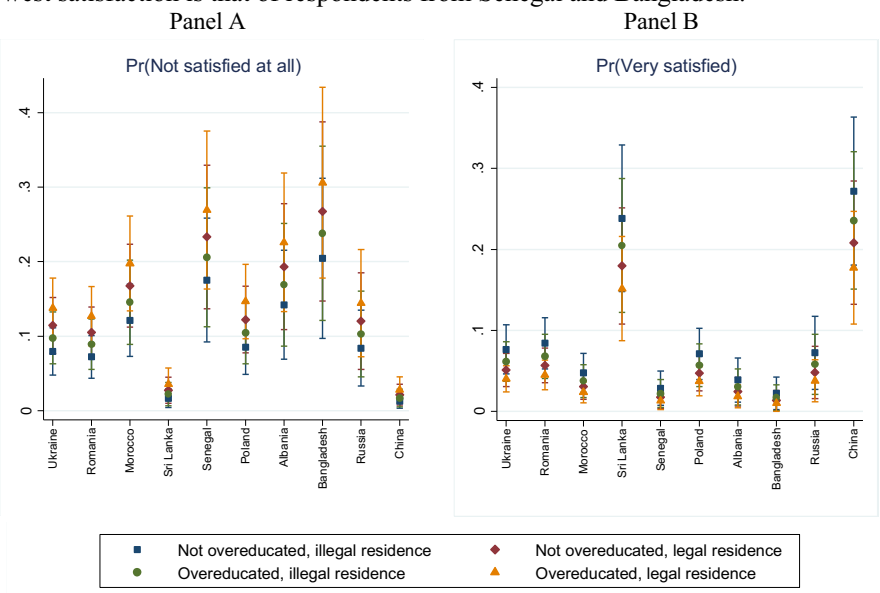


Figure 1: Estimated probability to be "not satisfied at all" (panel A) and "very satisfied" (panel B) for a man, not legally employed, with 40-49 weekly working hours, legally resident and continuous variable at their mean.

It could be interesting to compare different respondent profiles, considering two different migratory models. Table 3 collects the estimated probabilities for the two most frequent profiles in Campania: a caregiver Ukrainian woman, legally resident, overeducated, working more than 50 weekly hours, *vs* a general worker Moroccan man, legally resident, not overeducated, working 40-49 hours by week, distinguishing by legal/illegal employment status. In general, response behaviours are quite similar for such kind of respondents, although different patterns are likely to be expected in case of illegal residence status. In order to catch a possible heterogeneity of diverse migratory models more investigation is required.

Table 3. Level of satisfaction estimated probabilities for two immigrant profiles (standard error)

Estim. Probab.	Ukrainian woman		Moroccan man	
	Legal work	Illegal work	Legal work	Illegal work
Not at all	0.034 (0.007)	0.141 (0.021)	0.044 (0.010)	0.168 (0.028)
Not satisfied	0.258 (0.022)	0.438 (0.018)	0.288 (0.029)	0.455 (0.018)
Satisfied	0.553 (0.013)	0.382 (0.028)	0.539 (0.019)	0.347 (0.035)
Very satisfied	0.155 (0.020)	0.039 (0.008)	0.130 (0.023)	0.030 (0.008)

4 Conclusions

The likelihood of being employed clearly depends on gender, age, years since immigration and residence status. The level of education is also an important variable. The family condition of immigrants, on the other hand, appears less important. Job satisfaction is linked to certain demographic (gender and origin) and work (profession and hours worked) characteristics, as well as to the levels of integration. Particularly significant is the employment legal or illegal status. A possible deepening of the research could be in the direction of proposing analyses separately by gender and by legal or illegal working condition.

Acknowledgement: Grant from Ministry of Education, University and Research (MIUR), PRIN project 2017 titled "Immigration, integration, settlement. Italian-Style" (Grant No. 2017N9LCSC_004) is grateful acknowledged.

References

1. Baio, G., Blangiardo, G.C., Blangiardo, M.: Centre sampling technique in foreign migration surveys: a methodological note, Journal of Official Statistics, 27(3):451-465, (2011)
2. Blangiardo G. C., Migliorati S., Terzera L.: "Center sampling: from applicative issue to methodological aspects". Atti del Convegno della XLII Riunione Scientifica della Società Italiana di Statistica. Bari. 9-11 giugno (2004)
3. Blangiardo, G. C., Perez, M., Quattrociocchi, L., Zizza, R.: Employment and economic Conditions. In: Ministero dell'Interno, Istat (Eds.), Integrazione. Conoscere, Misurare, Valutare pp. 29-47, International Conference, Rome (2013)
4. Buonomo, A., Strozza, S., Gabrielli, G.: Immigrant youths: Between early leaving and continue their studies. In: Merrill, B., Padilla Carmona, M. T., González Monteagudo, J. (Eds.) Higher Education, Employability and Transitions to the Labour Market, pp. 131-148. Sevilla, España: EMPLOY Project: Universidad de Sevilla (2018)
5. Covington-Ward, Y: African Immigrants in Low-Wage Direct Health Care: Motivations, Job Satisfaction, and Occupational Mobility, J. of Immigrant and Minority Health, 19, 709–715 (2017)
6. de Filippo E., Strozza S. (Eds.): Gli immigrati in Campania negli anni della crisi economica. Condizioni di vita e di lavoro, progetti e possibilità di integrazione, FrancoAngeli, Milano (2015).
7. Strozza, S., De Santis, G., (Eds.): Rapporto sulla popolazione. Le molte facce della presenza straniera in Italia. Il Mulino, Bologna (2017)
8. Wang, Z., Jing, X.: Job satisfaction among immigrant workers: a review of determinants, Social Indicators Research, 139 (1). pp. 381-401 (2018)
9. Winship C, Mare RD. Models for Sample Selection Bias, Annual Review of Sociology, 18, pp. 327-350 (1992)

Social stratification of migrants in Italy: class reproduction and social mobility from origin to destination

Stratificazione sociale degli immigrati in Italia: riproduzione di classe e mobilità sociale dall'origine alla destinazione

Giorgio Piccitto, Maurizio Avola, Nazareno Panichella

Abstract In this work we aim at studying the social stratification of foreign workers in Italy adopting a class-based approach: this perspective allows us to shed light on the transmission of inequality and social mobility (both *inter*- and *intra*-generational) from country of origin to destination. We formulate the hypotheses that a low social origin and class position in home country is associated both with poor occupational achievement (*Hyp 1*) and with scarce chances of upward social mobility (*Hyp 2*) in the host country. Our results underline a clear class reproduction in the lowest strata of the occupational structure. Furthermore, occupational mobility to upper class (but not to petty bourgeoisie) is easier for who already has a high social position, while who lies at the bottom of the social structure is prevented from improving his\her position also in the host country.

Abstract *Questo lavoro studia la stratificazione sociale dei lavoratori stranieri in Italia con una prospettiva centrata sulla classe sociale: è così possibile analizzare la trasmissione delle disuguaglianze e la mobilità sociale (sia inter- che intra-generazionale) dal paese di origine a quello di destinazione. Si ipotizza che bassa origine sociale e posizione di classe nel paese d'origine siano associate a scarsi esiti occupazionali (Hyp 1) e a ridotte possibilità di mobilità ascendente (Hyp 2) nel paese di destinazione. I risultati sottolineano una chiara riproduzione di classe negli strati più bassi della struttura occupazionale. Inoltre, la mobilità occupazionale verso l'upper class (ma non verso la piccola borghesia) è più realizzabile per chi ha già un'elevata posizione sociale, mentre chi ha un basso status socio-economico incontra difficoltà a migliorare la propria posizione anche nel paese ospitante.*

Key words: international migration, labour market, social class reproduction, social mobility, social stratification

1 Introduction

In this work we study the social reproduction of migrants in Italy, combining the social stratification and mobility literature with the *migration studies*. In particular, we explicitly account for migrants' heterogeneity in terms of social class of origin, an issue which has been under-considered in the study of first generation migrants' integration (Panichella, Avola, Piccitto, *in press*). Hence, we adopt a class-based approach to analyse to what extent social inequalities are transferred across space and time from both an *inter*-generational and an *intra*-generational point of view.

The focus on social class allows to add some important contributions to this field of study. Social class of origin is likely to affect occupational outcomes in a host country since it provides migrants with different endowments of capitals, that could be distinguished in economic, cultural and social (Bourdieu, 1986). Other social background-related factors that affect the occupational achievement – over and above education – are motivations and aspirations, inheritance of parental business or material resources, different productivity, potential favouritism (Bernardi, Ballarino, 2016).

Social class in the country of origin has per se a direct effect on occupational achievement after migration: new-comers are indeed more at risk of entering in the secondary labour market (job positions characterized by lower chances of social improvement) (Portes, 1995). This is particularly true in a country like Italy, where migrants are over-represented among unskilled, not-standard and poorly rewarded jobs (Avola, 2015; 2018; Avola, Piccitto, 2020; Ballarino, Panichella, 2015; 2018; Fullin, Reyneri, 2011; Panichella, 2018; Panichella, Avola, Piccitto, *in press*), especially due to the large share of underground economy (Reyneri, 2003), a high turnover between (unskilled) employment and unemployment (Avola, 2015) and the stark processes of flexibilization at the margins (Barbieri, Cutuli, 2016). These characteristics of the Italian labour market make migrants more likely to be trapped in the so-called *3D* (dirty, dangerous and degrading) occupations (Piore, 1979) and to have very limited opportunities of upward social mobility (Fellini, Guetto, 2019; Avola, Piccitto, 2020; Panichella, Avola, Piccitto, *in press*).

On this ground, we hypothesize that migrants with the lowest occupational chances in Italy are those who already had weak social origin and class position in the country of origin: these subjects are the ones who get lower occupational achievements at the first job in Italy (*Hyp 1*) and are less likely to 'escape' the low-status jobs trap during the career (*Hyp 2*). Thanks to high-quality data, focused on migrant individuals living in Italy, we are able to highlight patterns of social class reproduction after migration.

2 Data and method

We base our analysis on the Social Condition and Integration of Foreign Citizens (SCIF) survey, which has been collected by the Italian National Institute of Statistics

(Istat) in 2011–2012. We define social class on the basis of the EGP class scheme (Erikson, Goldthorpe, 1992), identifying: 1) upper class (EGP I-II-IIIab); 2) petty bourgeoisie (EGP IVabc); 3) working class (EGP V-VI-VIIab). Within the working class, we further distinguish between: 3a) stable working class (permanent and full-time); 3b) unstable working class (fixed-term and/or part-time); this choice allows us to account for the growing within-classes heterogeneity of modern societies (Oesch, 2006; Panichella, Avola, Piccitto, *in press*). Individuals are classified as migrants if they have born abroad, except for people born in North America, Oceania and other high-income countries, since their occupational condition is in general similar to that of the native population (Avola, Piccitto, Vegetti, 2019). Our sample consists of males aged between 25 and 64. Since migration is a gendered process, we excluded women from the analysis: their inclusion should have required a more refined empirical strategy (Ballarino, Panichella, 2018). Analysis also excludes those migrants who migrated before being less than 15 years old, since they may migrate in order to join their parents (generation 1.5). After a list-wise deletion of missing observations, we end up with an analytical sample of 5,752 individuals.
Our empirical strategy consists of two sets of logit models, aiming at testing our two hypotheses. Three dependent variables are defined to test *Hyp 1*: a) entering the working class at first job; b) entering the stable working class at first job; c) entering the unstable working class at first job. In order to test *Hyp 2*, we define the following dependent variables: a) transiting from working class to upper class; b) transiting from working class to petty bourgeoisie; c) avoiding unstable working class or unemployment. We control for a number of covariates, namely: level and place of education (lower or less; upper/tertiary in country of origin; upper/tertiary in Italy); father's social class (upper class; petty bourgeoisie; working class); marital status (single, married, divorced); number of children (0; 1; 2; +3); age dummies (from 25 up to 64); macro-area of residence (North-west; North-east; Centre; South and Islands); years of residence dummies; direct migration (yes; no).

3 Preliminary findings

Table 1 shows the results for models testing *Hyp 1*, while in Table 2 are presented results for *Hyp 2*. Considering model (1), it emerges that being raised in a working class family increases the chances of getting a job within the working class after migration, net of the socio-economic position in the country of origin. Interestingly, social class before migration is a stronger predictor of social class at first job in Italy: indeed, individuals that were employed in the working class are more likely to maintain this very class in their first occupational episode in the host country; this is true for people with a previous job in their country of origin both in the stable working class and in the unstable one. Remarkably, also individuals not working in their home country have higher chances to have a working class job, with respect to former members of upper class or petty bourgeoisie. When 'unpacking' the working class in stable and unstable (models (2) and (3)), the reproduction of class position

from country of origin to destination becomes even more visible: indeed, people who belonged to the stable working class transit, after migration, to a permanent position within that class. Vice versa, who before migration was employed in a less secure and rewarding blue collar job, has more chances to find a similar job in the new country; these people is not even able to experience a pattern of horizontal mobility *within-working class*.

Table 1 – Probability of entering in the working class at first job in Italy. Average partial effects: logit models

	WC		Stable WC		Unstable WC	
	(1)		(2)		(3)	
	β	σ	β	σ	β	σ
Class of origin [Ref.: Upp]						
PB	0.01	(0.03)	-0.01	(0.04)	0.01	(0.04)
WC	0.03*	(0.02)	0.03	(0.02)	0.00	(0.02)
Class in country of origin [Ref.: Upp]						
PB	0.01	(0.03)	-0.00	(0.03)	0.01	(0.03)
WC (Stable)	0.17***	(0.02)	0.17***	(0.03)	-0.01	(0.03)
WC (Unstable)	0.12***	(0.02)	0.02	(0.02)	0.10***	(0.03)
Not working	0.07***	(0.02)	0.04*	(0.02)	0.03	(0.02)
(N)	5,752		5,752		5,752	

Table 2 analyses to what extent foreign workers succeed in: 4) transiting from working class to upper class; 5) transiting from working class to petty bourgeoisie; 6) avoiding unstable WC and unemployment. Looking at the two 'real' upward mobility processes, it emerges that social class does not have any effect on the transition from working class to petty bourgeoisie (5). It is likely that the non-financial resources related to the social status, like motivation and entrepreneurial attitude (but not the financial ones, geographically enrooted and not transferable), ease the access in this class in a new country only immediately after the process of migration: once 'trapped' in working class at first job, these resources cease being relevant for the transition to self-employment. Differently, coming from a job in upper class (4) confers higher chances on foreign workers of transiting from working class to upper class; indeed, coming from whatsoever other class is associated with less chances of experiencing this pattern of upward mobility. This finding may be lead from highly-educated people that accept a low-status first job in a new country, waiting for the recognition of their educational degree (Fellini, Guetto, 2019). Finally, when looking at the probability of downgrading from stable working class at first job in Italy to unstable working class or unemployment (6), it emerges that coming from a working class household is associated with fewer chances of dodging this pattern of mobility. Similarly, also the position in own home country is associated with the avoidance of this downward transition: people who were in the

unstable working class or not working are remarkably more at risk of incurring in this trajectory than people previously employed in other classes.

Table 2 – Probability of upward social mobility among migrants. Average partial effects: logit models

	From WC to UPP		From WC to PB		Avoiding unstable WC and unemployment	
	(4)		(5)		(6)	
	β	σ	B	σ	β	σ
Class of origin [Ref.: Upp]						
PB	0.00	(0.02)	-0.01	(0.02)	0.01	(0.06)
WC	-0.02*	(0.01)	-0.01	(0.01)	-0.10***	(0.03)
Class in country of origin [Ref.: Upp]						
PB	-0.07***	(0.02)	0.02	(0.03)	-0.05	(0.05)
WC (Stable)	-0.09***	(0.02)	-0.00	(0.03)	-0.03	(0.04)
WC (Unstable)	-0.07***	(0.02)	0.01	(0.02)	-0.12***	(0.04)
Not working	-0.08***	(0.02)	-0.01	(0.02)	-0.15***	(0.04)
(N)	4,887		4,887		2,703	

These first findings corroborate our hypotheses. With respect to *Hyp 1*, our analysis shows that people with weak social position in their country of origin is more likely to reproduce their low position in the new country. This process emerges also within working class, with the distribution of workers in core\peripheral occupations within this class mirrored in the country of destination. Also *Hyp 2* is to confirmed by our results: having had a job in the upper class in the home country eases the transition from working class to upper class, while who was placed in the periphery of the labour market in the own country finds severe difficulties in improving his\her social status also in the host country. Only the transition from working class to petty bourgeoisie is not affected by individuals' social background or by individual social position in his\her country of origin. Remarkably, despite the process of occupational upgrading which should trigger more and better jobs for everyone (Piccitto, 2019; Oesch, Piccitto, 2019), worker's social status continues to be an important factor of inequality within the labour market: the most fragile segments of the workforce bring their fragilities with them also after migration.

References

Avola, M.: The ethnic penalty in the Italian labour market: A comparison between the centre-north and south. J. Ethn. Migr. Stud. 41 (11–12): 1746–1768 (2015).

Avola, M.: Lavoro immigrato e dualismo territoriale nell'Italia della decrescita: Struttura della domanda e mutamenti dell'offerta. Stato e Mercato 113: 331–362 (2018).
Avola, M., Piccitto, G.: Ethnic penalty and occupational mobility in the Italian labour market. Ethnicities 20 (6): 1093-1116 (2020).
Avola, M., Piccitto, G., Vegetti, F.: Ethnic penalty in the European labour markets: a multilevel approach. Paper presented at European Consortium for Sociological Research Annual Conference, 12-14 September, Lausanne (2019).
Ballarino, G., Panichella, N.: The occupational integration of male migrants in western European countries: Assimilation or persistent disadvantage? Int. Migr. 53 (2): 338–352 (2015).
Ballarino, G., Panichella, N.: The occupational integration of migrant women in Western European labour markets. Acta Sociol. 61 (2): 126–142 (2018).
Barbieri, P., Cutuli, G.: Employment protection legislation, labour market dualism, and inequality in Europe, Eur. Sociol. Rev. 32 (4): 501-516 (2016).
Bernardi, F., Ballarino, G.: Education, occupation and social origin: A comparative analysis of the transmission of socio-economic inequalities. Cheltenham: Edward Elgar Publishing (2016).
Bourdieu, P. The forms of capital. In Richardson, J.G. (ed.) Handbook of Theory and Research for the Sociology of Education, pp.241.258. Greenwood, New York (1986).
Erikson, R., Goldthorpe J.H.: The constant flux: A study of class mobility in industrial societies. Clarendon Press, Oxford (1992).
Fellini, I., Guetto, R.: A "U-shaped" pattern of immigrants' occupational careers? A comparative analysis of Italy, Spain, and France. Int. Migr. Rev. 53 (1): 26-58 (2019).
Fullin, G., Reyneri, E.: Low unemployment and bad jobs for new immigrants in Italy. Int. Migr. 49 (1): 118-147 (2011).
Oesch, D.: Redrawing the Class Map: Stratification and Institutions in Britain, Germany, Sweden and Switzerland. Palgrave Macmillan, Basingstoke (2006).
Oesch, D., Piccitto, G.: The polarization myth: Occupational upgrading in Germany, Spain, Sweden, and the UK, 1992–2015. Work Occup 46 (4): 441-469 (2019).
Panichella, N.: Economic crisis and occupational integration of recent immigrants in Western Europe. Int. Sociol. 33 (1): 64–85 (2018).
Panichella, N., Avola, M., Piccitto, G.: (*in press*). Migration, class attainment and social mobility. An analysis of migrants' socio-economic integration in Italy. Eur. Sociol. Rev.
Piccitto, G.: Qualificazione o polarizzazione? Il mutamento della struttura occupazionale in Italia, 1992-2015. Polis 33 (1): 59-88 (2019).
Piore, M.: Birds of Passage: Migrant Labor and Industrial Societies. Cambridge University Press, New York (1979).
Portes, A.: The Economic Sociology of Immigration: Essays on Networks, Ethnicity, and Entrepreneurship. Russell Sage Foundation, New York (1995).
Reyneri, E.: Immigrants in a segmented and often undeclared labour market. J. Mod. Ital. Stud. 9 (1): 71-93 (2003).

3.19 Well-being, healthcare, integration measurements and indicators (SIEDS)

A Composite Index of Economic Well-being for the European Union Countries

Un indice sintetico di benessere economico per i Paesi dell'Unione Europea

Andrea Cutillo, Matteo Mazziotta, Adriano Pareto

Abstract The measurement of Equitable and Sustainable Well-being (BES) in Italy is one of the most appreciated monitoring tools by the Scientific Community. The focus on the Economic Well-being domain seems essential around the last serious economic crisis. The use of an innovative composite index can help to measure the multidimensional phenomenon and monitor the situation at European level.

Abstract *La misurazione del Benessere Equo e Sostenibile (BES) in Italia è uno degli strumenti di monitoraggio più apprezzati dalla Comunità Scientifica. Il focus sul dominio Benessere economico sembra indispensabile intorno all'ultima grave crisi economica. L'uso di un innovativo indice composito può aiutare a misurare il fenomeno multidimensionale e a monitorare la situazione a livello europeo.*

Key words: Composite index, ranking, economic well-being

1 Introduction

In this paper, the economic well-being in Europe is focused, taking as a reference point the economic domain of the project BES (Equitable and Sustainable Well-being in Italy) of the Italian National Institute of Statistics (Istat). The BES aims at evaluating the progress of societies by considering different perspectives through twelve relevant theoretical domains, each one measured through a different set of individual indicators. The BES project is inspired by the Global Project on Measuring the progress of Societies of the Oecd (2007), under the idea that the economic well-being is not enough for the developed Countries. However, since

Andrea Cutillo, Istat; email: cutillo@istat.it

Matteo Mazziotta, Istat; email: mazziott@istat.it

Adriano Pareto, Istat; email: pareto@istat.it

2007, the two crises have affected the households' economic well-being: the international economic crisis (2008-2009) derived from the Lehman Brothers failure; and the European crisis of the sovereign debt, whose effects were more intense in 2011-2012, and can be considered solved in 2014[1]. It is clear that the economic domain of the well-being still deserves particular relevance within the other dimensions. Following the timeliness described above, the longitudinal analysis is set at 4 relevant years: 2007, 2010, 2014 and 2019.

2 Theoretical framework

The economic well-being through 4 subdomains is measured, each one represented by an indicator coming from the Silc system.

1. Sub-domain *Purchasing Power*; indicator: *Median equivalised income in purchasing power standards (Pps)*. The Istat *average income per capita* is replaced for three reasons. First, the median is a better indicator of a monetary distribution, given its robustness to extreme values. Second, the equivalised form (through the modified Oecd scale) is better in order to consider the different sizes and needs of the households. Third, in the European context, it is essential to consider the different cost of life and purchasing powers in the Countries.

2. Sub-domain *Inequality*; indicator: *At risk of poverty rate (ARP)*. It is a relative measure of poverty: its threshold is set dependently on the income distribution and, therefore, it merely captures how many individuals are far from the others. That is, relative poverty is an inequality indicator rather than a poverty indicator (Sen, 1983). Istat measures inequality also through the *Disposable income inequality* (S80_S20 index). Since they are both representative of the same sub-domain, and it is a good practice to strictly select indicators, ARP is selected.

3. Sub-domain *Poverty*; indicator *Severe material deprivation (SMD)*, that is the share of population living in households lacking at least 4 items out of 9 economic deprivations. Far from being a perfect indicator, it is the most similar indicator to the concept of absolute poverty in the EU. Unwillingly, Istat *Absolute poverty rate is not used* since it is the "true" measure of poverty (the poverty lines are set independently on the monetary distribution, and also consider the different cost of life in different areas). However, the measure of the World Bank which does not fit for developed Countries is excluded (the absolute poverty is officially measured only in Italy and USA, because of the difficulties in the definition). And the European Commission project "Measuring and monitoring absolute poverty—ABSPO" is still in the phase of study (Cutillo et al., 2020).

4. Sub-domain *Subjective evaluation*; indicator *Index of economic distress*, that is the share of individuals in households that declare to get to the end of the month with great difficulty. The subjective sub-dimension is considered an important one, especially in the context of the BES.

[1] We can't forget the current crisis deriving from the Covid19 pandemic situation, even if the adopted indicators can't still measure its impact

The remaining Istat indicators are removed for the following reasons: *Per capita net wealth*: the sub-domain wealth is certainly a pillar of the households' monetary well-being. However, correctly measuring the value of wealth is extremely complex, since some types of wealth are statistically hidden (e.g., paintings, jewellery etc.), and attributing a value to wealth is arbitrary when some types of wealth are not sold/bought (e.g., houses). This is a very relevant issue in the European context, given the different weight between financial wealth and real estate wealth in the different countries. *People living in financially vulnerable households*: there is not such an indicator in the Eurostat database. Thus, in this stage, it is excluded from the set, reserving to search for an alternative indicator in the European Central Bank (ECB) data. *Severe housing deprivation* and *Low work intensity* measure important topics, but they can't be exactly considered as indicators of economic well-being under a theoretical point of view.

3 Methodological aspects

The composite index was constructed using the Adjusted Mazziotta-Pareto Index – AMPI (Mazziotta and Pareto, 2016). This aggregation function allows a partial compensability, so that an increase in the most deprived indicator will have a higher impact on the composite index (imperfect substitutability). Such a choice is advisable whenever a reasonable achievement in any of the individual indicators is considered to be crucial for overall performance (Chiappero-Martinetti and von Jacobi, 2012). The most original aspect of this index is the method of normalization, called "Constrained Min-Max Method" (Mazziotta and Pareto, 2021). This method normalizes the range of individual indicators, similarly to the classic Min-Max method, but uses a common reference that allows to define a 'balancing model' (i.e., the set of values that are considered balanced). Thus, it is possible to compare the values of the units, both in space and time, with respect to a common reference that does not change over time.

Let us consider the matrix $\mathbf{X}=\{x_{ijt}\}$ with 27 rows (countries), 4 columns (individual indicators), and 4 layers (years) where x_{ijt} is the value of individual indicator j, for country i, at year t. A normalized matrix $\mathbf{R}=\{r_{ijt}\}$ is computed as follows:

$$r_{ijt} = 100 \pm \frac{x_{ijt} - x_{j0}}{\max_{it}(x_{ijt}) - \min_{it}(x_{ijt})} 60$$

where $\min_{it}(x_{ijt})$ and $\max_{it}(x_{ijt})$ are, respectively, the overall minimum and maximum of indicator j across all times (goalposts), x_{j0} is the EU average in 2007 (reference value) for indicator j, and the sign ± depends on the polarity of indicator j.

Denoting with $\mathrm{M}_{r_{it}}$, $\mathrm{S}_{r_{it}}$, $\mathrm{cv}_{r_{it}}$, respectively, the mean, standard deviation, and coefficient of variation of the normalized values for country i, at year t, the composite index is given by:

$$\text{AMPI}_{it}^{-} = \text{M}_{r_{it}} - \text{S}_{r_{it}} \text{cv}_{r_{it}} .$$

Figure 1 shows the effect of normalization on three individual indicators with different shape. The first has an exponential distribution (Exp), the second has a normal distribution (Nor) and the third has a Beta distribution (Bet). In Figure 1a, indicators are normalized by the classic Min-Max method in the range [0, 1], and in Figure 1b, they are normalized by the constrained Min-max method with a reference (the mean) of 100 and a range of 60.

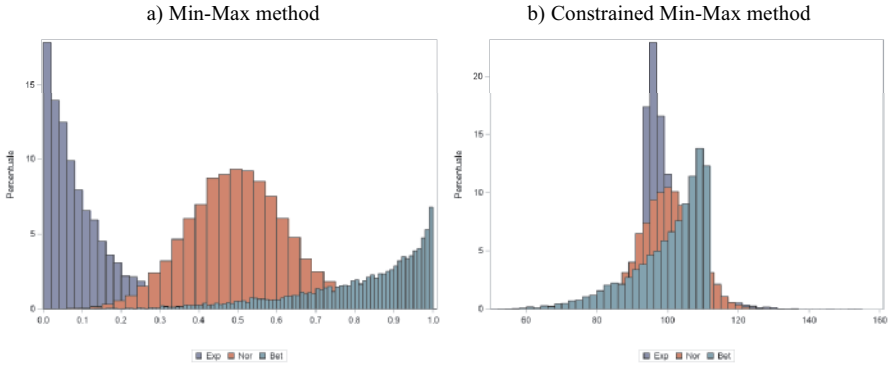


Figure 1: Comparing the classic and the constrained Min-Max method.

As we can see, the Min-Max method bring all values into a closed interval, but the distributions of indicators are not 'centred' and this leads to the loss of a common reference value, such as the mean. It follows that equal normalized values (i.e., balanced normalized values) can correspond to very unbalanced original values. For example, the normalized value 0.2 for the Exp indicator corresponds to a high original value; whereas for the Nor and Bet indicators it corresponds to a very low original value. Moreover, the normalized value 0.5 is the mean of the range, but not of distributions, and then it cannot be used as a reference for reading results (e.g., if the normalized value of a country is 0.3., we cannot know if its original value is above or below the mean). On the contrary, normalized values by the constrained Min-Max method are not forced into a closed interval, they are 'centred' with respect to a common reference, and they are easier to interpret: if the normalized value of a country is greater than 100, then it is above the reference value, else it is below the reference value. Finally, the comparability across time is maintained when new data become available (the goalposts do not need to be updated).

4 A longitudinal analysis

In the analysis, the reference value is the Eu27 in 2007 (=100). The Eu27 indicator is not far from 100 neither in 2010 (100.2) nor in 2014 (99.6). The last period, till 2019, shows instead an increase of the overall index of about 5 point (104.7 in 2019 versus 99.6 in 2014). The first period, the one corresponding to the international economic crisis, is the most stable. Indeed, the ranking, based on the AMPI, shows the lower degree of variability. The highest jump in the AMPI absolute value is observed for Poland, which passes from the 23rd position to the 19th. In the second period, corresponding to the crisis of the sovereign debt, there is a greater mobility in the ranking. Greece shows the highest jump, from 21st to 27th and last position. The Greek AMPI decreased dramatically from 89.2 to 74.6. This fall was mainly due to a dramatic fall in the purchasing power of the households (the median equivalized income in Pps decreased from 12,598 to 8,673 euros - if we exclude Cyprus, which obviously followed Greece, the second value was -272 euros of Ireland). Also, the SMD and the subjective economic distress greatly worsened, respectively from 11.6% to 21.5% and from 24.2% to 39.5%). Greece was the first country to be hit by the equity markets distrust on the debt sustainability, later followed by Portugal and Ireland and successively by Italy and Spain. In the 2010-2014 period, Ireland loses two positions (from 12 to 14th), Spain and Portugal one position (respectively, from 18th to 19th and from 21st to 22nd), while Italy gained one position (from 17th to 16th). However, also Italy showed a decrease in the synthetic index, from 96.2 to 94.2. The overall Italian situation was somewhat preserved by the fact that only the SMD indicator worsened (from 7.4 to 11.6%), while the other three were substantially unchanged. In this period, we can observe a new great advance of Poland (+4 in the ranking, from 19th to 15th). Finally, in the last period till 2019, the overall Eu27 index passed from 99.6 to 104.7, showing a general increase on the economic well-being of the households, and all the countries, but Sweden and Luxemburg, increased the value of the index. Some countries had a particularly great increase (Hungary, Cyprus, Croatia and Ireland, more than +10 points). As concerning the ranking, Hungary showed the greatest increase, +6 positions, especially due to an improvement in median purchasing power, SMD and subjective economic distress; Luxemburg and Sweden showed the greatest decrease, -5 positions, especially due to an increase of inequality as measured by the ARP rate in a context of general decrease in the European zone. Considering the entire period, 2007-2019, some Countries greatly increased their economic well-being, in particular, and somewhat obviously, the Countries that started from a disadvantaged situation: Bulgaria (+17.7), Poland (+15.6) and Romania (+13.3). In the case of Poland, this also pushed the ranking, from the 23rd to the 14th position; Bulgaria and Romania still remain at the bottom tail of the ranking, respectively 26th and 25th in 2019, +1 position for both the Countries, but strongly filled the gap in respect of the overall EU27. At the opposite, the Greek indicator has fallen down by 10.1 point (even if it is growing in the last sub-period), completely due to the 2011-2014 period. At present, Greece is in the last position of the ranking, 27th form the 22nd that occupied in 2007, while the first position is occupied by Finland. The other two

Countries with an important decrease in the MPI indicator are Luxemburg, -4.8 points, and Sweden, -3.5 points, which shifted, respectively, from 1st to 6th position and from 3rd to 12th position. Finally, Italy is 18th both in 2007 (index=95.4), and in 2019 (index=99.4).

5 Conclusions

In conclusion, the first two periods appear to be a unique long period of crisis, slightly softer and more diffuse in the first part; more intense and localized in a fewer number of Countries in the second part, especially Greece. The peculiarity of this second period is that even the Countries which didn't face the crisis of the sovereign debts didn't improve their economic well-being. This fact clearly show that the European response was not the right one, and the vexatious conditions imposed to Greece by EC, ECB and IMF highly worsened the household economic conditions and were badly used as a warning for other indebted Countries. Unsurprisingly, they were instead used by the stock markets' operators as a "go ahead" towards speculation, which quickly enlarged against the other Countries, and the entire Euro zone was put in doubt. Luckily, fiscal and monetary policies have completely changed since then. The IMF was only marginally involved; the Eurozone, even in a context of a formally stricter balance observation through the fiscal compact, contemplated several adjustments which allowed to keep in account different factors; and, mainly, the ECB completely changed its monetary policy. Indeed, the quantitative easing started in 2012 in order to support the financial system; somewhat enlarged its effects on the productivity system in 2014; and started its second phase in 2015, with more and more great intervention in the private sector. Such measures have permitted to relax the economic distress on the European households and all European countries have resumed the normal path towards the higher and generalized economic well-being that characterized the whole post-war period.

References

1. Chiappero-Martinetti, E., von Jacobi, N.: Light and shade of multidimensional indexes. How Methodological Choices Impact on Empirical Results. In: F. Maggino, G. Nuvolati (eds), Quality of life in Italy, Research and Reflections, pp. 69–103, Springer, Cham (2012).
2. Cutillo, A., Raitano, M., Siciliani, I.: Income-Based and Consumption-Based Measurement of Absolute Poverty: Insights from Italy. Social Indicators Research. https://doi.org/10.1007/s11205-020-02386-9 (2020).
3. Mazziotta, M., Pareto, A.: On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena. Social Indicators Research, 127, pp. 983-1003 (2016).
4. Mazziotta, M., Pareto, A.: Everything you always wanted to know about normalization (but were afraid to ask). Italian Review of Economics, Demography and Statistics, LXXV, 1, pp. 41–52 (2021).
5. Sen, A.: Poor, relatively speaking. Oxford Economic Series, *35*(2), 153–169 (1983).

Poverty orderings and TIP curves: an application to the Italian regions

Curve TIP e dominanza stocastica: un'applicazione alle regioni italiane

Francesco M. Chelli, Mariateresa Ciommi and Chiara Gigliarano

Abstract This paper contributes to the literature by analysing how poor the income poor are in the Italian regions. Using the data from the Italian version of the Statistics on Income and Living Conditions (IT-SILC), we apply TIP curves for representing the three different aspects of poverty: incidence, intensity and inequality. We study the evolution of poverty in Italy over the recent decades, by providing poverty orderings consistent with a large class of poverty indices and allowing different poverty lines. The main conclusion is an unambiguous increase in poverty levels from 2010 to 2015, both in the entire Italian population as well as in most of its regions.

Abstract *Scopo del lavoro é quello di studiare l'evoluzione della povertá in Italia negli ultimi decenni, fornendo ordinamenti di povertá coerenti con un'ampia classe di indici e diverse soglie di povertá. Usando i dati IT-SILC, vengono applicate le curve TIP per monitorare congiuntamente incidenza, intensitá e disuguaglianza dei poveri. I dati mostrano un rilevante aumento dei livelli di povertá dal 2010 al 2015, sia per l'intera popolazione italiana che per la maggior parte delle sue regioni.*

Key words: Poverty orderings, Regional analysis, TIP dominance test

Francesco M. Chelli
Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy, e-mail: f.chelli@univpm.it

Mariateresa Ciommi
Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy, e-mail: m.ciommmi@univpm.it

Chiara Gigliarano
Department of Economics, Insubria University, Varese, Italy e-mail: chiara.gigliarano@uninsubria.it

1 Introduction

In the literature there is quite a consensus that poverty should be measured by considering three specific aspects: the incidence, the intensity and the inequality among the poor (the so-called Three I's of Poverty proposed by [5]).

Poverty incidence is measured through the well-known *Headcount (H)* measure. It is defined as the proportion of the population below the poverty line. Formally, if $x = (x_1, x_2, \ldots, x_n)$ denotes the incomes of n individuals in a given society, such that $x_i \leq x_j$ for any $i \leq j$, $i, j : 1, \ldots, n$ and z denotes a fixed poverty line, we count the number of individuals whose income is below the threshold, say q. Hence the poverty headcount is defined as:

$$H = \frac{q}{n}.$$

Thanks to the simplicity in its calculation, this index has received a great attention and it is the most used poverty index. However, it is not without significant drawbacks. First of all, it just counts the number of people below the threshold and consequently it does not take into account how poor the poor are (intensity). Moreover, it does not change if the income of a poor individual increases but not enough to exceed the poverty line.

To account for *intensity*, the simplest measure is the *Income Gap Ratio (IGR)*, which is defined as the mean of the relative gap from the poverty line among the poor. Formally:

$$IGR = \frac{1}{q} \sum_{i=1}^{q} \frac{z - x_i}{z}.$$

The main advantage of *IGR* is that someone who falls just below the poverty line is counted less than someone who is much further below it.

Finally, the inequality among the poor can be measured using several different inequality indices (among which the well-known Gini index) and it could be computed looking at the inequality of incomes or of gaps.

As recommended by [6], any poverty measure should account for the three above mentioned components (Incidence, Intensity and Inequality) since they capture different aspects of poverty. In fact, monitoring these dimensions jointly may provide a better picture of the phenomenon. Moreover, if we focus just on one of the three I's, neglecting the importance of the other two, we may get an unrealistic and underestimated picture of level of poverty in a given country. In addition, the headcount and the income gap do not always provide the same ordering. We can compare a region that has a greater number of poor, but none of them extremely poor against another region with few poor but all very far from the poverty line. In this situation, which region is poorer? And, in general, how can poverty comparisons be made?

Aim of this work is to analyse poverty of the Italian households at regional level. We are interested in analysing how the Italian income distribution has changed over the period of time between 2005 and 2015 and whether there are substantial differences at local level. Thus, we are interested in examining whether the changes in national welfare are reflected at regional levels by conducting stochastic dominance

analysis among regions. To achieve our aim, we use a graphical approach to support the traditional methods based on the well-know poverty indices. In particular, following [5] proposal, we construct the so-called TIP curves for each region.

2 Methodology

TIP curves are a graphical representation of the cumulated proportion of population (x-axis) versus the cumulated normalized poverty gap (y-axis), where the gap is defined only for the poor and is calculated as the difference between income and the poverty line. Formally, if we indicate the relative poverty gap with

$$\left(1-\frac{x}{z}\right)\mathbf{1}(x\leq z),$$

where $\mathbf{1}(x\leq z)$ is the indicator function, then the TIP curve is obtained by integrating the relative poverty gap with respect to the income distribution $f(x)$ as follows (see [5] and [4]):

$$TIP(p,z)=\int_0^{F^{-1}(p)}\left(1-\frac{x}{z}\right)\mathbf{1}(x\leq z)f(x)dx,$$

where $F^{-1}(p)$ is the quantile function and p the proportion of individuals.

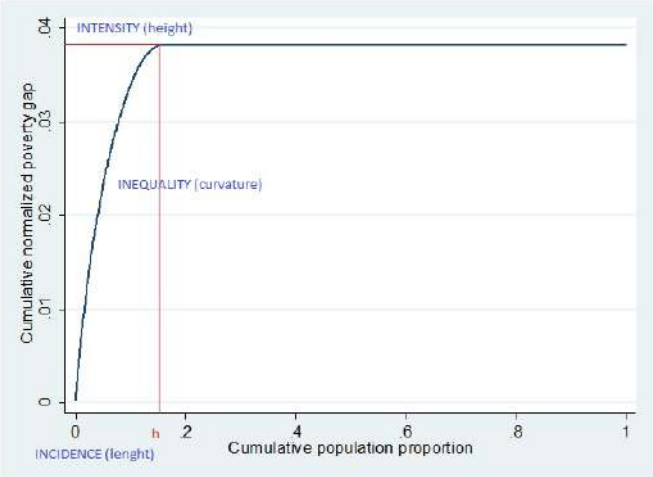
To construct the curve, gaps are ordered from largest to smallest. For values of p (horizontal axis) greater then the poverty incidence, the TIP curve becomes horizontal. At this point, the x-axis value corresponds to the incidence of poverty, while the y-axis value indicates the poverty intensity. Finally, the curvature indicates the degree of inequality among the poor. If the curve is a straight line, it means that all the poor are equally poor. The more the TIP curve deviates from linearity, the greater is the degree of inequality among the poor (see Figure 1).

If the TIP curve associated to a distribution A lies everywhere above the TIP curve associated to distribution B, we say that distribution A TIP-dominates distribution B. Consequently, TIP curves provide a dominance criterion that is robust to the choice of both the poverty line and the poverty measure. In case of crossing TIP curves, the ordering will be a partial order, and further investigation is required.

Although the TIP curves have been mainly used as a descriptive tool (see, among others, [3]), a recent attention has been growing on their inferential properties. In particular, [7] has proposed statistical inference in presence of stochastic survey weights, while [1] provided a non-parametric test for TIP dominance based on influence functions. Moreover, [4] proposed a Bayesian inference of TIP curves.

Here we follow [2] and use the procedure introduced by [8] for statistical inference about TIP dominance. In particular, let X_1 and X_2 be two income distributions and ϕ_1 and ϕ_2 be the corresponding $K\times 1$ vectors of TIP curve ordinates. We test the null hypothesis $H_0:\phi_1-\phi_2\geq 0$ against the alternative hypothesis $H_1:\phi_1-\phi_2<0$ Here dominance means that, given two income distributions X_1 and X_2 (namely two Italian Regions) and a common poverty line z (e.g. 60% of the Italian median in-

Fig. 1 An example of TIP curve



come), if the TIP curve associated with the income distribution X_2 lies always above the TIP curve of X_2, then in income distribution X_2 poverty is higher than in X_1 according to any index in the class of normalised poverty gap measures.

3 Data and main results

We use data from the Italian version of the Statistics on Income and Living Conditions (IT-SILC), which collects micro-data on income, poverty, social exclusion and living conditions of the Italians. Our variable of interest is the equivalised disposable household income (labelled HX090) for three different years, 2005, 2010, 2015. IT-SILC has been designed to ensure representativeness at the regional level, therefore we will compare the Italian regions in terms of poverty levels.

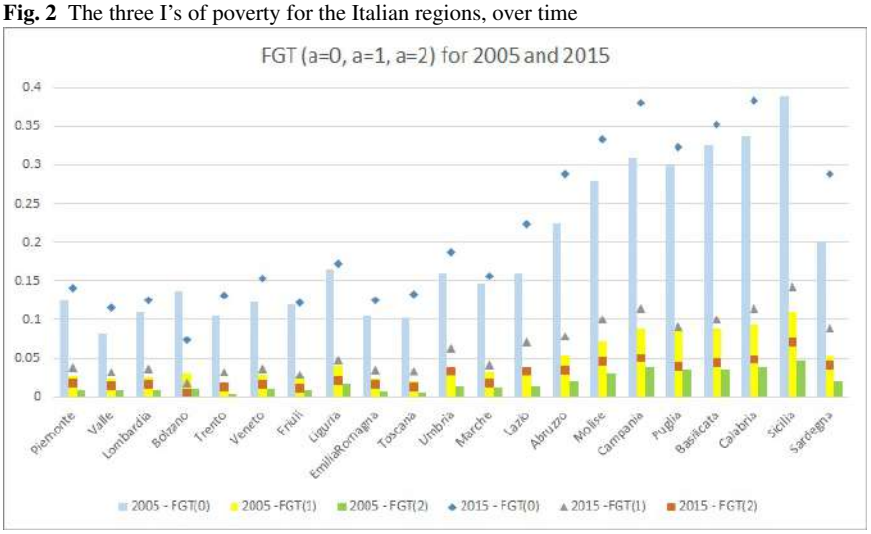
We first compute poverty incidence, intensity and inequality for each region and over time. Results, depicted in Figure 2, clearly show an increase in all the three I's of poverty from 2005 to 2015 in all the Italian regions. Moreover, we note that the regions are clearly bipolarized (North-Center vs South) in terms of poverty levels.

The TIP curves for Italy (Figure 3) confirm an increase in poverty from 2010 to 2015 while between 2005 and 2010 poverty remained quite stable. From the same Figure we note also a certain degree of variability across regions in terms of poverty, with several TIP curves strongly dominating other curves.

More details are provided in Table (1), which shows the results of the pairwise TIP dominance tests among the Italian regions for the year 2005 (significance level of 5%). The table should be read from row to column, as follows: symbol ">" denotes that Region A (row) dominates Region B (column), symbol "<" denotes that

Region *B* (column) dominates Region *A* (row). When two TIP curves intersect, the dominance criterion fails: this is the situation denoted in the table with the symbol "X". We note that most of the Southern regions dominates Northern regions, revealing that in the South not only the percentage of poor is higher but also the poor are poorer and more unequal than in the North of Italy.

Fig. 2 The three I's of poverty for the Italian regions, over time



References

1. Barrett, G. F., Donald, S. G., Hsu, Y.-C.: Consistent tests for poverty dominance relations. Journal of Econometrics,**191**, 360-373 (2016)
2. Berihuete, A., Ramos, C., Sordo, M.: Welfare, inequality and poverty analysis with rtip: An approach based on stochastic dominance. The R Journal **10**, 328-341 (2018)
3. Del Rio, C., Ruiz-Castillo, J.: TIPs for poverty analysis. The case of Spain, 1980-81 to 1990-91. Investigaciones Economicas **25**, 63-91 (2001)
4. Fourrier-Nicolai, E., Lubrano, M.: Bayesian inference for TIP curves: an application to child poverty in Germany. The Journal of Economic Inequality **18**, 91-111 (2020)
5. Jenkins, S. P., Lambert, P. J.: Three I's of poverty curves, with an analysis of UK poverty trends. Oxford economic papers **49**, 317-327 (1997)
6. Sen, A.: Poverty: an ordinal approach to measurement. Econometrica **44**, 219-231 (1976)
7. Thuysbaert, B.: Inference for the measurement of poverty in the presence of a stochastic weighting variable. The Journal of Economic Inequality **6**, 33-55 (2008)
8. Xu, K., Osberg, L.: A distribution-free test for deprivation dominance. Econometric Reviews **17**, 415-429 (1998)

Fig. 3 A comparison of the TIP curves across the Italian regions in 2015, and in Italy over time

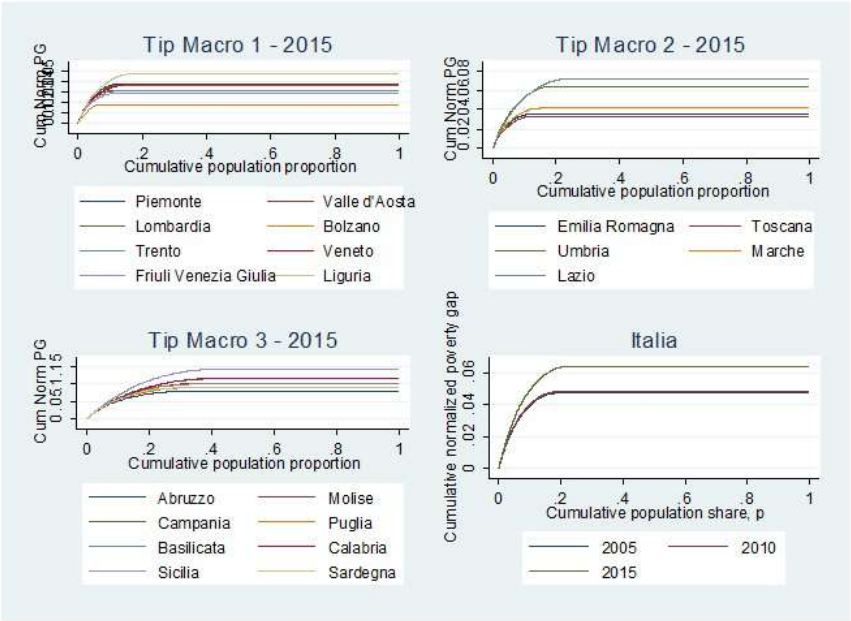


Table 1 Test of TIP dominance across the Italian regions, in year 2005

2005	10	20	30	41	42	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
10=Piemonte	-																				
20= Valle Aosta	<	-																			
30=Lombardia	>	>	-																		
41=Bolzano	<	>	<	-																	
42=Trento	<	>	<	<	-																
50=Veneto	<	>	<	X	>	-															
60=Friuli V.G.	<	>	<	>	>	X	-														
70=Liguria	>	>	>	>	>	>	>	-													
80=Emilia R.	>	>	X	>	>	>	>	<	-												
90=Toscana	<	X	<	<	>	<	<	<	<	-											
100=Umbria	<	X	<	>	>	X	X	<	X	>	-										
110=Marche	X	>	<	>	>	X	X	<	X	>	X	-									
120=Lazio	>	>	>	>	>	>	>	X	>	>	>	>	-								
130=Abruzzo	>	>	>	>	>	>	>	<	X	>	>	>	<	-							
140=Molise	>	>	>	>	>	>	>	>	>	>	>	>	X	>	-						
150=Campania	>	>	>	>	>	>	>	X	>	>	>	>	X	>	<	-					
160=Puglia	>	>	>	>	>	>	>	X	>	>	>	>	X	>	X	X	-				
170=Basilicata	X	X	X	X	>	X	X	<	X	>	>	X	<	X	<	<	<	-			
180=Calabria	>	>	>	>	>	>	>	X	>	>	>	>	>	>	X	>	>	>	-		
190=Sicilia	X	>	X	>	>	X	X	X	X	>	>	X	X	X	<	X	X	>	<	-	
200=Sardegna	>	>	X	>	>	>	>	X	>	>	>	X	X	X	X	X	X	>	<	>	-

Note: symbol ”>” means that region in the row dominates region in the column, ”<” means that region in the column dominates region in the row, symbol ”X” denotes no dominance.

4 Contributed Sessions

4.1 Advances in clinical trials

Quantitative depth-based [^{18}F]FMCH-avid lesion profiling in prostate cancer treatment

Profilazione quantitativa di lesioni in pazienti affetti da cancro alla prostata tramite misure di profondità

Lara Cavinato, Alessandra Ragni, Francesca Ieva, Martina Sollini, Francesco Bartoli, Paola A. Erba

Abstract Besides prognostic clinical factors, baseline risk assessment in personalized cancer research would benefit from quantitative disease characterization to inform therapy planning. Texture analysis of [^{18}F]FMCH PET/CT imaging is paving the way for such purposes but its potential is still braked by radiomic feature limitation, such as redundancy and lack of standardization. In this work, we provide a method for a robust assessment of intratumor heterogeneity in patients affected by prostate cancer, through a depth-based ranking quantifying the level of centrality/outlyingness of the lesion with respect to peers. We interpret the results in terms of clinical information of lesions.

Abstract *La terapia personalizzata per malattie tumorali si basa sull'individuare fattori prognostici, mirati a incasellare i pazienti in classi di rischio: ad oggi, oltre a considerare parametri clinici e biologici, si punta a inserire l'analisi quantita-*

Lara Cavinato
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, via Bonardi 9, Milan, Italy
e-mail: lara.cavinato@polimi.it

Alessandra Ragni
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, via Bonardi 9, Milan, Italy e-mail: alessandra.ragni@mail.polimi.it

Francesca Ieva
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, via Bonardi 9, Milan, Italy
CADS – Center for Analysis, Decision and Society, Human Technopole, Milan, Italy
e-mail: francesca.ieva@polimi.it

Martina Sollini
Humanitas University, Pieve Emanuele, Italy
Humanitas Clinical and Research Center, Rozzano, Italy

Francesco Bartoli
Regional Center of Nuclear Medicine, Department of Translational Research and New Technology in Medicine, University of Pisa, Pisa, Italy
Azienda Ospedaliero Universitaria Pisana, Pisa, Italy

Paola A. Erba
Regional Center of Nuclear Medicine, Department of Translational Research and New Technology in Medicine, University of Pisa, Pisa, Italy
Azienda Ospedaliero Universitaria Pisana, Pisa, Italy

tiva dei tessuti tumorali provenienti da dati di imaging, ad esempio la [^{18}F]FMCH PET/CT nel tumore alla prostata. La sfida è superare i problemi di ridondanza e standardizzazione che viziano questi dati non strutturati. Questo lavoro offre un metodo di valutazione robusta dell'eterogeneità intratumorale in ottica prognostica, tramite l'impiego di misure di profondità.

Key words: Depth Measures, PET/CT medical imaging, Precision medicine, Prostate cancer, Ranking, Radiomics

1 Contextual background

Prostate cancer is one of the leading causes of cancer death among men worldwide, with an estimate of over a million new cases of cancer and hundreds of thousands of deaths in 2018, a burden that is expected to grow in the upcoming years as population ages. Fortunately, death rate has been shown to have reduced in the past few years as extensive PSA-based screening programs have become available and been employed [1]. Indeed, treatment recommendation and risk factors currently relay upon patients' stratification, based on PSA, Gleason score, T-category which cluster men as low, intermediate and high risk patients who undergo increasingly aggressive treatments. Such therapies range from active surveillance to radiotherapy and radical prostatectomy.

Although most of the patients evolve well in long term [2], low-risk subjects may harbor more aggressive disease that remains undetected while resected patients may show occurrence of upgrading, upstaging or nodal metastases (i.e. biochemical progression). Consequently, risk stratification tools need to be improved and therapy pathways further optimized, in a personalized medicine fashion.

On the other hand, prostate cancer has been shown to exhibit spatial intratumor heterogeneity that can alter baseline risk assessment and behave as confounding factor in pre-treatment clinical-pathological prognosis [3]. Such knowledge could and should be exploited for improving treatment planning. Here comes the urge to assess and quantitively characterize intratumor heterogeneity in order to build an exhaustive representation of the disease. Indeed, the informed patient stratification will directly translate into improved patient treatments, wherein decisions regarding active surveillance or intensified therapy are made.

The role of [^{18}F]FMCH in patients with prostate cancer is well established, especially in ones with biochemical recurrence. Along with visual imaging inspection, the radiomic framework has spreaded in matter of quantitative PET/CT assessment, consisting on the extraction and evaluation of high dimensional advanced imaging features using high throughput methods [4]. Such features are referred as texture features and can be divided into conventional and higher-order parameters: more precisely, radiomic features include histogram-derived variables, shape-derived variables, GLCM matrix-derived variables, GLRLM matrix-derived variables, NGLDM matrix-derived variables and GLZLM matrix-derived variables. Each of these groups of variables is meant to capture distinct phenotypic differ-

ences of lesions resulting in quantitative measurements with potential prognostic power [5]]: lesions exhibiting similar radiomic profile are hypothesized to proxy similar physiologies and phenotypes, leading to similar outcomes. However, among radiomic features limitations, these variables suffer from redundancy and lack of standardization that prevent the radiomic workflow to significantly impact clinical practice.

In this work, we intend to propose a depth-based method for agnostic profiling of [^{18}F]FMCH-avid lesions in patients with recurrent prostate cancer, allowing a robust assessment of intratumor spatial heterogeneity. This could thus inform the yet unknown relationship between radiomic features and clinical outcomes, in terms of burden and biological aggressiveness.

2 Materials and Methods

92 patients (mean age 73 ± 7 years, median age 73 years, range 55-85) with multi-site, multi-lesion, recurrent prostate cancer have been retrospectively recruited (mean PSA at the time of [^{18}F]FMCH PET/CT 10,39 ng/ml) in the authors' institution. Clinical, biological and histology data as well as current treatment were recorded in all patients.

Whole-body PET/CT (GE Discovery ST) was acquired about 45 minutes after [^{18}F]FMCH (4 MBq/kg of body weight) administration. According to ISUP/WHO grading scale of prostate adenocarcinoma [6], patients were labeled as with mild and severe disease, having Gleason score ≤ 7 (n=53) and > 7 (n=31) respectively, except for 8 missing values.

A total of 370 lesions were found and classified according to TNM [14] in skeleton (n=221), distant lymph nodes (n=81) and regional lymph nodes (n=68). Lesions were semiautomatically segmented by experienced radiologists and radiomic texture features were extracted within regions of interest (lesions) using the LIFEx package (LIFEx website) [7]. Further statistical analysis has been implemented in R [8].

To overcome variable redundancy, the dataset was first filtered according to a correlation-based criterion. Specifically, Pearson pairwise correlation between radiomic variables were computed and highly correlated ($\geq$98%) variables were exclusively removed. No clinical rationale has been adopted in the choice of the variables to be kept, nevertheless attention has been payed to prefer conventional rather than higher-order variables for explainability reasons. Additionally, missing values of radiomic features have been filled according to median replacement rule. The resulting dataset variables has been standardized according to z-score method.

Data depths are a mathematical tool allowing for ranking multivariate objects, with respect to an underlying multivariate distribution [10]. They may be intended as the analogue of the quantiles for multivariate data: depth measures determine a centre-outward ordering of data points that are indeed geometrically ranked from the more central (median) to the more outsider (outlier) one [11].

Among data depths, several definitions are available in literature, such as Half-space (or Tukey) depth, Mahalanobis depth, Projection depth and Spatial depth. In

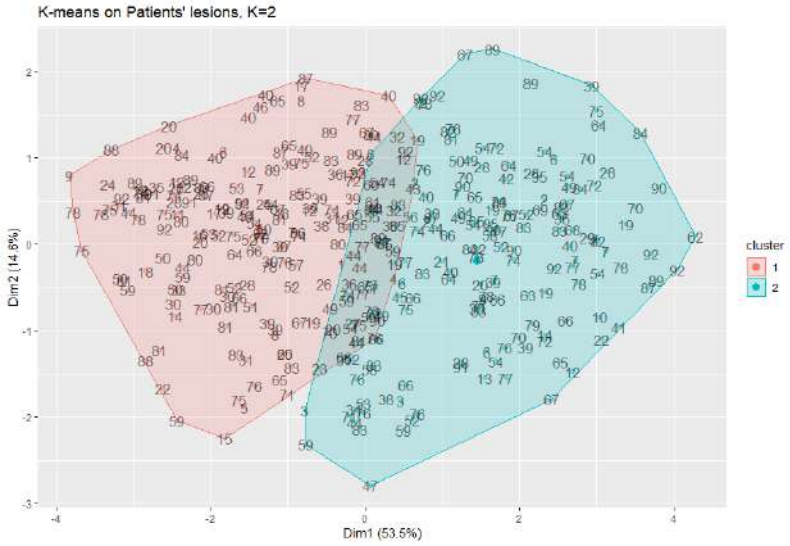


Fig. 1 Patients' lesions representation displayed in 2D latent space and labelled in two clusters.

the current analysis, Mahalanobis depth [9] is proposed as it leads to a *more spread out* distribution of points: in other words, the Mahalanobis depths distribution appears more dispersed about its center of symmetry than distributions stemming from other depths, leading to clearer results [12]. In this setting, Mahalanobis depth has been computed for each lesion, considering each group of radiomic variables separately. In this way, depth computation results in a variable reduction and transformation phase which allows to resume the 37 radiomic descriptors with 6 (one per semantic group of radiomic variables) agnostic ranking measures: indeed, the employment of a relative metric leads to overcome the lack of radiomic features standardization.

After that, the 370 6-dimensional vectors (one per lesion) were fed into a k-means clustering algorithm in order to find homogeneous classes of lesions, in an unsupervised setting. Optimal number of clusters was set equal to 2, according to exhaustive grid search implemented in [13]. Graphical inspection of such clusters is possible in Figure 1, where lesions, named with patient code, are plotted in 2D latent space and labelled in clusters. Evaluation was then performed in terms of clusters clinical characterization.

3 Results and Discussion

Kendall correlation coefficient was computed between each pair of the six depth measures, resulting in a 6x6 correlation matrix. According to Kendall, lesions appear not to show agreement between rankings: indeed, correlation coefficients are positive but never significant, with higher concordance for GLRLM and GLZLM depths, which show a correlation of 0.68. Being the depth of each lesion evalu-

ated for each radiomic group, the corresponding values depict the level of centrality of that lesion among the set of peers. Therefore, it is reasonable that concordance is not necessarily high for mainly two reasons: i) different radiomic groups capture different information about the lesion and the corresponding texture description, and ii) the lesion behaves differently over different descriptions provided by the groups. Therefore, such a dimensional reduction is able to globally capture different lesion's profiles, being agnostic in the way such diversity appears and may be evaluated. Ultimately, we end up with a six-dimensional depth vectors describing the radiomic lesion's profile (or fingerprint), and we focus on them in order to explore similarity patterns via unsupervised techniques. In fact, similar profiles were then grouped into homogeneous clusters, which can be interpreted as risk classes associated to different disease phenotypes and clinical outcomes. According to clusters characterization, as shown in table 1, membership to class 1 is coupled with no particular site, as 109 lesions can be found in skeleton and 81 in lymph nodes (36 distant + 45 regional); on the other hand, class 2 shows no prevalence for lesions to be in sites different from class 1 as well, with 112 of lesions being on skeleton and 68 on lymph nodes (45 distant + 23 regional). Of consequence, site may play a role in scoring disease severeness, however other factors may be further investigated to interpret the results.

For instance, along with lesion location, Standard Uptake Value (SUV) is assumed to differentiate malignant from benign processes. Correlated to conventional radiomic groups, SUV represents the lesion metabolic activity normalized over injected activity as highlighted by the tracer. Table 1 shows the number of lesions falling in the first $(-\infty,$-0.75], second (-0.75,-0.16], third (-0.16,0.58] and fourth $(0.58,\infty)$ SUV quartiles: accordingly, class 1 hosts lesions with SUV outlier values, since 125 lesions belong to first (59) and fourth (66) quartiles with respect to 65 lesions fitting in the second (32) and third (33) quartiles; on contrary, class 2 features lesions with SUV median values, as 119 lesions belong to second (60) and third (59) quartiles with the respect to 61 lesions fitting in the first (34) and fourth (27) quartiles. This assigns relevance to the tumor mutational burden, whose proxy is given by uptake values [16]: on one hand, average metabolically active lesions are labeled with class 2 risk score; on the other hand, both slightly and highly active lesions are given a distinct although unique risk score (class 1), as they might share higher-order descriptors' values.

Table 1 Characterization of clusters on the basis of clinical-physiological prognosticators.

	Lesions Site			**Lesions Standard Uptake Value (SUV)**			
	Regional Ln	Distant Ln	Skeleton	$(-\infty,$-0.75]	(-0.75,-0.16]	(-0.16,0.58]	$(0.58,\infty)$
Class 1	45	36	109	59	32	33	66
Class 2	23	45	112	34	60	59	27

4 Conclusion

Preliminary results showed that depth-based [^{18}F]FMCH-avid lesion profiling allows to overcome redundancy and lack of standardization issues in the radiomic framework. Such method could inform the investigation and the analysis of intratumor lesions heterogeneity, providing imaging biomarker for risk stratification to be proposed for a validation study to better characterize prostate cancer burden and biological aggressiveness, thus supporting imaging-based patients' treatment decision making.

Acknowledgements This work has been founded by AIRC IG 2017 Id. 20819 "Oligometastatic and Oligorecurrent Prostate Cancer: enhancing patients' selection by new imaging biomarkers".

References

1. Klotz, Laurence, et al. "Long-term follow-up of a large active surveillance cohort of patients with prostate cancer." J Clin oncol 33.3 (2015): 272-277.
2. Hayes, Richard B., et al. "Dietary factors and risks for prostate cancer among blacks and whites in the United States." Cancer Epidemiology and Prevention Biomarkers 8.1 (1999): 25-34.
3. Sala, E., et al. "Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging." Clinical radiology 72.1 (2017): 3-10.
4. Lambin, Philippe, et al. "Radiomics: extracting more information from medical images using advanced feature analysis." European journal of cancer 48.4 (2012): 441-446.
5. Aerts, Hugo JWL, et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." Nature communications 5.1 (2014): 1-9.
6. Humphrey, Peter A. "Gleason grading and prognostic factors in carcinoma of the prostate." Modern pathology 17.3 (2004): 292-306.
7. Nioche, C., et al. "LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity." Cancer research 78.16 (2018): 4786-4789.
8. Team, R. Core. "R: A language and environment for statistical computing." (2013): 201.
9. Pokotylo, Oleksii, Pavlo Mozharovskyi, and Rainer Dyckerhoff. "Depth and depth-based classification with R-package ddalpha." arXiv preprint arXiv:1608.04109 (2016).
10. Liu, Regina Y., Jesse M. Parelius, and Kesar Singh. "Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh)." The annals of statistics 27.3 (1999): 783-858.
11. Dong, Ye, and Stephen MS Lee. "Depth functions as measures of representativeness." Statistical Papers 55.4 (2014): 1079-1105.
12. Bickel, Peter J., and Erich L. Lehmann. "Descriptive statistics for nonparametric models IV. Spread." Selected Works of EL Lehmann. Springer, Boston, MA, 2012. 519-526.
13. Charrad, Malika, et al. "Determining the best number of clusters in a data set." J Stat Softw. (2014).
14. Brierley, James D., Mary K. Gospodarowicz, and Christian Wittekind, eds. "TNM classification of malignant tumours." John Wiley and Sons, 2017.
15. Gunderson, Leonard L., and Joel E. Tepper. Clinical radiation oncology. Elsevier Health Sciences, 2015.
16. Jahromi, Amin Haghighat, et al. "Relationship between tumor mutational burden and maximum standardized uptake value in 2-[18 F] FDG PET (positron emission tomography) scan in cancer patients." EJNMMI research 10.1 (2020): 1-7.

Modelling longitudinal latent toxicity profiles evolution in osteosarcoma patients

Modellazione dell'evoluzione di profili longitudinali latenti di tossicità in pazienti con osteosarcoma

Marta Spreafico, Francesca Ieva and Marta Fiocco

Abstract In cancer trials, the analysis of longitudinal chemotherapy data is a difficult task due to the complex registration and evolution of toxicity levels during treatment. Models to deal with both the longitudinal and the categorical aspects of toxicity level progression are necessary, still not well developed. In this work, a Latent Transition Analysis (LTA) procedure to identify and reconstruct the longitudinal latent profiles of toxicity evolution of each patient over time is proposed. The latent variables determining the progression of the observed toxicity levels can be thought of as the outcomes of an underlying latent process. This methodology has never been applied to osteosarcoma treatment and provides new insights for supporting decisions in childhood cancer therapy.

Abstract *Negli studi sul cancro, analizzare i dati longitudinali di chemioterapia è problematico a causa della complessa evoluzione dei livelli di tossicità durante il trattamento. Modelli in grado di tenere conto sia degli aspetti longitudinali che di quelli categoriali dell'evoluzione dei livelli delle tossicità sono necessari, ma non ancora ben sviluppati. In questo lavoro viene proposta una procedura di analisi a transizioni latenti per identificare e ricostruire i profili latenti longitudinali relativi all'evoluzione delle tossicità di ogni paziente nel tempo. Le variabili latenti che determina l'evoluzione dei livelli di tossicità osservati possono essere pensate come i risultati di un processo latente sottosante. Questo approccio non è mai stato applicato al trattamento dell'osteosarcoma e fornisce nuove intuizioni per lo studio del cancro infantile.*

Key words: latent markov models, latent transition analysis, longitudinal data, toxicity, osteosarcoma

Marta Spreafico[1,2,3] Francesca Ieva[1,3,4] Marta Fiocco[2,5,6]
[1]MOX – Department of Mathematics, Politecnico di Milano, Milan 20133, Italy
[2] Mathematical Institute, Leiden University, Leiden, The Netherlands
[3]CHRP, National Center for Healthcare Research and Pharmacoepidemiology, Milan 20126, Italy
[4]CADS, Center for Analysis Decisions and Society, Human Technopole, Milan 20157, Italy
[5] Dept. of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[6] Trial and Data Center, Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands
e-mail: marta.spreafico@polimi.it francesca.ieva@polimi.it m.fiocco@math.leidenuniv.nl

1 Introduction

In many clinical applications involving longitudinal data, the interest lies in the analysis of the evolution of similar latent profiles related to subgroups of individuals rather than in the study of their observed attributes [1, 2]. These latent characteristics may reflect patients' quality-of-life, including valuable information related to patient's health status and disease progression.

In cancer trials, the analysis of longitudinal chemotherapy data is a complex task due to the presence of negative feedbacks between exposure to cytotoxic drugs and the toxicities the latter provoke. Toxic adverse events are at the same time risk factors for mortality and predictors of future exposure levels, representing time-dependent confounders for the effect of chemotherapy on patient's status [3]. Toxicity data are usually considered in very simplistic ways in cancer studies, where they act as fixed covariate over treatment [4, 5], discarding substantial amount of information (e.g., isolated vs repeated events, single vs multiple episodes, toxic events timing). Methods for longitudinal adverse events have also been proposed [4, 5] but they improperly treated toxicity levels as numerical values, as a simplifying hypothesis due to the complexity of the problem. Indeed, since multiple types of adverse events with different extents of toxicity burden occur simultaneously, studying the toxicity evolution during treatment is a challenging problem in cancer research.

In this work, a novel procedure based on *Latent Transition Analysis* (LTA) [1], a special case of first-order *Latent Markov* (LM) *models for longitudinal data* [2], is proposed to identify and reconstruct the longitudinal latent profiles of toxicity evolution. LM models for longitudinal data have been successfully applied in several fields, such as social, economic and behavioural sciences, education and public health, criminology or marketing [1, 2]. Clinical examples include, among others, the evolution of psycho-physic conditions in elderly individuals, the course of emotions among anorectic patients, or the analysis of pneumococcal carriage in children to study interactions between co-colonizing serotypes [2]. The idea behind this approach is that the latent variables can be thought of as the outcomes of a latent process which determines the evolution of the observed toxicity levels. This approach properly allows to take into account both the longitudinal and categorical aspects of toxicity level progression and the corresponding toxic risk evolution in oncology.

This approach has never been applied to osteosarcoma treatment and provides new insights for childhood cancer therapy. The presented procedure is really flexible and appropriate to analyse cancer chemotherapy treatment in general. Data from the MRC BO06/EORTC 80931 randomized controlled trial for osteosarcoma [6], a malignant bone tumour mainly affecting children and young adults, are analysed.

2 Methods

Let us consider a set $\mathscr{J}$ of categorical response variables measured at $t = 1, \ldots, T$ occasions, with $J = |\mathscr{J}|$. For each $j = 1, \ldots, J$, let $Y_j^{(t)}$ denote the j-th response variable at time t, with set of possible categories $\mathscr{C}_j$. Let $\widetilde{\boldsymbol{Y}} = (\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)})$ be the

complete response vector, where $\boldsymbol{Y}^{(t)}$ is the observed multivariate response vector at time t. The general Latent Transition Analysis (LTA) formulation is a Latent Markov (LM) model that assumes the existence of a latent process which affects the distribution of the response variables. This latent process, denoted by $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(T)})$, follows a first-order Markov chain with state space $\{1, \ldots, k\}$, where the number of latent states k can be a priori defined or selected according to the Bayesian information criterion (BIC). Three different sets of model parameters $\boldsymbol{\theta}$ can be defined:

- the **item-response probability** $\phi^{(t)}_{jy|u}$, i.e., the probability of a particular observed response y on variable j at time t, conditional on latent class u membership:

$$\phi^{(t)}_{jy|u} = P(Y^{(t)}_j = y \big| U^{(t)} = u) \qquad y \in \mathscr{C}_j \quad j = 1, \ldots, J \quad u \in \{1, \ldots, k\};$$

- the **initial latent status prevalence** δ_u, i.e., the probability of membership in latent state u at time $t = 1$:

$$\delta_u = P(U^{(1)} = u) \qquad u \in \{1, \ldots, k\};$$

- the **transition probability** $\tau^{(t)}_{u|\bar{u}}$, i.e., the probability of a transition to latent state u at time t, conditional on membership in latent state $\bar{u}$ at time $t-1$:

$$\tau^{(t)}_{u|\bar{u}} = P(U^{(t)} = u \big| U^{(t-1)} = \bar{u}) \qquad t = 2, \ldots, T \quad u, \bar{u} \in \{1, \ldots, k\}.$$

Assuming *local independence*, i.e., the observed variables are independent conditional on the latent class, the *manifest distribution* of the response variables is:

$$P(\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{y}}) = \sum_{\boldsymbol{u}} P(\boldsymbol{U} = \boldsymbol{u}) \times P(\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{y}} \big| \boldsymbol{U} = \boldsymbol{u}) = \sum_{\boldsymbol{u}} \delta_{u^{(1)}} \prod_{t=2}^{T} \tau^{(t)}_{u^{(t)}|u^{(t-1)}} \times \prod_{t=1}^{T} \prod_{j=1}^{J} \phi^{(t)}_{jy^{(t)}_j|u^{(t)}}$$

where $\widetilde{\boldsymbol{y}}$ is a realization of $\widetilde{\boldsymbol{Y}}$ made by the subvectors $(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(T)})$, $\boldsymbol{y}^{(t)}$ is a realization of $\boldsymbol{Y}^{(t)}$ with elements $y^{(t)}_j$ and $\boldsymbol{u} = (u^{(1)}, \ldots, u^{(T)})$.
Parameters estimation $\hat{\boldsymbol{\theta}}$ is performed maximizing the log-likelihood for a sample of n independent units, i.e., $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log P(\widetilde{\boldsymbol{Y}}_i = \widetilde{\boldsymbol{y}}_i)$, using an Expectation-Maximization (EM) algorithm. For further details see [2, 7].

The EM algorithm also provides the estimated *posterior* probabilities of $U^{(t)}$ [2], which can be used to reconstruct the ***Longitudinal Latent Probability Profile*** (LLPP) for each latent state $u \in \{1, \ldots, k\}$ and subject $i \in \{1, ..., n\}$, as follows:

$$\boldsymbol{p}_{u,i} = \left\{ p^{(t)}_{u,i} = P\left(U^{(t)} = u \big| \widetilde{\boldsymbol{Y}}_i = \widetilde{\boldsymbol{y}}_i\right), \quad t = 1, \ldots, T \right\} \tag{1}$$

LLPP in Eq. (1) represents the probability over time t of being in latent state u for individual i, given the observed response $\widetilde{\boldsymbol{y}}_i$. Applying this procedure, a k-variate longitudinal latent profile $\left(\boldsymbol{p}_{1,i}, \ldots, \boldsymbol{p}_{k,i}\right)$ such that $\sum_u p^{(t)}_{u,i} = 1$ for each $t = 1, \ldots, T$ can be reconstructed for each subject i.

3 Data and Patients

Data from MRC BO06/EORTC 80931 randomised controlled trial for patients with non-metastatic high-grade osteosarcoma [6] were analysed. Patients were randomized at baseline between Conventional (*Reg-C*) or Dose-Intense (*Reg-DI*) regimens of six cycles of chemotherapy, with identical anticipated cumulative dose but different duration. Non-haematological chemotherapy-induced toxicity for nausea/vomiting (*naus*), infection (inf), mucositis (*oral*), cardiac toxicity (*car*), ototoxicity (oto) and neurological toxicity (*neur*) were registered at each cycle and graded according to the Common Terminology Criteria for Adverse Events (CTCAE) v3.0 [8], with grades ranging from 0 (none) to 4 (life-threatening). Additional details can be found in the primary analysis of the trial [6].

In the study cohort $n = 377$ patients that completed the chemotherapy within 180 days from randomization were included. Nausea/vomiting was reported at least once over cycles in 97.3% of patients (367/377), with a percentage that decreased over cycles from 84.9% (327/377) in cycle 1 to 52.5% (198/377) in cycle 6. The percentages of patients that reported oral mucositis or infections were more stable over cycles: 30.5%-43.3% for mucositis and 23.8%-31.3% for infection. Other toxicities were less frequent ($<10\%$), especially for grades above 1.

4 Application and Results

Latent Markov (LM) model formulation presented in Sect. 2 is now applied to chemotherapy-induced longitudinal categorical toxicity data presented in Sect. 3, as shown in Fig. 1. For each cycle $t = 1,\ldots,6$, let $\mathscr{J} = \{naus, inf, oral, car, oto, neur\}$ be the set of categorical response variables $Y_j^{(t)}$ with possible sets of response categories $\mathscr{C}_j = \{0 : none, 1 : mild, 2 : moderate, 3/4 : severe\}$ for $j = 1,2,3$ and $\mathscr{C}_j = \{0 : none, 1 : mild, 2/3/4 : mod/sev\}$ for $j = 4,5,6$. The item-response probabilities were assumed to be time homogeneous, i.e. $\phi_{jy|u}^{(t)} = \phi_{jy|u} \forall t$, that is a common parameters restriction in latent transition analysis [1]. Due to the multimodality of the log-likelihood function $\ell(\boldsymbol{\theta})$, different random initializations of EM algorithm were used and the final estimate $\hat{\boldsymbol{\theta}}$ was the one corresponding to the highest $\ell(\boldsymbol{\theta})$. Statistical analyses were performed using the `R` package `LMest` [7].

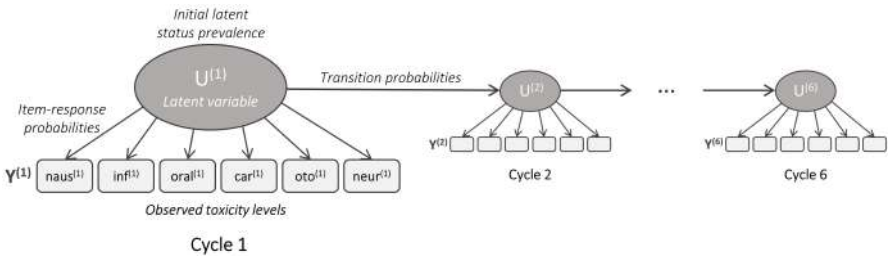


Fig. 1 Structure of the Latent Markov (LM) model for longitudinal data of toxicity levels.

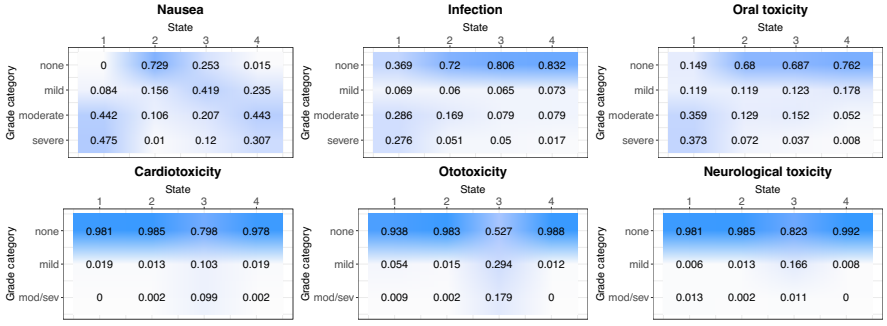


Fig. 2 Estimated item-response probabilities. Each panel refers to a different toxicity variable in $\mathscr{J} = \{naus, inf, oral, car, oto, neur\}$ with grade response categories $\{0 : none, 1 : mild, 2 : moderate, 3/4 : severe\}$ for $\{naus, inf, oral\}$ and $\{0 : none, 1 : mild, 2/3/4 : mod/sev\}$ for $\{car, oto, neur\}$.

According to minimum BIC, the selected number of latent states was $k = 4$. Fig. 2 shows the estimated item-response probabilities $\hat{\phi}_{jy|u}$ for each toxicity, which provided the basis for the interpretation of the latent states. From these results the following latent state labelling was derived:

- State 1 *Severe/Moderate a-specific toxic state*
- State 2 *Non-toxic state*
- State 3 *Mild nausea (with possible specific toxicity)*
- State 4 *Severe/Moderate nausea only*.

The estimated initial latent status prevalence showed that the prevailing state at time $t = 1$ was *Severe/Moderate nausea only* (61.8%), followed by *Non-toxic state* (19.5%), *Severe/Moderate a-specific toxic state* (12.1%) and *Mild nausea* (6.5%). In particular, the prevalence of *Severe/Moderate nausea only* decreased over cycles from 61.8% to 20.6% ($t = 6$), whereas the ones of *Non-toxic state* and *Mild nausea* increased from 19.5% to 49.3% and from 6.5% to 19.2%, respectively. Latent state prevalence of *Severe/Moderate a-specific toxic state* was more stable over cycles ranging in 10.8%-16.5%, with peaks in cycles 2 and 3.

Longitudinal latent probability profiles (LLPPs) were finally reconstructed according to Eq. (1). Fig. 3 shows the LLPPs $\boldsymbol{p}_{u,i}$ related to each latent state $u = \{1,2,3,4\}$ for seven random patients. LLPPs are able to capture the individual realisations of the latent process over cycles through a customized reconstruction, showing different patterns of toxicity evolution over treatment among patients.

5 Conclusion

The proposed approach allowed (i) to move from complex chemotherapy data to a set of different subgroups of individuals that exhibit similar patterns of toxicity grades progression over cycles by identifying the latent states and (ii) to reconstruct

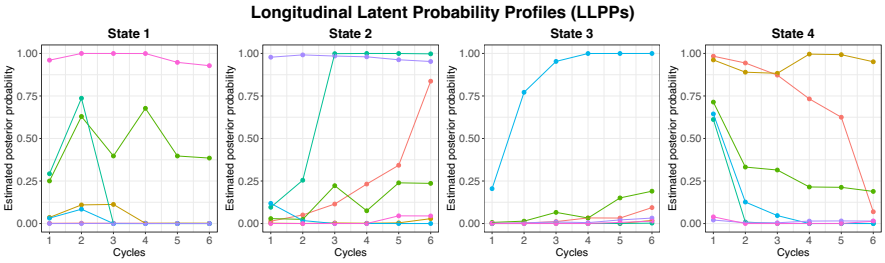


Fig. 3 Reconstructed longitudinal latent probability profiles $\boldsymbol{p}_{u,i}$ for seven random patients. Each panel refers to a different state $u = \{1,2,3,4\}$. Different colours refer to different patients.

and provide Longitudinal Latent Probability Profiles (LLPPs) related to toxicity evolution in a tailored way. This procedure represents a novelty for osteosarcoma treatment and, more generally, for cancer studies, providing new insights for childhood therapy.

This work opens doors for many further developments. The LM model could be enriched by considering other relevant clinical information as adjusting covariates. Moreover, it could be of clinical interest to study the association between LLPPs and time-to-event outcomes. This is a non-trivial modelling task, representing a challenging problem both for clinical and statistical research.

Acknowledgements The authors thank Medical Research Council for sharing the dataset used in this work.

References

1. Collins, L.M., Lanza, S.T.: Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. John Wiley and Sons Inc (2010).
2. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. Chapman & Hall/CRC, Boca Raton (2013).
3. Lancia, C. et al.: Marginal structural models with dose-delay joint-exposure for assessing variations to chemotherapy intensity. Stat. Methods Med. Res. **28**(9), 2787–2801 (2019).
4. Trotti, A. et. al.: TAME: development of a new method for summarising adverse events of cancer treatment by the Radiation Therapy Oncology Group. Lancet Oncol. **8**, 613–24 (2007).
5. Thanarajasingam, G. et al.: Longitudinal adverse event assessment in oncology clinical trials: the Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254. Lancet Oncol. **17**(5), 663–670 (2016).
6. Lewis, I.J. et al.: Improvement in Histologic Response But Not Survival in Osteosarcoma Patients Treated With Intensified Chemotherapy: A Randomized Phase III Trial of the European Osteosarcoma Intergroup. J. Natl. Cancer Inst. **99**(2), 112–128 (2007).
7. Bartolucci, F., Pandolfi, S., Pennoni, F.: LMest: An R Package for Latent Markov Models for Longitudinal Categorical Data. J. Stat. Soft. **81**(4), 1–38 (2017).
8. U.S. Department of Health and Human Services: Common Terminology Criteria for Adverse Events v3.0 (CTCAE). (2006) URL: https://www.eortc.be/services/doc/ctc/ctcaev3.pdf

Information borrowing in phase II basket trials: a comparison of different designs

Information borrowing negli studi clinici di fase II: un confronto tra differenti approcci

Marco Novelli

Abstract Traditional phase II oncology clinical trials are focussed on a single treatment for a subgroup of patients with specific histological characteristics. Recently, the advances in cancer biology and genomic medicine have led to the flourish of new molecularly targeted therapies and to a re-thinking of the design of phase II trials assessing clinical activity. In this work, we evaluate the potential benefit of information borrowing across tumor subgroups in basket trials by comparing the operating characteristics of the models presented in the literature. In addition, we suggest a modification of the model proposed in [2] which exhibits good performances in terms of both the average number of patients treated and the ability to preserve the type I error rate.

Abstract *I tradizionali studi clinici di fase II si focalizzavano su di un singolo trattamento specifico per un sottogruppo di pazienti con determinate caratteristiche istologiche. Recentemente, i progressi della medicina hanno portato allo sviluppo di nuove terapie molecolari mirate e ad una riprogettazione degli studi di fase II che valutano l'efficacia. In questo lavoro, si studiano i potenziali vantaggi della 'information borrowing' tra sottogruppi tumorali confrontando le caratteristiche operative dei modelli presentati in letteratura e si suggerisce una possibile modifica del modello presentato in [2] che mostra buone prestazioni sia in termini di numero medio di pazienti trattati che nella capacità di preservare l'errore di prima specie.*

Key words: Adaptive borrowing, Bayesian hierarchical models, clinical trials, heterogeneity, targeted therapy, type I error

Marco Novelli
Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, Bologna
e-mail: marco.novelli4@unibo.it

1 Introduction

One of the most difficult challenges in cancer research consists in dealing with the heterogeneity of patients. Until recently, the tumor histology has been considered the primary determinant of treatment effectiveness and, hence, the development of new cancer drugs has been conducted independently for each different histological type [4, 5, 7]. As a consequence, traditional phase II oncology clinical trials have been designed to focus on a specific single treatment for a subgroup of patients with particular molecular and/or histological characteristics. The recent advances in cancer biology and genomic medicine have led to a paradigm shift toward therapies which match patients with particular genetic or molecular aberration with molecularly targeted treatments [5]. A plethora of new clinical trial designs for biomarker-based cancer treatment development has flourished: basket, umbrella and platform trials (for a review see e.g., [7]). In this work, we will focus on the recent advances in basket trial designs, i.e., clinical trials where patients sharing the same molecular aberration but with different tumor histologies are placed in a common 'basket' and treated with a common therapy.

Although the patients enrolled in the trial share the same molecular aberration, their response to the treatment may still greatly vary depending on the tumor type. In such a situation, the two simplest and most intuitive approaches are either to conduct a single pooled analysis, by collecting all subgroups together, or to treat each arm independently. Clearly, both approaches suffers from drawbacks. A pooled analysis might not be able to distinguish the heterogeneity of the treatment effect across the subgroups, leading to inflated type I error and biased estimates. Treating each group independently instead, may not reach the required statistical power for reliably detecting the treatment efficacy, due to the limited sample size. In order to circumvent such limitations, a possible middle ground solution consists in 'sharing information' across different subgroups to improve inferential precision. To do so, Thall et al in [8] presented a Bayesian hierarchical model (BHM), able to adaptively borrow information across subgroups. This model has been recently adapted to basket trials in [1]. The main idea is to control the degree of information borrowing through a shrinkage parameter that is directly estimated from the data, in such a way that, as the trial unfolds, the amount of shared information is automatically adjusted. If the treatment is effective in all the subgroups (namely, the effect is homogeneous) the BHM borrows information and shrinks the estimate of the treatment effect in each group toward the overall mean, so as to gain statistical power. If instead the treatment turns out to be effective for some subgroups but not for others (i.e., the effect is heterogeneous) the BHM should not borrow information in order to preserve the type I error rate.

Since its introduction, the BHM has been greatly discussed: albeit conceptually appealing, the borrowing information approach suffers from a serious drawback related to the scarce available information to reliably estimate the shrinkage parameter. Unless the number of subgroups is 10 or greater, the information on the shrinkage parameter is very limited, leading to inflated type I error and to biased estimates of treatment effects [3]. This problem stems from the fact that the shrinkage parame-

ter represents the variance between subgroups, hence the observation units are the tumor subgroups, not patients. This means that if the number of subgroups is small, BHM is not able to reliably estimate the degree of information borrowing, even in the presence of large sample sizes [2, 3]. Such a problem is particularly relevant in the basket trial context where only a small/moderate number of subgroups (e.g., 3-10) is generally considered [2, 3]. As a possible solution, in [2] the authors propose a new calibrated Bayesian hierarchical model (CBHM) able to manage the borrowing strength by accounting for the homogeneity/heterogeneity among cancer types. Specifically, instead of a fully Bayesian approach, the authors propose to define the degree of information borrowing as a function of a similarity measure of the treatment effectiveness across subgroups. Once the function is properly calibrated, such an approach allows for a correct specification of the degree of information borrowing, while generally preserving the nominal type I error rate.

The aim of the present work is to extend the results of [3] through an extensive comparison of several models in order to better understand the role of information borrowing and its potential usefulness in basket trials. By introducing a different functional specification of the link between the homogeneity measure and the degree of information borrowing, we also show that the performance of the CBHM can be substantially improved.

2 A modification of CBHM

Following the notation in [2], let us consider a basket trial aimed at evaluating the effectiveness of a treatment in J different tumor subgroups by testing

$$H_0 : p_j \leq p_u \quad \text{vs} \quad H_1 : p_j \geq p_a, \ j = 1, \ldots, J$$

where $p_j (j = 1, \ldots, J)$ represents the response rate of group j and p_u and p_a are the cutoffs under which the drug is considered futile or promising, respectively. After $n_j (j = 1, \ldots, J)$ patients have been enrolled in each subgroup, an interim go/no-go analysis takes place. The number of patients in group j who responded favorably to the treatment, namely y_j, is assumed to follow a hierarchical model

$$\begin{aligned} y_j | p_j, n_j &\sim Bin(p_j, n_j) \\ \theta_j &= \log\left(\frac{p_j}{1-p_j}\right) \\ \theta_j | \theta, \sigma^2 &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\alpha_0, \omega_0) \end{aligned}$$

where θ_j is the log-odds of response in group j, α_0, ω_0 are hyperparameters and σ^2 is the parameter controlling the degree of shrinkage. In particular, a small (large) value of σ^2 induces strong (weak) information borrowing. In [2], the authors propose to use a chi-squared test statistic of homogeneity T as a measure of homogene-

ity/heterogeneity across groups, and to link σ^2 with T via a monotonically increasing function, i.e., $\sigma^2 = g(T)$. Through a simulation-based procedure, the function is then calibrated in such a way that strong information borrowing is enforced when the treatment effect is homogeneous, while little or none shrinkage occurs in the case of heterogeneous treatment effect.

Although the CBHM have shown fairly good performances, when the treatment effect is highly heterogeneous it may still suffer from inflated type I error rates [2]. In this work, we suggest a modified version of the function $g(\cdot)$ linking σ^2 with T that turns out to better preserve the type I error rate, while maintaining good operating characteristics. The proposed link function is defined as

$$\sigma^2 = \exp\{a + b\sqrt{T}\} - 1 \tag{1}$$

where the parameters a and b have to be calibrated through a procedure similar to that discussed in [2]. In what follows, we will refer to this model as the modified CBHM, i.e. mCBHM. Note that, by combining the functional form in (1) with the calibration strategy, the mCBHM can achieve full information borrowing when the treatment effect is homogeneous, and no information borrowing when the treatment is instead heterogeneous.

3 Numerical experiments

In this section, the operating characteristics of five different designs are compared, namely the Simon's Optimal Two-Stage design (STS) introduced in [6], the Bayesian independent model (BIM) and the BHM in [1], the CBHM in [2] and its modified version mCBHM presented in the previous section. We will consider a trial with $J = 5$ subgroups with a maximum sample size of 30 patients for each tumor type and 2 go/no-go interim analyses, after 10 and 20 patients enrolled respectively. Following [3], a response rate of 10% has been considered uninteresting, whereas a response rate of 30% is deemed promising. To facilitate comparisons, we set both type I and type II error rates equal to 0.10 for STS and each Bayesian model has been calibrated in order to have a type I error rate of 10% when all the treatment effects are equal to $p_u = 0.10$. All the remaining parameter values are set according to those reported in [2]. Three scenarios with various response rates are considered. Monte Carlo simulation is used to obtain 5,000 trials for each model and for each scenario. Table 1 summarizes the operating characteristics of the designs: rejection rates of the null hypothesis are displayed along with the average sample size of the trial, while the estimated response rates are in brackets.

In scenario 1, the treatment is not effective in the second and the third subgroup, only. All the designs show high power in detecting the treatment efficacy in subgroups 1, 4 and 5; the BHM and CBHM display inflated type I error. This is particularly evident for BHM which shows a value of 40%. The BHM also shrinks the estimated response rates to the overall mean: the values for the intermediate sub-

groups increase to 0.15 and those of the other subgroups register a decrease. The BHM enrolls the highest number of patients, followed by the Simon's design, and by the BIM and the CBHM, that have a similar performance. The mCBHM has the smallest average sample size.

Scenario	Design	Subgroup					Sample size
		1	2	3	4	5	
1		$p_1 = 0.35$	$p_2 = 0.1$	$p_3 = 0.1$	$p_4 = 0.35$	$p_5 = 0.4$	
	STS	0.95 (0.34)	0.10 (0.08)	0.10 (0.08)	0.95 (0.34)	0.98 (0.39)	142.3
	BIM	0.97 (0.35)	0.10 (0.08)	0.09 (0.08)	0.97 (0.34)	0.99 (0.40)	132.8
	BHM	0.99 (0.32)	0.40 (0.15)	0.40 (0.15)	0.99 (0.32)	1.00 (0.36)	146.7
	CBHM	0.98 (0.35)	0.16 (0.08)	0.16 (0.08)	0.98 (0.34)	0.99 (0.40)	132.9
	mCBHM	0.97 (0.35)	0.09 (0.08)	0.09 (0.08)	0.98 (0.34)	1.00 (0.40)	131.7
2		$p_1 = 0.3$	$p_2 = 0.3$	$p_3 = 0.3$	$p_4 = 0.45$	$p_5 = 0.1$	
	STS	0.89 (0.28)	0.91 (0.29)	0.90 (0.29)	0.99 (0.45)	0.10 (0.08)	153.8
	BIM	0.92 (0.29)	0.92 (0.29)	0.92 (0.29)	1.00 (0.45)	0.09 (0.08)	139.9
	BHM	0.98 (0.29)	0.98 (0.29)	0.98 (0.29)	1.00 (0.38)	0.58 (0.19)	149.2
	CBHM	0.95 (0.29)	0.94 (0.29)	0.95 (0.29)	1.00 (0.45)	0.16 (0.08)	140.0
	mCBHM	0.93 (0.29)	0.92 (0.29)	0.93 (0.29)	1.00 (0.44)	0.09 (0.08)	139.1
3		$p_1 = 0.3$	$p_2 = 0.15$	$p_3 = 0.1$	$p_4 = 0.2$	$p_5 = 0.3$	
	STS	0.90 (0.29)	0.33 (0.12)	0.10 (0.08)	0.60 (0.17)	0.90 (0.29)	139.2
	BIM	0.92 (0.29)	0.32 (0.13)	0.09 (0.08)	0.57 (0.18)	0.92 (0.29)	133.4
	BHM	0.96 (0.25)	0.75 (0.18)	0.60 (0.16)	0.86 (0.20)	0.96 (0.25)	144.9
	CBHM	0.94 (0.29)	0.44 (0.13)	0.17 (0.08)	0.68 (0.18)	0.94 (0.29)	133.1
	mCBHM	0.91 (0.28)	0.28 (0.13)	0.11 (0.09)	0.53 (0.18)	0.92 (0.28)	128.3

Table 1 Operating characteristics of the designs. Rejection rate of the null hypothesis, in brackets the estimated response rates.

In scenario 2, all the subgroups except for the last one are responsive. Here the over-shrinking effect of the BHM is even more pronounced with an inflated type I error up to 58% and 10 more patients enrolled with respect to BIM, CBHM and its modified version. The Simon's design preserves the type I and II error rates at the expenses of a higher average sample size. The mCBHM shows similar performance to those of BIM, preserves the type I error and has the smallest number of treated patients.

In the last scenario, the treatment effect is highly heterogeneous across subgroups, namely it is effective in group 1 and 5, not effective in group 3, group 2 and 4 have a response rate that is in between p_u and p_a. The BHM presents the same critical issues observed in the previous scenarios: highly inflated type I error (up to 60%) and biased estimates of the response rates; for example, that of the first group decreases from 0.30 to 0.25 and that of the third increases from 0.10 to 0.16. The CBHM exhibits good performance in terms of power with an inflated type I

error for the third group (17%). The STS and BIM designs perform quite similarly, but the latter requires on average smaller sample size (139.2 vs 133.4). The power of the mCBHM is comparable with those of the Simon's and the independent design, though with a loss up to $4-5\%$ with respect to the BHM. The modified calibrated model displays a slightly inflated type I error (11%) for the third group but requires the smallest sample sizes, with about 5 and 16 fewer patients with respect to the CBHM and the BHM, respectively. Note that the improvement of the mCBHM with respect to its original formulation is particularly evident in this complex scenario since it is able to both i) maintain a lower rejection rate for those subgroups that have a response rate in between the two desired thresholds and ii) to enroll fewer patients.

As stressed in [3], in phase II trials strong emphasis should be placed in controlling the false-positive error rate since it minimizes the number of negative phase III trials by screening out treatments that are ineffective. Our results confirm the crucial consequences that the heterogeneity of treatment effect has in conducting a solid statistical analysis. Indeed, when only few subgroups are responsive to the treatment, adopting a strong information borrowing approach severely undermines the statistical reliability of the analysis, also increasing the average number of patients treated. In such situations, the mCBHM seems to show performances similar to those of the independent approach while requiring smaller sample sizes at the same time. Future research should explore more the usefulness of sharing information across subgroups, in particular in the case where some response rates are in between the two considered thresholds. Moreover, although the discussed designs include the possibility to stop the treatment in one or more subgroups at the interim analysis, a proper study of how different stopping rules affect the operating characteristics of the designs is still an open problem.

References

1. Berry, S. M., Broglio, K. R., Groshen, S., Berry, D. A.: Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clin Trials*, 10(5), 720-734 (2013)
2. Chu, Y., Yuan, Y.: A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clin Trials* 15.2, 149-158 (2018).
3. Freidlin, B., Korn E. L.: Borrowing information across subgroups in phase II trials: is it useful?. *Clin. Cancer Res.* 19.6, 1326-1334 (2013).
4. Renfro, L. A., Ming-Wen A., Mandrekar S. J.: Precision oncology: a new era of cancer clinical trials. *Cancer Lett.* 387, 121-126 (2017).
5. Simon, R., Roychowdhury, S.: Implementing personalized cancer genomics in clinical trials. *Nat. Rev. Drug Discov.* 12.5, 358-369 (2013).
6. Simon, R.: Optimal two-stage designs for phase II clinical trials. *Control. Clin. Trials*, 10(1), 1-10 (1989).
7. Simon, R.: Critical review of umbrella, basket, and platform designs for oncology clinical trials. *Clin Pharmacol Ther* 102.6, 934-941 (2017).
8. Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., Benjamin, R. S. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Stat. Med.*, 22(5), 763-780 (2003).

Q-learning Estimation Techniques for Dynamic Treatment Regime

Q-learning per Problemi di Trattamento Dinamico

Simone Bogni and Debora Slanzi and Matteo Borrotti

Abstract Optimal individualized treatment estimation in single or multi-stage clinical trials is a breakthrough for personalized medicine. In this context, statistical methodologies can significantly improve the estimation over model-based methods. Furthermore it can help in selecting important variables in high-dimensional settings. In this work, we investigate the performance of an hybrid approach that combine Sequential Advantage Selection (SAS) method and Q-learning. In addition, Q-functions are estimated with linear regression model, random forest and neural network.

Abstract *L'individuazione del trattamento ottimale in studi clinici è un importante sviluppo della medicina personalizzata. In questo contesto, le metodologie statistiche possono contribuire a migliorare gli attuali approcci per la stima. Inoltre possono risolvere il problema della selezione delle variabili in contesti caratterizzati da alta dimensionalità. In questo lavoro, vengono analizzate le performance di un approccio sviluppato combinando le peculiarità della Sequential Advantage Selection (SAS) e del Q-learning. Inoltre, le funzioni Q sono stimate utilizzando regressione lineare, random forest e reti neurali.*

Key words: dynamic treatment regime, Q-learning, sequential advantage selection, non-linear models

Simone Bogni
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano (MI), e-mail: s.bogni@campus.unimib.it

Debora Slanzi
Ca' Foscari University of Venice, Dorsoduro, 3246, 30123 Venezia (VE), e-mail: debora.slanzi@unive.it

Matteo Borrotti
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano (MI), e-mail: matteo.borrotti@.unimib.it

1 Introduction

Personalized medicine refers to the tailoring of a medical treatment focusing on the patients based on their individual demographic, clinical and genetic characteristics [5]. In simple words, personalized medicine is an opportunity to take a *one-size-fits-all* approach to diagnostics and drug therapy and turn it into an individualized approach increasing the ability to predict which medical treatment will be safe and effective for each patient.

Within this framework, personalized treatments are defined as a set of decision rules that dictate which treatment to provide given a patient state and can be composed by a sequence of interventions that are made adaptive to the patient's time-varying or dynamic-state, namely Dynamic Treatment Regimes (DTRs)[2].

In the last years, there has been increasing interest on developing methodologies for estimating the DTRs [7, 12]. However, clinical trials and observational studies gather an incredible amount of patient information that are potentially useful for the optimization of the treatment decisions but only a small number of these features (i.e. variables, covariates) can actually impact the prediction of a target outcome (i.e response variable). Peto (1982) [8] defines as *prescriptive variables* those variables that have qualitative interactions with treatments, impacting the decision rules phase. Therefore, variable selection from a high-dimensional set of covariates targeted towards optimal decision making is an essential step in constructing a meaningful and practically useful DTR [13]. Several works go into this direction [3, 4, 13]. For example, Fan et al. (2016) develop a Sequential Advantage Selection (SAS) method based on a modified *S-score* method [4]. The proposed method sequentially selects variables with a qualitative interaction and can be applied in multiple decision-point settings [3].

In this work, we investigate the performance of an hybrid approach mainly based on two consecutive steps: (i) variable selection done by SAS method and (ii) optimal treatment identification based on Q-learning techniques [2]. Specifically, we develop a simulation study to evaluate the performance of the proposed approach in a two-stage DTR by comparing three different models for Q-function estimation: linear regression, random forest and neural network.

The paper is organised as follows. In Section 2 we briefly present the DTR's notations and assumptions and we introduce the SAS and Q-learning approaches. Section 3 describes the characteristics of the conducted simulations and presents some results. Finally in Section 4 we derive some concluding remarks.

2 Method

Consider a finite number of time points, t_k, $k = 1..,K$ where t_1 is the starting point or *baseline*. A single individual, namely a patient, is characterized by a set of information summarized by $(\mathbf{X}_1, A_1, \ldots, \mathbf{X}_k, A_k, \ldots Y)$, where $\mathbf{X}_k$ are the individual covariates and A_k is a treatment defined at each time point t_k. Lastly, Y is the outcome

or response, usually to maximize in order to assure the efficacy of the treatment. In this work, we assume that A_k can take values in $\{0,1\}$. Covariates, treatments and response are gathered for n individuals, resulting in n independent and identically distributed observations (*i.i.d.*). See Fan et al. (2016) [3].

A DTR is a set of rules that dictates how treatments are assigned to an individual over time based on past information. Specifically, a DTR is defined as $d = (d_1, .., d_K)$, where $d_k : \Gamma_k \rightarrow \mathscr{A}_k = \{0,1\}$. Γ_k represents the mapping of the information at time point t_k, where $\Gamma_k = \{(\hat{\mathbf{x}}_k, \hat{a}_{k-1}) \in \hat{\mathscr{X}_k} \times \hat{\mathscr{A}_{k-1}}\}$ is the set of historical information that includes both covariates and treatments. The potential outcomes, for a fixed treatment $\hat{a}_k \in \hat{\mathscr{A}_k}$, are given by Eq. 1.

$$W = \{\{\mathbf{X}_1, \mathbf{X}_2^*(\hat{a}_1), \ldots, \mathbf{X}_K^*(\hat{a}_{K-1}), Y^*(\hat{a}_K)\}, \forall \hat{a}_k \in \hat{\mathscr{A}_k}\}. \tag{1}$$

In Eq. 1, $\mathbf{X}_k^*(\hat{a}_{k-1})$ are the potential intermediate covariates that depend on the treatment history $\hat{a}_{(k-1)}$ and $Y^*(\hat{a}_k)$ represents the potential outcome given treatment $\hat{a}_k$. The optimal DTR is defined as $d^{opt} = \arg\max_{d \in \mathscr{D}} \mathbb{E}[Y^*(d)]$, where $\mathscr{D}$ is a set of candidate treatment regimes. Given a DTR, expected potential outcome can be estimated from observed data if two assumptions are satisfied: the stable united treatment assumption and the sequential randomization assumption. For a more detailed descriptions, see Robins (1997) [9] and Rubin (1978) [10].

When a large number of covariates is considered, a suitable variable selection approach should be used. Fan et al. (2016) [3] developed a framework for selecting variables having qualitative interactions with treatments by explicitly optimizing $\mathbb{E}[Y^*(d)]$ (SAS). In the case of multi-stage treatment decisions a combination of SAS method and Q-learning is used. A modified Q-learning via backward induction is used to estimate the optimal DTR. More precisely, the SAS algorithm is applied at each stage to select important variables for treatment decision making and then these variables are used to model Q-functions. See Fan et al. (2016) [3] for more details.

Following the description presented in Chakraborty et al. (2014) [2], **Q-learning** approach for two-stage studies is here illustrated. Suppose to have a two-stage study composed by longitudinal data where a single subject is characterized by $(\mathbf{X}_1, A_1, \mathbf{X}_2, A_2, \mathbf{X}_3)$. At each stage, the history is given by $H_1 \equiv \mathbf{X}_1$ and $H_2 \equiv (\mathbf{X}_1, A_1, \mathbf{X}_2)$. A random sample of n observations is used for estimation purpose. Consider a two-stage trial with two possible treatments at each stage, $A_j \in \{0,1\}$. Furthermore, consider a study with two rewards, Y_1 and Y_2, observed at end of each stage. A DTR consists on a set of decision rules, (d_1, d_2).
In this context, namely a two-stage study, Q-learning [11] is used to define the optimal treatments $d^{opt} = (d_1^{opt}, d_2^{opt})$ and it is based on the definition of Q-functions, as in Eq. 2:

$$\begin{aligned} Q_2^{opt}(H_2, A_2) &= \mathbb{E}[Y_2 | H_2, A_2], \\ Q_1^{opt}(H_1, A_1) &= \mathbb{E}[Y_1 + max_{a_2} Q_2^{opt}(H_2, a_2) | H_1, A_1]. \end{aligned} \tag{2}$$

The true Q-functions are not known and should be estimated from the data. Moodie et al. (2012) [6] pointed out that traditional Q-learning based on linear regression models is a simple and often reasonable approach, but in non regular settings can lead to bias estimates. In fact, when the outcome is a non linear function of the covariates, linear regression models can fail in capturing the dynamic of the problem. In such a context, there is the need to find valid alternatives in order to overcome the limitations of linear regression models. Some possible solutions are Random Forests [1] and Neural Networks [1]. In this work we contribute in the comparison of different Q-learning techniques based on non linear approaches.

3 Experimental Settings and Results

Three different Q-learning techniques are compared: Q-learning with linear regression model, Q-learning with Random Forest and Q-learning with Neural Networks. The SAS method is used at each stage of intervention to select prescriptive variables as input for each approach. Random Forest is applied with default settings of `randomForest` R package. Neural Network is composed by one layer with 64 nodes, the loss function is the mean absolute error and the identity function is used as activation function. In this preliminary study, one layer is considered only for computational reasons and a more careful analysis should be done in the future. The Neural Network is implemented with the `Keras` R package.

Following Fan et al. (2016) [3] and Zhang et al. (2018) [13], a simulation study is conducted in order to evaluate the performance of the proposed methods in a two-stage treatment decisions regimes. The simulation study is characterized by an high-dimensional setting based on $p = 1000$ covariates at each stage and $n = \{200, 500\}$ number of subjects.

More precisely, covariates at stage $k = 1$ are generated from a multivariate normal distribution with $\mu = 0$ and $\sigma = 1$. Treatments $A_k, k = 1, 2$ are generated from a Bernoulli distribution with success probability of 0.5. Covariates at stage $k = 2$ are generated as $X_{kj} = X_{(k-1)j} + \varepsilon, \forall j$ covariates and $\varepsilon \sim N(0, 0.25)$. Within this setting, we generate two models:

(1a) Model without interactions

$Y = 1 + \beta_1^T X_{k=1} + A_1 \gamma_1^T \tilde{X_{k=1}} + \beta_2^T X_{k=2} + A_2 \gamma_2^T \tilde{X_{k=2}} + \varepsilon$,
where $\tilde{X_{k=1}}$ and $\tilde{X_{k=2}}$ are $(1, X_{k=1}^T)^T$ and $(1, X_{k=2}^T)^T$, $\beta_1 = (1, -1, \mathbf{0}_{p-2})^T$, $\beta_2 = (1, -1, \mathbf{0}_{p-2})^T$ and $\gamma_1 = \gamma_2 = (0.1, 1, \mathbf{0}_7, -0.9, 0.8, \mathbf{0}_{10}, 1, 0.8, -1, \mathbf{0}_5, 1, -0.8, \mathbf{0}_{p-30})$. $\mathbf{0}_{p-2}$ stands for $p-2$ coefficients set to 0.

(1b) Model with interactions

$Y = 1 + \beta_1^T X_{k=1} + A_1 \gamma_1^T \tilde{X_{k=1}} + \beta_2^T X_{k=2} + A_2 \gamma_2^T \tilde{X_{k=2}} + A_1 A_2 + A_2 \alpha^T X_{k=1} + \varepsilon$,
where $\alpha = (1, -0.9, \mathbf{0}_{p-2})$.

Two performance metrics are computed for evaluating the performance of the different methods: (i) the Value Ratio (VR) [3] $= \frac{Q(\hat{d}^{opt})}{Q(d^{opt})}$, where $Q(\hat{d}^{opt})$ is the Q-value following the estimated optimal treatment regime and $Q(d^{opt})$ is the Q-value

following the true optimal treatment regime and (ii) $I^{opt} = \frac{Q(\hat{d}^{opt})-Q(d)}{Q(d)}$ that quantifies the improvement following the estimated optimal treatment regime instead of using a randomized treatment.
Table 1 shows the results of the simulation studies. Both Random Forest and Neural Network are always performing better than Linear Regression with $n = 200$. In terms of VR, Random Forest performs always better with respect to Neural Network in Stage 2 and $n = 200$. At Stage 1, Neural Network is found to be the best approach. Increasing the number of observations (patients) to $n = 500$, Linear Regression is the best model for the estimation of the Q-function even if the Neural Network reaches approximately the same results. Therefore, in situations with limited resources, non linear models with variable selection (*i.e.* SAS) seem to be more suitable than simpler approaches such as linear models.

Table 1 Results with $p = 1000$ and $n = \{200, 500\}$ for each setting considered.

			Stage 2		Stage 1	
n	Models	Methods	VR	I^{opt}	VR	I^{opt}
200	(1a)	Linear Regression	0.719	0.329	0.818	0.128
		Random Forest	**0.796**	**0.472**	0.838	0.155
		Neural Network	0.783	0.448	**0.842**	**0.161**
200	(1b)	Linear Regression	0.825	0.49	0.908	0.248
		Random Forest	**0.905**	**0.633**	0.833	0.145
		Neural Network	0.837	0.511	**0.912**	**0.253**
500	(1a)	Linear Regression	**1.000**	**0.902**	**0.997**	**0.486**
		Random Forest	0.841	0.600	0.902	0.345
		Neural Network	0.999	0.899	0.986	0.470
500	(1b)	Linear Regression	**1.000**	**0.959**	**0.983**	**0.432**
		Random Forest	0.893	0.749	0.909	0.325
		Neural Network	0.997	0.954	0.982	0.430

4 Conclusion and Future Works

In this work, we investigated the combination of a variable selection technique, SAS technique [3], together with a Reinforcement Learning technique, Q-learning [2], for DTRs in high-dimensional settings. We compared three different estimation models for the Q-functions: Regression model, Random Forest and Neural Network. We developed some simulation studies characterized by increasing complexity both in terms of number of available observations and underline type of model. From this preliminary work, Random Forest and Neural Network seem to be more suitable in presence of limited resources. If the number of patients is sufficient large and an

appropriate variable selection approach is used then Regression models can provide reliable results.

This work can be extended in different directions. First of all, a more careful analysis of the relation between p and n should be done in order to understand how p affects the performance of the considered approaches. We expect that more p diverges from n more the estimation problem will become challenging. Secondly, often statistical learning approaches are characterized by a high number of hyper-parameters that affect the performance of the techniques. For instance, a hyper-parameters tuning should be done for Neural Network in order to optimize the number of hidden layers, number of hidden nodes and type of activation functions. Further future researches are (i) exploit the results of Tao et al. (2018) [12] on more complex tree-based methods and (ii) extend the work of Murray et al. (2018) [7] based on Bayesian Machine Learning.

Acknowledgements We greatly acknowledge the DEMS Data Science Lab for supporting this work by providing computational resources.

References

1. Bradley, E., Hastie, T.: Computer age statistical inference. Algorithms, evidence and data science. Cambridge University Press, Cambridge UK (2016).
2. Chakraborty, B., Murphy, S. A.: Dynamic treatment regimes. Annu. Rev. Stat. Appl. **1**, 447–464 (2014).
3. Fan, A., Wenbin, L., Song, R.: Sequential advantage selection for optimal treatment regime. The Annals of Applied Statistics **10**(1), 32–53 (2016).
4. Gunter, L., Zhu, J., and Murphy, S.: Variable selection for qualitative interactions. Stat. Methodol., **8**(1), pp. 42–55 (2011).
5. Kosorok, M. R., Laber, E. B.: Precision Medicine. Annu. Rev. Stat. Appl. **6**(1), 263–286 (2019).
6. Moodie, E. E. M., Chakraborty, B., Kramer, M. S.: Q-learning for estimating optimal dynamic treatment rules from observational data. Can. J. Stat. **40**(4), 629–645 (2012).
7. Murray, T.A., Yuan, Y., Thall, P.F.: A Bayesian machine learning approach for optimizing dynamic treatment regimes. J. Am. Stat. Assoc. **113**(523), pp. 1255–1267 (2018).
8. Peto, R.: Statistical aspects of cancer trials. In: Halnan KE, editor. Treatment of Cancer, London, UK, Chapman, pp. 867–871 (1982).
9. Robins, J. M.: Casual Inference from complex longitudinal data. In: Berkane, M. (eds.) Latent Variable Modelling and Application to Casuality 1994, LNCS, vol. 120, pp. 69–117, Springer, New York (1997).
10. Rubin, D. B.: Bayesian inference for casual effects: The role of randomization. Ann. Stat. **6**, pp. 34–58 (1978).
11. Sutton, R.S., Barto, A.G.: Reinforcement learning: and introduction. MIT, Cambridge MA (1998)
12. Tao, Y., Wang, L., Almirall, D.: Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. Ann. Appl. Stat. **12**(3), pp. 1914–1938 (2018)
13. Zhang, B., Zhang, M.:. Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariates. Ann. Appl. Stat. **12**(4), pp. 2335–2358 (2018)

Sample Size Computation for Competing Risks Survival Data in GS-Design

Calcolo della Dimensione Campionaria per Dati di Sopravvivenza con Rischi Competitivi nei Disegni GS

Mohammad Anamul Haque and Giuliana Cortese

Abstract Competing-risks survival analysis is applied when the time to occurrence of events depends on multiple causes of failure. The objective of the paper is to estimate the sample size under group sequential (GS) design for a new treatment to justify the continuation or interruption of a trial in interim analyses, when we have competing risks data. We compared the GS-design under two popular approaches for regression with competing risks data: the cause-specific hazard (CSH) model and the sub-distribution hazard (SDH) model. We found that the conditional power is higher under the SDH approach as compared to the CSH.

Abstract *L'analisi di sopravvivenza per rischi competitivi riguarda lo studio del tempo di attesa ad un evento generato da cause multiple. L'obiettivo del contributo è quello di calcolare la dimensione campionaria nei disegni sequenziali a gruppi (GS) per un nuovo trattamento, per giustificare la continuazione o l'interruzione di un trial clinico, in presenza di rischi competitivi. Sono stati messi a confronto i disegni GS in due popolari approcci per la regressione: il modello per i tassi causa-specifici (CSH) ed il modello per i tassi 'sub-distribution' (SDH). Si è trovato che la potenza condizionata è maggiore nell'approccio SDH rispetto all'approccio CSH.*

Key words: competing risks, sample size, group sequential design.

1 Introduction

For designing a randomised clinical trial, an essential step is the calculation of the sample size or the number of patients to be recruited to detect the efficacy of treat-

Mohammad Anamul Haque
Department of Statistical Sciences, University of Padova,
e-mail: mohammadanamul.haque@studenti.unipd.it

Giuliana Cortese
Department of Statistical Sciences, University of Padova, e-mail: giuliana.cortese@unipd.it

ments with sufficient power. In a time-to-event study, the sample size is determined not by the number of patients accrued but rather by the number of events observed during a specific follow-up period. Furthermore, it is of great interest to study the effect of a main event accounting for the fact that patients can experience competing events. In this case, only part of the trial population will experience the main event, allowing subjects to be censored or to fail from competing events. Therefore, in determining the required sample size, we also need to estimate the probability Ψ of having the main event over time, which can be calculated by using the cumulative incidence function (CIF). The CIF is often of interest in medical research and can be estimated with different methods. In Section 2 we describe the theory of sample size calculation, related to the CIF, under two popular competing risks models: the cause-specific hazard (CSH) and sub-distribution hazard (SDH) approaches. We then extend results to group sequential design (*GS-design*) in Section 3.

2 Computing Sample Size in a Competing Risks Setting

Consider a single event of interest. Let D be the number of events required to be observed in the study, and T be the duration of a study. It is planned that $T = a + \tilde{f}$, where a is the first time period during which subjects are being enrolled into the study, while $\tilde{f}$ is the follow-up period during which subjects are under observation. A general formula for the required number of subjects is $N = D/\Psi = f(\alpha,\beta,\theta^*)/p\{S(t),G(t),H(t),a,T\}$, where D is obtained as a function $f(\cdot)$ of the type I error probability α, the power $1-\beta$ and the effect size θ^*, while Ψ is given as a function $p(\cdot)$ that depends on a, T, the survival function $S(t)$, the accrual distribution $G(t)$ and the distribution of the loss to follow up patterns $H(t)$. Effect size is usually expressed as either the hazard ratio (HR), $\theta = \lambda_2(t)/\lambda_1(t), \forall t,$. or $\log\theta$, or the regression coefficient in a Cox model.

Let P_E and P_C be the sample proportions assigned to, respectively, the experimental treatment group and control group, Z_1 and Z_2 be the standard normal quantiles at the desired one-sided significance level and power $1-\beta$, respectively. The required number of events can be further obtained as $D = \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2/[(\log\theta)^2 P_E P_C]$. and, moreover, we have $\Psi = \int_0^a P(\text{death} \mid \text{accrued at time } t) \times P(\text{accrued at time } t)dt$.

Consider now a competing risks setting with an event of interest (type 1) and a competing event (type 2). Our scope is here to derive the required sample size for event of type 1. Equivalent results can be obtained for event of type 2. Then, assuming a uniform distribution for the accrual of patients in $[0,a]$ for both control (C) and experimental (E) groups, the cumulative probabilities for the event of interest in each group ($j = C,E$) reduce to $\Psi_{1\mathrm{j}} = (1/a)\int_{\tilde{f}}^{a+\tilde{f}} CIF_{1j}(t)\,\mathrm{dt}$. Under the CSH approach, the function $CIF_1(t) = \int_0^t \lambda_1(u)e^{-\{\Lambda_1(u)+\Lambda_2(u)\}}du$, depends on both cause-specific hazard rates $\lambda_1(t)$ and $\lambda_2(t)$, where $\Lambda_k(t) = \int_0^t \lambda_k(s)ds$, for $k = 1,2$. For computing sample size, any parametric distribution can be assumed for the two cause-specific hazards. Under the SDH approach, the relation between $CIF_1(t)$ and

the survival function for the only event 1 is $CIF_1(t) = 1 - S_1(t)$. Then, by modifying the survival function according to Simpson's approximation, we can compute Ψ_{1C} and Ψ_{1E} as $\Psi_{1j} = \frac{1}{6}\left[CIF_{1j}\left(\tilde{f}\right) + 4\,CIF_{1j}\left(0.5a + \tilde{f}\right) + CIF_{1j}\left(a + \tilde{f}\right)\right]$ and combining these results we obtain $\Psi_1 = P_{\mathrm{C}} * \Psi_{1\mathrm{C}} + P_{\mathrm{E}} * \Psi_{1\mathrm{E}}$.

In the following we provide some practical guidelines for computing the required sample size in presence of competing risks. For this scope, the main quantities of interest are the hazard ratios and the probability Ψ_1 of observing an event of type 1.

Let us consider the CSH approach. The involved CSH ratios are $\theta_1 = \lambda_{1E}/\lambda_{1C}$ and $\theta_2 = \lambda_{2E}/\lambda_{2C}$. To obtain these ratios, we can derive the following CSHs by inverting the equations for $CIF_{1j}(t)$ and $CIF_{2j}(t)$ (see [7]):

$$\lambda_{\mathrm{kj}} = CIF_{\mathrm{kj}}(t) \times \frac{-\log\left(1 - CIF_{1\mathrm{j}}(t) - CIF_{2\mathrm{j}}(t)\right)}{t\left(CIF_{1\mathrm{j}}(t) + CIF_{2\mathrm{j}}(t)\right)}, \quad j = C,E \quad \text{and} \quad k = 1,2. \quad (1)$$

If we know $(\lambda_{1E}, \lambda_{2E})$ and then the ratios (θ_1, θ_2), we can calculate $(\lambda_{1C}, \lambda_{2C})$. Alternatively, if we do not know $(\lambda_{1E}, \lambda_{2E})$, we can calculate them from equations (1) assuming $CIF_{1E}(t)$ and $CIF_{2E}(t)$ are known. Finally, for each $j = C,E$, we can obtain the CIFs values by solving the system of equations in (1): $CIF_{1j}(t) = [1 - e^{-t\ \lambda_{1j}\ (1+\tilde{\lambda}_j)}]/(1+\tilde{\lambda}_j)$, where $\tilde{\lambda}_j = \lambda_{2j}/\lambda_{1j}$. Similarly we can compute $CIF_{2j}(t)$.

Under the SDH approach, we assume to know CIF_{1E}, CIF_{2E}, CIF_{1C}, CIF_{2C}, and need to compute the sub-distribution hazard rate $\lambda^*(t)$. This can be obtained by using the direct relationship $\int_0^t \lambda_k^*(u)du = -\log\{1 - CIF_k(t)\}$ [2]. Therefore, the SDH for event of type 1 can be computed as

$$\tilde{\theta}_1 = \frac{-\log\{1 - CIF_{1E}(t)\}}{-\log\{1 - CIF_{1C}(t)\}} = \frac{\int_0^t \lambda_{1E}^*(u)du}{\int_0^t \lambda_{1C}^*(u)du}. \quad (2)$$

Finally we can calculate the probability Ψ following the formulas for Simpson's approximation.

3 Sample Size under Group Sequential Design (GS-design)

It is of great interest to design and supervise a trial on an experimental treatment by performing interim analysis, referred to as *GS-design*. This type of design helps to reduce the number of allocated patients per treatment group which in turns save time and money. One of the main scope is to calculate the boundary limits after having adjusted for type I and type II errors (using the error spending functions proposed in [5, 6] and the conditional power in [3]). These limits help to make an early decision to stop a trial anytime before the final stage analysis is reached. The Data Monitoring Committee also requires this analysis for patients' ethical concern because of efficacy (treatment is found very effective earlier than expected) or futility (the trial shows adverse effects or the true effect is far away from the assumed H_1).

Calculation of boundaries limits

Suppose the interest is to test a parameter θ with $H_0 : \theta \geq 0$ vs $H_1 : \theta < 0$. Assume the estimator $\hat{\theta}$ is efficient, properly normalized and computed sequentially over time and has asymptotically a Gaussian independent increments process whose distribution depends only on θ and on the Fisher's information I [8]. For the interim analysis k, with $k = 1, \ldots K-1$, $\widehat{\theta}_k \sim N(\theta, \mathscr{I}_k^{-1})$, the standardized statistic is $Z_k = \widehat{\theta}_k \sqrt{\mathscr{I}_k}$ and the canonical joint distribution of $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K)$ implies that $(Z_1, \ldots, Z_K)$ is multivariate normal, where $Z_k \sim N(\theta\sqrt{\mathscr{I}_k}, 1)$ and $\mathrm{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathscr{I}_{k_1}/\mathscr{I}_{k_2}}$ for $k_1 < k_2$ [3].

For obtaining interim stage information (I_k), it is required to obtain the information I_K from final stage K. For this, two quantities, the fixed design event size (I_{fix}) and an inflation factor (IF), need to be calculated using the relation $I_K = I_{fix} * IF$ Then to obtain IF, it is required to calculate the Z-quantiles estimated at the interim stages using the so-called error spending function. The Z critical values and the IF are available from reference [3] or can be computed from the R package `gsDesign` [1]. Finally, I_{fix} is obtained by recalling formula for D from Section 2, $D = \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 / [(\log\theta)^2 P_E P_C] = I_{fix}/(P_E P_C) = 4\, I_{fix}$, when $P_1 = P_2 = 0.5$. Once the Z statistic values (efficacy or futility boundary limits) are obtained, then the decision at each interim analysis whether to continue or terminate the trial can be made, according to the rule

$$\begin{cases} Z_k \in \mathscr{M}_k = (l_k, u_k), \text{trial continue to the next stage} \\ Z_k \in \mathscr{U}_k = (u_k, +\infty), \text{stop the trial for efficacy and conclude } H_1 \\ Z_k \in \mathscr{L}_k = (-\infty, l_k), \text{stop the trial for futility and conclude } H_0 \end{cases}$$

where $\{u_k\}$ is the upper limit or the efficacy boundary and $\{l_k\}$ is the lower limit or the futility boundary, with $l_k \leq u_k$ until final stage and at final stage $l_K = u_K$.

Calculation of the error spending function

Since the same sample data are used in the clinical trial over the study period, we need to adjust the error rate by using flexible approaches of *error spending functions* which do not require the number and exact timing of interim to be fixed in advance. Define the proportion of events at interim k to be equal to the information fraction $t = I_k/I_K$. Then, two non-decreasing error spending functions for $\alpha(t)$ and $\beta(t)$ will be used to set α (under null, $\theta = 0$) and β (under alternative, $\theta = \theta^*$). Among several functional forms proposed by [5], we have used O'Brien-Fleming (spends very little α at the beginning), with $\alpha(t) = 2 - 2\phi(z_{1-\alpha/2}/\sqrt{t})$, and Pocock (spends α more evenly across the stages), with $\alpha(t) = \alpha\log\{1 + (e-1)t\}$. The β-spending function are calculated in a similar way. To have both these approaches together in a more flexible way, Wang-Tsiatis bounds can be used [9].

Calculation of conditional power (CP)

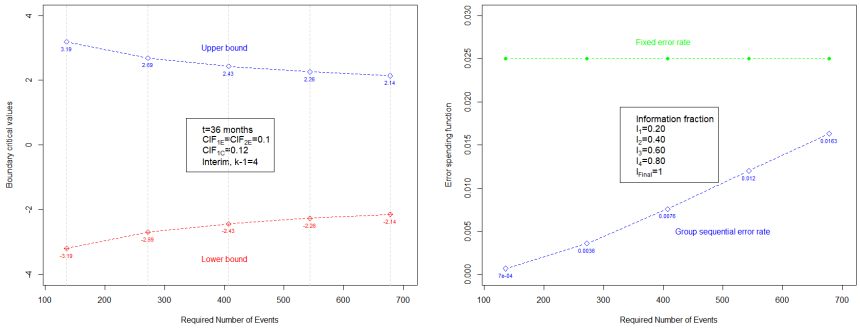
CP is the probability of rejecting H_0 (when H_1 is true) given the observed interim stage data. When the trial starts, the CP is actually equal to the unconditional power. Once we calculate the CP at each stage, we can decide to stop the trial for futility of efficacy, if CP is found to be very small or very large, respectively. Following [3],

for rejecting a null hypothesis about θ^* at the end of the study, the general lower one-sided CP at interim k, given the observed Z_k calculated using data information collected up-to $k-1$, is $P_k(\theta^*) = \Phi[(-Z_k\sqrt{I_k} - Z_K\sqrt{I_K} - (I_K - I_k)\theta^*)/\sqrt{I_K - I_k}]$ Here, let θ^* be the log hazard ratio or any other reparameterization at k under H_1.

Simulation studies

Simulations are performed to compare a *GS-design* with $K = 5$ interim stages under the CSH and SDH approaches. Assume that $\alpha = 0.025$, $\beta = 0.20$, $\theta_{1E} = \theta_{1C} = 0.8$, $P_E = P_C = 0.5$, $\tilde{f} = 3$ years and $\tau = 0.01$. From the theory in Section 2, we obtain $CIF_{1E} = CIF_{2E} = 0.1, CIF_{1C} = CIF_{2C} = 0.12$, $N = 8244$ and $D = 632$. The Wang-Tsiatis error spending function is used with $\omega = 0.25$ for symmetrical boundary values.

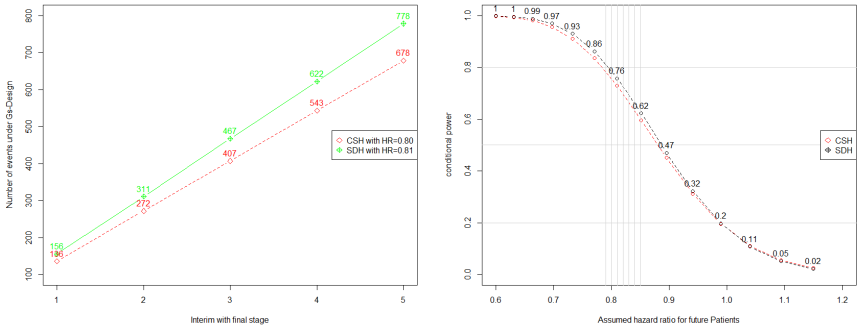
Fig. 1 The GS-Design for the CSH approach: Boundary values and Error spending function for fixed and GS designs.



Under the CSH approach, the boundary critical values are larger at the first interim stages and have a one third reduction at the final stage (Figure 1, left panel). Moreover, the *GS-design* error spending function increases with the stages, while the fixed-design error rate is flat and higher (Figure 1, right panel). Based on the interim results, the CP is calculated in Figure 2. For a treatment effect equal to $\log HR = \log 0.8$, at stage $k = 3$, we observe $Z = -2$, $var(\log HR) = 0.01$ and $D = 407$ (Figure 2, left panel). At the final stage we observe $Z = 2.14$ and $I_5 = 678$. Then, with the parameterization $\log \theta^* = -0.10$, we computed the CP as $P_3(\theta^*) = 0.75$. This indicates that, under H_1, the probability of rejecting the null if the experiment stops at stage 3, is reduced from 0.8 to 0.75, when compared with the trial completion. Changing the assumed $\log HR$, given the other parameters fixed, the CP changes accordingly, e.g., if we assume $HR = 0.9$, the CP reduces to 0.45.

Under the SDH approach, the computed SDH ratio is 0.81, a slightly higher value than the CSH ratio. However, for such a small amount of change in the hazard ratio $(0.81 - 0.80 = 0.01)$, the CP and D have changed considerably, this latter showing 60 additional events required, as compared to the CSH approach (Figure 2). We also observe that for larger changes of the hazard ratio, e.g., from 0.79 to 0.85, the CP reduces from ≈ 0.80 to 0.60 (Figure 2, right panel). Finally, the SDH model yields a higher D and a slight gain in power, as compared to the CSH approach (depending also on the discrepancy between the two hazard ratio).

Fig. 2 Comparison of GS-Design under the CSH and SDH models. Left: Estimated number of events at each stage. Right: Conditional power as a function of the assumed hazard ratio.



4 Conclusions and Remarks

Interim analyses (*GS-design*) are often needed to justify sample size in clinical trials as an ethical concern. We conducted simulation studies assuming the same *CIF* for treatment group under CSH and SDH competing risks models. Even for a negligible increase in hazard ratio (e.g. 0.01), we found that the SDH model yields a higher *D* at each interim as compared to CSH but with the advantage of slight gain in *CP*. As a general recommendation, the SDH approach can be preferred when the main attention is devoted to increase *CP*, while CSH is better in terms of reducing required *D*. The authors also give guidelines for the computation of fixed design.

References

1. Anderson, K.: R Package: gsDesign, version 3.0-1 (2016). https://cran.r-project.org/web/packages/gsDesign/gsDesign.pdf
2. Fine, J. P., Gray, R. J.: A proportional hazards model for the subdistribution of a competing risk. JASA **94**, 496–509 (1999)
3. Jennison, C., Turnbull, B. W.: Group Sequential Methods with Applications to Clinical Trials. Chapman and Hall/CRC, Boca Raton, FL (2000)
4. Latouche, A., Porcher, R., Chevret, S.: Sample size formula for proportional hazards modelling of competing risks. Stat. Med. **23**, 3263–3274 (2004)
5. Lan, K. G., DeMets, D. L.: Discrete sequential boundaries for clinical trials. Biometrika **70**, 659–663 (1983)
6. Pampallona, S., Tsiatis, A. A., Kim, K.: Spending functions for the type I and type II error probabilities of group sequential tests. J. Stat. Plan. Inference **42**, 1994–35 (1995)
7. Pintilie, M.: Competing risks: a practical perspective (Vol. 58). John Wiley & Sons, Ltd (2006)
8. Scharfstein, D. O., Tsiatis, A. A., Robins, J. M.: Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. JASA **92**, 1342–1350 (1997)
9. Wang, S. K., Tsiatis, A. A.: Approximately optimal one-parameter boundaries for group sequential trials. Biometrics **43**, 193–199 (1987)

4.2 Advances in neural networks

Linear models vs Neural Network: predicting Italian SMEs default.

Modelli lineari vs Reti Neurali nella previsione del default delle PMI.

Lisa Crosato, Caterina Liberati and Marco Repetto

Abstract As recently stated by EUROSTAT, Small and Medium Enterprises (SMEs) play a crucial role in the economy of European Union (EU). This is particularly true for Italy that has one of the largest share of SMEs in the Euro zone. In such a context, the assessment of the Italian firms creditworthiness is a priority, especially in the early stage of a SMEs life. The aim of this work is to explore the effectiveness of the Feedforward Neural Network for Italian firms defaults prediction in comparison with standard linear models (Logistic and Probit). Our analysis considers the Italian manufacturing SMEs in a short time span (2016-2017). Results, obtained via two different sampling strategies, are not clear-cut.

Abstract *Come affermato di recente da EUROSTAT, le piccole e medie imprese (PMI) svolgono un ruolo cruciale nell'economia dell'Unione Europea (UE). Ciò è particolarmente vero per l'Italia che una delle quote maggiori di PMI nella zona euro. In tale contesto, la valutazione del merito creditizio delle imprese italiane è una priorità, soprattutto nelle prime fasi di vita delle PMI. Lo scopo di questo lavoro è esplorare l'efficacia della Rete Neurale Feedforward nella predizione del default delle imprese italiane rispetto ai modelli lineari standard (Logistic e Probit). La nostra analisi considera le PMI manifatturiere italiane in un breve arco di tempo (2016-2017). I risultati, ottenuti in combinazione con due strategie di campionamento, non sono conclusivi.*

Key words: Default prediction, SMEs, Sampling methods, Neural Network

Lisa Crosato
Department of Economics, Ca' Foscari University of Venice, San Giobbe 873, 30121, Venezia
e-mail: lisa.crosato@unive.it

Caterina Liberati
DEMS, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano,
e-mail: caterina.liberati@unimib.it

Marco Repetto
DEMS, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano,
e-mail: marco.repetto@unimib.it

1 Introduction

As stated in a recent report by EUROSTAT[10], Small and Medium Enterprises are about 99.8% of the active firms of the Euro zone. They account for almost 60% of value-added within the non-financial economy, playing also a crucial role in the workforce of the European Union. This statement is particularly true for Italy which has one of the largest share of SMEs in Europe. For these reasons, more than in the past, assessing the firms creditworthiness remains a priority in the economic analysis of our Country.

The development of effective classification models, able to separate survived and defaulted firms, has been largely investigated in literature. If we focus on the Italian case, there are several papers that estimated SMEs default with linear prediction models based on both financial and/or non-financial information ([1], [2], [8] [5]). The results described are interesting especially on the economic interpretation side, although the usage of the linear models do not always provide accurate classification predictions. Other models, such as Binary Generalised Extreme Value Additive model [6], are more suited to treat unbalanced datasets with respect to the standard Logistic or Probit models and have been applied successfully (for a comparison between Italy and UK SMEs failure through BGEVA see [3]).

On the contrary the usage of Machine Learning algorithms, such as Artificial Neural Network, is still limited, although the work of [9] showed the effectiveness of this non-parametric approach.

In this paper we compare the performances of two standard linear models with the Feedforward Artificial Neural Network (FANN) [11]. In order to improve the percentage of the identified defaulters, we tested two sampling procedures: the robust version of the Random Over-Sampling Examples [4] and the Synthetic Minority Over-sampling Technique [7]. The ROC curve, together with sensitivity and specificity rates, is used to compare performances of the competing classification rules.

2 Research Design

Feedforward Artificial Neural Network has proven appropriate to predict Italian firms default [9], showing peculiar abilities to deal with non-linear patterns characterizing the data. FANN consists of a direct acyclic network of nodes organized in densely connected layers, where inputs after been weighted and shifted by a bias term are fed into the node's activation function and influence each successive layer. We opt for a two layers FANN composed of one layer for the inputs, one hidden layer with n hidden units, and a corresponding output layer. This promising application of machine learning, together with the increasing interest for that kind of models, could lead one to consider *old fashioned* classifiers such as Probit or Logistic regressions as overcome.

Another aspect worth considering regards default predictions being generally performed on unbalanced datasets, where the group of the defaulted firms does not exceed 1 to 3% of the total instances. This disequilibrium interferes with any model's ability to catch the inner patterns in the data [4], so that the classifiers tend to overclassify the largest class.

The literature proposed several strategies to cope with class imbalance. In our work we employed two resampling techniques, both of which generate synthetic observations from the minority class: the Synthetic Minority Over-sampling Technique, also referred to as SMOTE [7], where the generation is based on the k-nearest neighbor algorithm and the robust version of the Random Over-Sampling Examples (ROSE) [12], which uses instead a smoothed bootstrap approach [4].

3 Data

Our study focuses on the SMEs of the Italian manufacturing industry (NACE Rev. 2 codes 10-33) for 2016-2017. The firms were selected according to the European Commission definition of SME, which characterizes them by a turnover smaller than 50 million euros and a number of employees below 250. Our final sample is composed by 105.058 firms, with 1.72% of them defaulted (1.807 firms). We retrieved the synthetic indicators of the businesses balance sheets from Amadeus-Bureau Van Dijk database [14], which provides economic information on European private companies, selecting the financial ratios that were found to be relevant in previous research on SMEs default [2] [13], and precisely:

- Cash flow, th EUR
- GearingRatio= Total debt/ Total assets, Per cent
- ProfitMargin=P(L) before tax/ Operating revenue, Per cent
- ROCE=P(L) before tax/ (Total assets - Cur. liab.), Per cent
- ROE=P(L) before tax/ Shareholder funds, Per cent
- SolvencyRatio=Shareholders funds/Total assets, Per cent
- Number of employees
- Total assets, th EUR
- Revenues, th EUR

The target variable is a dummy scoring 1 if the last available data goes back to 2016 and the registered status in Amadeus is one of the following: Bankruptcy, Active (default of payment), Active (insolvency proceedings), Dissolved (bankruptcy), Dissolved (liquidation) or In liquidation.

4 Results

Here we show the results of our study. Our protocol partitioned the data into training (70%) and test (30%) sub-samples which are randomly selected from the original sample. The training set also serves for hyperparameter tuning. We addressed the robustness of the hyperparameters through Montecarlo Cross-Validation [15]. At each iteration, we fit different parametrized FANNs and collected the resulting metrics on the validation set. Then we average the metrics of these different models across the iterations, and we chose the FANN scoring the highest average metric. The metric we chose is the H-measure [16]. This procedure ensures a sound choice of the parameters. In order to get a reliable comparison among the competing models, we bootstrapped the dataset 100 times and calculated average rates of prediction of default (Sensitivity), survival (Specificity) and the Area Under the ROC Curve (AUC) (Table 1).

Table 1 Test set metrics table: average rates over 100 random samples

Model	Sampling	Sensitivity	Specificity	AUC
FANN	SMOTE	0.625	0.872	0.837
FANN	Rob ROSE	0.770	0.692	0.793
FANN	-	0.000	1.000	0.830
Logistic	SMOTE	0.662	0.808	0.811
Logistic	Rob ROSE	0.824	0.638	0.814
Logistic	-	0.010	0.999	0.796
Probit	SMOTE	0.627	0.809	0.799
Probit	Rob ROSE	0.120	0.987	0.554
Probit	-	0.003	1.000	0.795

Note that both sampling techniques help models to provide non-trivial predictions. In particular, employing SMOTE, the sensitivity rates of the competing models reach a minimum of 62%. According to the AUC value (0.837) the best classification is obtained via FANN.

On the contrary, using Rob ROSE sampling, the best performance is by Logistic regression both in the sensitivity rate and in AUC value. Specifically, this solution shows the largest percentage of defaulted firms correctly identified (0.824), together with the highest accuracy (AUC equal to 0.814).

We can provide no clear-cut conclusion about which model should be selected: as far as accuracy is concerned, FANN combined with SMOTE sampling outperforms logistic regression, but the opposite holds true when the focus is on correct classification of defaulted firms (and still keeping accuracy to a fair level). As to the present, we cannot declare logistic regression to be defeated.

References

1. Altman, E. I., Esentato, M., Sabato,G.:Assessing the credit worthiness of Italian SMEs and mini-bond issuers. Global Finance Journal. **143** 100450, (2020)
2. Altman, E. I., Sabato, G.: Modelling Credit Risk for SMEs: Evidence from the U.S. Market. Abacus, **43(3)**, 332-357 (2007).
3. Andreeva, G., Calabrese, R., Osmetti, S. A.: A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. European Journal of Operational Research, **249(2)**, 506-516 (2016).
4. Baesens, B., Höppner, S., Ortner, I., Verdonck, T.: robROSE: A robust approach for dealing with imbalanced data in fraud detection. arXiv:2003.11915 [Cs, Stat]. (2020). http://arxiv.org/abs/2003.11915
5. Bottazzi, G., Grazzi, M., Secchi, A., Tamagni, F.: Financial and economic determinants of firm default. Journal of Evolutionary Economics, **21(3)**, 373-406 (2011).
6. Calabrese, R., Osmetti, S. A.: Modelling small and medium enterprise loan defaults as rare events: The generalized extreme value regression model. Journal of Applied Statistics, **40(6)**, 1172-1188 (2013).
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, **16**, 321-357 (2002).
8. Ciampi, F.: Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. Journal of Business Research, **68**, 1012-1025 (2015).
9. Ciampi, F., Gordini, N.: Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises. Journal of Small Business Management. **51(1)**, 23-45 (2013).
10. EUROSTAT, Union Européenne, & Commission Européenne.: Key figures on European business: With a special feature on SMEs. Office for official publications of the European communities (2011).
11. Haykin, Simon S. Neural networks: a comprehensive foundation. 2nd ed. Upper Saddle River, N.J: Prentice Hall, (1999).
12. Lunardon, N., Menardi, G.,Torelli, N.: ROSE: A Package for Binary Imbalanced Learning. The R Journal, **6(1)**, 79, (2014).
13. Michala, D., Grammatikos, T., Filipe, S. F.: Forecasting distress in European SME portfolios (EIF Working Paper 2013/17). European Investment Fund (EIF) (2013).
14. van Dijk, B. M.: AMADEUS: A database of comparable financial information for public and private companies across Europe. Bureau Van Dijk (2010).
15. Xu, Qing-Song, e Yi-Zeng Liang: Monte Carlo Cross Validation. Chemometrics and Intelligent Laboratory Systems 56, n. 1 (2001).
16. Hand, David J.:Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve. Machine Learning 77, n. 1 (2009).

Network estimation via elastic net penalty for heavy-tailed data

Stima di reti con penalità elastic net in presenza di code pesanti

Davide Bernardini, Sandra Paterlini and Emanuele Taufer

Abstract We propose a 2-stage procedure relying on elastic net penalty to estimate a network based on partial correlations for heavy-tailed data from a multivariate t-Student distribution. Simulation analysis shows that the 2-stage estimator performs better both in terms of identification of sparsity patterns and numerical accuracy than two well-known penalized estimation techniques, namely *glasso* and *tlasso*. A real world application focuses on estimating the European banking network.

Abstract *In questo articolo viene proposta una procedura a 2 stadi che utilizza la penalità elastic net per la stima di una rete basata sulle correlazioni parziali di dati a code pesanti con distribuzione t-Student multivariata. Un'analisi mediante simulazioni evidenzia come la procedura a 2 stadi risulta migliore di due tecniche ben conosciute,* glasso *e* tlasso*, sia dal punto di vista dell'identificazione delle connessioni tra nodi, sia dal punto di vista dell'accuratezza numerica. Un'applicazione a dati reali si focalizza sulla stima della rete bancaria europea.*

Key words: Partial correlation, Elastic net penalty, Banking network

Davide Bernardini
University of Trento, Department of Economics and Management, Via Inama 5, 38122 Trento
e-mail: davide.bernardini@unitn.it

Sandra Paterlini
University of Trento, Department of Economics and Management, Via Inama 5, 38122 Trento
e-mail: sandra.paterlini@unitn.it

Emanuele Taufer
University of Trento, Department of Economics and Management, Via Inama 5, 38122 Trento
e-mail: emanuele.taufer@unitn.it

1 Introduction

Let $\boldsymbol{X} = [X_1, ..., X_p]^\top$ a p-dimensional random vector from the joint multivariate t-Student distribution $t_p(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1}, \nu)$ where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Psi^{-1}}$ is the positive definite dispersion matrix and ν are the degrees of freedom. The covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$ and its inverse, the precision matrix $\boldsymbol{\Theta}$, are equal to $\frac{\nu}{\nu-2}\boldsymbol{\Psi}^{-1}$ and $\frac{\nu-2}{\nu}\boldsymbol{\Psi}$, respectively.

Our goal is to estimate a sparse precision matrix $\boldsymbol{\Theta}$ from which to retrieve a sparse graph whose weights are the partial correlations between couples of variables. Partial correlation p_{jk} between components X_j and X_k can be computed by properly scaling the off-diagonal elements in $\boldsymbol{\Theta}$. Thus, we build a graph $\mathcal{G}(\mathbf{V}, \mathbf{E})$ with the set of nodes $\mathbf{V}=\{1, ..., p\}$ representing the elements in $\boldsymbol{X}$ and the set of edges $\mathbf{E} \subseteq \mathbf{V}\times\mathbf{V}$ and $(j,k) \in \mathbf{E}$ if $\mathrm{p}_{jk} \neq 0$ where (j,k) represents the edge between elements X_j and X_k.

Following the scale-mixture representation of a multivariate t-Student distribution as in Finegold and Drton [3], we have that:

$$\boldsymbol{X} = \boldsymbol{\mu} + \frac{\boldsymbol{Y}}{\sqrt{\tau}} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1}, \nu) \tag{1}$$

where $\boldsymbol{Y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi}^{-1})$ and $\tau \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$

By relying on this representation, it is possible to exploit the Expectation-Maximization (EM) algorithm (Dempster et al. [2]) to estimate the parameters of the multivariate t-Student distribution. Following closely the *tlasso* procedure proposed by Finegold and Drton [3], we introduce a similar EM algorithm to produce a sparse estimate of $\boldsymbol{\Theta}$. Differently from the *tlasso* that uses the *lasso*, or 1-norm, penalty (see Tibshirani [7]) to induce sparsity, we propose a 2-stage approach that explicitly considers the elastic net penalty, a linear combination of 1-norm and 2-norm (see Zou and Hastie [10]), to perform a penalized estimation of $\boldsymbol{\Theta}$.

2 Two-stage elastic net penalized EM algorithm *[2Stelnet]*

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be n p-vectors of observations drawn from the $t_p(\mu, \boldsymbol{\Psi}^{-1}, \nu)$ distribution, realizations of $\boldsymbol{X}$. The random variable τ in the mixture (1) is considered the hidden or latent variable whose value is updated given the current estimate of the parameters and the observed data. Let also τ_i be the value of the latent variable τ associated with observation $\mathbf{x}_i$. As in the *tlasso* of Finegold and Drton [3], we also assume that the degrees of freedom ν are known in advance. The EM algorithm is used to estimate unknown parameters. Let the superscripts $^{(t)}$ and $^{(t+1)}$ denote the t-th and $(t+1)$-th updates of the estimated parameters. In the E-step, the updated estimate

$\hat{\tau}_i^{(t+1)}$ of τ_i is obtained using the t-th estimates $\hat{\boldsymbol{\mu}}^{(t)}$ and $\hat{\boldsymbol{\Psi}}^{(t)}$ of $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$. In the M-step, updated estimates $\hat{\boldsymbol{\mu}}^{(t+1)}$ and $\hat{\boldsymbol{\Psi}}^{(t+1)}$ are evaluated using $\hat{\tau}_i^{(t+1)}$ from the E-step (see Finegold and Drton [3] for a more detailed description).

Here, for sake of brevity, we discuss only a penalized estimator of the matrix $\boldsymbol{\Psi}$, the precision matrix of the multivariate normal $\boldsymbol{Y}$ in the mixture (1), to use in the M-step. Finegold and Drton [3] rely on the graphical *lasso* (*glasso*) proposed by Friedman et al. [4], thus their algorithm is called *tlasso*. We propose a different estimator, *2Stelnet*, based on a 2-stage procedure and elastic net penalty. First, our procedure *2Stelnet* estimates the sparsity structure of $\hat{\boldsymbol{\Psi}}^{(t+1)}$ by using conditional regressions with elastic net penalty.

Let's transform the data as follow:

$$\widetilde{\mathbf{x}}_i = (\mathbf{x}_i - \hat{\boldsymbol{\mu}}^{(t+1)})\sqrt{\hat{\tau}_i^{(t+1)}} \tag{2}$$

Then, following the neighborhood selection idea of Meinshausen and Bühlmann [6], we reconstruct the graph representing the connections ($\mathrm{p}_{jk} \neq 0$) among the components of $\boldsymbol{X}$ (and also of $\boldsymbol{Y}$) relying on estimating the following conditional regressions:

$$\hat{\mathbf{b}}_k = \mathrm{argmin}_{\mathbf{b}_k}\left\{||\widetilde{\mathbf{X}}_k - \mathrm{a}_k - \widetilde{\mathbf{X}}_{-k}\mathbf{b}_k||_2^2 + \lambda[\alpha||\mathbf{b}_k||_1 + (1-\alpha)||\mathbf{b}_k||_2^2]\right\} \tag{3}$$

where $\widetilde{\mathbf{X}}_k$ is the k-th column of $\widetilde{\mathbf{X}}$ ($\widetilde{\mathbf{X}}$ is n by p matrix of transformed observations, such that the i-th row is equal to $\widetilde{\mathbf{x}}_i^\top$), $\widetilde{\mathbf{X}}_{-k}$ is $\widetilde{\mathbf{X}}$ without the k-th column, $\alpha \in [0,1]$ controls the convex linear combination of 1-norm and 2-norm and λ captures the overall strength of the penalty.

We include in the neighborhood of the node k the node j if the corresponding coefficient of the component j in the estimated vector of regression coefficients $\hat{\mathbf{b}}_k$ is different from 0. Then, through the reconstructed neighborhoods of all nodes, we can produce an estimate $\hat{\mathbf{E}}$ of the edge set $\mathbf{E}$. In the situations where an edge (j,k) is included in $\hat{\mathbf{E}}$ according to the neighborhood of j, $\mathrm{ne}(j)$, but not accordingly to the neighborhood of k, $\mathrm{ne}(k)$, we use the AND rule suggested by Meinshausen and Bühlmann [6].

After estimating $\hat{\mathbf{E}}$, we rely on it to set the zero elements constraints in the current update $\hat{\boldsymbol{\Psi}}^{(t+1)}$. In particular the update $\hat{\boldsymbol{\Psi}}^{(t+1)}$ is the maximizer of the following constrained optimization problem, with $\hat{\mathbf{S}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^n \hat{\tau}_i^{(t+1)}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}^{(t+1)})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}^{(t+1)})^\top$:

$$\begin{aligned} &\max_{\boldsymbol{\Psi}}\left\{\log(\det(\boldsymbol{\Psi})) - \mathrm{trace}(\hat{\mathbf{S}}^{(t+1)}\boldsymbol{\Psi})\right\} \\ &\text{s.t.} \\ &\psi_{jk} = \psi_{kj} = 0 \text{ if edge } (j,k) \notin \hat{\mathbf{E}} \end{aligned} \tag{4}$$

We use the algorithm proposed by Hastie et al. [5] to solve (4) and obtain an estimate of $\boldsymbol{\Psi}$, given the edge set $\hat{\mathbf{E}}$ estimated in the first step. We refer to

this 2 stage procedure used to obtain the sparse update $\hat{\mathbf{\Psi}}^{(t+1)}$ with the name *2Sgelnet*. Note that *2Sgelnet* can be used directly to estimate the precision matrix when we assume that data is multivariate Gaussian (see [1]).

3 Simulations

We compare the performance of *2Stelnet* with *tlasso*, *glasso* and *2Sgelnet* algorithms using simulations. We consider seven different network's structures embedded into the sparsity pattern of the theoretical precision matrix. We generate heavy-tailed data from multivariate t-Student with three degrees of freedom.

In order to measure the classification performances we considered the F_1-score, which is is a good measure when, like in our case, there is imbalance among classes (edge or missing edge). The closer the score is to 1, the better the identification of sparsity pattern. For the assessment of the numerical accuracy of the estimates, we compute the Frobenius distance between the theoretical and estimated partial correlation matrices.

We simulate 30 datasets with 1000 observations each and we search for the optimal value of λ using BIC criterion, exploring a sequence of 100 values exponentially spaced between e^{-6} and 1. For *2Sgelnet* and *2Stelnet*, we set α equal to 0.5 and 1, where 0.5 assigns equal weight to the 1-norm and 2-norm penalty while 1 results in only a 1-norm penalization.

Table 1 Average F_1-score - multivariate t-Student

	Scale-Free	Random	Hub	Cluster	Band	Small-World	Core-Periphery
2Sgelnet - $\alpha = 0.5$	0.327	0.512	0.322	0.609	0.602	0.497	0.508
2Sgelnet - $\alpha = 1$	0.333	0.524	0.326	0.625	0.615	0.503	0.519
2Stelnet - $\alpha = 0.5$	0.984	0.976	0.984	0.937	0.900	0.985	0.768
2Stelnet - $\alpha = 1$	0.967	0.982	0.977	0.962	0.940	0.985	0.793
glasso	0.222	0.353	0.221	0.413	0.393	0.345	0.377
tlasso	0.640	0.643	0.524	0.650	0.549	0.653	0.579

In Table 1 we report the average F_1-scores of optimally selected models in 30 runs, highlighting best values in red. It is evident that *2Stelnet* turns out to be the best approach by a large margin. *2Sgelnet* outperforms *glasso*, but in most of the cases it is outperformed by *tlasso*. Using $\alpha = 0.5$, which corresponds to the elastic net penalty, does not always lead to better estimates. In fact, it is only the case when the underlying networks have scale-free and hub topologies. Nonetheless, the classification performances with the two values of α considered are quite close in general.

In Table 2, we report the average Frobenius distance of the optimal models. Results are qualitatively similar to the ones obtained in classification performances, with few exceptions. The *2Stelnet* procedure performs always the

Table 2 Average Frobenius distance - multivariate t-Student

	Scale-Free	Random	Hub	Cluster	Band	Small-World	Core-Periphery
2Sgelnet - $\alpha = 0.5$	1.658	1.468	1.598	1.533	1.526	1.632	1.989
2Sgelnet - $\alpha = 1$	1.665	1.460	1.615	1.524	1.523	1.638	1.934
2Stelnet - $\alpha = 0.5$	0.207	0.283	0.179	0.383	0.409	0.291	0.756
2Stelnet - $\alpha = 1$	0.237	0.278	0.191	0.363	0.389	0.297	0.675
glasso	1.396	1.297	1.307	1.510	1.689	1.461	2.054
tlasso	0.444	0.593	0.342	0.907	1.137	0.581	1.971

best. However, there are few instances in which *glasso* outperforms *2Sgelnet*. Again the elastic net penalty ($\alpha = 0.5$) is not always the best choice. We notice that $\alpha = 0.5$ works well when considering scale-free, hub and small-world structures.

4 European banking network

Inspired by Torri et al. [8], we consider the daily stock prices of 36 large European banks as data input to estimate, using the *2Stelnet* algorithm, the European banking network in the period 2018-2020. To deal with autocorrelation and heteroskedasticity, two common characteristics of financial time series, we fit an AR(1)-GARCH(1,1) model for each of the 36 time series of log-returns. We then use residuals to estimate the partial correlation network both for single years (2018, 2019, 2020) and for the period 2018-2020, using a rolling window of 1 year with shifts of 1 month. We set the value of $\alpha = 0.5$ and considered a sequence of 100 exponentially spaced values between e^{-6} and 1.5 for λ and select its best value using the BIC criterion.

Table 3 Network measures (mean values)

	Degree	Eccentricity	Distance	Clustering	Strength	N°Edges
2018	6.889	3.528	2.146	0.465	0.828	124
2019	7.000	3.694	2.235	0.471	0.870	126
2020	7.444	4.083	2.292	0.484	0.913	134

In Table 3, we report the mean values of some common network measures and the total number of edges detected. Small values of eccentricity and distance suggest that a shock could diffuse quickly in the network. The degree tells us about the average number of connections a bank has, while the clustering coefficient measures how many connections the banks connected to a specific bank have among themselves, pointing out how much they tend to form highly connected subgraphs. Small values of average distance and not too small mean values of clustering coefficient suggest that the underlying structures have features of small-world graphs [9]. The strength instead sug-

gests how intense the relationships among nodes are, becoming an indicator of potential crisis periods.

Looking at Figure 1, we detect a rising trend of the average strength even before the Covid-19 pandemic, nonetheless it is still visible the possible effect of the pandemic on the network strength. In fact, notice a sharp increase between January and April 2020 and a subsequent stabilization at an higher level for the entire 2020. The proposed approach can then be used to detect the presence and intensity of the recent pandemic crisis.

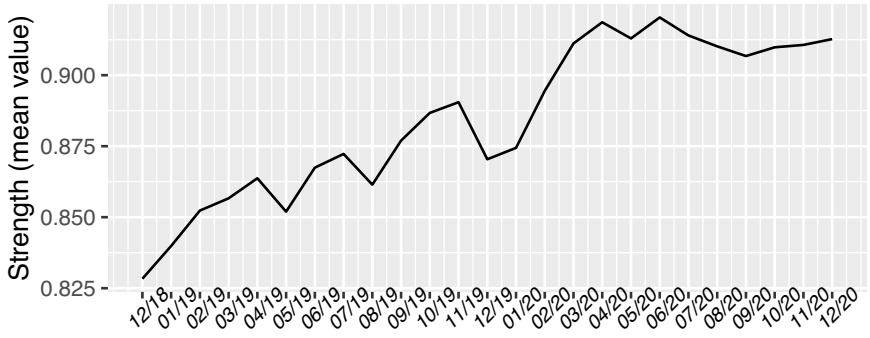


Fig. 1 Strength estimated using a rolling window of 1 year, with shifts of 1 month

References

1. Bernardini, D., Paterlini, S., Taufer, E., New estimation approaches for graphical models with elastic net penalty. arXiv:2102.01053 (2021)
2. Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society **39**, 1–38 (1977)
3. Finegold, M., Drton, M., Robust graphical modeling of gene networks using classical and alternative t-distributions. The Annals of Applied Statistics **5**, 1057–1080 (2011)
4. Friedman, J., Hastie, T., Tibshirani, R., Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**, 432–441 (2008)
5. Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2009)
6. Meinshausen, N., Bühlmann, P., High-dimensional graphs and variable selection with the lasso. The Annals of Statistics **34**, 1436–1462 (2006)
7. Tibshirani, R., Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society **58**, 267–288 (1996)
8. Torri, G., Giacometti, R., Paterlini, S., Robust and sparse banking network estimation. European Journal of Operational Research **270**, 51–65 (2018)
9. Watts, D. J., Strogatz, S. H., Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)
10. Zou, H., Hastie, T., Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society **67**, 301–320 (2005)

Neural Network for statistical process control of a multiple stream process with an application to HVAC systems in passenger rail vehicles

Rete neurale per il controllo statistico di un processo a flussi multipli con un'applicazione per sistemi HVAC nei veicoli ferroviari a trasposto passeggeri

Gianluca Sposito, Antonio Lepore, Biagio Palumbo, Giuseppe Giannini

Abstract A multiple stream process (MSP) is a process at a point in time that generates several streams of output with quality variable of interest and specifications that are identical in all streams. When the process is in control, the sources or streams are assumed to be identical, or, more in general, stationary of any kind. To enhance the monitoring of a MSP and the detection of changes in individual streams, a new control charting procedure is proposed based on artificial neural networks (NN). A wide Monte Carlo simulation is performed to assess the performance of the proposed approach and to compare it with the traditional Mortell and Runger's MSP control chart. The proposed approach is illustrated by means of a real-case study from a heating, ventilation and air conditioning (HVAC) systems installed on-board of all modern trains.

Abstract *Un processo a flussi multipli (MSP) è un processo descritto nel tempo da flussi concorrenti che misurano la stessa variabile di interesse, con identiche specifiche. Quando il processo è in controllo, i flussi sono ipotizzati identici, o più in generale, stazionari. Per migliorare il monitoraggio di un MSP e rilevare eventuali anomalie nei singoli flussi, viene presentato un nuovo approccio basato su reti neurali artificiali (NN). Mediante simulazione Monte Carlo, vengono valutate le prestazioni dell'approccio proposto e confrontate con la tradizionale carta di controllo per MSP di Mortell e Runger. L'approccio proposto viene applicato a dati reali provenienti da sistemi di riscaldamento, ventilazione e condizionamento dell'aria (HVAC) istallati a bordo di tutti i moderni veicoli ferroviari.*

Key words: Neural network, Multilayer perceptron, Multiple stream process, Statistical process control

---

Gianluca Sposito, Antonio Lepore, Biagio Palumbo
Department of Industrial Engineering, University of Naples Federico II, Naples, Italy, e-mail: gianluca.sposito@unina.it; antonio.lepore@unina.it; biagio.palumbo@unina.it

Giuseppe Giannini
Head of Operation Service and Maintenance Product Evolution, Hitachi Rail Group, Naples, Italy, e-mail: giuseppe.giannini@hitachirail.com

1 Introduction

Rail transport in Europe is as a viable alternative to other means of transport, and naturally leads to a fierce competition between operators. The comfort of thermal environment of passenger rail coaches, especially for long trips, is one of the most challenging and relevant aspects. European standards, such as UNI EN 14750 [9], have been established over the past few years and settle operational requirements of passenger rail coaches in terms of air quality and comfort level. Urged by these regulations and by passenger thermal comfort demand, in the last years, railway companies have been involved in gathering and storing data to track on-board heating, ventilation and air conditioning (HVAC) systems and to improve reliability and maintenance programs by applying predictive maintenance. Usually, each train is composed of more than one coach and each coach is equipped with a dedicated HVAC system. The sensor signals coming from each HVAC system can be assumed, under standard conditions, to be identical at least in terms of some summary statistics, i.e., stationary, and with identical specifications. Therefore, such setting can be regarded as a multiple stream process (MSP), i.e, a process at a point in time that generates several streams of output with quality variable of interest and specifications that are identical in all streams. When a MSP process is in control (IC), the sources or streams are assumed to be identical, or more in general, stationary of any kind. This paper examines the possibility of exploiting the nice properties of NNs by proposing a control charting procedure for the statistical process control (SPC) of a MSP, which is inspired by this industrial context and follows the latest promising applications of NNs to SPC [1]. The ultimate goal is setting up a procedure to track changes in one or a few process streams rather than in the overall process mean. In what follows, we aim (i) to explore the potentiality of using a NN in the SPC of a MSP; (ii) to train the proposed NN through a wide Monte Carlo simulation and compare it with the traditional MSP control chart based on the Mortell and Runger's range statistic [6], which is hereinafter referred to as R_t control chart; (iii) to apply the proposed approach to a real-case study concerning the monitoring of the HVAC systems installed on board of passenger railway vehicles. The data were acquired during lab tests and made available by the rail transport company Hitachi Rail based in Italy.

2 Materials and Methods

Let us consider the model [6]

$$Y_{tjk} = \mu + A_t + e_{tjk}, \tag{1}$$

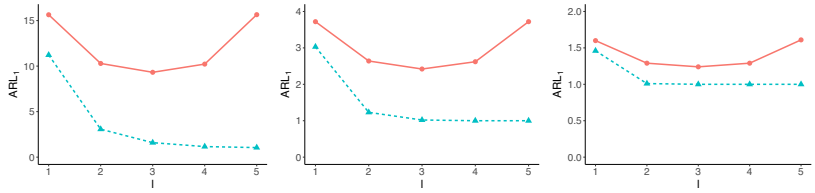
for $t = 1,2...,T$, $j = 1,2,...,s$ and $k = 1,2,...,n$, where Y_{tjk} is the measurement k of the quality variable from the stream j at time t, and μ is the process mean. The value Y_{tjk} can be expressed as the sum of each stream common component A_t, and

individual components e_{tjk}. The terms A_t and e_{tjk} are independent and distributed as standard normal random variables with variance σ_p^2 and σ_e^2, respectively. Note that in real MSPs, the term A_t can be however affected by autocorrelation. To mitigate this issue, it is convenient to monitor the residual $X_{tj} = Y_{tj} - Y_t$ [8] in place of the Y_{tj}, that has zero mean and variance σ^2. For the sake of simplicity, when we omit a subscript this means that the variable is averaged over that subscript, i.e., Y_t is the average over the subgroup means k across all the streams j at a sample time. A control charting procedure based on X_{tj} in place of Y_{tj} has the additional advantage of improving the sensitivity in the monitoring of individual stream shifts. Therefore, for each stream, we generate pseudorandom observations of X_{tj} from a standard normal random variable, and, without loss of generality, inspired by the real-case study, we set $s = 6$ and $n = 5$.

The multilayer perceptron (MLP) [3] is one of the most widely used NN to solve non-linear problems and has already been shown to achieve better performance than traditional Shewhart control charts, in terms of average run length (ARL) [4, 7, 10]. In this paper, we aim to extend the use of MLPs to the SPC of a MSP, that is, to solve the binary classification problem of detecting whether the MSP at hand is IC or out of control (OC). A sample drawn from an IC process is said *negative sample*, as usually associated to class zero. In contrast, a sample drawn from an OC process is said *positive sample* and associated to class one. As long as we deal with a binary problem classification, the activation function in the output layer is the sigmoid function and the cross-entropy is the loss function [3]. The input layer of the NN has $s+2$ neurons to represent the s residual means X_{tj} at time t for each stream $j = 1, 2, ..., s$, the grand average X_t over all the streams at time t and the range statistic R_t. The MLP is trained through back-propagation algorithm [3].

In order to train the MLP, a proper training set must be generated by balancing the number of *negative samples* with the total number of *positive samples* generated from $\sum_{l=1}^{s-1} \binom{s}{l}$ OC scenarios in which $l = 1, \ldots, s-1$ streams shift off the target at the same time by as much as $\Delta\mu = 1.0\sigma, 2.0\sigma, 3.0\sigma$. To be specific, for each of the latter 3 severity levels, 333 pseudorandom *positive samples* of size $n = 5$ for $s = 6$ streams are generated from $\sum_{l=1}^{5} \binom{6}{l} = 62$ OC scenarios and as many *negative samples*. Note that, even if in this paper we do not train explicitly the considered NN to signal which stream is OC, it is still crucial that training set is balanced with *positive samples* from all possible OC scenarios in order to allow the MLP to classify with the same probability as OC a MSP with l out of s streams shifted off target, regardless of their order.

In order to design the MLP, i.e., to analyze the sensitivity of the proposed procedure with respect to typical MLP hyperparameters (viz., output neuron threshold, number of hidden neurons and type of activation function in the hidden layer) [3], we use the so-called area under the receiver operating characteristic (ROC) curve (AUC) [2] as performance measure, known to be independent from the choice of the particular threshold in the output neuron. The ROC curve is calculated on a separate validation set, with size set equal to $1/2$ of that of the training data. Different NN architectures have been explored in terms of number of hidden neurons and activation functions. Even if the results are not shown here, a MLP with one hidden



(a) (b) (c)

Fig. 1: ARL_1 of the R_t control chart (solid line) and of the proposed NN approach (dotted line) based on 100 000 simulations of a MSP with $s = 6$ streams and $n = 5$, at different number $l = 1, ..., s-1$ of streams that shift off the target and mean shift size $\Delta\mu = 1.0\sigma(a), 1.5\sigma(b), 2.0\sigma(c)$

layer with five neurons and the rectified linear unit (ReLU) activation function turns out to have the best performance in terms of the area under the ROC curve (AUC). Furthermore, to set a fair comparison with the traditional SPC, the cut-off value (CV) of the output neuron is properly fixed in the extent of controlling the Type-I and Type-II error rates of the resulting control charting procedure, usually denoted by α and β, respectively. Then, further 100 000 pseudorandom *negative samples* are generated to set the CV threshold that corresponds to the desired IC average run length $ARL_0 = 1/\alpha$. By simulation, α is calculated as the proportion of misclassified *negative samples*. As an example, we set $ARL_0 = 370$. The OC average run length ARL_1 performance is finally evaluated through 100 000 additional simulations in the case that at least one stream of the MSP shifts off target (see Fig. 1). Analogously, $ARL_1 = 1/(1-\beta)$, and β is calculated as the number of misclassified *positive sample.* According to the existing SPC literature [5], the proposed NN based control charting procedure is compared with the competitor R_t control chart by means of the ARL_1, at given $ARL_0 = 370$. The upper control limit (UCL) for the R_t control chart, corresponding to an $ARL_0 = 370$, is found by simulation to be 2.35, which is coherent with values reported in the original paper [6]. Additional 100 000 simulations are needed to evaluate the ARL_1.

3 Conclusion

Fig. 1 displays the ARL_1 performance of the R_t control chart and the proposed one at different number $l = 1, 2..., s-1$ of streams that shift off target by as much as $\Delta\mu = 1.0\sigma, 1.5\sigma, 2.0\sigma$. It is clear that the proposed MLP method outperforms the R_t control chart in all simulated scenarios.

In addition, data mentioned in the introduction are used to test the practical applicability of the proposed method on a real-case study in the monitoring of the performance of HVAC systems that are installed on board of passenger trains. Table 1 summarises the variables used to describe HVAC operating conditions. The

monitoring variable is the difference between target temperature T_{set}, which is set automatically by the HVAC central unit on the basis of the other variables reported in Table 1, and T_{in}, which represents the attained interior temperature. Acquisition frequency on each of the 6 coaches of each train is equal to two minutes, but acquired data are made available to the service and maintenance department every 10 minutes, only. Thus, we suitably set $s = 6$ and $n = 5$, as already done in the simulation study. Also by means of this real-case study, even though results are not shown for data confidentiality reasons, the proposed NN based control charting procedure successfully has proven to be capable of enhancing the detection power of OC streams and the real-time prognosis of faults, with respect to the industrial common practice of triggering signals based on traditional Shewart's control charts.

Table 1: Operational variables available for each of the $s = 6$ train coaches

Variable	Description
T_{in}	Interior temperature
T_{out}	Outdoor Temperature
T_{set}	Target temperature
T_{supply}	Air flow temperature measured at the exit of the HVAC system

Acknowledgements This work has been done in the framework of the R&D project of the multi-regional investment programme "REINForce: REsearch to INspire the Future" (CDS000609) with Hitachi Rail S.p.A., supported by the Italian Ministry for Economic Development (MISE) through the Invitalia agency.

References

1. Bersimis, S., Psarakis, S., Panaretos, J.: Multivariate statistical process control charts: an overview. Qual. Reliab. Eng. Int. **23**(5), 517-543 (2007)
2. Fawcett, T: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861-874 (2006)
3. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning Cambridge. MIT press (2016)
4. Hwarng, H. B.: Neural Networks in Statistical Process Control. Wiley StatsRef: Statistics Reference Online (2014)
5. Montgomery, D. C.: Statistical quality control. Wiley Global Education (2012)
6. Mortell, R. R., Runger, G. C.: Statistical process control of multiple stream processes. J. Qual. Technol. **27**(1), 1-12 (1995)
7. Niaki, S. A., Abbasi, B.: Detection and classification mean-shifts in multi-attribute processes by artificial neural networks. Int. J. Prod. Res. **46**(11), 2945-2963 (2008)
8. Ott, E. R., Snee, R. D.:Identifying useful differences in a multiple-head machine. J. Qual Technol. **5**(2), 47-57 (1973)

9. UNI EN: 14750-1 Railway applications—air conditioning for urban and suburban rolling stock. Part 1: Comfort parameters (2006)
10. Zorriassatine, F., Tannock, J. D. T.: A review of neural networks for statistical process control. J. Intell. Manuf. **9**(3), 209-224 (1998)

Forecasting air quality by using ANNs

L'uso delle ANNs per la previsione della qualità dell'aria

Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Francesco Bucci

Abstract The artificial neural networks (ANNs) have been extensively used in air pollution prediction because of their flexibility to deal with processes involving non linear and complex data and/or to solve articulate problems in which a priori knowledge is incomplete or noisy. In this study, we trained different ANNs in assessing the capability of models for the prediction of air quality. The air pollution data from two monitoring stations in Pescara (Central Italy), along with some meteorological parameters, were used in forecasting Nitrogen Dioxide (NO_2) levels, one day in advance, in the area of interest. The evaluation of obtained results shows that the degree of success in forecasting NO_2 is promising.

Abstract *Le reti neurali sono state ampiamente utilizzate nella previsione dell'inquinamento atmosferico per la loro flessibilità nel trattare processi che coinvolgono dati non lineari e complessi. In questo studio, diverse reti neurali sono state allenate per la previsione della qualità dell'aria. I dati sull'inquinamento atmosferico, provenienti da due stazioni di monitoraggio di Pescara (Centro Italia), insieme ad alcuni parametri meteorologici, sono stati utilizzati per prevedere, con un giorno di anticipo, le concentrazioni di biossido di azoto nell'area di interesse. La valutazione dei risultati ottenuti mostra che il grado di successo nella previsione di NO_2 è promettente.*

Key words: Air quality, Artificial Neural Network, Multilayer perceptron, Forecast

---

Annalina Sarra
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: annalina.sarra@unich.it

Adelia Evangelista
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: adelia.evangelista@unich.it

Tonio Di Battista
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: tonio.dibattista@unich.it

Francesco Bucci
Independent researcher , e-mail: frabucci@gmail.com

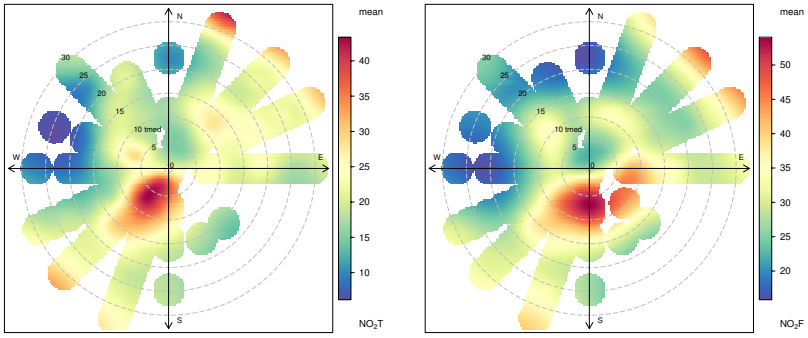
1 Introduction

In recent years, understanding the status and development of air quality levels in urban areas has become of paramount importance worldwide due to the severe effects of atmospheric pollution on human health [4]. In order to obtain accurate and comparable air quality information of a specific area, countries around the world have established air quality monitoring networks. By relying on data retrieved from monitoring stations, many studies are aimed at characterizing the underlying dynamic interaction of pollutant time series in different sites of the monitored areas. Over the years, the implementation of artificial intelligence techniques for air pollution time series modeling and air pollution-concentration forecasting as well, has significantly increased. Among them, Artificial Neural Networks (ANNs) have been widely used in air pollution prediction (see, among others, Hadjiiski and Hopke [3] and Boznar et al. [1]). In this paper, a modeling framework based on ANNs is explored to forecast Nitrogen Dioxide (NO_2) levels, one day in advance, in the urban area of Pescara (Central Italy). Air pollution data from two monitoring stations, along with some meteorological parameters, are included in the analysis. The benefit of using ANNs in our research is twofold. Firstly, ANNs help to establish which meteorological variables have a strong impact on the behavior of the target air pollutant. Secondly, by focusing on prediction of air pollution in each station, ANNs might represent an effective way to investigate redundancy and optimize the layout of air quality monitoring networks. The rest of the paper is arranged as follows: Section 2 illustrates the study area and the available data, Section 3 briefly describes the ANNs framework used and Section 4 is devoted to present the main results and some concluding remarks.

2 Study area and data

Air pollution data for this study consist of measurements of NO_2 obtained from Pescara hourly air quality reporting platform, run by Regional Agency for the Environmental Protection of Abruzzo Region (ARTA). Daily measurements of NO_2 pollutant have been collected from January 1, 2015 to December 31, 2017. Since weather strongly influences pollutants formation and transport, the analyzed data set also includes daily metereological variables, such as wind, rain, temperature.
NO_2 measurements were taken at two monitoring stations: one designed of *urban traffic type* and the other deemed as *urban background station*, representative of the population average exposure. The urban traffic (UT) station is located next to an one-way street (*Via Firenze*), characterized from constant vehicular traffic. The urban background (UB) station, close to *Teatro d'Annunzio*, is located away from urban traffic or other direct pollution sources. A first glance of the trend and variability of the NO_2 concentrations observed in the two monitoring stations reveals that there are, on average, higher values at the urban traffic station rather than in the background one, whereas the average difference between the concentrations measured

in the two stations is less marked in the summer semester (April- September). The highest concentrations of this pollutant are recorded at the urban traffic station, with an overall average of 34.2 μg /m^3. Wind vector data were used to help verifying the effects of synoptic meteorological conditions on NO_2 pollution. The bivariate polar plots of wind-direction and temperature contribution (Fig.1) show that in both stations the highest concentrations occur to a greater extent at low temperatures and this can be ascribed to domestic heating systems and combustion. However, there are higher concentrations observed even at higher temperatures which indicate the influence of road transport throughout the year.



(a) *Urban Background station*. (b) *Urban Traffic station*.

Fig. 1: The bivariate polar plots with wind direction and temperature presented on a radial scale.

3 Artificial Neural Networks model for Nitrogen Dioxide Prediction

Artificial neural network (ANN) is a computational model inspired by the human brain functioning [5]. ANNs may be defined as structure comprising of a group of interconnected basic processing units, called neurons, associated with a learning rule. The neurons provide a parallel processing of the data. In a commonly used ANN architecture, known as the multilayer perceptron, the neurons are arranged in layers. Here, we used the *Multi-Layer Perceptron Neural Network architecture (MLPNN)*. The chosen architecture is made up of only two layers of multiple neurons (the input and hidden layer) and of a single neuron in the output layer. The elemental structure of the adopted MLPNN is described in Fig.2.

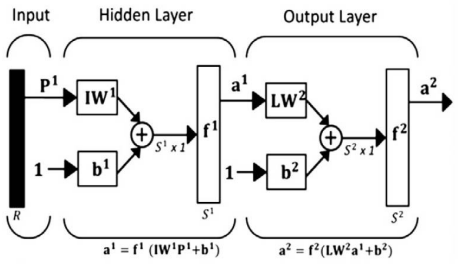


Fig. 2: Sketch of the MultiLayer Perceptron Neural Network (MLPNN) architecture

In this architecture, $\mathbf{P}^1$ is the input data vector, $\mathbf{IW}^1$, $\mathbf{LW}^1$ and $\mathbf{b}^i$ are the input and layer weight matrix and bias vectors that should be calculated by a training and validation procedure; $\mathbf{S}^1$ is the number of neurons in the hidden layer that should be optimized; $\mathbf{a}^i$ are the output vectors and $\mathbf{f}^i$ are the chosen transfer functions (tan-sigmoid and pureline). The input variables chosen for the considered built neural network consist of some meteorological parameters (i.e. average temperature, temperature max, temperature min, rain, wind speed, wind high, wind direction) and NO_2 observed concentrations.

4 Results and conclusions

ANNs are supervised learning techniques involving a training step to create a mathematical model and a prediction step to compute the output for a given set of input values using the model created in the training step. To define the best numbers of neurons and optimize the network we follow the procedure describe below [2].
For each monitoring station, the three years data 2015-2017 are divided into two data sets: the data for the years 2015 and 2016 constitute the set used to train the network whereas the data of the last year were employed in the validation phase. Hourly data have been aggregated into daily observations, and from each dataset missing values for more than 48 hours have been deleted. The Levenberg-Marquardt technique was used to train the chosen net, coupled with the repeated random sample validation procedure. The net structure was identified through an optimization process that provided the most favorable number of neurons in the hidden layer (S) through the

Mean Square Error (MSE) minimization procedure. For the training of the ANNs we used the Matlab Neural Networks Toolbox. In order to detect the optimum number of neurons for the hidden layer, we have built as many networks as those obtained by varying the number of neurons. In this respect, it worth noting that the optimum number of interior neurons was searched between 1 and 30. Each network has been trained 300 times and among them we have chosen the best one by looking at the minimum MSE. Subsequently, the selected structure has been trained to optimize the number of input variables, stopping the convergence process again in the minimum MSE. We built all the networks resulting from the all possible combinations of neurons and input variables. The better MLPNN networks in term of performance metrics are displayed in Table 1. According to the considered model indicators, the ANN with one neuron and all variables as input (model 1) has resulted the best architecture for the air quality prediction in the urban background (UB) monitoring station. As shown in Fig.3, this network has a very high accuracy in forecasting NO_2 concentrations. Looking at the values for the urban traffic (UT) monitoring station, from the aspect of absolute error (measured by RMSE and MAE) and min MSE, the most accurate prediction is achieved through the architecture of model 3. However, from Fig.3 it is evident that this model exhibits a poor performance in mimic the observed data at the beginning of time window analysed. Probably, the higher variability in the traffic volumes and the limited wind exposition of this monitoring site are at the basis of the initial poor accuracy of ANN that achieve better results for smooth data. Nonetheless, overall, the correlation coefficient for measured versus predicted NO_2 concentrations for both selected ANN models was shown to be greater of 0.70 (0.81 in the UB *vs* 0.71 in UT) over the span of 1 year.

Table 1: Models performance main results

Test days	Station	Model	Inputs	Neurons	minMSE	RMSE	CVRMSE	nMAE	CORR
707	UT	1	all	5	85.64	9.25	26.91	19.83	0.75
707	UT	2	NO_2, tmed, tmax, ws	16	84.73	9.5	27.99	21.72	0.69
707	UT	3	NO_2, tmed, tmax, ws, wdir	23	73.82	8.6	25	18.74	0.79
707	UT	4	NO_2, tmed, rain, wh, wdir	4	84.42	8	29.55	23.12	0.7
707	UT	5	NO_2, tmed, tmin, rain, wh, wdir	9	74	10.04	29.6	22.72	0.65
709	UB	1	all	1	53.85	7.34	30.76	22.29	0.81
709	UB	2	NO_2, tmax, ws, wh	12	51.68	7.8	28.76	22.52	0.71
709	UB	3	NO_2, tmed, tmax, tmin, rain, ws, wd, wdir	16	43.81	7.58	27.97	22.24	0.71
709	UB	4	NO_2, tmax, tmin, ws, wh, wdir	9	43.91	8	26.55	23.12	0.7

Legend:

minMSE = minimum Mean Square Error;
RMSE = Root Mean Square Error;
CVRMSE = Coefficient of Variation of RMSE;
nMAE= normalized Minimum Absolute Error;
CORR =Correlation coefficient

These are promising results: managers, authorities of urban air quality, practitioners and decision makers could efficiently exploit the toolkit of ANN modeling to estimate the temporal profile of pollutants and air quality indices. Additionally,

the different factors involved in the input stages in the two monitoring sites could provide useful insights in ascertaining which variables affect the spatial distributions of Nitrogen Dioxide (NO_2) concentrations and to what extent redundancy information can be detected in the monitoring stations.

(a) *Urban Background station*. (b) *Urban Traffic station*.

Fig. 3: Measured and predicted NO_2 concentrations at urban background and traffic stations.

References

[1] M. Boznar, M. Lesjak, and P. Mlakar. A neural network-based method for short-term predictions of ambient SO_2 concentration in highly polluted industrial areas of complex terrain. *Atmos. Environ.*, 27:221–230, 1993.

[2] C. Cornaro, F. Bucci, M. Pierro, F. Del Frate, S. Peronaci, and A. Taravat. Solar radiation forecast using neural networks for the prediction of grid connected PV plants energy production (DSP project). *Proceedings of 28th European Photovoltaic Solar Energy Conference and Exibition*, pages 3992–3999, Sept 30–Oct 4, 2013.

[3] L. Hadjiiski and P. Hopke. Application of artificial neural networks to modeling and prediction of ambient ozone concentrations. *J. Air Waste Manag. Assoc.*, 50(5):894–901, 2000.

[4] R. Kelishadi and P. Poursafa. Air pollution and non-respiratory health hazards for children. *Arch. Med. Sci.*, 6:483–495, 2010.

[5] R. Lippman. An introduction to computing with neural nets. *IEEE ASSP Mag.*, 4(2):4–22, 1987.

4.3 Advances in statistical methods

Robustness of Fractional Factorial Designs through Circuits

Piani Fattoriali Frazionari Robusti attraverso i Circuiti

Roberto Fontana and Fabio Rapallo

Abstract Given a model we define the robustness of an experimental design as a function of the number of estimable minimal sub-fractions of it. We show how the circuit basis of the design matrix can be used to see if a minimal fraction is estimable or not and we describe an algorithm for finding robust fractions.
Abstract *Dato un modello si definisce la robustezza di un piano sperimentale in funzione del numero dei suoi sottoinsiemi minimali per cui tale modello risulta stimabile. Si dimostra che è possibile determinare se una frazione minimale è stimabile usando i circuiti della design matrix e si descrive un algoritmo per la ricerca di frazioni robuste.*

Key words: algebraic statistics, design of experiments, robust fractions

1 Robustness of a Design

In Design of Experiments, the choice of a design from a candidate set of runs is probably the most relevant problem, with a number of open questions from the point of view of both theory and applications. When searching for an efficient experimental designs, we aim to select a design in order to produce the best estimates of the relevant parameters for a given sample size. There are a lot of criteria for the selection of a design, both in the area of model-based designs (e.g., D-optimality and related criteria), and model-free designs (e.g., orthogonal arrays, space filling designs). In this work we focus on the model-based setting and we limit the analysis

Roberto Fontana
Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, e-mail: roberto.fontana@polito.it

Fabio Rapallo
Dipartimento di Economia, Università di Genova, via Vivaldi 5, 16126 Genova, e-mail: fabio.rapallo@unige.it

to Fractional Factorial Designs. In particular, we consider the notion of robustness of a design. This property is particularly important when the design may be incomplete, for instance when for some reasons one does not have the measurement for all planned design points. Fractional Factorial Designs with removed runs are studied in, e.g., [1], [7], [11]. A combinatorial analysis of the problem is introduced in [2], but in a model-free context.

Following the works by Ghosh ([4] and [5]), we define here the robustness in terms of the estimability of a given model on the basis of incomplete designs.

Let $\mathscr{D}$ be a large discrete set in $\mathbb{R}^m$ from which a small set $\mathscr{F}$, usually referred to as design or fraction, is to be selected. One standard example is to consider as candidate set $\mathscr{D}$ the cartesian product of the level sets of the m factors. The set $\mathscr{D}$, when thought of as a design in its own right, is referred to as a full factorial.

We point out that in our theory the coding of the level set is irrelevant, so that for the level set of a factor with s levels we can use $\{0,\dots,s-1\}$ or the complex coding or any other coding that is considered appropriate.

A linear model on the candidate set $\mathscr{D}$ is written as:

$$\mathbf{y} = X_{\mathscr{D}}\beta + \varepsilon\,,$$

where $X_{\mathscr{D}}$ is the full-design model matrix with dimensions $K \times p$, β is the p-dimensional vector of the parameters, $\mathbb{E}(\mathbf{y}) = X_{\mathscr{D}}\beta$. In this work we assume that the design matrix $X_{\mathscr{D}}$ is full rank to simplify the presentation.

In the model-based approach to experimental design, the quality of the chosen design $\mathscr{F}$ is expressed by some properties of $X_{\mathscr{F}}$. Here we focus our attention on the notion of robustness. Robustness measures how many minimal fractions (i.e., fractions with as many points as the number of parameters) are estimable. A minimal estimable fraction is named as a saturated fraction.

Definition 1. We define the *robustness* of a fraction $\mathscr{F}$ with design matrix $X_{\mathscr{F}}$ as

$$r(X_{\mathscr{F}}) = \frac{\#\text{ saturated }\ \mathscr{F}_p}{\#\mathscr{F}_p} = \frac{\#\text{ saturated }\ \mathscr{F}_p}{\binom{n}{p}} \tag{1}$$

where $\mathscr{F}_p$ denotes a fraction with p runs and $\#\{\cdot\}$ denotes the cardinality of the set $\{\cdot\}$.

In particular a design $\mathscr{F}$ is *robust* if its robustness is equal 1, $r(X_{\mathscr{F}}) = 1$.

2 Circuits of the Design Matrix and Robustness

Let $X = X_{\mathscr{F}}$ be a model matrix on $\mathscr{F}$, and assume that X has integer entries. To simplify the notation, we drop the subscript $\mathscr{F}$ if there is no ambiguity. The matrix X has dimensions $K \times p$. Moreover, in order to match the common notation in Statistics with the notation in Commutative Algebra, we consider the matrix $A = X^T$, the transpose of the model matrix.

We are interested to use a special basis of the kernel of A as a lattice in $\mathbb{Z}^K$. A vector $\mathbf{u} = (u(1), \ldots, u(K))$ belongs to $\ker(A)$ if $A\mathbf{u} = 0$, or equivalently if $\mathbf{u}$ is orthogonal to $A^T = X$. The *support* of a vector $\mathbf{u} \in \mathbb{Z}^K$ is the set of indices i $(i = 1, \ldots, K)$ such that $u(i) \neq 0$. We denote the support of $\mathbf{u}$ with $\mathrm{supp}(\mathbf{u})$.

The circuits of A are the integer vectors of $\ker(A)$ with relatively prime entries and with minimal support. This means that if $\mathbf{u}$ is a circuit, then there does not exist another circuit $\mathbf{v}$ with $\mathrm{supp}(\mathbf{v}) \subset \mathrm{supp}(\mathbf{u})$. We denote the set of all circuits of A with $\mathscr{C}(A)$ (or $\mathscr{C}(X)$ if we refer to a design matrix $X = A^T$). The set $\mathscr{C}(A)$ is called the circuit basis of the matrix A. Among the properties of the circuit basis, we make use of the following:

1. $\mathscr{C}(A)$ is a basis of $\ker(A)$ as vector space.
2. Every vector $\mathbf{v} \in \ker(A)$ can be written as a non-negative rational combination of $(K - p)$ circuits
$$\mathbf{v} = \sum c_j \mathbf{u}_j \qquad c_j \in \mathbb{Q}^+,\ \mathbf{u}_j \in \mathscr{C}(A)$$
and each circuit in the decomposition above is sign-compatible with $\mathbf{v}$.
3. The support of a circuit has cardinality at most $(p+1)$.

For the proofs and for a detailed introduction to circuits in the context of Commutative Algebra and Combinatorics, see [8].

The circuit basis of an integer matrix A can be computed through several packages for symbolic computation. The computations presented in the present paper are carried out with `4ti2`, see [9]. `4ti2` can be used as an independent executable program or as a package of the Computer Algebra System `Macaulay2`, see [6]. For small designs the computations are performed in a few seconds at most, and the circuit basis in the output can be easily analyzed. For instance, if we want to compute the circuit basis for the full factorial 2^4 design with main effects and first order interactions, it is enough to run `4ti2`, input the design matrix and the circuit basis with 140 elements is computed in less than 0.1 seconds on a standard PC.

From the minimal support property in the definition of the circuits we can use the circuits to see if a minimal fraction (i.e., a fraction with exactly p points) is saturated or not. Given a set $\mathscr{F}$ of p column-indices of A, the sub-matrix $A_{\mathscr{F}}$ is non-singular if and only if $\mathscr{F}$ does not contain any of the supports of the circuits $\mathbf{u} \in \mathscr{C}(A)$. This result is proved and applied to the analysis of Fractional Factorial Designs in [3].

Now, we extend the analysis to fractions with more than p points. The key property here is that the circuit basis is consistent under selection of sub-fractions. Consider two fractions $\mathscr{F}_1$ and $\mathscr{F}_2$ with k_1 and k_2 design points respectively, such that $\mathscr{F}_1 \subset \mathscr{F}_2$. Without loss of generality, the matrix $X_{\mathscr{F}_2}$ can be partitioned into

$$X_{\mathscr{F}_2} = \begin{pmatrix} X_{\mathscr{F}_1} \\ X_{\mathscr{F}_2 - \mathscr{F}_1} \end{pmatrix}$$

and each vector $\mathbf{u} \in \mathbb{Z}^{k_2}$ can be written as

$$\mathbf{u} = (\mathbf{u}_{|\mathscr{F}_1}, \mathbf{u}_{|\mathscr{F}_2 - \mathscr{F}_1}) \qquad \text{with } \mathbf{u}_{|\mathscr{F}_1} \in \mathbb{Z}^{k_1}.$$

Theorem 1. *If $\mathscr{F}_1$ and $\mathscr{F}_2$ are two fractions with $\mathscr{F}_1 \subset \mathscr{F}_2$, then the circuits in $\mathscr{C}(X_{\mathscr{F}_1})$ are*

$$\{\mathbf{u}_{|\mathscr{F}_1} \,:\, \mathbf{u} \in \mathscr{C}(X_{\mathscr{F}_2}) \;\; \textit{with} \;\; \mathrm{supp}(\mathbf{u}) \subset \mathscr{F}_1\}\,.$$

The result above lead us to the construction of an algorithm for finding robust fractions. The strategy is as follows. First, a good fraction should avoid as much as possible the circuits with support on p points or less. Second, small circuits are worse than large circuits, since they are contained in a larger number of minimal fractions, leading to a higher loss in robustness. Thus, at each step a loss function is computed for each point R of the current fraction as the number of minimal fractions becoming non-estimable when removing the point R. In formulae:

$$L(R) = \sum_{\mathbf{u}} \binom{n - \#\mathrm{supp}(\mathbf{u})}{p - \#\mathrm{supp}(\mathbf{u})} \tag{2}$$

where the sum is taken over all the circuits $\mathbf{u}$ in the current fraction containing the point R. Notice that the formula in Eq. (2) does not guarantee that the relevant minimal fractions are all distinct. The formula should be viewed as a first-order approximation of the inclusion-exclusion formula.

Therefore, we can define a selection algorithm as detailed below. It works like an exchange algorithm as introduced for optimal designs in [10]. To run such an algorithm we only need the circuit basis $\mathscr{C}(X_{\mathscr{D}})$, and we extract the circuits of $\mathscr{C}(X_{\mathscr{D}})$ with support on p points or less. We denote this set of circuits by $\mathscr{C}^p(X_{\mathscr{D}})$.

1. Starting with an arbitrary fraction $\mathscr{F}$ of a specified size n;
2. Repeat:
   a. Consider the circuits of $\mathscr{C}^p(X_{\mathscr{D}})$ which are contained in $\mathscr{F}$;
   b. For each point R in $\mathscr{F}$ compute its associated loss $L(R)$ as the (weighted) number of circuits which include R;
   c. Take all the points with the highest loss and build up all possible pairs with one point not in $\mathscr{F}$. Make the exchange using the pair which reduces as much as possible the number of circuits contained in the fraction.
   d. If no reduction is possible, then break.

3 Examples and Final Remarks

We describe the use of the algorithm on some examples where the candidate sets are full factorial designs. We consider four and five 2-level factors with the main-effect model and two mixed-level cases. In both mixed-level cases we consider three factors, with 2, 3 and 4 levels. In the first case we work with the main-effect model without interactions and in the second one with the main-effect model plus the interaction between the second and the third factor. The candidate sets are the 2^4, the 2^5 and the $2 \times 3 \times 4$ full factorial designs respectively. The algorithm is used for finding

robust fractions with different sizes. For each case the algorithm has been used starting from 500 randomly selected fractions. For each case, in Tables 1 and 2 the mean value $\bar{r}_B$ of the robustness of the randomly selected fractions, the mean value $\bar{r}_A$ and the maximum value $\max r_A$ of the robustness of the fractions which are the output of the algorithm are reported. It is worth noting that in all cases the mean of the final robustness ($\bar{r}_A$) is greater than the mean of the robustness of the starting fractions ($\bar{r}_B$), and the efficiency of the algorithm is strong especially in the binary cases. In Fig. 1 the 500 pairs of robustness of the starting fraction and robustness of the final fraction are reported in the case of 5 factors and $n = 10$ as pre-specified size of the fraction. We observe that the distribution of the robustness of the final fractions is better than that of the starting fractions both in terms of mean and dispersion.

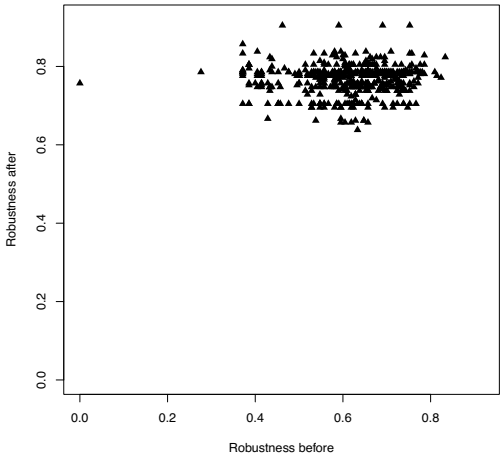


Fig. 1 2^5 design and fractions of size $n = 10$. Robustness after *vs* robustness before the algorithm on a set of 500 fractions.

	4 factors			5 factors		
n	$\bar{r}_B$	$\bar{r}_A$	$\max r_A$	$\bar{r}_B$	$\bar{r}_A$	$\max r_A$
8	0.6889	0.8214	0.8214	0.6134	0.8163	0.9643
10	0.6909	0.7619	0.7619	0.6203	0.7700	0.9048
12	0.6887	0.7222	0.7222	0.6164	0.7589	0.8171
14	0.6887	0.6933	0.6933	0.6134	0.7162	0.7586

Table 1 Mean and maximum values of the robustness: 2^4 and 2^5 designs under the main-effect model.

The computation of the circuit basis is actually feasible only for small designs, with at most 6 to 9 factors, depending on the number of interactions included in the model. We are working on an algorithm that uses only the circuits with mini-

	2×3×4 no interaction			2×3×4 with interaction		
n	$\bar{r}_B$	$\bar{r}_A$	$\max r_A$	$\bar{r}_B$	$\bar{r}_A$	$\max r_A$
14	0.3818	0.4482	0.5216	0.0086	0.2857	0.2857
16	0.3793	0.3985	0.4399	0.0097	0.0571	0.0571
18	0.3838	0.3854	0.3986	0.0102	0.0224	0.0224
20	0.3833	0.3853	0.3906	0.0098	0.0132	0.0132

Table 2 Mean and maximum values of the robustness: $2 \times 3 \times 4$ model under the main-effect and the first-order-interaction models.

mal support. This new version of the algorithm could be used also in the case of large designs, since in most cases the circuits with minimal support can be defined theoretically, without any computation.

Acknowledgements

This paper is part of a joint project with Henry Wynn, Emeritus Professor of Statistics, London School of Economics. Roberto Fontana gratefully acknowledges financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

References

1. Butler, N.A., Ramos, V.M.: Optimal additions to and deletions from two-level orthogonal arrays. J. R. Stat. Soc. Ser. B **69**(1), 51–61 (2007)
2. Fontana, R., Rapallo, F.: On the aberrations of mixed level orthogonal arrays with removed runs. Stat. Pap. (Berl.) **60**(2), 479–493 (2019)
3. Fontana, R., Rapallo, F., Rogantin, M.P.: A characterization of saturated designs for factorial experiments. J. Stat. Plan. Inference **147**, 205–211 (2014)
4. Ghosh, S.: On robustness of designs against incomplete data. Sankhya Ser. B **40**(3/4), 204–208 (1979)
5. Ghosh, S.: Robustness of BIBD against the unavailability of data. J. Stat. Plan. Inference **6**(1), 29–32 (1982)
6. Grayson, D.R., Stillman, M.E.: Macaulay2, a software system for research in algebraic geometry (2019). URL http://www.math.uiuc.edu/Macaulay2/
7. Street, D.J., Bird, E.M.: *D*-optimal orthogonal array minus *t* run designs. J. Stat. Theory Pract. **12**(3), 575–594 (2018)
8. Sturmfels, B.: Gröbner bases and convex polytopes, *University Lecture Series*, vol. 8. American Mathematical Society, Providence, RI (1996)
9. 4ti2 team: 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces (2018). URL https://4ti2.github.io
10. Wynn, H.P.: The sequential generation of *D*-optimum experimental designs. Ann. Math. Stat. **41**(5), 1655–1664 (1970)
11. Xampeny, R., Grima, P., Tort-Martorell, X.: Which runs to skip in two-level factorial designs when not all can be performed. Qual. Eng. **30**(4), 594–609 (2018)

Multi-objective optimal allocations for experimental studies with binary outcome

Allocationi ottime multi-obiettivo per studi sperimentali con risposta binaria

Alessandro Baldi Antognini, Rosamarie Frieri, Marco Novelli and Maroussa Zagoraiou

Abstract The design of multi-arm clinical trials should take into account several objectives. From the one hand, in order to maximize patients' care, the treatment assignments should be unbalanced in favour to the superior experimental arm. On the other hand, according to the inferential goal of deriving correct statistical conclusions with high precision, the design strategies should be aimed at optimizing a suitable inferential criterion. Following a multi-purpose design methodology, optimal designs for testing the efficacy of several treatments have been derived in [3] in unified framework for heteroscedastic experimental groups that encompasses the general ANOVA set-up. Starting from these results, the objective of this work is to assess the performance of the proposed optimal allocations for binary outcomes in terms of power, estimation efficacy and ethics, also performing comparisons with other designs presented in the literature.

Abstract *Il disegno di prove cliniche per confrontare trattamenti sperimentali dovrebbe tener contro di diversi obiettivi. Da una parte, per massimizzare i benefici dei pazienti, le assegnazioni dovrebbero essere sbilanciate in modo da favorire l'allocazione al trattamento migliore. Dall'altra, con l'obiettivo inferenziale di ottenere conclusioni statisticamente corrette, il disegno dovrebbe essere derivato in modo tale da ottimizzare un opportuno criterio inferenziale. Seguendo una metodologia multi-obiettivo, in [3] sono stati derivati disegni ottimi per testare l'efficacia di diversi trattamenti usando un framework unificato per gruppi sperimentali eteroschedastici che comprende l'analisi della varianza ad un fattore. Partendo da questi risultati, il lavoro ha l'obiettivo di valutare le prestazioni delle allocazioni ottime proposte nel caso di un modello binario, anche considerando confronti con altri disegni presentati in letteratura.*

Key words: unbalanced allocations, binary trials, power of the Wald test, ethics.

Alessandro Baldi Antognini, Dept of Statistical Sciences, University of Bologna ·
Rosamarie Frieri, Dept of Statistical Sciences, University of Bologna (rosamarie.frieri2@unibo.it) ·
Marco Novelli, Dept of Statistical Sciences, University of Bologna ·
Maroussa Zagoraiou, Dept of Statistical Sciences, University of Bologna.

1 Introduction

The design of randomized clinical trials is a complex issue where multiple experimental objectives, often conflicting, should be simultaneously considered. Nowadays, it is clear that the popular balanced design is unsuitable in many set-ups [2, 4, 8]. The equal allocation has been widely used as it mirrors the condition of equipoise at the beginning of the trial and it is usually considered optimal for the estimation of treatment effects. However, multi-arm trials (i.e., trials comparing $K \geq 2$ treatments) have been increasingly adopted to speed-up the pharmaceutical development and, in the presence of several experimental groups, balancing the assignments may be not efficient from both ethical and statistical viewpoints. Indeed, (i) the demand of patient's benefit often induces to skew the allocations in favour of the most efficacious treatment(s) and ii) the equal allocation does not coincide with the optimal design for hypothesis testing [2, 11].

Besides the inferential problem of estimating the treatment effects as precisely as possible, the task of designing an experiment to maximize the power of the test of homogeneity among the treatment effects has recently interested many authors [2, 3, 4, 11]. Note that the overall null hypothesis in a multi-arm context allows to compare many competing treatments at once, which could be particularly useful in trials for new infectious diseases [6, 7].

The aim of this work is to elaborate on the results of [3] in which optimal designs for testing in an unified framework have been derived. Such designs include also the general ANOVA set up. We focus on clinical experiments with dichotomous outcomes and we show that the proposed optimal allocations present good results in terms of power, ethics and also estimation efficiency.

2 Notation and model

Let us consider a clinical trial with binary outcomes in which patients join sequentially the study and are assigned to one of $K \geq 2$ treatments. The success probabilities on each treatment group are $p_1, \ldots, p_K$, (so that we assume that higher values are more desirable for patients), with $\boldsymbol{p} = (p_1, \ldots, p_K)^\top$ and the failure probabilities are $q_k = 1 - p_k$ for $k = 1, \ldots, K$. At each step j, the treatment assignment indicator $\delta_{kj} = 1$ if patient j is allocated to treatment k ($\delta_{kj} = 0$ otherwise), for $k = 1, \ldots, K$ with $\sum_{i=1}^{K} \delta_{ij} = 1$. The corresponding response of subject j is denoted by Y_j with $E(Y_j|\delta_{kj} = 1) = p_k$ and $V(Y_j|\delta_{kj} = 1) = p_k q_k$. Moreover, let $\pi_{kj} = j^{-1} \sum_{i=1}^{j} \delta_{ki}$, then $\boldsymbol{\pi}_j = (\pi_{1j}, \ldots, \pi_{Kj})^\top$ is the vector of the allocation proportions to the treatments so far, where $\sum_{k=1}^{K} \pi_{kj} = 1$. We also denote by $\hat{\boldsymbol{p}}_j$ the vector of MLEs of $\boldsymbol{p}$ after j treatment assignments.

For ease of notation and without loss of generality we assume that $p_1 \geq p_2 \geq \cdots \geq p_K$, i.e. the treatment with the highest probability of success is denoted with label 1 and the one with the lowest is denoted with label K. Note that the treatment

ranking is a priori unknown but it can be sequentially estimated by using response adaptive randomization procedures [1, 5, 9].

As in many multi-arm clinical trials, the inferential focus is on the contrasts. Therefore, by letting $\mathbf{A}^\top = [\mathbf{1}_{K-1} | -\mathbf{I}_{K-1}]$ ($\mathbf{1}_w$ and $\mathbf{I}_w$ are the w-dim vector of ones and the identity matrix, respectively), we denote by $\boldsymbol{p}_c = \mathbf{A}^\top \boldsymbol{p} = (p_1 - p_2, \ldots, p_1 - p_K)^\top$ the vector of contrasts wrt the first treatment and by $\hat{\boldsymbol{p}}_{cj} = \mathbf{A}^\top \hat{\boldsymbol{p}}_j$ their MLEs (at step j). If $\boldsymbol{\Sigma} = \text{diag}\left(\pi_{kj}^{-1} p_k q_k\right)_{k=1,\ldots,K}$ is the Fisher information matrix for $\boldsymbol{p}$ and $\boldsymbol{\Sigma}_c = (\mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A})^{-1}$, by well-known results $\hat{\boldsymbol{p}}_{cj} \xrightarrow{a.s.} \boldsymbol{p}_c$ and $\sqrt{n}(\hat{\boldsymbol{p}}_{cj} - \boldsymbol{p}_c) \xrightarrow{d} \text{N}(\mathbf{0}_{K-1}, \boldsymbol{\Sigma}_c)$.

In this setting, the typical question of experimental design is how to assign n subjects to K treatments. In the presence of multiple experimental objectives, the compromise among the conflicting goals can be formalized through suitable optimization problems. The solution is a target allocation $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)^\top$, with $\rho_k \geq 0$ and $\sum_{i=1}^K \rho_i = 1$.

3 Optimal designs for binary model

By taking into account the inferential goal of estimating treatment contrasts, adopting the well-known A-optimal criterion, Sverdlov and Rosenberger [10] derived the tr_A optimal target $\boldsymbol{\rho}^A$ by minimizing the trace of $\boldsymbol{\Sigma}_c$, i.e.

$$\rho_1^A = \frac{\sqrt{p_1 q_1 (K-1)}}{\sqrt{p_1 q_1 (K-1)} + \sum_{i=2}^K \sqrt{p_i q_i}} \quad \text{and} \quad \rho_k^A = \frac{\sqrt{p_k q_k}}{\sqrt{p_1 q_1 (K-1)} + \sum_{i=2}^K \sqrt{p_i q_i}} \tag{1}$$

for $k = 2, \ldots, K$.

However, in experimental studies for treatment comparisons, many authors have recently considered the problem of hypothesis testing. For multi-arm binary trials, in [11] the optimal target $\boldsymbol{\rho}^R$ has been proposed. It is obtained by maximizing the power of the Wald test subject to the constraint $\rho_k^R \geq T$, $\forall k = 1, \ldots, K$ i.e. the allocation proportion to each treatment should be at least T, where $T \in [0, 1/K]$ is a user-selected threshold. Along the lines of [2, 4], in [3] the authors extended the previously obtained results by providing optimal targets maximizing the power of the Wald test for a general heteroscedastic model. In particular, by taking into account $\text{H}_0 : \boldsymbol{p}_c = \mathbf{0}_{K-1}$ vs $\text{H}_1 : \boldsymbol{p}_c \neq \mathbf{0}_{K-1}$ (where $\mathbf{0}_{K-1}$ is the $(K-1)$-dimensional vector of zeros) the optimal target $\boldsymbol{\rho}^*$ is defined as follows. Given $p_1 = \cdots = p_r \geq p_{r+1} \geq \cdots \geq p_{K-s} > p_{K-s+1} = \cdots = p_K$ with r, s positive integers such that $r + s \leq K$, every allocation such that $\sum_{i=1}^r \rho_i^* = (\sqrt{p_1 q_1} + \sqrt{p_K q_K})^{-1} \sqrt{p_1 q_1} = 1 - \sum_{i=K-s+1}^K \rho_i^*$ is optimal. Note that in the presence of a single superior and inferior treatment $\boldsymbol{\rho}^*$ is unique and it is a generalization of the Neyman allocation with non-zero allocation proportions only to treatments 1 and K.

In addition, taking into account ethical considerations, in [3] the target $\tilde{\boldsymbol{\rho}}$ has been derived by maximizing the power of Wald's test under the constraint $\tilde{\rho}_1 \geq \cdots \geq \tilde{\rho}_K$,

which resembles the need for patients' care to receive the more effective treatments. The optimal constrained design $\tilde{\boldsymbol{\rho}}$, in the presence of a single superior treatment (i.e. $p_1 > p_2$) can be restated as

$$\tilde{\boldsymbol{\rho}} = \begin{cases} (1-(K-1)x, x, \ldots, x)^\top & \text{if } x \leq 1/K, \\ \boldsymbol{\rho}^B & \text{if } x > 1/K, \end{cases} \tag{2}$$

where $\boldsymbol{\rho}^B$ is the balanced allocation and

$$x = \frac{[(\sum_{i=1}^K \frac{p_1-p_i}{q_1 q_i})(\sum_{i=1}^K \frac{p_1-p_i}{p_1 p_i})]^{-1/2} \sum_{i=1}^K \frac{p_1-p_i}{p_i q_i} - 1}{\sum_{i=1}^K \frac{p_1 q_1}{p_i q_i} - K}.$$

In the presence of a group of superior treatments, i.e. $p_1 = \cdots = p_r > p_{r+1} \geq \cdots \geq p_K$ (with $r \in \{2, K-1\}$) when $x > 1/K$ then $\tilde{\boldsymbol{\rho}} = \boldsymbol{\rho}^B$. Whereas for $x \leq 1/K$, by letting $\tilde{\boldsymbol{\rho}}_{(i)}$ be the target such that $\sum_{h=1}^i \tilde{\rho}_h = 1-(K-i)x$, $\tilde{\rho}_{i+1} = \cdots = \tilde{\rho}_K = x$ and clearly $\tilde{\rho}_1 \geq \cdots \geq \tilde{\rho}_i \geq x$, every convex combination of $\tilde{\boldsymbol{\rho}}_{(1)}, \ldots, \tilde{\boldsymbol{\rho}}_{(r)}$ is optimal. For instance, if $\boldsymbol{p} = (0.6, 0.6, 0.2, 0.1)^\top$ then $r = 2$ and e.g. $\tilde{\boldsymbol{\rho}} = (0.37, 0.21, 0.21, 0.21)^\top$ and all the other targets such that $\tilde{\rho}_1 + \tilde{\rho}_2 = 0.58$ are optimal.

In Table 1 we report the behaviour of $\tilde{\boldsymbol{\rho}}$ for $K = 4$ treatments and increasing p_1 (scenario I-II), increasing p_2 (scenario III-IV) and increasing p_3 (scenario IV-V).

scenario	p_1	p_2	p_3	p_4	$\tilde{\rho}_1$	$\tilde{\rho}_2$	$\tilde{\rho}_3$	$\tilde{\rho}_4$
I	0.33	0.30	0.28	0.10	0.25	0.25	0.25	0.25
II	0.60	0.30	0.28	0.10	0.52	0.16	0.16	0.16
III	0.80	0.50	0.40	0.20	0.43	0.19	0.19	0.19
IV	0.80	0.70	0.40	0.20	0.34	0.22	0.22	0.22
V	0.80	0.70	0.50	0.20	0.31	0.23	0.23	0.23

Table 1 Behaviour of $\tilde{\boldsymbol{\rho}}$ for $K = 4$ treatments and several values of $\boldsymbol{p}$.

Starting from the balanced design (obtained in scenario I), the allocation proportion to the superior treatment increases as its success probability grows (scenario II). Conversely, by comparing scenario III and IV we observe how the proportion of subjects assigned to treatment 2 increases when p_2 increases, leading clearly to a smaller value for ρ_1. When p_3 grows (scenario IV-V) also $\tilde{\rho}_3$ behaves accordingly.

4 Performance in terms of power, estimation efficacy and ethics of optimal targets

As correctly stated by many authors (see e.g. [10]), there exists a complex interplay between the efficiency criteria adopted to assess the design performances. More specifically, measures of power, estimation efficiency or ethical considera-

tions lead to different choices of the best design. In this work, we consider a measure of efficiency in terms of power i.e. $\mathscr{E}_P(\boldsymbol{\rho}) = \phi(\boldsymbol{\rho})/\phi(\boldsymbol{\rho}^*)$ and the tr_A efficiency, $\mathscr{E}_A(\boldsymbol{\rho}) = \frac{\text{tr}[\mathbf{A}^\top \boldsymbol{\Sigma}(\boldsymbol{\rho}^A)\mathbf{A}]}{\text{tr}[\mathbf{A}^\top \boldsymbol{\Sigma}(\boldsymbol{\rho})\mathbf{A}]}$. As a global measure of ethics, we take into account the normalized total expected outcome $\mathscr{E}_E(\boldsymbol{\rho}) = p_1^{-1}\sum_{i=1}^K p_i\rho_i$. We compare the above-mentioned optimal targets $\boldsymbol{\rho}^A$, $\boldsymbol{\rho}^*$, $\tilde{\boldsymbol{\rho}}$ and $\boldsymbol{\rho}^R$ for $T = 0.2$. As a benchmark we also included the balanced design $\boldsymbol{\rho}^B$. Figure 1 displays the operating characteristics of the targets for $K = 4$ treatments, for $p_2 = 0.2$, $p_3 = 0.15$, $p_4 = 0.1$ and p_1 ranging from 0.3 to 0.8.

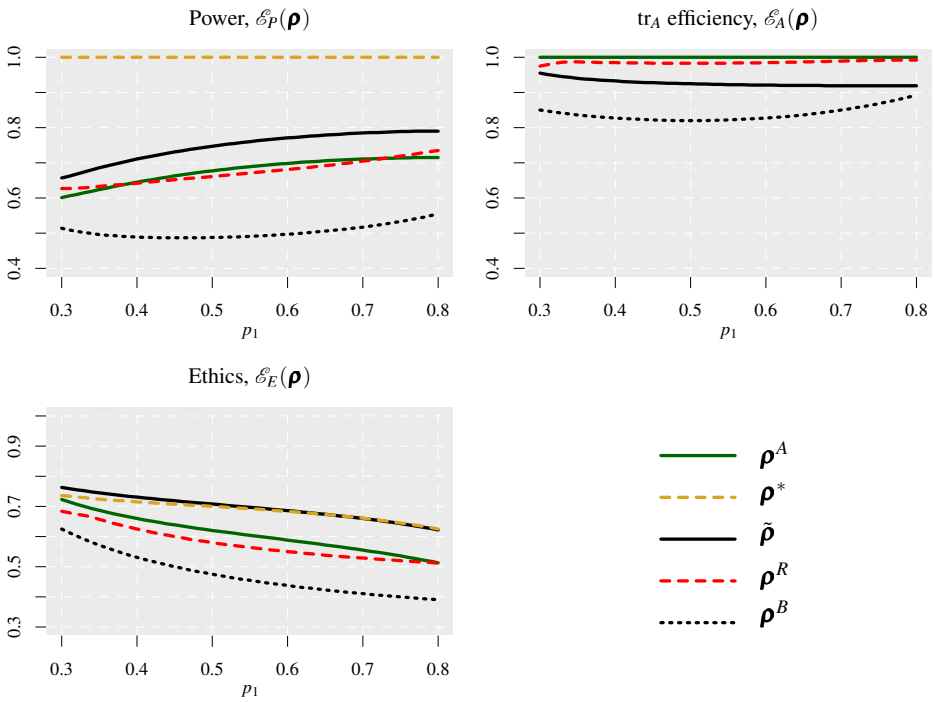


Fig. 1 Comparisons of $\boldsymbol{\rho}^A$, $\boldsymbol{\rho}^*$, $\tilde{\boldsymbol{\rho}}$, $\boldsymbol{\rho}^R$ for $T = 0.2$ and $\boldsymbol{\rho}^B = (0.25, 0.25, 0.25, 0.25)^\top$ in terms of normalized power, tr_A efficiency and ethics where $p_1 \in [0.3, 0.8]$, $p_2 = 0.2$, $p_3 = 0.15$ and $p_4 = 0.1$.

In terms of power, it does not exist a design outperforming $\boldsymbol{\rho}^*$. As regards other targets, $\tilde{\boldsymbol{\rho}}$ shows the highest values of normalized power with a maximum gain (i.e. 7.3% and 9% on $\boldsymbol{\rho}^A$ and $\boldsymbol{\rho}^R$ respectively) when p_1 is around 0.6. Note that $\boldsymbol{\rho}^*$ is unsuitable from the viewpoint of estimation precision, as it does not collect information on the intermediate treatments. On the other hand, $\mathscr{E}_A(\tilde{\boldsymbol{\rho}})$ is higher than 0.92, showing that the optimal constrained design reaches good performance even in terms of estimation efficiency. Slightly higher values are observed for $\boldsymbol{\rho}^R$, while the balanced design exhibits inferior tr_A efficiency. The targets whose values of $\mathscr{E}_E(\boldsymbol{\rho})$ are closest to 1 are $\boldsymbol{\rho}^*$ and $\tilde{\boldsymbol{\rho}}$. More specifically the optimal constrained design

outperforms all the other targets for $p \leq 0.6$ while for higher success probabilities the ethical efficiencies of $\boldsymbol{\rho}^*$ and $\tilde{\boldsymbol{\rho}}$ tend to coincide. The loss in terms of normalized total expected outcome induced by $\boldsymbol{\rho}^A$ varies from 4% up to 11% while adopting $\boldsymbol{\rho}^R$ the loss ranges from 8% to 14%.

In summary, the results obtained in this work emphasizes that a good trade-off between power, estimation efficiency and ethics is achieved by $\tilde{\boldsymbol{\rho}}$. Indeed, the optimal constrained target is a valid tool in designing randomized multi-purposes binary trials comparing several treatments, especially when drugs for life-threatening or rare diseases are involved.

References

1. Baldi Antognini, A., Giovagnoli, A.: Adaptive designs for sequential treatment allocation. Chapman and Hall/CRC (2015)
2. Baldi Antognini, A., Novelli, M., Zagoraiou, M.: Optimal designs for testing hypothesis in multiarm clinical trials. *Stat. Meth. Med. Res.*, **28**, 3242-3259 (2019)
3. Baldi Antognini, A., Frieri, R., Novelli, M., Zagoraiou, M.: Optimal designs for testing the efficacy of heterogeneous experimental groups *Submitted 2021*.
4. Frieri, R., Zagoraiou, M. : Optimal and ethical designs for hypothesis testing in multi-arm exponential trials. *Stat. Med.* In press. (2021)
5. Hu, F., Zhang, L. X.: Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Ann. Stat.*, *32*, 268-301 (2004)
6. Jung, S., George, S.: Between-arm comparisons in randomized Phase II trials. *J. Biopharm. Stat.*, *19*, 456-468 (2009)
7. Magaret, A., Angus, D., Adhikari, N. et al. : Design of a multi-arm randomized clinical trial with no control arm. *Contemp. Clin. Trials.*, *46*, 12-17 (2016)
8. Peckham, E., Brabyn, S., Cook, L. et al. : The use of unequal randomisation in clinical trials: An update. *Contemp. Clin. Trials.*, *45*, 113-122 (2015)
9. Rosenberger, W. F., Lachin, J. M.: Randomization in clinical trials: theory and practice. John Wiley and Sons (2015)
10. Sverdlov, O., Rosenberger, W. F.: On recent advances in optimal allocation designs in clinical trials. *J. Stat. Theory Prac.*, **7**, 753-773 (2013)
11. Tymofyeyev, Y., Rosenberger, W. F., Hu, F.: Implementing optimal allocation in sequential binary response experiments. *J. Am. Stat. Assoc.*, **102**, 224-234 (2007)

Analysis of three-way data: an extension of the STATIS method

Analisi dei dati a tre vie: un'estensione del metodo STATIS

Laura Bocci and Donatella Vicari

Abstract In order to explore differences and similarities in three-way data (units × variables × occasions), an extension of the STATIS method is presented. STATIS is a generalization of PCA to analyze several sets of variables collected on the same units. As a novelty, the method proposed here searches for a "compromise" of the occasions by assuming that the variables can be differently weighted to capture the similarity structure between the corresponding variables across the occasions. An application to real data is presented to illustrate the potentiality of the method.

Abstract *Il modello proposto è un'estensione del metodo STATIS per l'analisi dei dati a tre vie (unità × variabili × occasioni). STATIS è una generalizzazione dell'analisi in componenti principali il cui obiettivo è analizzare diversi insiemi di variabili osservate sulle stesse unità. Il metodo proposto definisce un "compromesso" delle occasioni assumendo che le variabili siano pesate per tener conto della struttura di similarità tra le variabili corrispondenti nelle diverse occasioni. Viene presentata un'applicazione a dati reali per illustrare le potenzialità del metodo.*

Key words: multiway analysis, compromise, dimensional reduction.

1 Introduction

A three-way three-mode data set is a data set pertaining to three different sets of

[1] Laura Bocci, Department of Social and Economic Sciences, Sapienza University of Rome; email: laura.bocci@uniroma1.it

Donatella Vicari, Department of Statistical Sciences, Sapienza University of Rome; email: donatella.vicari@uniroma1.it

entities (i.e., units, variables and occasions). In this paper, we refer to the case where the available information consists of a number of variables collected on the same set of units in different occasions. Such data are usually stored in a three-way array, say $\mathbf{X}$, of size $(I \times J \times K)$ where the generic element x_{ijk} contains the value of the j-th variable observed on the i-th unit at the k-th occasion. A three-way array $\mathbf{X}$ can be always viewed as a cube where multiple data matrices $\mathbf{X}_k$ $(k = 1, \ldots, K)$ of order $(I \times J)$ are stacked as slices along the third dimension which represents the occasions.

Generally, the goal in studying such kind of data is: a) to compare and analyse the relationships between occasions and variables; b) to represent the data in a low-dimensional space to visualize communalities and discrepancies between observations. Several methods have been developed to summarize the information of three-way data from different standpoints. Three-way component analysis techniques (see for a review, Kroonenberg, 1983; Kiers and Van Mechelen, 2001; De Roover et al., 2012) summarize the entities of each mode through few components and describe their relations. In the dimensional reduction context, STATIS - acronym which stands for the French expression "Structuration des Tableaux à Trois Indices de la Statistique" (Escoufier, 1980; see, Abdi et al., 2012 for a comprehensive review of its main developments) - is a generalization of the Principal Component Analysis (PCA) for three-way data. Its goal is to analyse the relationships between occasions and obtain a representation of the units in a space of low dimensions which is common to all occasions. Generally speaking, STATIS can be applied even when the number and/or the nature of the variables observed on the same set of I units vary from one occasion to the other; therefore, the two-way data sets $\mathbf{X}_k$ $(k = 1, \ldots, K)$ may have a different number of columns. The general idea behind STATIS is to analyse the structure of the data array performing two main steps:

1) the *inter-structure* analysis, which consists in deriving an optimal set of weights from the analysis of the similarities between occasions: weights are used to get an optimal consensus (the so-called *compromise*) of the K data matrices as representative as possible of all occasions;
2) the *intra-structure* analysis, where a generalized Principal Component Analysis (PCA) of the *compromise* is performed to obtain the best representation of the units in a common space.

In this paper, an extended STATIS is presented which, in the inter-structure step, searches for a compromise of the occasions which explicitly takes into account how similarly behave the corresponding variables across occasions. Instead of giving the same weights to all variables within each occasion, as in the ordinary STATIS, a different compromise matrix is defined here where the variables themselves are assumed to be differently weighted to capture the similarity structure. An application to real data is presented to illustrate the potentiality of the method.

2 Methodology

Our concern here is three-way three-mode data, where K frontal slices consist of $(I \times J)$ matrices $\mathbf{X}_k$ $(k = 1, \ldots, K)$ containing the values of the same J variables

measured on a set of I units at K different occasions. Each data matrix $\mathbf{X}_k$ is, in general, preprocessed so that each variable is column centered (i.e., the mean of each variable is zero).

Let $\underline{\mathbf{X}} = [\mathbf{X}_1, \dots, \mathbf{X}_k, \dots, \mathbf{X}_K]$ be the $I \times JK$ matrix formed by collecting the K matrices $\mathbf{X}_k$ next to each other and $\mathbf{Z}$ the $I^2 \times JK$ matrix formed by the (column-wise) Khatri–Rao product (Rao and Mitra, 1971) of $\mathbf{X}$ with itself, i.e., $\mathbf{Z} = \underline{\mathbf{X}} | \otimes | \underline{\mathbf{X}} = (\mathbf{x}_{1k} \otimes \mathbf{x}_{1k}, \dots, \mathbf{x}_{jk} \otimes \mathbf{x}_{jk}, \dots, \mathbf{x}_{Jk} \otimes \mathbf{x}_{Jk})$ where $\mathbf{x}_{jk}$ is the j-th column of $\mathbf{X}_k$ $(k = 1, \dots, K)$ and $\otimes$ denotes the Kronecker product.

In the first step, the STATIS method, starting from the K cross-product matrices $\mathbf{S}_k = \mathbf{X}_k \mathbf{X}_k'$ $(k = 1, \dots, K)$, derives the optimal set of weights $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_k^*, \dots, \alpha_K^*)'$ to compute the *compromise* $\mathbf{S}_+^*$ as the optimal linear combination of the matrices $\mathbf{S}_k$ with weights α_k^*,

$$\mathbf{S}_+^* = \sum_{k=1}^K \alpha_k^* \mathbf{S}_k. \tag{1}$$

The optimal weights are chosen so that the compromise provides the best representation of the K single matrices in the least-squares sense. Therefore, $\boldsymbol{\alpha}^*$ is obtained from the eigendecomposition of the matrix whose generic entry is the cosine between any two matrices $\mathbf{X}_k$ and $\mathbf{X}_{k'}$. Actually, it comes to performing a PCA and computing $\boldsymbol{\alpha}^*$ as the first eigenvector which is then rescaled to sum to one for the second step of the method.

In such a respect, data matrices $\mathbf{X}_k$ which most agree with the other matrices have the largest weights: this implies that, in building the compromise (1), the same weight α_k^* is assigned to each variable of $\mathbf{X}_k$ $(k = 1, \dots, K)$.

Differently from the ordinary STATIS we may consider to take into account how the *same* variables measured across the occasions differently contribute to the compromise and highlight how they influence the similarity structure between occasions. Therefore, in order to take into account not only the similarity structure between occasions but also between variables across occasions, a different weight is assumed to be assigned to each variable. The *compromise* matrix becomes

$$\mathbf{S}_+ = \sum_{k=1}^K \alpha_k \mathbf{X}_k \mathbf{W} \mathbf{X}_k' \tag{2}$$

where α_k is the weight assigned to all variables within occasion k $(k = 1, \dots, K)$ and $\mathbf{W}$ is a diagonal weight matrix of size J which differently weighs the variables, regardless of the occasions.

Let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k, \dots, \alpha_K)'$ be the K-column vector whose elements are the weights of the occasions and $\mathbf{w} = (w_1, \dots, w_j, \dots, w_J)'$ the J-column vector containing the diagonal elements of $\mathbf{W}$ (i.e., the weights of the variables). The two sets of weights $\boldsymbol{\alpha}$ and $\mathbf{w}$ can be estimated by solving the following least-squares problem

$$\max_{\boldsymbol{\alpha}, \mathbf{w}} g(\boldsymbol{\alpha}, \mathbf{w}) = \|\mathbf{S}_+\|^2 \quad \text{subject to} \quad \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1 \ \text{ and } \ \mathbf{w}'\mathbf{w} = 1 \tag{3}$$

Note that, as solutions of problem (3), the optimal sets of weights maximize the variance of the compromise matrix $\mathbf{S}_+$.

Problem (3) can be solved using an Alternating Least-Squares (ALS) algorithm which alternates the estimation of a set of parameters when all the other are kept fixed.

The algorithm proposed here estimates in turn:

a) the vector of the occasion weights $\boldsymbol{\alpha}$, given $\mathbf{w}$, by maximizing $\boldsymbol{\alpha}'\mathbf{M}\boldsymbol{\alpha}$ subject to $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$, where $\mathbf{M} = (\mathbf{I}_K \otimes \mathbf{w})'\mathbf{Z}'\mathbf{Z}(\mathbf{I}_K \otimes \mathbf{w})$ and $\mathbf{I}_K$ is an identity matrix of order K;
b) the vector of the variable weights $\mathbf{w}$, given $\boldsymbol{\alpha}$, by maximizing $\mathbf{w}'\mathbf{N}\mathbf{w}$ subject to $\mathbf{w}'\mathbf{w} = 1$, where $\mathbf{N} = (\boldsymbol{\alpha} \otimes \mathbf{I}_J)'\mathbf{Z}'\mathbf{Z}(\boldsymbol{\alpha} \otimes \mathbf{I}_J)$ and $\mathbf{I}_J$ is an identity matrix of order J.

The solution of the two steps a) and b) of the algorithm is achieved by taking the first eigenvector of $\mathbf{M}$ and $\mathbf{N}$, respectively. The two steps are alternated and iterated until convergence.

It is worth noting that when $\mathbf{W} = \mathbf{I}_J$, (2) reduces to (1) and the maximization of (3) gives the optimal solution of STATIS.

As in the intra-structure analysis (second step) of the ordinary STATIS, the structure of the unit space is investigated by performing a PCA of the compromise to analyse the space spanned by the first principal components. Here, this second step can be applied by performing the PCA of the compromise $\mathbf{S}_+$ defined in (2) instead of (1) as in the ordinary method (Abdi et al., 2012).

Note that, since the compromise (2) is less constrained than (1), $\mathbf{S}_+$ is more representative of $\mathbf{X}$ and better accounts for the variability of the three-way array. Consequently, the second step better allows to represent the relationships between units and variables in the common space.

3 Application

The data used in this application have been taken from the OECD (Organization for Economic Co-operation and Development) Employment Database, which collects data annually through the OECD Labour Force Questionnaire, and are based on information drawn from national labour force surveys. The three-way data array consists of 6 variables observed on 20 countries (units) from 2002 to 2017 (16 occasions). The observed variables report the percentages of employed men and employed women across three hours bands for usual weekly working hours: 1) 1-34 hours per week (variables V1 for men and V4 for women); 2) 30-34 hours per week (variables V2 for men and V5 for women); 3) 40 hours or more per week (variables V3 for men and V6 for women). The countries are the 20 founding members of the OECD: Austria (AT), Belgium (BE), Canada (CA), Denmark (DK), France (FR), Germany (DE), Greece (EL), Iceland (IS), Ireland (IE), Italy (IT), Luxembourg (LU), Netherlands (NL), Norway (NO), Portugal (PT), Spain (ES), Sweden (SE), Switzerland (CH), Turkey (TR), United Kingdom (UK) and United States (US).

The proposed method was applied to the $(20 \times 6 \times 16)$ data matrix $\mathbf{X}$ to explore the differences and similarities of the countries taking into account the different role of the variables along time and then to define the position of the countries in a common space across years. At first, each of the $K = 16$ data matrices $\mathbf{X}_k$ has been preprocessed by column centering each variable and then by normalizing the matrix

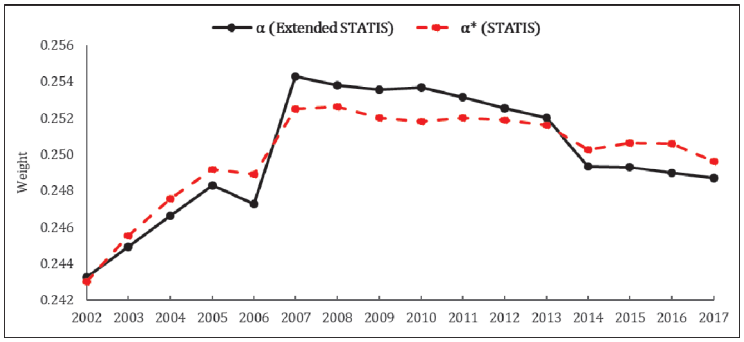
such that the corresponding cross-product matrix $\mathbf{S}_k$ has Euclidean norm equal to 1.

The inter-structure step is then performed applying both the ordinary STATIS method and its extension proposed here and deriving the optimal weights $\boldsymbol{\alpha}^*$and $\boldsymbol{\alpha}$, respectively (Figure 1). The common optimal weights $\mathbf{w}$ assigned to the variables regardless of the occasions are reported in Table 1.

Both the compromise $\mathbf{S}_+^*$, computed with weights $\boldsymbol{\alpha}^*$, and the compromise $\mathbf{S}_+$, computed with weights $\boldsymbol{\alpha}$, explain the 96.83% of the "between variance". Actually, the solution from STATIS only accounts for the 53.85% of the total variability of the variables across years ($\|\mathbf{Z}\|^2$), while the compromise $\mathbf{S}_+$ accounts for the 80.81% which indicates that it better summarizes the information on the variables.

Both weights $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}$ show the same pattern (Figure 1): they reveal that even though years are not very different in terms of weekly working hours, the first five years (2002 - 2006) result more dissimilar from the others than the central years (2007 to 2013). Nonetheless, the two sets of weights differ because $\boldsymbol{\alpha}$ amplifies the differences and the similarities between years: actually, lower weights are assigned to less similar years while more similar years have higher weights.

Figure 1: Occasion weights from extended and ordinary STATIS



Looking at the weights of the variables (Table 1), it is possible to evaluate the contribution of the variables to the similarity structure across the 16 years.

Table 1: Variable weights from extended STATIS

Variable		
	Weekly hours band	**w**
% of Employed men	1 to 34 hours (V1)	0.0071
	35 to 39 hours (V2)	0.5067
	40 hours or more (V3)	0.6263
% of Employed women	1 to 34 hours (V4)	0.0637
	35 to 39 hours (V5)	0.2525
	40 hours or more (V6)	0.5321

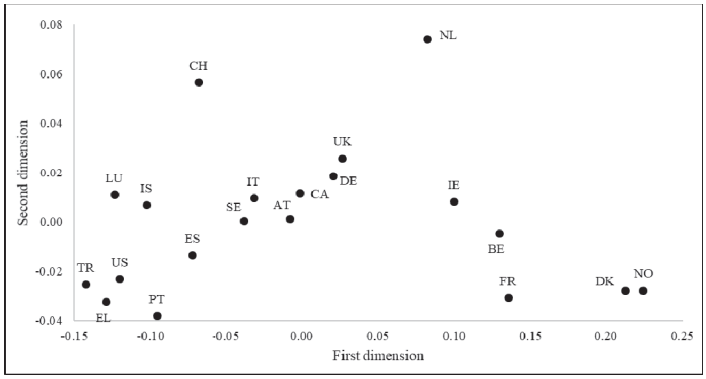
Variables V3 and V6 have the largest weights because in most OECD countries, the most common hours-band for both male and female full-time workers is 40 hours or more per week. The weights of the corresponding hours bands V1 and V4 for male and female workers reflect the differential incidence of part-time work which is

known to be higher for women than men, but in different ways in OECD countries.

Furthermore, by combining weights $\boldsymbol{\alpha}$ and $\mathbf{w}$ it is possible to highlight the evolution of the variables over years: in many countries, the share of the workforce that normally works at least 40 hours a week is declining even if such weekly hours band remains the most frequent.

The analysis of the compromise $\mathbf{S}_+$ computed with the optimal weights $\boldsymbol{\alpha}$ and $\mathbf{w}$ is then performed in the intra-structure step of the method. Figure 2 displays the plot of the OECD countries in the plane spanned by the first two principal components of $\mathbf{S}_+$ which accounts for the 97.1% of its variability. The first component can be interpreted as the opposition of countries, such as Denmark and Norway, where a majority of both male and female workers usually work 35-39 hours per week, to countries, such as Turkey and Greece, where a majority of full-time workers usually work at least 40 hours per week. The second component contrasts the Netherlands, where the large majority of female workers usually work less than 34 hours per week, to Portugal where the lowest proportion of women are part-time workers.

Figure 2: Analysis of the compromise: plot of the OECD countries in the plane of the first two principal components



A comparison with existing alternatives in the literature will be presented and discussed.

References

1. Abdi, H., Williams, L. J., Valentin, D., Bennani-Dosse, M.: STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. WIREs Computational Statistics **4** (2), 124-167 (2012).
2. Escoufier, Y.: L'analyse conjointe de plusieurs matrices de données. In M. Jolivet (Ed.), Biométrie et Temps. Paris: Société Française de Biométrie, 59-76 (1980).
3. De Roover, K., Ceulemans, E., Timmerman, M.E.: How to perform multiblock component analysis in practice. Behavior Research Methods **44**, 41-56 (2012).
4. Kiers, H.A.L., Van Mechelen, I.: Three-Way Component Analysis: Principles and Illustrative Application. Psychological Methods **6** (1), 84-110 (2001).
5. Kroonenberg, P.M.: Three-mode principal component analysis. Theory and applications. Leiden: DSWO Press (1983).
6. Rao, C. R., Mitra, S.: Generalized inverse of matrices and its applications. New York: Wiley (1971).

KL-optimum designs to discriminate models with different variance function

Disegni KL-ottimi per discriminare tra modelli con diversa funzione di varianza

Alessandro Lanteri, Samantha Leorato and Chiara Tommasi

Abstract In many applications, researchers are interested in models that can be defined by interpretable statistics, such as mean and variance. Kullback-Leibler criterion is one of the best known optimum criteria to select designs to discriminate between two competing models. We provide a simple closed form formula to obtain the optimal KL-design to discriminate between regression models with different variance structures and common response mean and we conduct numerical experiments to compare its performance with other benchmark designs in terms of statistical power.

Abstract *In molte applicazioni, i ricercatori sono interessati a modelli che possono essere definiti tramite statistiche interpretabili, come media e varianza. Il criterio di Kullback-Leibler e uno dei migliori criteri di ottimalità per selezionere disegni atti a discriminare tra due modelli alternativi. In questo lavoro, forniamo l'espressione in forma chiusa del disegno KL-ottimo per discriminare tra modelli di regressione con diversa struttura di varianza e stessa funzione media. Inoltre, attraverso uno studio di simulazione, abbiamo confrontato il disegno KL-ottimo con altri disegni di riferimento in termini di potenza statistica.*

Key words: Heretoschedasticity, KL-divergence, Optimal Design

1 Introduction

In many modern sciences, despite the development of new technologies, to gather information and empirical evidence about specific hypotheses can be very expensive, not only from a strictly economical standpoint. The time cost of the data-

Alessandro Lanteri[1] - e-mail: alessandro.lanteri@unimi.it
Samantha Leorato[2] - e-mail: samantha.leorato@unimi.it
Chiara Tommasi[3] - e-mail: chiara.tommasi@unimi.it
[123]Università degli studi di Milano, DEMM.

collection process could render the information gathered obsolete while ethical costs could overwhelm the scientific benefits of an experiment. For these and many other reasons, optimal or efficient experimental design is important in the scientific research. The T-criterion [1] is one of the most widely used methods to obtain optimal design when the goal is to discriminate between two rival regression models with homoschedastic Gaussian errors. T-criterion has been generalized with weaker assumptions in subsequent works [2, 5, 8]. A general criterion for discriminating between models, based on the Kullback-Leibler (KL) divergence, has been introduced in [3] and extended in [6]. The problem of discriminating between homoschedastic and heteroschedastic regression models with the same regression function has not been less considered in the literature. To handle this problem we consider the KL-criterion that generalizes the T-criterion (as it discriminates between any two rival statistical models). A KL-optimum design, as well as T-designs, depends on the nominal values of the parameters of the true model, that in the case of nested models is the larger model. When the values of the parameters are unknown, but it is available a prior information about such parameters, it has been proposed a Bayesian approach, for the T-criterion [4] and for the KL-criterion [7]. Therefore, beside the KL-optimal design, we compute the Bayesian KL-optimal design to handle the problem of dependence on unknown parameters. In particular, in Section 2 we state a theorem which provides a closed form for a KL-optimal design for discriminating between homoschedastic and non-homoschedastic Gaussian models. In Section 3 we conduct a numerical experiment to analyze how good are different designs in terms of statistical power when we use the log-likelihood ratio test to discriminate between the two nested models.

2 Discriminating between different variance functions

One way to discriminate between two models is with the use of the KL-optimality criterion [3] which is based on the well known Kullback Leibler divergence. In real data applications, practitioners are often interested in models where the distribution of the response $y \in \mathscr{Y}$ can be defined by some interpretable statistics somehow linked with the covariates $x \in \chi \subseteq \mathbb{R}^p$. Let us recall that a continuous design with K design points is denoted as

$$\xi = \left\{ \begin{array}{c} x_1, \ \ldots, \ x_K \\ \omega_1, \ \ldots, \ \omega_K \end{array} \right\}; \qquad 0 \leq \omega_k \leq 1; \qquad \sum_{k=1}^{K} \omega_k = 1$$

where the domain χ of any experimental point x is assumed to be compact. A proportion of ω_k of responses are observed at the experimental point x_k, $k = 1, \ldots, K$. Let $f_1\left[y, \mu_1(x, \beta_1), \sigma_1^2(x, \theta_1)\right]$ and $f_2\left[y, \mu_2(x, \beta_2), \sigma_2^2(x, \theta_2)\right]$ be two competing statistical models, where μ_j and σ_j^2, $j = 1, 2$, are the mean and the variance functions, respectively. Let the two models be nested and let the first be the "true" and completely known largest model. The KL-optimality criterion function is

$$I_{21}(\xi) = I_{21}(\xi;\theta_1,\beta_1) = \min_{(\theta_2,\beta_2)\in\Omega_2} \int_{\chi}\int_{\mathscr{Y}} f_1\left[y,\mu_1(x,\beta_1),\sigma_1^2(x,\theta_1)\right] \log\left\{\frac{f_1\left[y,\mu_1(x,\beta_1),\sigma_1^2(x,\theta_1)\right]}{f_2\left[y,\mu_2(x,\beta_2),\sigma_2^2(x,\theta_2)\right]}\right\} dy\,\xi(dx)$$

and thus, a design $\xi^{KL}_{\theta_1,\beta_1}$ which maximize $I_{21}(\xi)$ is called KL-optimal. The subscripts θ_1,β_1 underline that in general a KL-optimum design depends on the assumed value for the parameters of the true model.

Let us assume that we are interested in discriminating between two rival Gaussian models with the same regression function and different variance structures so that $y_i = \mu(x_i;\beta_j) + \varepsilon_i$ with $\varepsilon_i \sim N(0,\sigma_j^2(x_i))$ for $i = 1,\ldots,n$ and $j = 1,2$. Consider the specific case where $\sigma_1^2(x_i) = \varsigma_1 h(x_i;\theta_1)$ and $\sigma_2^2(x_i) = \varsigma_2$, where $h : \mathbb{R} \to \mathbb{R}_+$ is a continuous positive function in χ. Let also $\tilde{\theta}$ be a specific value for θ_1 such that $h(x;\tilde{\theta}) = 1$, this implies that model 2 is nested in model 1. Then the following theorem, which allows to compute analytically the KL-optimal design, can be proved.

Theorem 1. *Let* $\underline{h} = \inf_x h(x) > 0$ *and* $\bar{h} = \sup_x h(x) < \infty$. *Let* $\chi_l = \{x : h(x) = \underline{h}\}$ *and* $\chi_u = \left\{x : h(x) = \bar{h}\right\}$. *Then*

$$\xi^* = \left\{ \begin{matrix} x_l, & x_u \\ \omega, & 1-\omega \end{matrix} \right\}, \quad \textit{with} \quad \omega = \left(\frac{\bar{h}}{\bar{h}-\underline{h}} - \frac{1}{\log\bar{h}-\log\underline{h}}\right)$$

is a KL-optimal design, where $x_l \in \chi_l$ *and* $x_u \in \chi_u$.

Theorem 1 is quite interesting because it is uncommon to find KL-optimum design in a closed form. Note that if χ_l or χ_u contain more than one point, then any design with more support points in χ_l or χ_u is KL-optimal, provided that the sum of the weights corresponding to the points in χ_l or χ_u is ω and $1-\omega$, respectively. From this theorem we can deduce that since the design points are the ones which provide the most extreme values of $h(x)$, then, when the variance function is strictly monotone in the compact $\chi \subseteq \mathbb{R}$, the design points are necessarily the edge points of χ. This implies that, in this setting, KL-optimal designs will have the same design points, independently on the values of the true parameters, and only the design weights will vary. On the other hand, when $h(x)$ is non-monotone, KL-optimal designs might differ greatly also with regards to the design points for different values of the parameters. To discriminate between the homoschedastic and the heteroschedastic model, and thus when the hypotheses are

$$\begin{cases} H_0 : \sigma^2 = \varsigma_2 \\ H_1 : \sigma^2 = \varsigma_1 h(x;\theta_1) \end{cases} \quad \text{or equivalently} \quad \begin{cases} H_0 : \theta_1 = \tilde{\theta} \\ H_1 : \theta_1 \neq \tilde{\theta} \end{cases},$$

we use the log-likelihood ratio test.

So far we assumed that all the parameters of the true model are known, but in real applications they might be unknown. It is less stringent to assume that an approximate range of possible values of the parameters is available and it is possible to build a prior probability distribution π on the unknown parameters. Taking the expected value of KL-optimality criterion over the prior distribution of the parameters we can define the partially Bayesian KL-optimality (PBKL-optimality)

[7] as $I_{21}^{PB}(\xi,\pi) = \mathbb{E}_\pi[I_{21}(\xi)]$ and, consequently, the design ξ_π^{PB} which minimizes $I_{21}^{PB}(\xi,\pi)$ is called PBKL-optimum design under the distribution π.

3 Numerical Experiment

In this section we conduce some numerical experiments in order to compare the performance of the proposed optimal design criterion with other designs. Consider two normal models, $f_1\left[\mu_1(x),\sigma_1^2(x)\right]$ and $f_2\left[\mu_2(x),\sigma_2^2(x)\right]$, with $\sigma_1^2(x) = \varsigma_1 h(x;\theta_1)$, $\sigma_2^2 = \varsigma_2$ and the same mean structure $\mu_1(x) = \alpha_1 + \beta_1 x$ and $\mu_2(x) = \alpha_2 + \beta_2 x$. In each scenario we vary the sample size n and the variance function parameter θ_1. In the study we consider two different variance functions against homoschedasticity $h_1(x;\theta_1) = \frac{1}{1+\theta_1 x}$ and $h_2(x;\theta_1) = 1+\sin(9\theta_1 x)+\theta_1 x$. Note that the first is monotone while the second is not. From Figure 3 we can appreciate how, for different values of θ_1, h_1 reaches its minimum and maximum value always at the extremes of χ while h_2 reaches its extreme values in scattered locations. Note that, with both choices

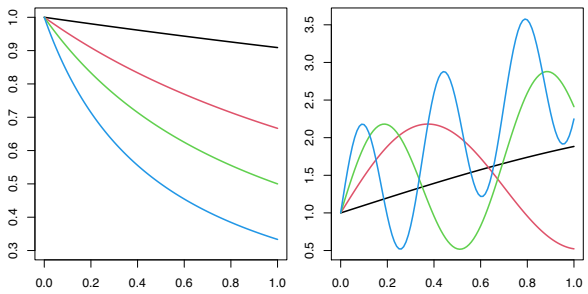


Fig. 1: Variance functions in $\chi = [0,1]$. $h_1(x;\theta_1)$, left plot, and $h_2(x;\theta_1)$, right plot, for $\theta_1 = 0.1, 0.5, 1, 2$, represented by black, red, green and blue curves, respectively.

of variance functions, the two rival models (homoschedastic and heteroschedastic) become more and more similar as θ_1 goes to $\tilde{\theta} = 0$. We let $\chi = [0,1]$ and set the nominal parameter $\alpha_1 = \beta_1 = \varsigma_1 = 1$. Table 1 displays the estimated power of the likelihood ratio test for different designs, in different experimental scenarios and using the variance function $h_1(x;\theta_1)$. Each table entry is the average over 10000 repetition of the experiment in the same setting. The notation $\xi_{\theta_1}^{KL}$ represents the KL-optimal design for a specific value of θ_1, such designs can be easily obtained using Theorem 1. For $\theta_1 = 0.5, 1, 2$ we obtain:

$$\xi_{0.5}^{KL} = \begin{pmatrix} 0 & 1 \\ 0.44 & 0.56 \end{pmatrix}; \xi_1^{KL} = \begin{pmatrix} 0 & 1 \\ 0.47 & 0.53 \end{pmatrix}; \xi_2^{KL} = \begin{pmatrix} 0 & 1 \\ 0.41 & 0.59 \end{pmatrix}.$$

We compare the performance of KL-optimal designs with two uniform designs, which consist of a fixed number of equidistant and equally weighted design points that cover all the domain χ. We denote $U3$ and $U4$ the uniform designs with three and four points, respectively.

Table 1: Estimated power obtained in different designs for different scenarios using the variance function $h_1(x;\theta_1)$

θ_1	n	$\xi_{0.5}^{KL}$	ξ_1^{KL}	ξ_2^{KL}	$U3$	$U4$
0.5	30	0.1309	0.1175	0.1203	0.1013	0.0511
	50	0.1721	0.1757	0.1682	0.1117	0.1173
	100	0.2962	0.2895	0.2863	0.2175	0.1824
1	30	0.2676	0.2579	0.2581	0.1889	0.0735
	50	0.4062	0.4030	0.3989	0.2014	0.2573
	100	0.6814	0.6756	0.6563	0.5177	0.4342
2	30	0.5460	0.5266	0.5232	0.4032	0.1362
	50	0.7560	0.7613	0.7481	0.4512	0.5431
	100	0.9698	0.9631	0.9598	0.8838	0.8195

From Table 1 we can appreciate, as expected, how the power increases with the sample size n and decreases as θ_1 gets smaller, that is because for smaller values of θ_1 the two models become more and more similar an thus it is more difficult to discriminate between the two. From this numerical experiment, we can see how the KL-optimal designs outperform the uniform designs in all settings, even when they are assuming a wrong θ_1. We also notice that the design obtained from $U3$ provides better results than $U4$, this is because $U3$ is incidentally more similar to the KL-optimal designs than $U4$.

We perform a similar experiment using the non-monotone variance function $h_2(x;\theta_1)$. The KL-optimal designs, obtained with the application of Theorem 1 for $\theta_1 = 0.5, 1, 2$, are:

$$\xi_{0.5}^{KL} = \begin{pmatrix} 0.37 & 1.00 \\ 0.38 & 0.62 \end{pmatrix}; \xi_1^{KL} = \begin{pmatrix} 0.51 & 0.88 \\ 0.64 & 0.36 \end{pmatrix}; \xi_2^{KL} = \begin{pmatrix} 0.26 & 0.79 \\ 0.65 & 0.35 \end{pmatrix}.$$

Differently from the case with monotone variance, here the designs are very different from each other.

In Table 2 we show, for each value of $\theta_1 = 0.5, 1, 2$, the KL-efficiency of a design ξ, $\text{Eff}_{KL}(\xi) = I_{21}(\xi)/I_{21}(\xi_{\theta_1}^{KL})$, which is a measure of the goodness of ξ with respect to $\xi_{\theta_1}^{KL}$ for discrimination purposes. From Table 2 we can appreciate how the difference between KL-optimum designs determines a poor KL-efficiency when a KL-optimum design with a wrong value of θ_1 is used. Uniform designs are more robust but far from been efficient.

Table 2: KL-Efficiency of different designs with respect to $\xi_{\theta_1}^{KL}$ for different values of θ_1 and variance function $h_2(x;\theta_1)$

	$\text{Eff}_{KL}(\xi_{0.5}^{KL})$	$\text{Eff}_{KL}(\xi_1^{KL})$	$\text{Eff}_{KL}(\xi_2^{KL})$	$\text{Eff}_{KL}(U3)$	$\text{Eff}_{KL}(U4)$
$\xi_{0.5}^{KL}$	1.0000	0.4550	0.2201	0.6009	0.5118
ξ_1^{KL}	0.1689	1.0000	0.0184	0.5482	0.1442
ξ_2^{KL}	0.0003	0.0043	1.0000	0.1604	0.0994

In order to obtain more efficient and robust designs we rely on the PBKL-optimality criterion. We use two different prior distribution describing two different type of prior knowledge. The first prior distribution, π_1, assigns uniform weights to the values of θ_1 that we might consider to be true, in our experiment $\theta_1 = 0.5, 1, 2$. To represent the case where the candidate values of θ_1 are not known, but it is

available a range of possible values, say $\theta_1 \in [0.5, 2]$, we use a second prior distribution, π_2, which is a discrete uniform distribution with several equidistant points with maximum distance between each other. The PBKL-optimal designs, that we have obtained computationally using a first order algorithm, are:

$$\xi_{\pi_1}^{PB} = \begin{pmatrix} 0.000 & 0.255 & 0.486 & 0.828 & 1.000 \\ 0.0002 & 0.3848 & 0.3118 & 0.0586 & 0.2446 \end{pmatrix}; \xi_{\pi_2}^{PB} = \begin{pmatrix} 0.000 & 0.293 & 0.571 & 0.842 & 1.000 \\ 0.0001 & 0.4839 & 0.2182 & 0.2125 & 0.0853 \end{pmatrix}.$$

From Table 3 we can appreciate that the KL-optimal designs provide the best results in terms of statistical power when they are used for the correct value of θ_1, although they can be very inefficient when they are improperly adopted. As we commented before, uniform designs seem to be more robust than KL-optimal designs but generally provide a low power. On the other hand, PBKL-designs seem to combine the qualities of the other two kinds of designs, providing a high power in most settings.

Table 3: Estimated power obtained in different designs for different scenarios using the variance function $h_2(x;\theta_1)$

θ_1	n	$\xi_{0.5}^{KL}$	ξ_1^{KL}	ξ_2^{KL}	$U3$	$U4$	$\xi_{\pi_1}^{PBKL}$	$\xi_{\pi_2}^{PBKL}$
0.5	30	0.7349	0.5014	0.2753	0.6013	0.4612	0.5984	0.4237
	50	0.9117	0.7029	0.4066	0.7340	0.7715	0.8285	0.7122
	100	0.9975	0.9517	0.6850	0.9662	0.9451	0.9732	0.9376
1	30	0.2402	0.8461	0.0373	0.6587	0.2896	0.7034	0.5478
	50	0.4268	0.9721	0.0615	0.8603	0.4382	0.9052	0.7424
	100	0.7164	0.9999	0.1173	0.9893	0.6690	0.9932	0.9242
2	30	0.0275	0.0232	0.9038	0.4084	0.3111	0.8524	0.6356
	50	0.0324	0.0187	0.9878	0.5861	0.4012	0.9561	0.8280
	100	0.0345	0.0246	1.0000	0.8010	0.5993	0.9969	0.9620

References

1. AC Atkinson and VV Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975.
2. AC Atkinson and VV Fedorov. Optimal design: Experiments for discriminating between several models. *Biometrika*, 62(2):289–303, 1975.
3. J López-Fidalgo, C Tommasi, and PC Trandafir. An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):231–242, 2007.
4. AC Ponce de Leon and AC Atkinson. Optimum experimental design for discriminating between two rival models in the presence of prior information. *Biometrika*, 78(3):601–608, 1991.
5. AC Ponce de Leon and AC Atkinson. The design of experiments to discriminate between two rival generalized linear models. In *Advances in GLIM and Statistical Modelling*, pages 159–164. Springer, 1992.
6. C Tommasi. Optimal designs for discriminating among several non-normal models. In *mODa 8-Advances in Model-Oriented Design and Analysis*, pages 213–220. Springer, 2007.
7. C Tommasi and J López-Fidalgo. Bayesian optimum designs for discriminating between models with any distribution. *Computational Statistics & Data Analysis*, 54(1):143–150, 2010.
8. D Ucinski and B Bogacka. T-optimum designs for multiresponse dynamic heteroscedastic models. In *mODa 7—Advances in Model-Oriented Design and Analysis*, pages 191–199. Springer, 2004.

Riemannian optimization on the space of covariance matrices

Ottimizzazione riemanniana nello spazio delle matrici di covarianza

Jacopo Schiavon, Mauro Bernardi and Antonio Canale

Abstract In many modern statistical applications the data complexity may require techniques that exploit the geometrical properties of the objects of interest. For example, if the parameter of interest is a covariance matrix, the parameter space is non-Euclidean. In this work we focus on this notable example and study the Riemannian manifold of symmetric and positive definite matrices. Specifically an optimization procedures which takes into account such geometrical properties is described and tested via simulations.

Abstract *In varie applicazioni statistiche moderne, la complessità dei dati può richiedere tecniche che sfruttino le proprietà geometriche degli oggetti d'interesse. Ad esempio, se il parametro di interesse è una matrice di covarianza, lo spazio parametrico è non-Euclideo. In questo lavoro ci concentriamo su questo esempio notevole e studiamo la varietà Riemanniana delle matrici simmetriche e definite positive. Nello specifico, una procedura di ottimizzazione che tenga conto di queste proprietà geometriche verrà descritta e testata attraverso delle simulazioni.*

Key words: Riemannian Optimization, Symmetric Positive Definite matrices, multivariate Student-*t* distribution, Skew-Normal distribution.

1 Introduction

In different modern statistical applications, mostly enabled by the recent availability of new data sources, the main difficulties are not only related the large scale of the problem but also from the complex constrains that the data or the model's param-

Jacopo Schiavon, Mauro Bernardi and Antonio Canale
Department of Statistical Science, University of Padova, Italy,
e-mail: jacopo.schiavon.1@phd.unipd.it
e-mail: mauro.bernardi@unipd.it
e-mail: antonio.canale@unipd.it

eters require. A primer example is the necessity to study the space of covariance matrices, namely the manifold of Symmetric and Positive Definite matrices (SPD) both as a sample space [2] or as a parameter space.

In this work we will focus on the latter, and more specifically we will focus on how to solve the general problem of optimization

$$\min_{\Sigma \in \mathscr{S}^+} f(\Sigma) \tag{1}$$

where $\mathscr{S}^+$ is the SPD manifold. To solve this problem, an Euclidean approach is to transform the matrix Σ with a Cholesky decomposition such that $\Sigma = LL^\top$ and, after a vectorization procedure to create a $p(p+1)$ dimensional vector by stacking the columns below the diagonal, to perform a standard Euclidean optimization. Of course, depending on the specific problem to be solved other procedures might be used, such as the Expectation-Maximization algorithm or (in some cases) a fixed point procedure.

In this work instead we will describe and implement a general alternative approach that leverages the properties of the Riemannian manifold $\mathscr{S}^+$. This procedure is comparable (and an alternative to) the Cholesky-Euclidean approach broadly described above, which is generally applicable whenever it is required to optimize a function of a covariance matrix. In the next section we describe the required differential geometry and some implementation detail of the algorithms, while in § 3 a simulation study on the performance of this algorithm compared to the Cholesky-Euclidean one is presented.

2 Riemannian optimization

The basic strategy to solve problems of the general form of equation (1) when the space is Euclidean is simply to perform an iteration scheme with the following two general steps:

1. find a descent direction (usually related to the gradient of the objective function $f()$);
2. move in that direction with a step length that satisfies some sufficient decrease conditions.

Hereafter, when discussing Riemannian optimization, we will follow the same strategy, but applying the appropriate corrections that allow us to take into account the actual manifold properties.

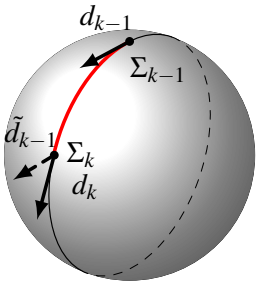


Fig. 1 The parallel transport map allows to compare vectors that originally belong to different vector spaces, namely the tangent space at different points $T_{\Sigma_k}\mathscr{S}^+$ and $T_{\Sigma_{k+1}}\mathscr{S}^+$.

2.1 Differential geometry of the covariance matrices manifold

The main results for the differential geometry of the SPD manifold are well established and many of the properties here discussed can be found in [3].

The first step in order to build a Riemannian manifold is to endow the space with a metric, which can be equivalently expressed as the inner product between elements of the tangent space in a point. In the case of the SPD manifold, the most widely adopted metric is the *affine-invariant* one, which is obtained from the Frobenius (Euclidean) one and that can be expressed as

$$g_\Sigma(V,W) = \mathrm{Tr}\left(\Sigma^{-1}V\Sigma^{-1}W\right). \tag{2}$$

From the metric, all the relevant quantities (that we collected in table 1 for ease of conciseness) can be derived, such as the expression of the retraction map (which is the map that advances a point on the manifold along the geodesic tangent to an element of the tangent space) or the construction of the Riemannian gradient starting from the Euclidean one. Finally, for the most advanced quasi-Newton optimization

Table 1 Geometrical quantities for the SPD manifold.

Description	Expression
Inner product between two tangent vectors	$g_\Sigma(V,W) = \mathrm{Tr}\left(\Sigma^{-1}V\Sigma^{-1}W\right)$
Riemannian gradient	$\Sigma\left(\nabla_\Sigma f + \nabla_\Sigma f^\top\right)\Sigma$
Retraction of Σ with direction W	$R_{\Sigma W}(t) = \Sigma^{1/2}\exp\left[t\Sigma^{-1/2}W\Sigma^{-1/2}\right]\Sigma^{1/2}$
Parallel transport of $U \in T_\Sigma\mathscr{S}$ with direction W	$\mathscr{T}_{\Sigma,W}(U) = R_{\Sigma W}(\frac{d}{2})\Sigma^{-1}U\Sigma^{-1}R_{\Sigma W}(\frac{t}{2})$

techniques (which require an approximation of the Hessian matrix built by comparing the gradient at various point on the manifold surface) the parallel transport of a vector along the geodesic is required, as exemplified in Figure 1. It is worth noting that in the actual implementation of our algorithm we used a second order approximation of both the retraction expression and of the vector transport. Indeed, in both maps, the exponential of a matrix (defined as $\exp(X) = U\exp(\Lambda)U^\top$ that leverages

the spectral decomposition (Λ, U) of X) needs to be computed, which is extremely expensive. Moreover, the full spectral decomposition is computationally expensive (it requires $O(p^3)$ operations) and numerically unstable if eigenvalues large in magnitude appears during the intermediate steps of the optimization algorithm. The second order approximation is used because it is the smallest order approximation still preserves positive definiteness of the updated matrix.

2.2 Riemannian optimization algorithms

We have implemented two quasi-Newton optimization algorithms, namely a Riemannian Conjugate-Gradient (R-CG) and a Riemannian Limited memory BFGS, (R-L-BFGS). Both methods are presented for the Euclidean setting in [4], and both implementations feature a Wolfe line search condition to obtain the optimal step length, as described, for the Euclidean case, in [4]. Starting from an initial value Σ_0, for $k = 0, 1, \ldots$ both algorithms follow the same steps:

1. obtain a descent direction W_k using information on the gradient at point Σ_k and from the previous descent direction (after parallel transporting it);
2. define $\phi(\alpha) = R_{\Sigma_k W_k}(\alpha)$ and obtain the optimal α_k that satisfy the Wolf conditions to obtain $\Sigma_{k+1} = R_{\Sigma_k W_k}(\alpha_k)$;
3. store the relevant quantities to compute the descent direction at the next iteration.

Specifically, for the Conjugate-gradient procedure we define the descent direction at the k-th step as

$$d_k = G_k - \beta_k d_{k-1} \qquad \text{with} \quad \beta_k = \frac{\langle G_k, G_k - G_{k-1} \rangle_{\Sigma_k}}{\langle G_k, G_k \rangle_{\Sigma_k}}, \tag{3}$$

while the L-BFGS require a double loop as prescribed in [4, § 7 algorithm 7.4 and 7.5]. As can be see, both techniques require to compare vector from the tangent space in different points, thus requiring the implementation of the vector transport.

3 Simulation study

We performed an extensive simulation study to check the performance of our method. All the simulations were done by generating a sample $\{y_i\}_{i=1}^n$ of a p-dimensional random variable Y distributed according to various distributions and then estimating the maximum likelihood estimator (MLE) of the variance-covariance parameter by leveraging our optimization techniques. We considered two examples. First, a simple p-variate Gaussian distribution, for which the closed-form expression for the MLE exists. As a second case-study, we estimated the parameters (scale and asymmetry) of a Skew-Normal (SN) distribution, see [1]. To tackle the optimization of the likelihood function for the SN case, we split the problem into two-steps:

Table 2 Simulation results for the three algorithms (*Riemannian Conjugate Gradient*, *Riemannian L-BFGS* and *Cholesky-Euclidean*). For each problem $R = 50$ replications of sample size $n = 1000$ were generated from the postulated distributions. The algorithms were required to reach a tolerance 10^{-5} on the gradient norm before stopping. In parenthesis, the 25% and 75% quartiles of the number of iterations and the time needed is reported, with the indication of the median. Note that we fixed the maximum number of iterations to 1000 to limit the computational resources required by the Cholesky-Euclidean method, thus some element in the table are censored from the right and are shown as -.

			Gaussian			
	R-CG		R-L-BFGS		C-E	
p	Iterations	Time [s]	Iterations	Time [s]	Iterations	Time [s]
2	12 (10,16)	0.99 (0.7,1.5)	11 (8,13)	0.79 (0.7,0.9)	29.5 (18,142)	0.40 (0.3,0.5)
5	18 (15,21)	1.55 (1.2,2.2)	16 (15,19)	1.50 (1.2,1.7)	88 (53,173)	0.68 (0.4,0.9)
10	25 (23,28)	2.68 (2.4,3.3)	24.5 (21,27)	2.66 (2.3,3.2)	183 (122,318)	1.02 (0.9,1.3)
25	30.5 (27,33)	10.24 (9.1,11.9)	27 (26,32)	10.03 (9.5,11.8)	372.5 (239,661)	2.25 (1.6,3.3)
50	32.5 (29,38)	12.29 (10.7,16.1)	28 (27,31)	11.87 (10.9,12.8)	811 (509,–)	6.57 (4.4,7.1)
75	34.5 (33,40)	13.97 (12.8,16.8)	32 (29,35)	13.65 (12.9,14.4)	- (766,–)	16.24 (12.9,17.1)
100	32.5 (30,37)	13.5 (11.6,15.3)	29.5 (27,34)	13.53 (11.9,15.5)	- (–,–)	22.00 (21.5,22.5)
			Skew-Normal			
2	4 (3,5)	1.1 (0.9,1.3)	4 (3,4)	1.1 (0.9,1.2)	21 (16,34)	0.59 (0.5,0.7)
3	5 (4,6)	1.4 (1.1,1.6)	4 (4,5)	1.3 (1.1,1.5)	31 (23,68)	0.77 (0.7,0.9)
5	6 (5,7)	1.8 (1.5,2.0)	5 (5,6)	1.5 (1.3,1.8)	72 (42,112)	0.91 (0.7,1.1)
10	11 (9,15)	3.6 (2.8,4.9)	9 (8,12)	3.1 (2.6,3.8)	98 (75,164)	1.7 (1.4,2.9)
25	31 (23,42)	12.9 (8.8,16.7)	25 (17,32)	9.6 (6.6,12.5)	178 (107,290)	4.55 (3.4,6.3)
50	44 (31,99)	22.6 (15.9,46.3)	65 (37,77)	30.3 (19.3,40.1)	190 (125,350)	4.82 (3.7,7.2)

for every iteration we first step updates the variance-covariance geodesic, while the second consider the conditional optimization of the objective function with respect to the skewness parameter. Without loss of generality we considered the location parameter as fixed and known.

For all the settings and for various dimensions of the problem p, we generated $R = 50$ *true* values for the parameters, checking the performance of the two algorithms and comparing them to the standard Cholesky-Euclidean optimization performed with the `scipy.optimize` package. The results collected in Table 2 shows both the number of iteration needed to reach convergence and the time required, to better highlight the scaling properties of the competing algorithms. As it can be seen, the number of iterations required for the convergence scale very badly for the Cholesky-Euclidean optimization. This is due to the fact that nonlinear optimization in high dimensional spaces rapidly become very slow, a problem that our methods avoid due to their abilty to account for the geometrical properties of the manifold. It is worth noting that the Cholesky-Euclidean procedure employs pre-compiled and heavily optimized `Fortran` routines available from the `scipy` package, that we are not able to fully exploit with our current implementation. This may explain the large overhead shown for small problem sizes in Table 2.

4 Discussion

In this work we have implemented two algorithms for the numerical optimization of a function over the space of SPD (Symmetric and Positive Definite) matrices, exploiting the geometrical properties of the manifold that correspond to the domain of the function. Since this procedure can be considered a *plug-and-play* technique that can replace any optimization of this kind, from the estimation of the covariance parameter of a statistical model to the computation of the Fréchet mean of SPD matrices, we are working toward releasing a lightweight `Python` package that will allow users to simply replace their optimization routine with the Riemannian one.

References

1. Azzalini, A., Capitanio, A.: The Skew-Normal and Related Families. Cambridge University Press (2014)
2. Barachant, A., Bonnet, S., Congedo, M., Jutten, C.: Multiclass brain-computer interface classification by Riemannian geometry. IEEE Trans. Biomed. Eng. **59**, 920–928 (2012)
3. Bhatia, R.: Positive definite matrices. Princeton University Press (2007)
4. Nocedal, J., Wright, S.J.: Numerical optimization. Springer, New York (2006)
5. Ormerod, J.T., Wand, M.P.: Gaussian Variational Approximate Inference for Generalized Linear Mixed Models. Journal of Computational and Graphical Statistics **21**, 2–17 (2012)

4.4 Advances in statistical methods and inference

Estimation of Dirichlet Distribution Parameters with Modified Score Functions

Funzioni di Punteggio Modificate per la Stima dei Parametri della Distribuzione Dirichlet

Vincenzo Gioia and Euloge Clovis Kenne Pagui

Abstract The Dirichlet distribution, also known as multivariate beta, is the most used to analyse frequencies or proportions data. Maximum likelihood is widespread for estimation of Dirichlet's parameters. However, for small sample sizes, the maximum likelihood estimator may shows a significant bias. In this paper, Dirchlet's parameters estimation is obtained through modified score functions aiming at mean and median bias reduction of the maximum likelihood estimator, respectively. A simulation study and an application compare the adjusted score approaches with maximum likelihood.

Abstract *Abstract in Italian* La distribuzione di Dirichlet, anche nota come beta multivariata, è la distribuzione più usata per analizzare dati nella forma di proporzioni o frequenze relative. I parametri della distribuzione di Dirichlet sono comunemente stimati in massima verosimiglianza. Tuttavia, per piccoli campioni, lo stimatore di massima verosimiglianza può esibire una notevole distorsione. In questo articolo, la stima dei parametri della Dirichlet è ottenuta mediante funzioni di punteggio modificate in grado di ridurre, rispettivamente, la distorsione in media e in mediana dello stimatore di massima verosimiglianza. Gli approcci basati sulle funzioni di punteggio modificate vengono confrontati con quello della massima verosimiglianza attraverso uno studio di simulazione e una applicazione.

Key words: compositional data, likelihood, bias reduction.

Vincenzo Gioia
University of Udine, Department of Economics and Statistics, e-mail: gioia.vincenzo@spes.uniud.it,

Euloge Clovis Kenne Pagui
University of Padova, Department of Statistical Sciences, e-mail: kenne@stat.unipd.it

1 Introduction

Proportions data, also referred as compositional data, are very pervasive in many disciplines, ranging from natural sciences to economics. Dirichlet distribution, that is a multivariate generalization of the beta distribution and belongs to the exponential family, is the simplest choice to handle with proportions. Inference on parameters is easily carried out with maximum likelihood (ML). However, for small sample size and large number of parameters, the ML estimator exhibits a relevant bias, as is apparent in simulation results of Narayanan (1992).

In Bayesian framework, the Dirichlet distribution is commonly used as a prior, leading to a conjugate prior of the categorical and multinomial distributions. Moreover, as exponential family the Dirichlet distribution has a conjugate prior. Unfortunately, direct Bayesian inference is not analytically tractable. To our knowledge, there are no works in that direction, apart the following conference (Ma, 2012) and working (Andreoli, 2018) papers.

This paper aims to improve the ML estimates by using modified score functions. Following Firth (1993), the mean bias reduced (mean BR) estimator is obtained as solution of a suitable modified score equation. An alternative modified score function, proposed by Kenne Pagui et al. (2017), aims at median bias reduction (median BR). Mean BR estimator has smaller mean bias than ML and equivariant under linear transformations of the parameters, whereas median BR estimator is componentwise third-order median unbiased in the continuous case and equivariant under componentwise monotone reperameterizations. We study the proposed adjusted score methods through a simulation study and an application, comparing their performance with respect to ML.

2 Dirichlet Distribution

Let $y_i = (y_{i1}, \ldots, y_{im})^\top$, $i = 1, \ldots, n$, be independent realizations of the m-dimensional Dirichlet random vectors parameterized by $\alpha = (\alpha_1, \ldots, \alpha_m)^\top$, with $\alpha_k > 0$, $k = 1, \ldots, m$. The probability density function of $Y_i \sim Dir(\alpha)$ is

$$f_{Y_i}(y_i; \alpha) = \frac{\Gamma(\sum_{j=1}^m \alpha_j)}{\prod_{j=1}^m \Gamma(\alpha_j)} \prod_{j=1}^m y_{ij}^{\alpha_j - 1}$$

with $y_{ik} > 0$, $k = 1, \ldots, m$, and $\sum_{j=1}^m y_{ij} = 1$. The log-likelihood is

$$\ell(\alpha) = n\left\{ \log\Gamma(s) - \sum_{j=1}^m \log\Gamma(\alpha_j) + \sum_{j=1}^m \alpha_j z_j \right\},$$

where $z_j = (\sum_{i=1}^n \log y_{ij})/n$. The log-likelihood is globally concave and the ML estimate needs to be obtained numerically. Parameter estimation is usually carried out

through a Fisher scoring-type algorithm with a sensible choice of the starting value. Wicker et al. (2008)'s proposal seems to be a stable initialisation.

3 Modified Score Functions

For a general parametric model with m-dimensional parameter α and log-likelihood $\ell(\alpha)$, based on a sample of size n, let $U_r = U_r(\alpha) = \partial\ell(\alpha)/\partial\alpha_r$ be the r-th component of the score function $U(\alpha)$, $r = 1,\ldots,m$. Let $j(\alpha) = -\partial^2\ell(\alpha)/\partial\alpha\partial\alpha^\top$ be the observed information and $i(\alpha) = E_\alpha\{j(\alpha)\}$ the expected information.

In order to reduce the bias of the ML estimator, Firth (1993) proposes a suitable modified score aiming at mean BR, of the form

$$\tilde{U}(\alpha) = U(\alpha) + A^*(\alpha),$$

where the vector $A^*(\alpha)$ has components $A^*_r = \frac{1}{2}\mathrm{tr}\{i(\alpha)^{-1}[P_r + Q_r]\}$, with $P_r = E_\alpha\{U(\alpha)U(\alpha)^\top U_r\}$ and $Q_r = E_\alpha\{-j(\alpha)U_r\}$, $r = 1,\ldots,m$. The resulting estimator, $\hat{\alpha}^*$, has a mean bias of order $O(n^{-2})$, less than $O(n^{-1})$ of the ML estimator. Since α is the canonical parameter of the full exponential family, $\hat{\alpha}^*$ corresponds to the mode of the posterior distribution obtained using Jeffreys invariant prior (Firth, 1993).

A competitor estimator, $\tilde{\alpha}$, with accurate median centering property is obtained as solution of the estimating equation based on the modified score (Kenne Pagui et al., 2020)

$$\tilde{U}(\alpha) = U(\alpha) + \tilde{A}(\alpha),$$

with $\tilde{A}(\alpha) = A^*(\alpha) - i(\alpha)F(\alpha)$. The vector $F(\alpha)$ has components $F_r = [i(\alpha)^{-1}]_r^\top \tilde{F}_r$, where $\tilde{F}_r$ has elements $\tilde{F}_{r,t} = \mathrm{tr}\{h_r[(1/3)P_t + (1/2)Q_t]\}$, $r,t = 1,\ldots,m$, with the matrix h_r obtained as $h_r = \{[i(\alpha)^{-1}]_r[i(\alpha)^{-1}]_r^\top\}/i^{rr}(\alpha)$, $r = 1,\ldots,m$. Above, we denoted by $[i(\alpha)^{-1}]_r$ the r-th column of $i(\alpha)^{-1}$ and by $i^{rr}(\alpha)$ the (r,r) element of $i(\alpha)^{-1}$.

In the continuous case, each component of $\tilde{\alpha}$, $\tilde{\alpha}_r$, $r = 1,\ldots,m$, is median unbiased with error of order $O(n^{-3/2})$, i.e. $\Pr_\alpha(\tilde{\alpha}_r \le \alpha_r) = \frac{1}{2} + O(n^{-3/2})$, compared with $O(n^{-1/2})$ of ML estimator. Both $\hat{\alpha}^*$ and $\tilde{\alpha}$ have the same asymptotic distribution as that of the ML estimator, that is $\hat{\alpha} \dot{\sim} \mathcal{N}_m(\alpha, i(\alpha)^{-1})$.

4 Simulation Study

Through a simulation study, with small sample size settings, we compared the performance of the ML, mean and median BR estimators, $\hat{\alpha}$, $\hat{\alpha}^*$ and $\tilde{\alpha}$, respectively. The estimators are compared in terms of empirical probability of underestimation (PU), estimated relative mean bias (RB), and empirical coverage of the 95% Wald-

Table 1 Estimation of parameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$. Simulation results for ML ($\hat{\alpha}$), mean BR ($\hat{\alpha}^*$) and median BR ($\tilde{\alpha}$) estimators.

		$n = 10$			$n = 20$			$n = 40$		
	α	PU	RB	WALD	PU	RB	WALD	PU	RB	WALD
S1	$\hat{\alpha}_1$	40.89	20.89	96.34	43.19	9.23	95.69	44.40	4.39	95.63
	$\hat{\alpha}_1^*$	60.87	-0.17	90.25	56.75	0.01	92.75	54.30	0.05	94.09
	$\tilde{\alpha}_1$	50.26	10.39	94.31	49.54	4.69	94.75	49.11	2.27	95.04
	$\hat{\alpha}_2$	40.77	21.08	96.12	43.21	9.39	95.79	45.16	4.48	95.48
	$\hat{\alpha}_2^*$	60.32	-0.03	89.67	57.29	0.16	92.92	55.09	0.13	94.11
	$\tilde{\alpha}_2$	50.04	10.56	94.07	49.84	4.84	94.76	49.96	2.35	95.03
	$\hat{\alpha}_3$	39.93	21.13	96.54	43.40	9.24	95.82	45.32	4.50	95.19
	$\hat{\alpha}_3^*$	60.55	0.02	90.35	57.71	0.02	92.97	54.87	0.15	93.84
	$\tilde{\alpha}_3$	49.50	10.61	94.36	50.19	4.70	94.67	49.97	2.37	94.64
S2	$\hat{\alpha}_1$	38.22	33.48	96.57	40.27	14.68	96.11	44.13	6.70	95.84
	$\hat{\alpha}_1^*$	63.91	-0.61	86.97	58.66	0.40	91.61	56.60	0.15	93.70
	$\tilde{\alpha}_1$	49.94	16.12	93.30	49.16	7.51	94.53	50.24	3.43	95.11
	$\hat{\alpha}_2$	40.40	23.22	96.23	42.71	10.15	95.88	44.03	4.92	95.23
	$\hat{\alpha}_2^*$	61.35	-0.08	89.16	57.35	0.13	92.94	54.38	0.22	93.90
	$\tilde{\alpha}_2$	50.20	11.27	93.73	50.24	5.04	95.08	49.34	2.54	94.77
	$\hat{\alpha}_3$	42.84	15.08	96.01	45.15	6.84	95.46	46.63	3.23	95.51
	$\hat{\alpha}_3^*$	59.75	-0.04	91.10	56.75	0.02	93.12	54.26	-0.02	94.26
	$\tilde{\alpha}_3$	49.77	8.26	94.54	50.02	3.80	94.81	49.99	1.79	95.23
S3	$\hat{\alpha}_1$	33.06	26.14	96.03	38.48	11.28	95.47	42.29	5.37	95.40
	$\hat{\alpha}_1^*$	59.07	0.25	89.37	56.72	-0.14	92.14	54.32	-0.03	93.67
	$\tilde{\alpha}_1$	49.75	9.06	92.88	50.12	3.73	93.95	50.01	1.80	94.61
	$\hat{\alpha}_2$	33.88	25.49	95.79	38.46	11.05	95.62	42.69	5.26	95.29
	$\hat{\alpha}_2^*$	58.98	0.16	89.29	56.15	-0.13	92.31	54.24	-0.02	93.52
	$\tilde{\alpha}_2$	50.28	8.91	93.13	49.98	3.73	94.15	50.21	1.80	94.49
	$\hat{\alpha}_3$	35.06	23.68	96.05	39.47	10.19	95.58	42.96	4.79	95.32
	$\hat{\alpha}_3^*$	58.61	0.26	89.79	56.26	-0.13	92.39	54.55	-0.10	93.90
	$\tilde{\alpha}_3$	49.31	8.81	93.52	49.96	3.66	94.38	50.02	1.70	94.50
S4	$\hat{\alpha}_1$	33.22	25.32	96.32	38.12	10.92	95.54	41.66	5.19	95.69
	$\hat{\alpha}_1^*$	58.13	0.32	89.37	56.70	-0.12	92.27	53.96	-0.04	94.04
	$\tilde{\alpha}_1$	49.43	8.78	93.34	50.34	3.61	94.06	49.75	1.73	94.70
	$\hat{\alpha}_2$	33.26	25.32	96.34	38.43	10.98	95.34	41.50	5.18	95.17
	$\hat{\alpha}_2^*$	58.25	0.32	89.46	56.33	-0.07	92.35	54.77	-0.05	93.81
	$\tilde{\alpha}_2$	49.16	8.78	93.31	50.15	3.67	94.08	50.21	1.72	94.59
	$\hat{\alpha}_3$	33.25	25.45	96.31	38.62	10.98	95.64	41.91	5.18	95.36
	$\hat{\alpha}_3^*$	58.65	0.43	89.55	56.35	-0.07	92.65	54.85	-0.05	94.01
	$\tilde{\alpha}_3$	49.00	8.90	93.21	50.14	3.67	94.27	50.09	1.71	94.71

type confidence interval (WALD). The three performance measures are expressed in percentages.

We consider the sample sizes $n = 10, 20, 40$, and, for each of 10000 replications, we draw samples of independent observations from 3-dimensional Dirichlet random vector, with true parameter value α_0. Combination of small and large true parameter values with equal and different values are considered. In particular, we perform the study under the settings $\alpha_0 = (0.25, 0.25, 0.25)$ (S1), $\alpha_0 = (0.6, 0.3, 0.1)$ (S2), $\alpha_0 = (12, 6, 2)$ (S3), and $\alpha_0 = (40/3, 40/3, 40/3)$ (S4).

Table 1 shows the numerical results of the simulations. For all settings, mean and median BR estimators proved to be remarkably accurate in achieving their own goals, respectively, and are preferable to ML estimators. The poor coverage of the mean BR estimator is implied by the strong shrinkage effect of the estimator, whereas median BR shows empirical coverage closer to nominal values. The good performances of the ML estimator in terms of empirical coverages, especially when compared with mean BR, are overwhelmed by very large estimated relative mean bias and a noteworthy overestimation of the true parameter.

5 Application

We consider the serum-protein data of Pekin-ducklings analysed in Ng et al. (2011), coming from Mosimann (1962). Data concerns blood serum proportions of $n = 23$ sets of Pekin-ducklings, characterized by having the same diet in each set. For the i-th set, $i = 1, \ldots, 23$, the proportion of pre-albumin (y_{i1}), albumin (y_{i2}) and globulin (y_{i3}), are reported. Ternary plot, in Figure 1, shows in two-dimensions the distibution

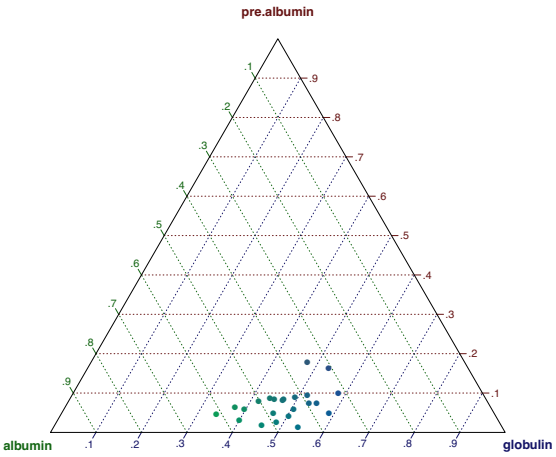


Fig. 1 Serum-protein data of Pekin-ducklings. Ternary plot.

of $y_i = (y_{i1}, y_{i2}, y_{i3})^\top$ on the simplex. Data shows that for a small amount of prealbumina there is about a 50/50 composition of albumin and globulin.

Table 2 Serum-protein data of Pekin-ducklings. Estimates of parameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, estimated standard errors and 95% Wald-type confidence intervals (95% Wald CI) using ML, mean and median BR.

α	Estimate	Standard error	95% Wald CI
$\hat{\alpha}_1$	3.22	0.68	1.89 - 4.54
$\hat{\alpha}_1^*$	2.95	0.62	1.73 - 4.17
$\tilde{\alpha}_1$	3.04	0.64	1.79 - 4.30
$\hat{\alpha}_2$	20.38	4.32	11.91 - 28.86
$\hat{\alpha}_2^*$	18.59	3.95	10.84 - 26.33
$\tilde{\alpha}_2$	19.19	4.08	11.20 - 27.18
$\hat{\alpha}_3$	21.69	4.60	12.67 - 30.70
$\hat{\alpha}_3^*$	19.77	4.20	11.54 - 28.01
$\tilde{\alpha}_3$	20.41	4.34	11.92 - 28.91

Table 2 reports point and interval estimates of the parameters, by using ML, mean and median BR. It is noteworthy the shrinkage effect of the mean BR estimator. Median BR estimates are intermediate between those of mean BR and ML estimates, as well as for the estimated standard errors. As a result of the shrinkage effect of the mean and median BR estimators, the 95% Wald-type confidence intervals for mean BR and median BR are narrower than those of ML.

References

1. Andreoli, J. M. A conjugate prior for the Dirichlet distribution. arXiv:1811.05266, available at https://arxiv.org/abs/1811.05266 (2018)
2. Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika **80**, 27–38 (1993)
3. Kenne Pagui, E. C., Salvan, A. and Sartori N.: Median bias reduction of maximum likelihood estimates. Biometrika **104**, 923–938 (2017)
4. Kenne Pagui, E. C., Salvan, A. and Sartori N.: Efficient implementation of median bias reduction with applications to general regression models. arXiv: 2004.08630, available at https://arxiv.org/abs/2004.08630 (2020)
5. Ma, Z.: Bayesian estimation of the Dirichlet distribution with expectation propagation. Proceedings of the 20th European Signal Processing Conference (EUSIPCO). (2012)
6. Mosimann, J. E.: On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika **49**, 65–82 (1962)
7. Narayanan, A.: A note on parameter estimation in the multivariate beta distribution. Comput. Math. with Appl. **24**, 11–17 (1992)
8. Ng, K. W., Tian, G. L., and Tang, M. L.: Dirichlet and Related Distributions: Theory, Methods and Applications. Chichester: Wiley (2011)
9. Wicker, N., Muller, J., Kalathur, R. K. R., and Poch, O: A maximum likelihood approximation method for Dirichlet's parameter estimation. Comput. Stat. Data Anal. **52**, 1315–1322 (2008)

Confidence distributions for predictive tail probabilities

Distribuzioni di confidenza per probabilità di previsione sulle code

Giovanni Fonseca, Federica Giummolè and Paolo Vidoni

Abstract In this short paper we propose the use of a calibration procedure in order to obtain predictive probabilities for a future random variable of interest. The new calibration method gives rise to a confidence distribution function which probabilities are close to the nominal ones to a high order of approximation. Moreover, the proposed predictive distribution can be easily obtained by means of a bootstrap simulation procedure. A simulation study is presented in order to assess the good properties of our proposal. The calibrated procedure is also applied to a series of real data related to sport records, with the aim of closely estimate the probability of future records.

Abstract *In questo lavoro proponiamo l'utilizzo di una procedura di calibrazione per determinare probabilità predittive per una variabile futura di interesse. Il metodo proposto fornisce distribuzioni di confidenza le cui probabilità si avvicinano a quelle vere con un buon ordine di approssimazione. Le distribuzioni predittive proposte si possono ottenere facilmente attraverso una procedura di bootstrap. Un primo studio di simulazione mostra le buone proprietà delle distribuzioni predittive ottenute. Il nuovo metodo viene anche applicato all'analisi di un insieme di dati reali riguardanti record sportivi, con lo scopo di stimare la probabilità di un nuovo record mondiale.*

Key words: athletic records, asymptotics, bootstrap, calibration, confidence distributions, generalised extreme value distribution, prediction.

Giovanni Fonseca
University of Udine, Via Tomadini 30/A, 33100 Udine (UD), Italy, e-mail: giovanni.fonseca@uniud.it

Federica Giummolè
Ca' Foscari University Venice, Via Torino 155, 30172 Mestre (VE), Italy, e-mail: giummole@unive.it

Paolo Vidoni
University of Udine, Via Tomadini 30/A, 33100 Udine (UD), Italy, e-mail: paolo.vidoni@uniud.it

1 Introduction

Consider the problem of predicting the value of a future or not yet observed random variable, using a sample generated by the same random mechanism. In the frequentist approach, prediction usually requires the specification of a suitable estimate for the (conditional) distribution of the interest random variable, based on the available data, which can be viewed as a confidence distribution ([4], [6]). In particular, this predictive distribution is considered for defining prediction intervals or more simply prediction quantiles, requiring that the associated coverage probability corresponds, exactly or approximately, to the prescribed target probability. Several papers have addressed this problem and, in particular, we mention the calibration approach introduced in [1], and the related bootstrap-based procedure proposed in [3]. In this paper we focus on the different, albeit related, problem of defining a predictive distribution giving well calibrated probabilities for the future random variable. The bootstrap calibration procedure, introduced for the quantiles, is applied in this dual framework, giving a new calibrated distribution in order to obtain predictive probabilities. This new proposal is briefly compared with the existing ones by considering an example involving normal distributed samples. Finally, a real data application, related to sport records and based on the GEV distribution, is presented.

2 Calibrated distributions for prediction probabilities

Let us define the notation and the general assumptions that we require for obtaining the result. Suppose that $\{Y_i\}_{i\geq 1}$ is a sequence of continuous random variables with probability distribution specified by the unknown d-dimensional parameter $\theta \in \Theta \subseteq \mathbf{R}^d$, $d \geq 1$; $Y = (Y_1, \ldots, Y_n)$, $n > 1$, is observable, while $Z = Y_{n+1}$ is a future or not yet available observation. For simplicity, we consider the case of Y and Z being independent random variables and we indicate with $G(z;\theta)$ and $Q(\alpha;\theta)$ the distribution function and the quantile function of Z, respectively. Given the observed sample $y = (y_1, \ldots, y_n)$, we look for a predictive distribution $\hat{G}(z;y)$, with corresponding quantile function $\hat{Q}(\alpha;y)$, that fullfills some good requirements for prediction.

As far as we know, modern literature has mainly focused on the problem of finding a predictive distribution which quantiles satisfy

$$E_Y\{G(\hat{Q}(\alpha;Y);\theta)\} = \alpha, \tag{1}$$

for all $\alpha \in (0,1)$, at least with a high approximation. In this work, we concentrate on the dual problem, that is finding a predictive distribution function $\hat{G}(z;y)$ such that, exactly or approximately,

$$E_Y\{Q(\hat{G}(z;Y);\theta)\} = z, \tag{2}$$

for every $z \in \mathbf{R}$. As it can be noted, instead of assessing the quantile function of Z we are trying to estimate the distribution function itself. In order to solve this problem, we simply apply the same procedure proposed by [3] to the quantile function $Q(\alpha;\theta)$ of Z instead of the distribution function itself. This easily lead to the definition of a new calibrated predictive distribution that may be useful for the calculation of probabilities for Z.

Consider the maximum likelihood estimator $\hat{\theta} = \hat{\theta}(Y)$ for θ, or an asymptotically equivalent alternative, and the estimative predictive distribution and quantile function, $G(z;\hat{\theta})$ and $Q(\alpha;\hat{\theta})$, respectively. The mean of quantiles of level equal to $G(z;\hat{\theta})$ is

$$E_Y[Q\{G(z;\hat{\theta});\theta\}] = A(z,\theta)$$

and, although its explicit expression is rarely available, it is well-known that it does not match the target value z even if, asymptotically, $A(z,\theta) = z + o(1)$, as $n \to +\infty$. It is easy to see that the function

$$Q_c(\alpha;\hat{\theta},\theta) = A\{Q(\alpha;\hat{\theta}),\theta\}, \tag{3}$$

which is obtained by substituting z with $Q(\alpha;\hat{\theta})$ in $A(z,\theta)$, is a proper quantile function, provided that $A(\cdot,\theta)$ is sufficiently smooth. Furthermore, the corresponding distribution function $G_c(z;\hat{\theta},\theta) = G\{A^{-1}(z,\theta);\hat{\theta}\}$ satisfies (2) for every $z \in \mathbf{R}$. Indeed,

$$\begin{aligned} E_Y\{Q(G_c(z;\hat{\theta},\theta);\theta)\} &= E_Y[Q\{G(A^{-1}(z,\theta);\hat{\theta});\theta\}] \\ &= A\{A^{-1}(z,\theta),\theta\} = z. \end{aligned}$$

The calibrated predictive quantile function (3) and the corresponding predictive distribution are not useful in practice, since they depend on the unknown parameter θ. However, a suitable parametric bootstrap estimator for $Q_c(\alpha;\hat{\theta},\theta)$ may be readily defined. Let y^b, $b = 1,\ldots,B$, be parametric bootstrap samples generated from the estimative distribution of the data and let $\hat{\theta}^b$, $b = 1,\ldots,B$, be the corresponding estimates. We can thus write

$$Q_c^{boot}(\alpha;\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} Q\{G(z;\hat{\theta}^b);\hat{\theta}\}|_{z=Q(\alpha;\hat{\theta})}. \tag{4}$$

The associated distribution function allows to estimate the target probability $G(z;\theta) = P(Z \leq z)$, for each $z \in \mathbf{R}$, with an error term which depends on the efficiency of the bootstrap simulation procedure. Indeed, the estimate is the value α such that $Q_c^{boot}(\alpha;\hat{\theta}) = z$.

3 The Normal distribution: a simulation study

Let us first consider the case of prediction for a normally distributed random variable. If we use $\bar{Y} = \sum_i Y_i/n$ and $S = \sqrt{\sum_i (Y_i - \bar{Y})^2/(n-1)}$ as estimators for the unknown parameters, then $T = \sqrt{n/(n+1)}(Z-\bar{Y})/S$ is a pivotal quantity having a Student t distribution with $n-1$ degrees of freedom. Its quantiles satisfy (1) exactly. In spite of this, it could be also interesting to consider the calibrated procedure proposed in [3], which satisfies (1) approximately. Indeed, in some situations the sample mean and standard deviation may not be the most convenient estimators for the parameters and, thus, a pivotal quantity may not be easily available.

In the following we compare the estimative distribution function (Est), the exact distribution function obtained from the pivotal quantity (Piv), the quantile calibrated distribution function of [3] (Qcal) and our proposal, that we name probability calibrated distribution function (Pcal). Figure 1 represents an example of the different predictive distributions obtained from a particular sample y.

We have performed a simulation study in order to assess the properties of the different predictive distributions. Tables 1 and 2 show the results of a Monte Carlo simulation based on $M = 1000$ replications. The bootstrap procedure is based on $B = 500$ replications. The sample size is $n = 10$ and the true parameter values are $\mu = 0$ and $\sigma = 1$. We have compared the different predictive distributions on the basis of the corresponding coverage probability for $\alpha = 0.9, 0.95, 0.99$ (Table 1) and the mean quantiles of levels $\hat{G}(z;y)$ for $z = 1.5, 2, 2.5$ (Table 2). As expected, the pivotal and the quantile calibrated predictive distributions perform better with respect to criterion (1) whereas the probability calibrated predictive distribution outperforms the others with respect to criterion (2).

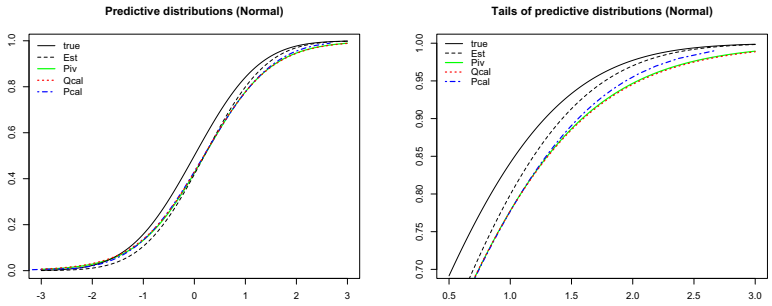


Fig. 1 Normal case: predictive distribution functions (left) and upper tails of predictive distribution functions (right).

Target	Piv	Est	Qcal	Pcal
$\alpha = 0.9$	0.902	0.876	0.902	0.895
$\alpha = 0.95$	0.946	0.919	0.945	0.936
$\alpha = 0.99$	0.990	0.972	0.990	0.982

Table 1 Normal model: coverage probabilities (standard errors are always smaller than 0.0025).

Target	Piv	Est	Qcal	Pcal
$z = 1.5$	1.40	1.63	1.40	1.47
$z = 2$	1.83	2.24	1.83	1.96
$z = 2.5$	2.16	2.75	2.16	2.30

Table 2 Normal model: mean quantiles of level $\hat{G}(z;y)$ (standard errors are always smaller than 0.025).

4 The GEV distribution: an application to athletic records

As a further example, we consider the case of prediction for the generalised extreme value (GEV) distribution, which is usually applied to model maxima of a process over certain time intervals; see for instance [2]. The GEV distribution has three parameters: location, scale and shape. It is important noticing that when the shape parameter is positive (Fréchet distribution) or equal to 0 (Gumbel distribution) the support of the distribution is not limited from above. We have collected annual records in the period 2001 to 2019 for female long jump from the web site of the World Athletics (formerly known as International Association of Athletics Federations (IAAF)) [5].

Using the proposed probability calibrated predictive distribution, we can properly compute probabilities related to the variable Z which represents the best performance in the year to come. In particular we can evaluate the probability of having a new world record in the next year as $\alpha_{WR} = P(Z > WR)$, where WR represents the present world record. This probability can also be used to evaluate the goodness of the world record: the smaller α_{WR} the better the world record. Moreover, from α_{WR} we can calculate the expected number of years for the next record, $T_{WR} = 1/\alpha_{WR}$.

In our example the estimate of the shape parameter of the GEV distribution is positive, thus the estimative GEV distribution function is a Fréchet distribution with no upper bound. Though, the confidence interval for the shape parameter includes 0 and hence, from an inferential point of view, the specification procedure indicates the Gumbel model as the obvious candidate. However, prediction can be quite affected by such a choice, as it can be seen from the results presented in Table 3. Figure 2 shows the estimative (solid), the quantile (dashed) and the probability (dotted) calibrated GEV (red) and Gumbel (black) distribution functions for women's long jump data. The bootstrap procedures are based on 1000 replications. The present world record (solid) is also represented.

The present world record, $WR = 7.52$ m, dates back to 1988 and is not included in the data. Using the GEV probability calibrated distribution instead of the Gumbel one, we take into account for the uncertainty related to the shape parameter

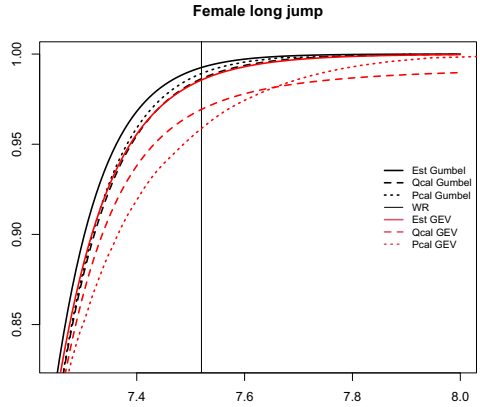


Fig. 2 Women's long jump: GEV and Gumbel predictive distribution functions.

estimation and we can properly assess the probability of improving the current world record: $\alpha_{WR} = P(Z > WR) = 0.041$. Notice also that both the GEV estimative and quantile calibrated predictive distributions wrongly estimate this probability to 0.014 and 0.031, respectively. The expected time for improving the current world record is about 24.4 years.

$WR = 7.52$	Est Gumbel	Qcal Gumbel	Pcal Gumbel	Est GEV	Qcal GEV	Pcal GEV
Probability	0.007	0.014	0.011	0.014	0.031	0.041
Expected time	134.1	73	90.9	69.5	32.5	24.4

Table 3 Probabilities of improving the current world record (WR) with corresponding mean waiting times.

Acknowledgements This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS 003.

References

1. R. Beran. Calibrating prediction regions. *Journal of the American Statistical Association*, 85:715–723, 1990.
2. S. Coles. *An introduction to statistical modeling of extreme values*. Springer-Verlag, London, 2001.
3. G. Fonseca, F. Giummolè, and P. Vidoni. Calibrating predictive distributions. *Journal of Statistical Computation and Simulation*, 84:373–383, 2014.
4. N.L. Hjort and T. Schweder. Confidence distributions and related themes. *Journal of Statistical Planning and Inference*, 195:1–13, 2018.
5. World Athletics (I.A.A.F.). *https://www.worldathletics.org/*.
6. J. Shen, R. Liu, and M. Xie. Prediction with confidence: A general framework for prediction. *Journal of Statistical Planning and Inference*, 195:126–140, 2018.

Impact of sample size on stochastic ordering tests: a simulation study

Impatto della dimensione campionaria sui test per lo stochastic ordering: uno studio di simulazione

Rosa Arboretti, Riccardo Ceccato, Luca Pegoraro, Luigi Salmaso

Abstract Evaluating the presence of a stochastic order among C populations in terms of a certain variable of interest X represents quite a complex problem requiring dedicated statistical solutions. A number of non-parametric techniques have been proposed in the literature to address stochastic ordering problems and this paper investigates the impact of group sizes on the power of some of these methods, in particular considering imbalanced scenarios, by means of a simulation study.

Abstract *Valutare la presenza di un ordinamento stocastico tra C popolazioni in termini di una variabile risposta X rappresenta un problema piuttosto complesso che richiede degli approcci statistici appropriati. Svariate soluzioni non-parametriche sono state proposte in letteratura per affrontare questi problemi di stochastic ordering ed in questo paper si valuta l'impatto della dimensione campionaria di ciascun gruppo sulla potenza di questi metodi, considerando in particolar modo degli scenari sbilanciati, attraverso uno studio di simulazione.*

Key words: stochastic ordering, non-parametric, sample size

Rosa Arboretti
Department of Civil, Environmental and Architectural Engineering, University of Padova, Padova, Italy e-mail: rosa.arboretti@unipd.it

Riccardo Ceccato
Department of Management and Engineering, University of Padova, Vicenza, Italy e-mail: riccardo.ceccato.3@phd.unipd.it

Luca Pegoraro
Department of Management and Engineering, University of Padova, Vicenza, Italy e-mail: pegoraro@gest.unipd.it

Luigi Salmaso
Department of Management and Engineering, University of Padova, Vicenza, Italy e-mail: luigi.salmaso@unipd.it

1 Introduction

Stochastic ordering refers to a problem in which the main interest lies in evaluating the presence of a stochastic order among C populations in terms of a certain variable of interest X.

This type of problem presents a number of possible practical applications. Let us suppose that a manufacturing company is interested in evaluating the performances of three different machines used in their production process. Let us suppose they have an old machine A, a new standard machine B and a new top-of-the-range machine C. We would expect machine C to outperform machine B, which in turn should outperform machine A. In other words, an evident order should emerge from the analysis of the performances of these 3 machines and we may be interested in applying an appropriate test to verify this behavior.

From a mathematical point of view, C populations are stochastically ordered if $X_1 \overset{d}{\geq} \ldots \overset{d}{\geq} X_C$ and at least one strict inequality $\overset{d}{>}$, where the symbol $\overset{d}{>}$ denotes stochastic dominance and X is the variable of interest. Let us note that X_j stochastically dominates X_i if and only if $F_i(x) \leq F_j(x), \forall x$ and $\exists I : F_i(x) < F_j(x), x \in I$ with $Pr(I) > 0$, where F_i and F_j are the cumulative distribution functions of population i and population j. For this reason, using the cumulative distribution functions, a stochastic ordering problem can be formulated using the following system of hypotheses:

$$\begin{cases} H_0 : F_1 = F_2 = \cdots = F_{(C-1)} = F_C \\ H_1 : F_1 \leq F_2 \leq \cdots \leq F_{(C-1)} \leq F_C \text{ and at least one strict inequality,} \end{cases} \tag{1}$$

where the null hypothesis is equality in distribution.

Several non-parametric methods can be found in the literature to address such a problem. Among them are the Jonckheere-Terpstra test (Jonckheere, 1954; Terpstra, 1952), Cuzick's test (Cuzick, 1985) and the permutation-based solutions involving Non-Parametric Combination (NPC) (Pesarin and Salmaso, 2010; Klingenberg et al., 2009; Finos et al., 2007, 2008).

The aim of this paper is to investigate the impact of the size of each group on the power of these common testing procedures for stochastic ordering. We therefore perform a simulation study focusing on imbalanced scenarios, where the size of the samples drawn from each of the C populations is different.

In section 2 we briefly describe the considered methods and in section 3 we conduct the simulation study. In section 4 we draw some conclusions about the results achieved in the simulation study.

2 Methodology

The Jonckheere-Terpstra test (Jonckheere, 1954; Terpstra, 1952) is a non-parametric test based on the adoption of multiple Mann-Whitney tests (Mann and Whitney, 1947). Where C is the number of populations or groups to be compared, $C \times (C-1)/2$ pairwise comparisons are performed using the Mann-Whitney test. If $MW_{ij} = \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} [\mathbb{I}(X_{ik} < X_{jl}) + 0.5\mathbb{I}(X_{ik} = X_{jl})]$ is the test statistic used to compare group i and group j, for $i, j = 1, \ldots, C$ and $i \neq j$, the Jonckheere-Terpstra test statistic is calculated as follows:

$$T^{JT} = \sum_{i=1}^{(C-1)} \sum_{j=i+1}^{C} MW_{ij}, \tag{2}$$

where $\mathbb{I}(z)$ is the indicator function which is 1 if condition z is satisfied and 0 otherwise, and n_i and n_j are the sample sizes of group i and j respectively. Critical values are provided in the literature for small sample sizes, while for large samples a normal approximation is adopted (Jonckheere, 1954).

The implementation proposed in R package *clinfun* (Seshan, 2018) is used.

Cuzick's test (Cuzick, 1985) is another non-parametric solution for testing ordered alternatives proposed as an extension of the Wilcoxon rank sum test. Firstly, scores w_j are given to the C groups according to their ordering: 1 to the first group, 2 to the second, and so on. Ranks are then calculated in each group j and their sum S_j is retrieved. Cuzick's test statistic is calculated as follows:

$$Z^C = (T^C - \mu^C)/\sigma^C \tag{3}$$

where $T^C = \sum_{j=1}^{C} w_j S_j$, μ^C is its expected value and σ^C is its standard error. Under H_0, the distribution of Z^C is approximated to a standard normal distribution.

The version implemented in R package *PMCMRplus* (Pohlert, 2021) is adopted.

The NPC-based solution, described at length in Pesarin and Salmaso (2010), takes advantage of the adoption of permutation tests. For $k = 1, \ldots, C-1$, the first k and the last $(C-k)$ samples are pooled to achieve the pooled samples X_1^k and X_2^k with sizes N_1 and N_2. Therefore, for each k the following sub-problem is addressed:

$$\begin{cases} H_{k0} : X_1^k \overset{d}{=} X_2^k \\ H_{k1} : X_1^k \overset{d}{>} X_2^k. \end{cases}$$

by using appropriate permutation tests. The adopted test statistic is:

$$T^{NPC} = \sum_{i=1}^{N} [\hat{F}_2(X_i^k) - \hat{F}_1(X_i^k)]/\{\bar{F}(X_i^k)[1 - \bar{F}(X_i^k)]\}^{\frac{1}{2}} \tag{4}$$

where $X^k = \{X_1^k, X_2^k\}$ is the pooled sample, $\hat{F}_1(t) = \sum_{i=1}^{N_1} \mathbb{I}(X_{i1}^k \leq t)/N_1$, $\hat{F}_2(t) = \sum_{i=1}^{N_2} \mathbb{I}(X_{i2}^k \leq t)/N_2$, $\bar{F}(t) = \sum_{i=1}^{N} \mathbb{I}(X_i^k \leq t)/N$, and $t \in \mathscr{R}^1$.

$C-1$ partial p-values λ_k and their simulated distribution $\lambda_{kb}^*, b = 1, \ldots, B$, achieved permuting the original data set B times, are therefore computed. A combination

step is then performed to address the global stochastic ordering problem (see System 1). The partial p-values λ_k are combined using Fisher's combining function $T_F'' = -2 \cdot \sum_{k=1}^{C-1} \log(\lambda_k)$ to form a second-order test statistic. At the same time, the distribution of the achieved test statistic is simulated by combining the B vectors $\lambda_{kb}^*, k = 1, \ldots, C-1$. It is therefore possible to calculate a global p-value λ'' to assess the stochastic ordering problem.

The whole procedure is implemented in R (R Core Team, 2020) and codes are available upon request.

3 Simulation study

To investigate the impact of the size of each group on the aforementioned testing procedures, we conducted a simulation study.

We considered a number of settings for sample sizes of the 3 different groups considered:

$Se1$: $n_1 = n_2 = n_3 = 15$
$Se2$: $n_1 = n_2 = 20$ and $n_3 = 5$
$Se3$: $n_1 = n_3 = 20$ and $n_2 = 5$
$Se4$: $n_2 = n_3 = 20$ and $n_1 = 5$
$Se5$: $n_1 = n_2 = 5$ and $n_3 = 35$
$Se6$: $n_1 = n_3 = 5$ and $n_2 = 35$
$Se7$: $n_2 = n_3 = 5$ and $n_1 = 35$

The total sample size remained fixed at 45.

For each setting we simulated data from 3 different distributions, namely Student's t-distribution (with non-centrality parameter μ_S and 2 degrees of freedom), the Log-Normal distribution (with the logarithm of the distribution having mean μ_L and standard deviation $\sigma_L = 1$), and the Cauchy distribution (with location parameter μ_C and scale parameter $\sigma_C = 1$).

For each setting-distribution pair we investigated 2 different scenarios:

- $Sc1$, where H_0 is true and $\mu_G = \mu_S = \mu_L = \mu_C = 10$ for each of the $C = 3$ groups (i.e. $F_1 = F_2 = F_3$)
- $Sc2$, where H_0 is false and $\mu_G = \mu_S = \mu_L = \mu_C = 10$ in the first group, but equal to 8 and 6 in groups 2 and 3 respectively (i.e. $F_1 < F_2 < F_3$)

Both the number of simulation runs and the number of permutations B were set to 2000. For each setting, distribution and scenario 2000 p-values were therefore achieved and a rejection rate was calculated as the proportion of p-values less than or equal to a chosen α-level.

Table 1 shows the type I error rates (i.e. the rejection rates under H_0) for each test and setting under scenario $Sc1$. The considered tests appear mainly to maintain the nominal α-level fixed at 5%, with all the rates being between 0.04 and 0.06.

Let us now focus on the achieved rejection rates (see Table 2) under the second scenario (i.e. $Sc2$). With an α-level equal to 5%, it appears that the NPC-based solution and Cuzick's test are generally more powerful than the Jonckheere-Terpstra test when the Student-t distribution and the Log-Normal distribution are considered.

All the solutions seem to underperform when both the first and the last group are particularly small (i.e. $Se6$: $n_1 = 5$, $n_2 = 35$ and $n_3 = 5$). In the case of the NPC-based solution, this happens because it relies on a sequential pooling of the groups, so that when the observations from the middle group tend to represent the majority of the observations in the pooled groups, the method fails to detect the substantial difference in mean between the first group and the third group. The weights of these two strongly different groups tend to be low and also this affects the performances of the two remaining methods.

Finally, all the methods show their best performance when the middle group is the smallest (i.e. $Se3$: $n_1 = 20$, $n_2 = 5$ and $n_3 = 20$).

Table 1 Scenario $Sc1$. Type I error rates.

Method	Scenario	Stud	Log	Cau
NPC test	15-15-15	0.049	0.040	0.044
NPC test	20-20-5	0.052	0.043	0.052
NPC test	5-20-20	0.051	0.052	0.046
NPC test	20-5-20	0.049	0.048	0.048
NPC test	35-5-5	0.046	0.044	0.055
NPC test	5-35-5	0.048	0.050	0.048
NPC test	5-5-35	0.054	0.050	0.047
JT test	15-15-15	0.054	0.052	0.052
JT test	20-20-5	0.056	0.057	0.047
JT test	5-20-20	0.048	0.054	0.045
JT test	20-5-20	0.052	0.048	0.046
JT test	35-5-5	0.044	0.050	0.048
JT test	5-35-5	0.057	0.048	0.048
JT test	5-5-35	0.050	0.040	0.056
Cuzick test	15-15-15	0.055	0.052	0.052
Cuzick test	20-20-5	0.060	0.052	0.046
Cuzick test	5-20-20	0.049	0.053	0.046
Cuzick test	20-5-20	0.054	0.050	0.044
Cuzick test	35-5-5	0.049	0.047	0.048
Cuzick test	5-35-5	0.060	0.050	0.048
Cuzick test	5-5-35	0.052	0.042	0.057

Table 2 Scenario $Sc2$. Rejection rates.

Method	Scenario	Stud	Log	Cau
NPC test	15-15-15	0.270	0.146	0.754
NPC test	20-20-5	0.218	0.110	0.624
NPC test	5-20-20	0.200	0.122	0.606
NPC test	20-5-20	0.334	0.165	0.827
NPC test	35-5-5	0.218	0.117	0.628
NPC test	5-35-5	0.148	0.096	0.416
NPC test	5-5-35	0.194	0.118	0.605
JT test	15-15-15	0.254	0.126	0.603
JT test	20-20-5	0.189	0.119	0.585
JT test	5-20-20	0.188	0.119	0.589
JT test	20-5-20	0.298	0.150	0.582
JT test	35-5-5	0.194	0.122	0.564
JT test	5-35-5	0.136	0.082	0.449
JT test	5-5-35	0.198	0.136	0.567
Cuzick test	15-15-15	0.270	0.130	0.743
Cuzick test	20-20-5	0.202	0.120	0.652
Cuzick test	5-20-20	0.194	0.128	0.661
Cuzick test	20-5-20	0.312	0.156	0.740
Cuzick test	35-5-5	0.215	0.128	0.638
Cuzick test	5-35-5	0.140	0.087	0.468
Cuzick test	5-5-35	0.204	0.130	0.646

4 Conclusions

In this paper we investigated the impact of size of different groups on a collection of dedicated non-parametric methods in a stochastic ordering problem. In particular we considered the Jonckheere-Terpstra test (Jonckheere, 1954; Terpstra,

1952), Cuzick's test (Cuzick, 1985) and a permutation-based solution involving Non-Parametric Combination (NPC) (Pesarin and Salmaso, 2010).

The conducted simulation study showed that the size of the first and last groups should be high in order to enhance the power of the considered methods, while only a few observations are needed for the middle groups.

These results could provide useful guidelines for practitioners when collecting data to address a stochastic ordering problem. A more complete simulation study has been planned in order to provide further insights on the topic.

References

Bonnini, S., Prodi, N., Salmaso, L., and Visentin, C. (2014). Permutation approaches for stochastic ordering. *Communications in Statistics-Theory and Methods*, 43(10-12):2227–2235.

Cuzick, J. (1985). A Wilcoxon-type test for trend. *Statistics in Medicine*, 4(1):87–90.

Finos, L., Salmaso, L., and Solari, A. (2007). Conditional inference under simultaneous stochastic ordering constraints. *Journal of statistical planning and inference*, 137(8):2633–2641.

Finos, L., Pesarin, F., Salmaso, L., and Solari, A. (2008). Exact inference for multivariate ordered alternatives. *Statistical Methods and Applications*, 17(2):195–208.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41(1/2):133–145.

Klingenberg, B., Solari, A., Salmaso, L., and Pesarin, F. (2009). Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics*, 65(2):452–462.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Pesarin, F. and Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Pohlert, T. (2021). *PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended*. R package version 1.9.0.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Seshan, V. E. (2018). *clinfun: Clinical Trial Design and Data Analysis Functions*. R package version 1.0.15.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14(3):327–333.

On testing the significance of a mode

Verifica della significatività di una moda

Federico Ferraccioli and Giovanna Menardi

Abstract We propose a nonparametric test for the significance of a mode, with the aim of evaluating whether a region of relatively high observed density reflects the actual presence of a mode in the true distribution underlying a set of data. The method leverages on Morse theory to characterize the local properties of the modes and the gradient. This allows the definition of an asymptotic test, based on the concept of gradient ascent paths and relying on resampling methods, to approximate the distribution of the test statistic under the null hypothesis. The performances of the proposed test statistic and the control of the Type-I error are shown via multiple simulation studies.

Abstract Al fine di valutare se una regione ad alta densità osservata riflette la presenza di una moda nella reale distribuzione sottostante i dati, in questo lavoro si propone un test di verifica della significatività di una moda. La procedura proposta sfrutta la teoria Morse per caratterizzare le proprietà locali delle mode e del gradiente di una funzione di densità. In questo modo, è possibile definire una procedura asintotica basata sull'ascesa del gradiente e che sfrutta una tecnica di ricampionamento per approssimare la distribuzione della statistica test sotto l'ipotesi nulla. Il comportamento del test è valutato rispetto alla probabilità di commettere un errore di I-tipo via simulazione.

Key words: bootstrap, mode, nonparametric inference, modal clustering

Federico Ferraccioli
Department of Statistical Sciences, University of Padova e-mail: federico.ferraccioli@unipd.it

Giovanna Menardi
Department of Statistical Sciences, University of Padova e-mail: menardi@stat.unipd.it

1 Introduction

Inference on the modes of a distribution has been historically overlooked with respect to other common location measures such as mean and median. In fact, especially when data exhibit non-Gaussian features as skewness or heavy tails, or some unlabeled heterogeneity occurring in the form of multimodal structures, modes represent useful tools to summarize distributions. Additionally, their understanding may represent a fundamental step to aid deciding how to subsequently approach the analysis the most fruitfully.

A first, well established, branch of literature on this topic addresses the problem of building statistical tests or confidence bounds on the true number of modes [see, for a review, Chacón, 2020, and reference therein]. Alternatively, one may be interested in evaluating the position, rather than the number of modes, to understand whether the regions of relatively high observed density reflect the actual clustering of data in subpopulations; similarly, the observation of somewhat clumped data in the tails of an empirical distribution may induce to wonder if they are real or just a spurious effect of sample variability. These problems can be formalized in the - relatively neglected and fairly complicated - aim of testing the significance of a mode. The few contributions in this direction mostly rely on the study of density features like the gradient or the curvature. See Godtliebsen et al. [2002], Duong et al. [2008] and more recently Genovese et al. [2016]. Consistently with this latter aim, we propose an asymptotic test to evaluate if a specific point is a true mode of the - unknown - probability density function underlying an observed set of data. The procedure borrows some tools from both the theory and the operational means addressing the modal formulation of the clustering problem and it is here applied by following a nonparametric approach. Specifically, we leverage on Morse theory to characterize the local properties of the modes, viewed as local maxima of a function, and their gradient. This formalization allows us to approximate the bootstrap distribution of a mode estimator based on the gradient ascent paths of the density, and used to define an asymptotically chi-squared test statistic.

After framing the problem in the context of Morse theory (Section 2), in the following we illustrate the test and its underlying rationale (Section 3), and show its behaviour with respect to the probability of type-I error via some simulations.

2 Modes as critical points of the density

While intuitively clear, the problem of testing mode significance is firstly definitional. The concept of mode itself is, indeed, ambiguous, as for example the Uniform distribution can be regarded to as both unimodal or without modes. To overcome this problem and formalize our framework without any elusiveness, we shall restrict the analysis to smooth distributions, and exclude non-standard ones as, for example, functions with plateaux. For our purpose, we resort to the framework provided by Morse Theory, a branch of differential topology which draws the relationship be-

tween the stationary points of a smooth real-valued functions on a manifold, and the global topology of the manifold. See Matsumoto [2002] for an introduction.

Given a continuous random variable X, with probability density function $f : \mathbb{R}^d \to \mathbb{R}$, we then assume that f is a Morse function, i.e. a function having non-degenerate critical points. For any $\mathbf{x} \in \mathbb{R}^d$ it is possible to define the *integral curve* of the negative density gradient $-\nabla f$, as the path $\nu_{\mathbf{x}} : \mathbb{R} \mapsto \mathbb{R}^d$ such that

$$\begin{cases} \nu'_{\mathbf{x}}(t) & = -\nabla f(\nu_{\mathbf{x}}(t)) \\ \nu_{\mathbf{x}}(0) & = \mathbf{x}. \end{cases}$$

With that in mind, we identify the set of the local maxima, or modes, of f as

$$\Theta = \{\mathbf{x} \in \mathbb{R}^d : \lim_{t\to\infty} \nu_{\mathbf{x}}(t) = \mathbf{x}\},$$

i.e. the set of points whose integral curve is degenerate at $\nu_{\mathbf{x}}(0)$.

A standard result in Morse theory is that there is a unique gradient ascent path starting at a point that eventually arrives at one of the modes (except for a set of measure 0). Hence, the set of integral curves of the negative gradient allows us to define a partition of $\mathbb{R}^d$ in "domains of attraction" of each mode, to be intended as the sets of points for which $\nu_{\mathbf{x}}(t)$ converges to that mode:

$$\mathscr{D}(\theta) = \{\mathbf{x} \in \mathbb{R}^d : \lim_{t\to\infty} \nu_{\mathbf{x}}(t) = \theta \in \Theta\}.$$

The problem of finding the integral curve $\nu_{\mathbf{x}}(\cdot)$ and its limit $\lim_{t\to\infty} \nu_{\mathbf{x}}(t)$, can be approximated by the iterative scheme

$$\begin{cases} \mathbf{x}_{(0)} & = \mathbf{x}, \\ \mathbf{x}_{(s+1)} & = \mathbf{x}_{(s)} + A\frac{\nabla f(\mathbf{x}_{(s)})}{f(\mathbf{x}_{(s)})}, \end{cases} \tag{1}$$

where A is a $d \times d$ positive definite matrix chosen to guarantee the convergence. Operationally, the function f is unknown, and mode estimation is then performed by plugging in (1) a suitable estimate of both f and its gradient, built from a a sample $\mathscr{X} = (X_1, \ldots, X_n)$ of i.i.d realizations of X, so that the recurrence in (1) becomes

$$\mathbf{x}_{(s+1)} = \mathbf{x}_{(s)} + A\frac{\widehat{\nabla f}(\mathbf{x}_{(s)}; \mathscr{X})}{\hat{f}(\mathbf{x}_{(s)}; \mathscr{X})}. \tag{2}$$

The convergence properties of this gradient-ascent algorithm have been studied in Arias-Castro et al. [2016] and Chen et al. [2016]. A possible choice is to estimate the density and its gradient with a kernel estimator, leading to a particularly convenient iteration scheme known as the mean-shift [see Chacón and Duong, 2018, Ch. 6, for a more detailed derivation]. In the next Section, we will use these properties to define a test statistic for the modes of a density function.

3 Methodology

In the lack of information about the true modal structure of f, testing the significance of a mode recasts to defining the system of hypotheses

$$H_0 : \theta_0 \in \Theta \quad \text{vs} \quad H_1 : \theta_0 \notin \Theta, \tag{3}$$

for some $\theta_0 \in \mathbb{R}^d$. While apparently composite, the null hypothesis is fact a simple one, as the - yet unknown - partition of $\mathbb{R}^d$ in the set $\{\mathscr{D}(\theta)\}_{\theta\in\Theta}$ allows us to intend H_0 as "θ_0 is the mode of the domain $\mathscr{D}(\theta)$ where it belongs".

Building on the sample $\mathscr{X}$, we first obtain an estimate $\hat{f}(\cdot;\mathscr{X})$ and $\widehat{\nabla f}(\cdot;\mathscr{X})$ of the density function and its gradient. For the subsequent developments, we consider a nonparametric kernel density estimator, which has been proven to provide consistent estimates of f under some regularity conditions on the function and the selected amount of smoothing [see, e.g. Chacón et al., 2015]. In fact, other methods for density estimation - not necessarily nonparametric - with good general properties and producing differentiable estimates can be used.

To test (3) we then build an estimate $\hat{\theta}$ of the mode. This is obtained as the convergence point of the iteration scheme (2), with $\mathbf{x}_{(0)} = \theta_0$, and represents the mode of $\hat{f}$ associated with the domain $\mathscr{D}(\theta)$ to which θ_0 belongs. Afterwards, we obtain an approximation of the distribution of $\hat{\theta}$ under the null hypothesis. Here we propose to approximate such distribution with a resampling procedure, such as bootstrap or subsampling [Politis et al., 1999], together with the iteration scheme in (2). In particular, let $\mathscr{X}^*$ be a resampled version of the original data $\mathscr{X}$. With the obtained sample, using the initial condition $\mathbf{x}_{(0)} = \hat{\theta}$, we compute

$$\theta^* = \hat{\theta} + A\frac{\widehat{\nabla f}(\hat{\theta};\mathscr{X}^*)}{\hat{f}(\hat{\theta};\mathscr{X}^*)}.$$

Here we consider $A = \alpha I_d$, with $0 < \alpha < 1$, to guarantee the convergence. The underlying rationale is that, under the null hypothesis, since $\hat{\theta} \to \theta_0$ and $\widehat{\nabla f} \to \nabla f$, we expect $\widehat{\nabla f}(\hat{\theta};\mathscr{X}^*)$ to be close to zero. Hence, by iterating the process B times, we obtain a set $\{\theta_1^*,\dots,\theta_B^*\}$ of realizations from the bootstrap distribution of $\hat{\theta}$ under H_0. With that in mind, we define

$$\hat{\mu} = \frac{1}{B}\sum_{b=1}^{B}\theta_b^* \quad \text{and} \quad \hat{\Sigma} = \frac{1}{B}\sum_{b=1}^{B}(\theta_b^* - \hat{\mu})^2.$$

From the multivariate central limit theorem [Van der Vaart, 2000] it follows that under the null hypothesis $\sqrt{n}(\hat{\mu} - \theta_0) \dot{\sim} \mathscr{N}(0,\hat{\Sigma})$. We can therefore define a test statistic

$$T = (\hat{\mu} - \theta_0)^\top \hat{\Sigma}^{-1}(\hat{\mu} - \theta_0) \dot{\sim} \chi^2_d,$$

and reject H_0 for large values of T.

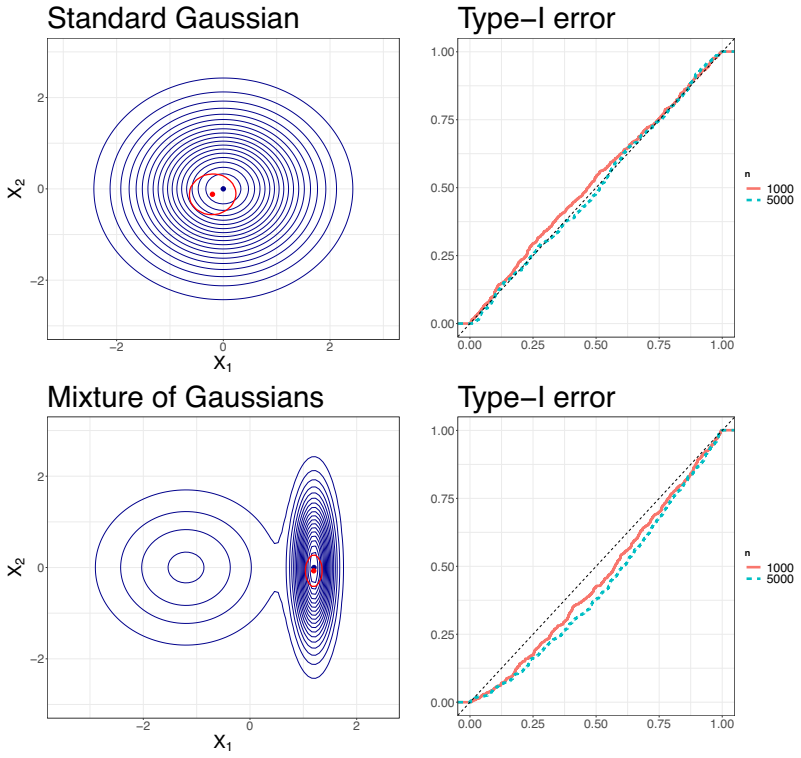


Fig. 1 In the first row, two dimensional standard Gaussian. On the second row, the Gaussian mixture. The left panels show the resampled distribution of the modes, with the black dot corresponding to θ_0. The right panels show the control of the Type-I error for two different sample sizes, $n = 1000$ and $n = 5000$.

4 Empirical study

To check the control of the Type-I error probability of the proposed test statistic, we have conducted a simulation study. For brevity, we report here the results of two bivariate settings of different complexity only, illustrated in the left panels of Figure 1, and referred to the mode of a standard Gaussian distribution and the most prominent mode of a balanced mixture of two Gaussian distributions with even variance components.

In both cases we generated 500 samples of size $n = 1000$ and $n = 5000$, and we compared the p-value curves vs increasing values of Type-I error probabilities.

In the first scenario, where the distribution is unimodal and isotropic, the test shows very good performances and the control of the Type-I error is almost perfect, even with a smaller sample size. Although this case is fairly simple, it is nonetheless informative on the behaviour of the proposed test in a benchmark setting. In the second, more complex, scenario, we focused on the right-most mode. As clear in

Figure 1, the region of interest is highly anisotropic in the vertical direction, with very steep gradients in the horizontal direction. In this case the true distribution of the mode might have a smaller variability in the horizontal direction with respect to the resampled distribution, thus leading to a more conservative test.

The proposed test shows fairly good performances and control of the Type-I error in both scenarios. Moreover, due to the small number of iterations in the gradient procedure, it is computationally efficient even in higher dimensions and with larger sample sizes. Future research will focus on a more thorough analysis on the control of the Type-I error and the power of the test in more complicated scenarios. It would also be of interest to better understand the theoretical and asymptotic properties of the proposed procedure.

References

Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *The Journal of Machine Learning Research*, 17(1):1487–1514, 2016.

José E Chacón. The modal age of statistics. *International Statistical Review*, 88(1): 122–141, 2020.

José E Chacón and Tarn Duong. *Multivariate kernel smoothing and its applications*. CRC Press, 2018.

José E Chacón et al. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015.

Yen-Chi Chen, Christopher R Genovese, Larry Wasserman, et al. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016.

Tarn Duong, Arianna Cowling, Inge Koch, and Matt P Wand. Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, 52(9):4225–4242, 2008.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 99–126, 2016.

Fred Godtliebsen, JS Marron, and Probal Chaudhuri. Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics*, 11(1):1–21, 2002.

Yukio Matsumoto. *An introduction to Morse theory*, volume 208. American Mathematical Soc., 2002.

Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Hommel BH: an adaptive Benjamini-Hochberg procedure using Hommel's estimator for the number of true hypotheses

Hommel BH: una procedura Benjamini-Hochberg adattativa basata sullo stimatore di Hommel per il numero di ipotesi vere

Chiara G. Magnani, Aldo Solari

Abstract We propose an adaptive Benjamini and Hochberg procedure for control of the false discovery rate (FDR) by using Hommel's estimator for the number of true hypotheses. We show that the proposed procedure has FDR control under independence and under the assumption that p-values corresponding to true null hypotheses satify Simes' inequality. We illustrate the new method with an application to two well-known examples from the literature.

Abstract *Con l'intento di controllare il false discovery rate (FDR) proponiamo una procedura Benjamini e Hochberg adattativa, che utilizza lo stimatore di Hommel per il numero di ipotesi vere. Mostriamo che la procedura proposta garantisce il controllo del FDR, supponendo che i p-value siano indipendenti e che quelli corrispondenti alle ipotesi vere soddisfino la disuguaglianza di Simes. Illustriamo il nuovo metodo applicandolo a due noti esempi in letteratura.*

Key words: adaptive procedure, Benjamini-Hochberg procedure, false discovery rate, Hommel's method, multiple testing, Simes' inequality.

1 Introduction

When testing many hypotheses simultaneously, it is essential to control, or at least to quantify, the flood of type I errors. In their seminal 1995 paper [1], Benjamini and Hochberg introduced the False Discovery Rate (FDR) – the proportion of type I errors among the rejections – and argued that in large-scale testing it is preferable to control FDR because it is a scalable criterion as opposed to familywise error rate. The Benjamini and Hochberg procedure, BH for short, has since become the

Chiara G. Magnani
Department of Economics, Management and Statistics, e-mail: c.magnani9@campus.unimib.it

Aldo Solari
Department of Economics, Management and Statistics, e-mail: aldo.solari@unimib.it

standard for multiple hypothesis testing in many fields, evidenced by more than 70000 citations as of February 27, 2021.

Given m ordered p-values $p_1 \leq \ldots \leq p_m$ for m null hypotheses, the BH(α) procedure rejects the $R(\alpha)$ hypotheses with smallest p-values, where $R(\alpha) = \max\{i \in [m] : p_i \leq i\alpha/m\}$ with $[m] = \{1,\ldots,m\}$, and $R(\alpha) = 0$ if this maximum does not exist. Under the assumption of positive regression dependence on the subset of p-values of true null hypotheses (PRDS) [3], the BH(α) procedure controls the FDR at level α, i.e.

$$\mathrm{FDR}_{\mathrm{PRDS}}(\mathrm{BH}(\alpha)) = \mathbb{E}\Big(\frac{V(\alpha)}{R(\alpha) \vee 1}\Big) \leq \pi_0 \alpha \leq \alpha, \tag{1}$$

where $V(\alpha)$ is the number of type I errors of BH(α) procedure, m_0 is the number of true hypotheses and $\pi_0 = m_0/m$ is the proportion of true hypotheses.

If π_0 were known, one could use the more powerful BH(α/π_0) and still control FDR at level α, i.e. $\mathrm{FDR}(\mathrm{BH}(\alpha/\pi_0)) \leq \alpha$. Since π_0 is usually not known, several authors have suggested adaptive procedures that first estimate π_0 by $\hat{\pi}_0$, and subsequently use BH($\alpha/\hat{\pi}_0$). However, FDR control of most adaptive procedures, i.e. $\mathrm{FDR}(\mathrm{BH}(\alpha/\hat{\pi}_0)) \leq \alpha$, has been proved only under the assumption of independent p-values [4, 5]. An adaptive procedure that is a tiny, but uniform improvement over the original BH procedure and which is valid under exactly the same conditions has been proposed by [13]. This procedure, named Minimally Adaptive BH (MABH), only admits the estimates $\hat{\pi}_0 = 1$ if $R(\alpha) = 0$ and $\hat{\pi}_0 = (m-1)/m$ otherwise. Because of this limitation, the gain in power relative to the BH procedure is negligible when m is large.

In this short contribution we propose Hommel BH (HBH), an adaptive procedure that uses Hommel's estimator [9] for π_0. HBH can be seen as an iterative version of MABH, although it is not a uniform improvement of BH. The procedure is presented in Section 2, and FDR control (i) under the assumption of independent p-values or (ii) under the assumption that p-values corresponding to true null hypotheses satify Simes' inequality [12] is discussed in Sections 3 and 4, respectively.

2 The HBH procedure

The method of Hommel [9] is a well-known multiple testing procedure that controls the familywise error rate: it guarantees no type I error with probability at least $1-\alpha$. Hommel's method is uniformly more powerful than the methods of Bonferroni, Holm, and Hochberg, and the gain in power comes from making a certain assumption on the dependence structure of the p-values. Hommel's procedure assumes that p-values corresponding to true null hypotheses satisfy Simes' inequality, i.e.

$$q_i \geq i\alpha/m_0, \quad i = 1,\ldots,m_0 \tag{2}$$

with probability at least $1-\alpha$, where $q_1 \leq \ldots \leq q_{m_0}$ denote the ordered null p-values. Simes' inequality is necessary but not sufficient for the validity of the BH procedure.

A key element in Hommel's method is the estimator for the number of true hypotheses m_0, i.e.

$$\hat{m}_0(\alpha) = \max\left\{i \in [m] : p_{m-i+j} > j\alpha/i \text{ for } j = 1,\ldots,i\right\} \tag{3}$$

It has been shown [8] that Hommel's estimator $\hat{m}_0$ is an upper $(1-\alpha)$-confidence bound for m_0, i.e. $m_0 \leq \hat{m}_0(\alpha)$ with probability at least $1-\alpha$ if (2) holds for the null p-values.

Hommel's estimator for π_0, $\hat{\pi}_0(\alpha) = \hat{m}_0(\alpha)/m$, results in $\hat{\pi}_0(\alpha) = 1$ if and only if $R(\alpha) = 0$, as the MABH estimator. Otherwise, Hommel's estimator $\hat{\pi}_0$ quantifies the largest proportion of p-values satisfying Simes' inequality, as opposed to MABH estimator that always indicates a proportion of $(m-1)/m$.

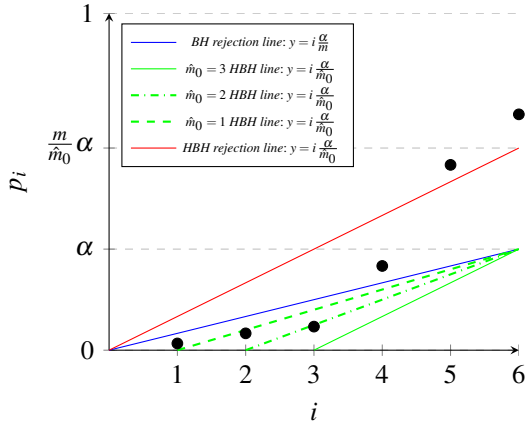
The HBH(α) procedure is defined as follows:

1. If $p_i > i\alpha/m$ for all $i = 1,\ldots,m$, HBH(α) rejects 0 hypothesis, or if $p_m \leq \alpha$ HBH(α) rejects all m hypotheses;
2. Otherwise, HBH(α) rejects

$$R(\alpha/\hat{\pi}_0(\alpha)) = \max\left\{i \in [m] : p_i \leq i\alpha/\hat{m}_0(\alpha)\right\}$$

where $\hat{m}_0(\alpha)$ is defined in (3).

The following Figure displays a geometrical representation of the steps involved in Hommel's estimator $\hat{m}_0$ and a comparison of BH and HBH rejection lines. In the example, BH(α) rejects 3 hypotheses, but HBH(α) rejects an additional hypothesis by using the estimate $\hat{m}_0(\alpha) = 3$,



3 FDR control under independence

In [10] we proved that by assuming independent p-values, the least favorable parameter configuration (LFC) for HBH is the Dirac-Uniform (DU) configuration of p-values, meaning that non-null p-values are equal to zero and the null p-values are Uniform(0,1) [7, 11]. Then

$$\mathrm{FDR}_{\mathrm{IND.}}(\mathrm{HBH}(\alpha)) \leq \max_{1\leq m_0\leq m} \mathrm{FDR}_{\mathrm{DU}(m,m_0)}(\mathrm{HBH}(\alpha)) = \overline{\mathrm{FDR}}_{\mathrm{IND.}}(m,\alpha) \quad (4)$$

and in order to control FDR we can find a corrected significance level $\alpha^* \leq \alpha$ such that $\overline{\mathrm{FDR}}_{\mathrm{IND.}}(m,\alpha^*) = \alpha$. For example, $\overline{\mathrm{FDR}}_{\mathrm{IND.}}(5,\alpha) = \mathrm{FDR}_{\mathrm{DU}(5,2)} = \alpha + 3\alpha^2/8$ implies $\alpha^* = (-8+\sqrt{64+96\alpha})/6$. The following is a general expression of the FDR for HBH assuming independence and DU configuration of p-values:

$$\begin{aligned} \mathrm{FDR}_{\mathrm{DU}(m,m_0)}(\mathrm{HBH}(\alpha)) &= \mathrm{FDR}_{\mathrm{DU}(m,m_0)}(\mathrm{BH}(\alpha/\hat{\pi}_0)) \quad (5)\\ &= \sum_{v=1}^{m_0} \frac{v}{m_1+v} \sum_{b=0}^{v} \mathbb{P}(V(\alpha/\hat{\pi}_0)=v, \hat{m}_0(\alpha)=m_0-b), \end{aligned}$$

where $m_1 = m - m_0$ is the number of false hypotheses and $\mathbb{P}$ is under $\mathrm{DU}(m,m_0)$. For more details see [10], but it's important to underline that we've been able to express all the events $(V=v, \hat{m}_0 = m_0 - b)$ depending on ordered null p-values, which under DU are order statistics of the uniform distribution.

Calculation of $\mathbb{P}(V=v,\hat{m}_0=m_0-b)$ requires iterated integrals and may be computationally hard especially for large values of m and m_0. To make the algorithm faster we decided to implement an upper bound $\overline{\mathrm{FDR}}_{\mathrm{DU}(m,m_0)}$ for $\mathrm{FDR}_{\mathrm{DU}(m,m_0)}$, which happened to be very tight [10].

Figure 1 shows $\overline{\mathrm{FDR}}_{\mathrm{DU}(m,m_0)}$ as a function of m_0 for $m=15$ and $\alpha=0.1$. The decreasing trend after the maximum point is common for many values of m. Figure 2 shows $\overline{\overline{\mathrm{FDR}}}_{\mathrm{IND.}}(m) = \max_{1\leq m_0\leq m} \overline{\mathrm{FDR}}_{\mathrm{DU}(m,m_0)}$ as a function of m for $\alpha=0.1$.

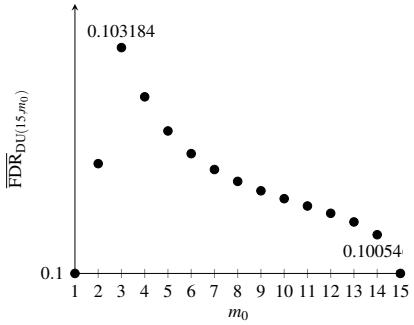


Fig. 1

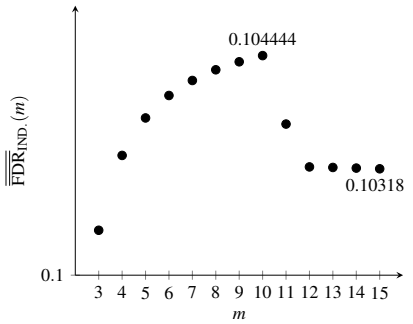


Fig. 2

Based on the previous results, we consider the 15 p-values example presented in [2, 4] and the 34 p-values example presented in [3]. The number of rejected hypotheses of BH and HBH, when they both control the FDR at level α, are reported in the following Table. For the data sets considered, HBH rejects at least as many hypotheses as BH does, and possibly more.

Data set	m	α	Number of rejections BH	HBH
[2, 4]	15	5%	4	5
		10%	9	9
[3]	34	5%	11	12
		10%	12	21

4 FDR control under dependence

A procedure is compliant at level α if every rejected p-value p_i satisfies $p_i \leq \alpha R(\alpha)/m$ [6]. Control of the FDR for compliant procedures under the Positive Regression Dependence within Nulls (PRDN) has been recently proved by [14].

Theorem 1 (Su, 2018). *Assume that the null p-values satisfy the PRDN property. Then for any compliant at level α multiple testing procedure*

$$\mathrm{FDR}_{\mathrm{PRDN}}(\mathrm{COMPLIANT}(\alpha)) \leq \pi_0\alpha + \pi_0\alpha \ \log \ \frac{1}{\pi_0\alpha} \leq \alpha + \alpha \ \log \ \frac{1}{\alpha} \quad (6)$$

It is easy to prove that BH and HBH are compliant respectively at level α and $\alpha m/\hat{m}_0$. PRDN is a sufficient condition for Simes' Inequality [9] and necessary for PRDS.

Corollary 1. *Assume that the null p-values satisfy the Simes' inequality. Then*

$$\mathrm{FDR}_{\mathrm{SIMES}}(\mathrm{HBH}(\alpha)) \leq 2\alpha + \alpha \ \log \ \frac{1}{\alpha} \quad (7)$$

The proof is available in [10]. The price to pay in order to use HBH is α, which is low considering the values usually assumed by the significance level. Taking $\alpha =$ 0.0072 and 0.0163 is sufficient to ensure FDR control of HBH at level 0.05 and 0.1 under PRDN, compared to 0.0087 and 0.0204 for BH [14].

5 Discussion

FDR control of the BH procedure has been proved under the following assumptions on the p-values:

$$\text{INDEPENDENCE} \Rightarrow \text{PRDS} \Rightarrow \text{PRDN} \Rightarrow \text{ANY DEPENDENCE}$$

We have shown that the HBH procedure controls FDR under INDEPENDENCE and PRDN with a small correction of the significance level. The relevant case of FDR control under PRDS will be addressed in future research.

References

1. Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. **57**, 289–300 (1995).
2. Benjamini, Y. and Hochberg, Y.: On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Stat. **25**, 60–83 (2000)
3. Benjamini, Y. and Yekutieli, D.: The control of the False Discovery Rate in Multiple Testing under dependency. Ann. Stat. **29**, 1165–1188 (2001)
4. Benjamini, Y., Krieger, A. and Yekutieli, D.: Adaptive linear step-up procedures that control the false discovery rate. Biometrika **93**, 491–507 (2006)
5. Blanchard, G. and Roquain, E.: Adaptive false discovery rate control under independence and dependence. J. Mach. Learn. Res. **10**, 2837–2871 (2009)
6. Dwork, C., Su, Weijie J. and Zhang, L.: Differentially private false discovery rate control. (2018) arXiv preprint arXiv:1807.04209.
7. Finner, H., Dickhaus, T. and Roters, M.: On the False Discovery Rate and an asymptotically optimal rejection. Ann. Stat. **37**, 596–618 (2009)
8. Goeman, J.J., Meijer, R.J., Krebs, T. J. and Solari, A.: Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. Biometrika **106**, 841-856 (2019)
9. Hommel, G.: A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika **75**, 383–6 (1988)
10. Magnani, C. G.: False Discovery Rate in a particular adaptive Benjamini and Hochberg procedure. Master's Thesis, University of Milano-Bicocca (2021)
11. Roquain, E. and Villers, F.: Exact Calculations for False Discovery Proportion with applications to least favorable configurations. Ann. Stat. **39**, 584-612 (2011)
12. Simes, R. J.: An improved Bonferroni procedure for multiple tests of significance. Biometrika **73**, 751–754 (1986)
13. Solari, A. and Goeman, J.J.: Minimally adaptive BH: A tiny but uniform improvement of the procedure of Benjamini and Hochberg. Biom. J. **59**, 776–780 (2017)
14. Su, Weijie J.: The FDR-linking theorem. (2018) arXiv preprint arXiv:1812.08965

4.5 Advances in statistical models

Specification Curve Analysis: Visualising the risk of model misspecification in COVID-19 data

Analisi della Curva di Specificazione: Visualizzare il rischio di specificazione nei modelli per l'analisi dei dati COVID-19

Venera Tomaselli, Giulio Giacomo Cantone, and Vincenzo Miracula

Abstract The study aims at exploring COVID-19 data through the Specification Curve Analysis. Checking a multiverse of regression relationships allows the researcher to minimise the risk misspecification in the models.

Abstract *Questo studio mira ad esplorare i dati COVID-19 attraverso l'analisi della Curva di Specificazione. Indagare un multiverso di relazioni di regressione consente al ricercatore di ridurre al minimo il rischio di errore di specificazione nei modelli di analisi.*

Key words: Specification Curve Analysis, model misspecification, COVID-19, Multiverse Analysis.

1 Introduction

According to Kenneth Arrow, model misspecification is when "uncertainty is induced by the approximate nature of the models under consideration to use in assigning probabilities", a direct consequence of "ambiguity among models, where the uncertainty is about which alternative model, or convex combination of such models, should be used to assign the probabilities" (p. 511) [7] .

Social scientists have debated about how overconfidence in singular statistics (i.e., p-values), paired with excessive ambiguity in the social sciences models, actually

Venera Tomaselli
Department of Political and Social Sciences, University of Catania, 8, Vittorio Emanuele II, 95131 Catania, e-mail: venera.tomaselli@unict.it

Giulio Giacomo Cantone
Department of Physics and Astronomy "E. Majorana", University of Catania, 64, S. Sofia, 95123 Catania, e-mail: giulio.cantone@phd.unict.it

Vincenzo Miracula e-mail: vincenzomiracula13@gmail.com

increases uncertainty about the outcome of decision-making. Finance, for example, has an established recognition of the risk of model misspecification [15], [16].

Contemporary society is defined as a risk society [1]. Industrial and technological progress is often a threat to health and environment. Unlike in the past, when the dangers were evident and easy to spot, today the risk is invisible and widespread. The topic of risk is a typical feature in modern science. The perception of a threat from nature is associated with a growing awareness of the limits of expert knowledge[5]. Theory of risk of model misspecification was linked to overconfidence in assumptions of policy-makers in environmental issues: the role of science in the society of risk can be both cause and source for risk management [19].

The paper is focused on the methodological discussion about employment of multiversal statistics and specification curve analysis to explore COVID-19 data.

2 Specification Curve Analysis

The first proponents of the analysis of the multiverse made through specifications of a conceptual regressive model are Steegen *et al.* [18]. The issue they were facing through the theory of 'multiverse of regressions' was the so-called 'replication crisis' or 'crisis of p-values': the recognition that a singular significant p-value is not sufficient to justify the belief that the relationship between two constructs of a model is effective [6]. Their work was inspired from reflections about scientific practices discussed by scholars as Gelman and Loken [4] and Ioannidis [11].

This methodology asks researchers to produce a 'multiverse' of all the combinations ('specifications') of the conceptually plausible variables, subsets, and typologies of regression and then to observe the distribution of p-values of the first independent variable ('regressor') in all the specifications. Simonsohn *et al.* [17] identified five components of a specification shown in the table 1.

Table 1 The five components of specification.

Component	Research Questions
Subset	Which cases (e.g., outliers) should be removed from the dataset?
Operationalising of input	How to define, quantify, and choose regressors among alternatives, and why?
Operationalising of output	How to define, quantify, and choose the dependent variable? Should it be observable or a latent concept?
Type of regression model	Linear, Generalised Linear, Non Linear, etc.
Functional form	Should we apply a function to some variables (e.g., *log*)? etc.

Simonsohn *et al.* [17] suggest to employ non-parametric 'multiversal' statistics (median and percentage of significant p-values among all the specifications) to estimate the robustness of the regressor in the multiverse of specifications. This intuition is very helpful to overcome statistical issues, as the Simpson's Paradox of

cumulation of time-phases. Since timed observations can be subset (see Subset in Table 1) in different phases making different specifications of the model, one can employ a 'multiversal' statistic to test the robustness of a more general hypothesis.

Since the number of plausible model specifications within a scientific theory can be large, this methodology falls not very distant from those proposals of "*just running a lot of regressions*" [14] and the recent work in [9]. Multiverse analysis was expanded into Specification Curve Analysis in [17] and was employed in [13] and [2].

The innovative idea of specification curve was to rank all the specifications (or a representative sample) through the estimate of the first regressor and then shading the area of the confidence intervals. This helps to visualise the robustness of the latent concept through the observation of the slopes of the curve. Ideally, a flat curve lying distant from 0 indicates that the latent concept is robust: no matter about how the researchers elicit the five components, all the regressors keep their effect on the dependent variable. Instead, curves intersecting with x-axis or with slopes distant from 0 are problematic and indicate that different nominal specifications of the conceptual model actually produce divergent outcomes, meaning that likely the theory is weak.

3 Data and concepts

COVID-19 affects the responsiveness and the availability of healthcare systems. Among the most monitored and analysed COVID-19 variables there are population features, impact of airport passengers mobility [3], pollution, industrial and business districts [12]. Levels of morbidity were associated with the risk of infection [10]. Since public health services play an important role in both controlling the epidemic and preventing new infections, it is essential to understand the effects of so many variables on the capacity and efficiency of healthcare services.

The present study is aimed at testing the effect of the variables related to the spread of COVID-19 in Italy in 3 different epidemic phases (Table 2). All the variables are observed at regional territorial level (19 regions + 2 autonomous provinces) in order to avoid missing data often observed at provincial level.

As shown in Table 2, the demographic variables enter also as controls for multiple regressions, generating new specifications of the model. All the specifications dependent ~ regressor(s) are linear. Benefits to observe a multiverse of regressions, even with very distinct conceptual regressors on the territorial spread of COVID-19 in the epidemic phases, are discussed in Section 2. The data analysis is carried out by software Spec*R* from Masur and Scharkow [8].

Table 2 Variables in Multiverse Model.

Concept	Description of Variable	Role in Regression	Source[a]	Year
Covid-19	N. hospitalised	Dependent	ICP Dept. (GitHub)	
Covid-19	N. intensive care	Dependent	ICP Dept. (GitHub)	
Covid-19	N. quarantined	Dependent	ICP Dept. (GitHub)	
Time	Epidemic phase[b]	Subsets of Dependent		
Demography	N. over 65+ years	Control & Regressor	ISTAT	2020
Demography	Population Density Index	Control & Regressor	ISTAT	2020
Elderly Care	Care Capacity Index	Regressor	ISTAT	2020
Pollution	PM10 Index	Regressor	SNPA	2019
Pollution	NO2 Index	Regressor	SNPA	2019
Mobility	N. aeroportual passengers	Regressor	AssoAeroporti	2019
Economy	N. companies (Primary Sector)	Regressor	Infocamere	2019
Economy	N. companies (Manifacturies)	Regressor	Infocamere	2019
Economy	N. companies (Services)	Regressor	Infocamere	2019
Health	% high blood pressure	Regressor	EpiCentro (ISS)	2019
Health	% obeses	Regressor	EpiCentro (ISS)	2019
Health	% hypercholesterolaemia	Regressor	EpiCentro (ISS)	2019
Health	% smokers	Regressor	EpiCentro (ISS)	2019
Health	N. drug stores	Regressor	Health Department	2020

[a] ICP: Italian Civil Protection; ISTAT: Italian Statistical Institute; SNPA: National Environmental Protection of Italy; AssoAeroporti: National Association of Aeroports of Italy; InfoCamere: Italian Boards of Trade; ISS: Italian Institute of Health.
[b] 1st phase: 02/24/2020 - 05/03/2020, 2nd phase: 05/04/2020 - 10/23/2020, 3nd phase: 10/24/2020 - 01/31/2021, Cumulative: 02/24/2020 - 01/31/2021.

4 Discussion of results, limitations, and developments

The model from the combination of components of Table 2 results in a multiverse of 624 specifications:

$$(4S \times 3D) \times (12R \times 4F + 2R \times 2F) = 624 \text{ specifications} \quad (1)$$

where:

- S = N. subsets (3 epidemic phases + 1 cumulative) of the dependent variable
- D = N. dependent variables
- R = N. regressors
- F = N. regressive forms (controlled / uncontrolled)

Figure 1 shows that the regressive variables have different effects on dependent variable. The multiverse of regressions (Table 3) provides relevant insights:

- there are some regressors very tied with the selected indicators of territorial diffusion of COVID-19: *Manifacturies* and *age*
- some variables are mischievous. High blood pressure shows clearly how misleading can be a singular p-value $< .05$ because half of p-values of specifications of this regressor would have been candidates for statistically significant effects, but

no case < .01 would have been observed. Half of the specifications of *Aeroportual Passengers* are surely associated with COVID-19. However the other half have a higher *p*-value because some specifications were more effective than others

- finally, health and pollution do not seem correlated with territorial diffusion of COVID-19 from the perspective of an observational study.

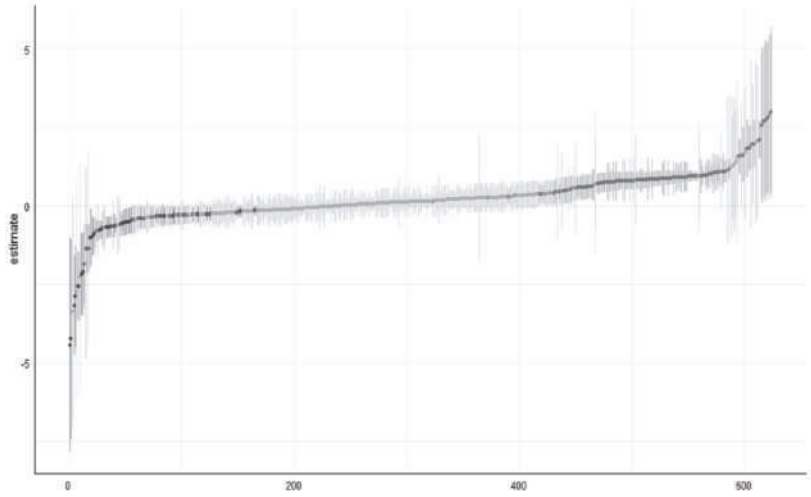


Fig. 1 Specification Curve: the black dotted estimates have *p*-value < .05.

Table 3 Multiversal statistics of regressors.

Regressor	Median *p*-value	*p*-values < 0.01 (%)
Manifacturies	0	85.41
Over 65+ years	0	79.1
Aereoportual Passengers	0	47.9
Drug Stores	.011	47.9
Primary Sector	.019	45.8
High blood pressure	.053	0
Services	.123	39.5
Population Density Index	.081	41.7
Obesity	.172	0
Hypercholesterolaemia	.185	0
Level of NO2	.213	0
Care Capacity Index	.226	4.2
Smokers	.410	0
Level of PM10	.417	4.2

A regression analysis on only 21 units is a limitation, which is partially overcome presenting 4 values (one for each subset) *per* dependent variable.

Results can lead towards substantial insights, overcoming issues of risk mentioned in Section 1. For example, for future research may be interesting to iterate the multiverse on the dimensions of *manifacturies* and *age*: collecting new data, these concepts can further be specified with more empirical indicators (i.e., revenues and n. of employees) with the aim at testing the robustness of the underlying regressive dynamics.

References

1. Beck U.: Risk Society: Towards a New Modernity. SAGE Publications Ltd (1992).
2. Beijers, L., Hanna, M., van Loo, H.M., Romeijn, J.W., Lamers, F. Schoevers, R.A., Wardenaar, K.J.: Investigating Data-Driven Biological Subtypes of Sychiatric Disorders Using Specification-Curve Analysis. Psychol. Med. (2020), doi:10.1017/S0033291720002846.
3. Chang, S., Pierson, E., Koh, P.W., Gerardin, J., Redbird, B., Grusky, D., Leskovec, J.: Mobility network models of COVID-19 explain inequities and inform reopening. Nature **589**, 82–87 (2021), doi.org/10.1038/s41586-020-2923-3.
4. Gelman, A, Loken, E.: Ethics and Statistics: The AAA Tranche of Subprime Science. Chance **27**, 51–56 (2014).
5. Giddens A.: Risk and Responsibility. Mod. Law Rev. **62**(1), 1–10 (1999).
6. Halsey, L.G., Curran-Everett, D., Vowler, S.L., Drummond, G.B.: The fickle P value generates irreproducible results. Nat. Methods **12**, 179–185 (2015).
7. Hansen, L.P., Marinacci, M: Ambiguity Aversion and Model Misspecification: An Economic Perspective. Stat. Sci. **31**(4), 511–515 (2016).
8. Masur, P., Scharkow, M.: SpecR: Statistical functions for conducting specification curve analyses (2020). https://cran.r-project.org/package=specr.
9. Muñoz, J., Young, C.: We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness, Sociol. Methodol. **48**(1), 1–33 (2018).
10. Parohan, M., Yaghoubi, S., Seraji, A., Javanbakht, M. H., Sarraf, P., Djalali, M. (2020): Risk factors for mortality in patients with Coronavirus disease 2019 (COVID-19) infection: a systematic review and meta-analysis of observational studies. Aging Male (2020). doi:10.1080/13685538.2020.1774748.
11. Patel, C.J., Burford, B., Ioannidis, J.P.: Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J. Clin. Epidemiol. **68**(9), 1046–1058 (2015).
12. Pluchino A., Biondo, A.E., Giuffrida N., Inturri, G., Latora V., Le Moli R., Rapisarda A., Russo G., Zappalà, C.: A Novel Methodology for Epidemic Risk Assessment: the case of COVID-19 outbreak in Italy (2020). arXiv:2004.02739.
13. Rohrer, J.M., Egloff, B., Schmukle, S.C.: Probing Birth-Order Effects on Narrow Traits Using Specification-Curve Analysis. Psychol. Sci. **28**(12), 1821–1832 (2017).
14. Sala-i-martin, X.X.: I Just Ran Two Million Regression, Am. Econ. Rev. **87**, 178–183 (1997).
15. Schmeiser, H., Siegel, C., Wagner, J: The risk of model misspecification and its impact on solvency measurement in the insurance sector, J. Risk Finance **13**(4), 285–308 (2012).
16. Seitshiro, M.B., Mashele, H.P.: Quantification of model risk that is caused by model misspecification. J. Appl. Stat. (2020). doi:10.1080/02664763.2020.1849055.
17. Simonsohn, U. Simmons, J.P., Nelson, L.D.: Specification Curve: Specification curve analysis. Nat. Hum. Behav. **4**(11), 1208–1214 (2020).
18. Steegen, C., Tuerlinckx, F., Gelman, A., Vanpaemel, W: Increasing Transparency Through a Multiverse Analysis. Perspect. Psychol. Sci. **11**(5), 702–712 (2016).
19. Zinn, J.O., Social Theories of Risk and Uncertainty: An Introduction. Blackwell (2008).

Semiparametric Variational Inference for Bayesian Quantile Regression

Inferenza Variazionale Semiparametrica per Regressione Quantilica Bayesiana

Cristian Castiglione and Mauro Bernardi

Abstract Variational approximations are promising methods for fast and efficient Bayesian inference and represent a valid alternative to the computational demanding simulation-based algorithms such as Markov chain Monte Carlo. Though, mean field variational Bayes requires strong assumptions on the local conjugacy structure of the posterior distribution, which can often be achieved through a data augmentation strategy, as in case of Bayesian quantile regression. This approach do not scale well in high dimension, since the number of observations and the number of parameters grow in parallel. Here we present an alternative semiparametric variational Bayes approach to approximate the posterior distribution in mixed quantile regression models that does not rely on expensive data augmentation techninques.
Abstract *I metodi di inferenza variazionale costituiscono una valida alternativa ad algoritmi Markov chain Monte Carlo per stimare modelli Bayesiani. L'assunzione chiave per l'implementazione di un algoritmo "mean field variational Bayes" è che le distributioni a priori e la verosimiglianza siano almeno localmente coniugate, cosa che talvolta può essere garantita introducendo variabili stocastiche ausiliarie nel modello, come nel caso di regressione quantilica Bayesiana. Questo approccio si rivela di diffcile applicazione quando il numero di osservazioni è molto elevato, poiché ciò comporta parallelamente un aumento dei parametri da stimare. In questo articolo introduciamo un algoritmo variazionale semiparametrico per la stima di modelli quantilici gerarchici indipendente da strategie di "data augmentation".*

Key words: Quantile regression, Mean field variational Bayes, Semiparametric variational Bayes.

Cristian Castiglione, Mauro Bernardi
Department of Statistical Sciences, University of Padova, Italy,
e-mail: cristian.castiglione@studenti.unipd.it,
e-mail: mauro.bernardi@unipd.it

1 Introduction

Quantile regression is a flexible distribution-free tool introduced by [5] to estimate the conditional quatiles of a response variable given a set of covariates. The contributions of [14] and [6] made possible to deal with quantile regression even in the Bayesian framework, exploiting the conditional Gaussian representation of the Laplace distribution. Stochastic data augmentation technique for the Laplace distribution permits to implement efficient Gibbs sampling [6] and mean field variational Bayes (MFVB) algorithms [12, 7], but constrains the dimension of the parametric space to be at least equal to the number of observed data.

Variational message passing [13] and semiparametric MFVB [9] give an alternative implementation of the mean field principle that applies even when the local posterior conjugacy of the Bayesian factor graph is not satisfied. In particular, the updating scheme for approximate Gaussian variational Bayes pointed out by [11] provides an effective machinery to implement variational inference with closed form updates in all those situations in which an analytic expression for the lower bound to the marginal log-likelihood is available, that is very common, for example, in generalized linear mixed model learning.

Our aim in this article is to present a semiparametric variational Bayes (SVB) scheme to make approximate posterior inference for a Bayesian mixed quantile regression model avoiding data augmentation strategies.

2 Model

Consider the Bayesian mixed quantile regression model specified as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon, \quad \varepsilon \sim \mathrm{AL}_n(\tau\mathbf{1}_n, \mathbf{0}_n, \sigma_\varepsilon^2\mathbf{I}_n), \tag{1}$$

where $\mathbf{y}$ is a $n \times 1$ vector of continuous response variables, β is a $p \times 1$ vector of fixed effects, $\mathbf{u}$ is a $d \times 1$ vector of random effects, $\mathbf{X}$ and $\mathbf{Z}$ are the corresponding design matrices of dimension $n \times p$ and $n \times d$. The $n \times 1$ vector of random errors ε has an asymmetric-Laplace (AL) distribution with shape parameter $\tau \in (0,1)$, which is the desired quantile level that we want to estimate, location parameter centered in 0 and scale parameter σ_ε^2. The notation $\mathbf{0}_n$, $\mathbf{1}_n$, $\mathbf{I}_n$ denote a $n \times 1$ vector of zeros, a $n \times 1$ vector of ones and the $n \times n$ identity matrix, respectively. In the following we will also denote by $\mathbf{O}_{n\times m}$ a $n \times m$ matrix of zeros. Further, we assume multivariate Gaussian (N) prior distributions for β and $\mathbf{u}$, being

$$\beta \sim \mathrm{N}_p(\mu_\beta, \Sigma_\beta), \quad \mathbf{u} \sim \mathrm{N}_d(\mathbf{0}_d, \sigma_u^2\mathbf{R}), \tag{2}$$

and conjugate inverse-Gamma (IG) prior distributions for the scale parameters σ_ε^2 and σ_u^2:

$$\sigma_\varepsilon^2 \sim \mathrm{IG}(A_\varepsilon, B_\varepsilon), \quad \sigma_u^2 \sim \mathrm{IG}(A_u, B_u). \tag{3}$$

The constants $\mu_\beta \in \mathbb{R}^p$, $\Sigma_\beta \in \mathbb{S}^p_{++}$, $\mathbf{R} \in \mathbb{S}^d_{++}$, $A_\varepsilon > 0$, $B_\varepsilon > 0$, $A_u > 0$, $B_u > 0$ are user-specified prior parameters. We adopt the notation $\mathbb{S}^n_{++}$ to indicate the space of $n \times n$ symmetric positive definite real matrices. Hereafter we will denote with $\mathbf{C}$ the column stacked design matrix $\mathbf{C} \equiv [\mathbf{X}, \mathbf{Z}]$, and with $\mathbf{c}_i^\top$ its i-th row, for $i = 1, \ldots, n$.

The local dependence structure induced by model (1) and its priors (2) and (3) is described in Figure 1.

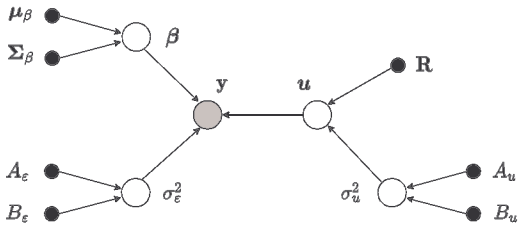


Fig. 1 Directed acyclic graph representation of model (1)–(3). The empty knots indicate the latent variables to be estimated, the shaded knot indicates the observed data, the black knots indicate the user-specified prior parameters.

3 Variational Inference

The main intuition behind the variational Bayes principle is to perform approximate inference on the parameter vector $\theta \in \Theta$ by substituting an intractable posterior distribution $p(\theta \mid \mathbf{y}) = p(\theta)p(\mathbf{y} \mid \theta)/p(\mathbf{y})$ with a simpler density function $q(\theta)$ assumed to belong to a specific space of functions $\mathscr{Q}$. The optimal q-distribution, $q^\star(\theta) \in \mathscr{Q}$, is then chosen so that it is the maximizer of the variational problem

$$\max_{q \in \mathscr{Q}} \int_\Theta q(\theta) \log \frac{p(\mathbf{y}, \theta)}{q(\theta)} d\theta. \tag{4}$$

The right hand side integral in equation (4) is called the evidence lower bound, $\log \underline{p}(\mathbf{y}; q)$, which corresponds to $\log \underline{p}(\mathbf{y}; q) = \log p(\mathbf{y}) - \mathrm{KL}(q \| p)$, where $\mathrm{KL}(q \mid p)$ is the Kullback-Leibler divergence between $q(\theta)$ and $p(\theta \mid \mathbf{y})$. In particular, under the mean field restriction, i.e. $\mathscr{Q} = \{q : q(\theta) = \prod_{j=1}^m q(\theta_j)\}$, where $(\theta_1, \ldots, \theta_m)$ is a partition of θ, the variational optimization problem in equation (4) have an exact coordinate-wise solution [8, 1], that is $q^\star(\theta_j) \propto \exp\left\{\mathrm{E}_{-q(\theta_j)}[\log p(\theta_j \mid \text{rest})]\right\}$. Here $\mathrm{E}_{-q(\theta_j)}(\cdot)$ denotes the expected value calculated with respect to $\prod_{k \neq j} q(\theta_k)$, $p(\theta_j \mid \text{rest})$ is the full-conditional distribution of θ_j and rest is the set of data and model parameters excluding θ_j. Under local conjugacy, $q^\star(\theta_j)$ has a closed form expression belonging to the exponential family. Hereafter we will denote with $\mu_{q(f(\theta_j))} \equiv \mathrm{E}_{q(\theta_j)}[f(\theta_j)]$ the expected value of $f(\theta_j)$ calculated with respect to $q(\theta_j)$, and, similarly, $\Sigma_{q(\theta_j)} \equiv \mathrm{Var}_{q(\theta_j)}(\theta_j)$ will be the variance-covariance matrix of θ_j calculated with respect to $q(\theta_j)$. Now, referring to model (1)–(3), consider the

product restriction

$$p(\beta, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2 \mid \mathbf{y}) \approx q(\beta, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2) = q(\beta, \mathbf{u}) q(\sigma_u^2) q(\sigma_\varepsilon^2). \tag{5}$$

Then, the optimal q-densities are

- $q^\star(\beta, \mathbf{u}) \propto \exp\Big\{ -\frac{1}{2} \Big(\mu_{q(\beta,u)} - \begin{bmatrix} \mu_\beta \\ \mathbf{0}_d \end{bmatrix} \Big)^\top \begin{bmatrix} \Sigma_\beta^{-1} & \mathbf{O}_{p\times d} \\ \mathbf{O}_{d\times p} & \mu_{q(1/\sigma_u^2)} \mathbf{R}^{-1} \end{bmatrix} \Big(\mu_{q(\beta,u)} - \begin{bmatrix} \mu_\beta \\ \mathbf{0}_d \end{bmatrix} \Big)$
  $- \frac{1}{2} \mathrm{trace} \Big(\begin{bmatrix} \Sigma_\beta^{-1} & \mathbf{O}_{p\times d} \\ \mathbf{O}_{d\times p} & \mu_{q(1/\sigma_u^2)} \mathbf{R}^{-1} \end{bmatrix} \Sigma_{q(\beta,u)} \Big) - \mu_{q(1/\sigma_\varepsilon^2)} \sum_{i=1}^n \mathrm{E}_{q(\beta,u)}[\rho_\tau(\varepsilon_i)] \Big\}$,
  where $\rho_\tau(\varepsilon)$ is the quantile check function, defined as $\rho_\tau(\varepsilon) = \varepsilon[\tau - \mathbb{I}_{(-\infty,0]}(\varepsilon)]$ and $\mathbb{I}_{(a,b]}(\varepsilon)$ is the indicator function equal to 1 if $\varepsilon \in (a,b]$, an 0 otherwise;
- $q^\star(\sigma_u^2) = \mathrm{IG}(A_u + \frac{d}{2}, B_{q(\sigma_u^2)})$, with $B_{q(\sigma_u^2)} = B_u + \frac{1}{2} \mathrm{trace}\big[\mathbf{R}^{-1}(\Sigma_{q(u)} + \mu_{q(u)} \mu_{q(u)}^\top)\big]$;
- $q^\star(\sigma_\varepsilon^2) = \mathrm{IG}(A_\varepsilon + \frac{3}{2} n, B_{q(\sigma_\varepsilon^2)})$, with $B_{q(\sigma_\varepsilon^2)} = B_\varepsilon + \sum_{i=1}^n \mathrm{E}_{q(\beta,u)}[\rho_\tau(\varepsilon_i)]$.

Even though the Gaussian prior for $(\beta, \mathbf{u})$ is not conjugate with the likelihood, it is possible to find an analytic expression for $\mathrm{E}_{q(\beta,u)}[\rho_\tau(\varepsilon_i)]$, obtaining so a closed form solution for the whole set of approximating distributions. Derivations of these results and the explicit expression for $\mathrm{E}_{q(\beta,u)}[\rho_\tau(\varepsilon_i)]$ are not provided here for brevity.

Notice that $q^\star(\beta, \mathbf{u})$ is not a standard distribution and in practice it is difficult to derive its properties. A very common solution is to project $q^\star(\beta, \mathbf{u})$ onto the space of exponential family distributions using a non-conjugate variational message passing strategy [13]. In particular, we will choose a multivariate Gaussian projection, such that $q^\star(\beta, \mathbf{u}) \approx q^\star(\beta, \mathbf{u} \mid \mu_{q(\beta,u)}, \Sigma_{q(\beta,u)}) = \mathrm{N}_{p+d}(\mu_{q(\beta,u)}, \Sigma_{q(\beta,u)})$, leveraging the results of [11] and [9], which proved that the optimal updating rule for $\mu_{q(\beta,u)}$ and $\Sigma_{q(\beta,u)}$ is given by

$$\Sigma_{q(\beta,u)} \leftarrow \big\{ -\mathsf{H}_\mu S(\mu, \Sigma) \big\}^{-1}_{\mu=\mu_{q(\beta,u)}, \Sigma=\Sigma_{q(\beta,u)}} \tag{6}$$

$$\mu_{q(\beta,u)} \leftarrow \mu_{q(\beta,u)} + \Sigma_{q(\beta,u)} \big\{ \mathsf{D}_\mu S(\mu, \Sigma) \big\}_{\mu=\mu_{q(\beta,u)}, \Sigma=\Sigma_{q(\beta,u)}}, \tag{7}$$

where $S(\mu_{q(\beta,u)}, \Sigma_{q(\beta,u)}) \equiv \mathrm{E}_q[\log p(\beta, \mathbf{u} \mid \text{rest})]$ is the posterior contribution of β and $\mathbf{u}$ to the evidence lower bound. Here D_μ and H_μ are, respectively, the gradient and Hessian operators calculated with respect to μ, while $\leftarrow$ is the assignment operator.

4 Algorithm

The iterative update of $q^\star(\beta, \mathbf{u})$, $q^\star(\sigma_u^2)$ and $q^\star(\sigma_\varepsilon^2)$ gives rise to a coordinate ascent scheme summarized in Algorithm 1, which converge to the optimal set of q-distributions. The convergence of the algorithm is assessed by monitoring the relative change of parameters and lower bound, then the execution ends when both fall below a given threshold, that we fix equal to 10^{-5}.

In Algorithm 1, $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the probability density function and cumulative density function of a univariate Gaussian random variable, $\text{diag}(\cdot)$ is a column vector equal to the main diagonal of its argument, $\text{Diag}(\cdot)$ is a diagonal matrix, whose diagonal is equal to its argument, $\odot$ is the Hadamard element-wise product between two arrays.

Algorithm 1: SVB algorithm for approximate inference in model (1)–(3)

Data: $\mathbf{y}, \mathbf{X}$

Input: $\mu_\beta, \Sigma_\beta, \mathbf{R}, A_u, B_u, A_\varepsilon, B_\varepsilon$

Output: $\mu_{q(\beta,u)}, \Sigma_{q(\beta,u)}, B_{q(\sigma_u^2)}, B_{q(\sigma_\varepsilon^2)}$

while *convergence is not reached* **do**

$$B_{q(\sigma_u^2)} \leftarrow B_u + \tfrac{1}{2}\left\{\mu_{q(u)}^\top \mathbf{R}^{-1}\mu_{q(u)} + \text{trace}[\mathbf{R}^{-1}\Sigma_{q(u)}]\right\};$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \textstyle\sum_{i=1}^n [\mathbf{c}_i^\top \Sigma_{q(\beta,u)}\mathbf{c}_i]^{1/2}\left\{\phi(z_i) + [\tau - 1 + \Phi(z_i)]z_i\right\};$$

$$\mu_{q(1/\sigma_u^2)} \leftarrow (A_u + \tfrac{d}{2})/B_{q(\sigma_u^2)}; \quad \mu_{q(1/\sigma_\varepsilon^2)} \leftarrow (A_\varepsilon + \tfrac{3}{2}n)/B_{q(\sigma_\varepsilon^2)};$$

$$\mathbf{s} \leftarrow \text{diag}(\mathbf{C}\Sigma_{q(\beta,u)}\mathbf{C}^\top)^{-1/2}; \quad \mathbf{z} \leftarrow (\mathbf{y} - \mathbf{C}\mu_{q(\beta,u)}) \odot \mathbf{s};$$

$$\Sigma_{q(\beta,u)} \leftarrow \left\{ \begin{bmatrix} \Sigma_\beta^{-1} & \mathbf{O}_{p\times d} \\ \mathbf{O}_{d\times p} & \mu_{q(1/\sigma_u^2)}\mathbf{R}^{-1} \end{bmatrix} + \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}\,\text{Diag}[\phi(\mathbf{z}) \odot \mathbf{s}]\,\mathbf{C}^\top \right\}^{-1};$$

$$\mu_{q(\beta,u)} \leftarrow \mu_{q(\beta,u)} + \Sigma_{q(\beta,u)}\left\{ \begin{bmatrix} \Sigma_\beta^{-1}\mu_\beta \\ \mathbf{0}_d \end{bmatrix} + \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top[\tau - 1 + \Phi(\mathbf{z})] \right\};$$

5 Application

In this section we present a real data application where the mixed quantile regression model specified in equations (1)–(3) is applied to the Italian general electricity market prices [3]. It is well known that electricity prices are highly dominated by multiple seasonal cycles of different lengths like daily, weekly and monthly patterns [10], and by the presence of heteroskedasticity [4]. The mixed quantile regression model specified in equations (1)–(3) provides a flexible approach for modelling the τ-th conditional quantile as a function of the seasonal cycles by means of B-splines regression [2]. The mixed quantile regression model has been fitted to the data on daily average electricity prices for the Italian market over the period from January 1, 2013 to December 31, 2020. The data and the fitted quantiles are represented in Figure 2.

6 Conclusions

We proposed a coordinate ascent algorithm for fitting Bayesian mixed quantile regression models within the variational approximation framework. Our method does not require data augmentation strategies to enforce the conjugacy between prior dis-

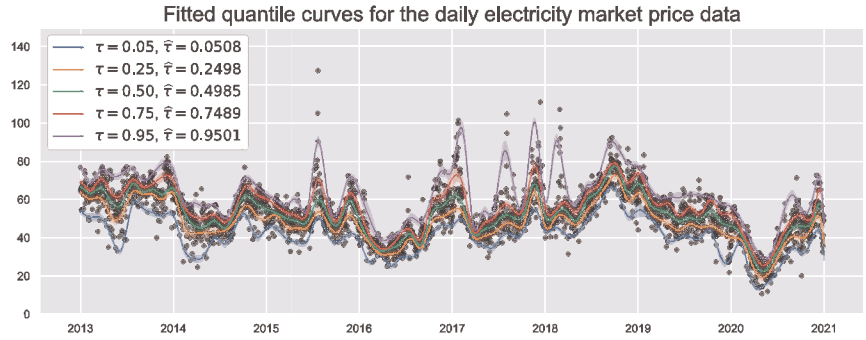


Fig. 2 Daily electricity prices for the Italian market and fitted quantiles for the period from January 1, 2013 to December 31, 2020. The legend provides the theoretical quantile level τ associated to each curve and gives also the corresponding estimated empirical quantile level, i.e. $\widehat{\tau}$.

tribution and likelihood, and this feature permits to apply our algorithm even in high-dimensional contexts. Future research directions include, but are not limited to, generalizations for online learning and dynamical quantile models.

References

1. Blei, M. D., Kucukelbir, A. and McAuliffe, J. D.: Variational Inference: A Review for Statisticians. J. Am. Stat. Assoc. **112**, 859-877 (2017)
2. de Boor, C.: A Practical Guide to Splines. Springer-Verlag, New York (2001)
3. Gestore Mercati Energetici: Dati storici MGP. `https://www.mercatoelettrico.org/It/download/DatiStorici.aspx`
4. Karakatsani, N. V., Bunn, D. W.: Forecasting electricity prices: The impact of fundamentals and time-varying coefficients. Int. J. Forecast. **24**, 764–785 (2004)
5. Koenker, R., Bassett, G. Jr.: Regression quantiles. Econometrica. **46**, 33–50 (1978)
6. Kozumi, H. and Kobayashi, G.: Gibbs sampling methods for Bayesian quantile regression. J. Stat. Comput. Simul. **81**, 1565–1578 (2011)
7. McLean, M. W., Wand, M. P.: Variational message passing for elaborate response regression models. Bayesian Anal. **14**, 371–398 (2019)
8. Ormerod, J. T., Wand, M. P.: Explaining variational approximations. Am. Stat. **64**, 140153 (2010)
9. Rohde, D, Wand, M. P.: Semiparametric mean field variational Bayes: general principles and numerical issues. J. Mach. Learn. Res. **17**, 5975-6021 (2016)
10. Taylor, J. W., Snyder, R. D.: Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. Omega. **40**, 748–757 (2012)
11. Wand M. P.: Fully simplified multivariate normal updates in non-conjugate variational message passing. J. Mach. Learn. Res. **15**, 1351–1369 (2014)
12. Wand, M. P., Ormerod, J. T., Padoan, S. A., Frührwirth, R.: Mean field variational Bayes for elaborate distributions. Bayesian Anal. **6**, 847–900 (2011)
13. Winn, J., Bishop, C. M.: Variational message passing. J. Mach. Learn. Res. **6**, 661–694 (2005)
14. Yu, K. and Moyeed, R. A.: Bayesian quantile regression. Stat. Probab. Lett. **54**, 437–447 (2001)

Searching for a source of difference in undirected graphical models for count data - an empirical study

Ricerca di una sorgente per la differenza di modelli grafici adirezionati con dati di conteggio: uno studio empirico

Federico Agostinis, Monica Chiogna, Vera Djordjilović, Luna Pianesi, Chiara Romualdi

Abstract A study is presented for exploring the possibility of applying the source set approach ([3]), developed under the assumption of normality, to count data, after data transformation. Some explanations about the source set approach, data transformations and the simulation setting are provided. The suggestion is given that the deviance-based or quantile randomized residuals could provide a better basis for data transformation when coupled with source set analysis, along with standard trasformations such as log transformation or square root transformation.

Abstract *Si presenta uno studio volto ad esplorare la possibilità che l'approccio source set ([3]), sviluppato sotto l'assunzione di normalità, possa essere applicato al caso di modelli grafici adirezionati per dati di conteggio, dopo opportuna trasformazione dei dati. Si forniscono alcuni dettagli sull'approccio source set, sulle trasformazioni dei dati, sull'impianto di simulazione. Emerge il suggerimento che i residui di devianza, oltre che trasformazioni tradizionali come la trasformazione logaritmica e radice quadrata, producano risultati migliori quando usati all'interno dell'approccio source set.*

Key words: Undirected graphical models, Source set, count data, negative binomial distribution, RNA-Seq data.

Federico Agostinis
Department of Biology, University of Padova, Italy, e-mail: federico.agostinis@studenti.unipd.it

Monica Chiogna
Department of Statistical Sciences, University of Bologna, Italy, e-mail: monica.chiogna2@unibo.it

Vera Djordjilović
Department of Economics, Ca' Foscari University, Italy, e-mail: vera.djordjilovic@unive.it

Luna Pianesi
Department of Information Engineering and Computer Science, University of Trento, Italy e-mail: luna.pianesi@studenti.unitn.it

Chiara Romualdi
Department of Biology, University of Padova, Italy e-mail: chiara.romualdi@unipd.it

1 Background

In many genomics studies, the expression of the set of genes is measured in two conditions, and the main objective is to identify all genes showing differing behaviour between two conditions. Given the interconnectedness between genes, it is useful to go beyond this first level differential analysis, and to try to distinguish the site of original perturbation – the so-called primary dysregulation – from the elements affected by perturbation through dysregulation propagation, i.e. secondary dysregulation. A subset of authors has recently proposed a novel statistical approach called SourceSet [8] that aims to identify the source of primary dysregulation on the basis of observations from the control and perturbed condition.

SourceSet models data from two experimental conditions under study, for instance measurements of gene expression in patients affected by a certain disease and in healthy controls, as realizations of two Gaussian graphical models sharing the same graphical structure. More formally, we assume that the data from the i-th condition, where $i = 1,2$, arise as independent draws from a p-dimensional normal distribution $\mathsf{N}(\mu^{(i)}, \Sigma^{(i)})$, with $\mu^{(i)} \in R^p$ and Σ a $p \times p$ positive definite matrix such that the zeros in the concentration matrix $\Omega^{(i)} = (\Sigma^{(i)})^{-1}$ correspond to the missing edges in a given undirected graph $G = (V,E)$. Here, each node of the set of nodes $V = \{1,\ldots,p\}$ is associated to a single variable, i.e. a gene under study, and $E \subset V \times V$ is a set of gene-gene edges obtained from pathway topology conversion.

If we denote by $X^{(i)}$ a random vector distributed according to $\mathsf{N}(\mu^{(i)}, \Sigma^{(i)})$, then the novel entity termed *source set* introduced in [8] is a set $D \subset V$ such that the distributions of $X_D^{(1)}$ and $X_D^{(2)}$ differ, but conditional distributions of $X_{V\setminus D}^{(i)}$ given $X_D^{(i)}$ coincide for $i = 1,2$. Here, X_A denotes a subvector of X induced by a subset $A \subset V$. In words, D contains genes in V that have different marginal distributions in the two conditions, but, conditionally on their realization, the distribution of the remaining genes is unaltered. Thus, assuming no confounding factors, genes in D may be considered the starting point of the dysregulation process, while elements in $V \setminus D$, if affected by the dysregulation, are affected through the process of network propagation.

The problem of estimating D from data has been addressed by means of an efficient procedure based on exploiting the modular structure of Gaussian graphical models; we refer the interested reader to [8] and [3] for a detailed exposition. The proposed estimating procedure has been implemented in the `SourceSet R` package, where the input consists of a matrix of data from two conditions and an underlying graphical structure. The output is an estimate of the source set, also termed the primary set, and a secondary set composed of nodes affected by dysregulation in the process of network propagation. See package vignette for some examples and [4] for a description of the package.

SourceSet approach is based on the assumption that the data from the two experimental conditions follow a multivariate normal distribution. While this assumption has become fairly standard in the analysis of microarray experiments, it is not suitable in the context of count data coming from RNA-Seq experiments. RNA-seq

technology is a type of next generation sequencing technology for estimating the expression level of genes in whole-genome scale studies and has become the standard technology for the study of genomics. RNA-Seq data are usually represented as a matrix of counts, with genes in rows and biological samples in columns. In this work, we investigate, via simulations, the possibility of applying SourceSet to count data, after a proper trasformation. We consider some standard transformations, such as the logarithm or the square root, and transformations resulting as side effects of modelling of the data through generalized linear models.

2 The proposal

We are concerned with finding a transformation which will justify, in practice, the use of the source set approach outside normality. Numerous transformations for using discrete data in Gaussian settings have been examined in the literature, some of which are considered here. Moreover, we adopt a regression perspective and consider residuals from generalized linear modelling. Residuals measure discrepancy between a statistical model and observed data and are typically used to evaluate the quality of the model either through summary statistics or directly, e.g., via plots. Here, the aim is to consider residuals that are approximately normally distributed, to be used as input of the source set analysis. We first specify a null model of constant gene expression in the two conditions, assuming a valid distribution for the counts at hand. Next, we fit the models and compute appropriately chosen residuals. Typically, when dysregulation is present, the differences remain stored in the residuals from the null model and should therefore be captured in a downstream analysis. The residuals are thus interpreted as the z-scores to be used in the following source set analysis. The use of residuals enables a fast transformation to normality, allowing also for potential adjustment for other covariates of interest, if present.

3 The experiment

3.1 Modelling RNA-Seq measurements

The most obvious choice for modelling count data is a Poisson distribution. However, RNA-Seq data exhibit significant overdispersion, rendering the Poisson model inadequate. Among many alternatives proposed over the years, the negative binomial distribution – lending itself to a straightforward biological interpretation – stands out. In this section, we briefly review a negative binomial model for modelling gene expression X of a single gene.

To allow for overdispersion, the Poisson distribution model can be modified as

$$X \mid \lambda, \varepsilon \sim Poi(\theta), \text{ with } \theta = \lambda\varepsilon$$

where $\varepsilon \sim \text{Gamma}(\alpha, \alpha)$, for $\alpha > 0$, is a nonnegative multiplicative random-effect term to model individual heterogeneity (biological variability). Marginally, we thus have a negative binomial distribution, $X \sim NB(\alpha, p)$, parameterized by a probability parameter $p = \frac{\lambda}{\lambda+\alpha}$ and dispersion parameter α. For any $\alpha \geq 0$ and $p > 0$, the probability mass function of the negative binomial distribution is

$$f_{NB}(x; \alpha, p) = \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)} (1-p)^{\alpha} p^{x}, \quad \forall x \in N.$$

It is

$$\text{E}[X] = \lambda, \quad \text{Var}[X] = \lambda + \phi\lambda^2,$$

where $\phi = 1/\alpha$ is the dispersion parameter. In this way, the model entails both the biological variation (Gamma distribution) and the technical variation due to the sequencing process (Poisson distribution).

3.2 Simulating observations from a negative binomial graphical model

In the simulation study here reported, we will use a small undirected graph consisting of five nodes shown in Figure 1. The difference between two conditions that we will aim to uncover with the method of Source set, is the edge between nodes A and B present in condition 1 and absent in condition 2. This perturbation of the model leads to a source set $D = \{A, B\}$

While generating observations from marginal gene-wise models described in the previous section is straightforward, generating data from a (multivariate) negative binomial graphical model is far from trivial. We adapt the procedure proposed for simulating data from a Poisson graphical model in [6] that we briefly describe here. Let $X \in R^{n \times p}$ be the data matrix, i.e., the set of n independent observations of random vector $\mathbb{X} \in R^p$ in a single condition. In our example, $\mathbb{X} = (A, B, C, D, E)^T$, $p = 5$ and $n = 400$. Then, X is obtained from the model $X = YW + \varepsilon$, where $Y = (y_{st})$ is an $n \times p$ matrix whose entries y_{st} are realizations of independent random variables $Y_{st} \sim Poi(\lambda_{true})$ and $\varepsilon = (e_{st})$ is the noise, i.e., an $n \times p$ matrix with entries e_{st} which are realizations of random variables $E_{st} \sim Poi(\lambda_{noise})$. Matrix W is the adjacency matrix of the given true graph shown in Figure 1. To simulate observations from a negative binomial model, instead of sampling each element from the same Poisson distribution with mean λ_{true}, we sample the elements of the i-th row from a Poisson distribution with mean $\lambda_{true}\varepsilon_i$, where $\varepsilon_i \sim \text{Gamma}(\alpha, \alpha)$, for $i = 1, \ldots, n$. In other words, we first generate a sample of size n from $\text{Gamma}(\alpha, \alpha)$, and then proceed by sampling the elements of $\mathbb{Y}$ from row specific Poisson distributions.

To simulate data from a second condition, we modify the matrix W by removing an edge between nodes A and B and then proceed as above. As in [1], we simulate data at two levels of signal-to-noise ratio (SNR). We set $\lambda_{true} = 1$ and $\alpha = 0.1$ with $\lambda_{noise} = 0.1$ for the high SNR level, and $\lambda_{noise} = 0.5$ for the low SNR level.

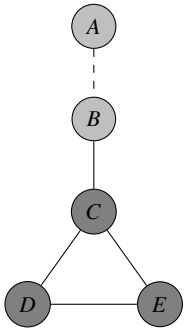


Fig. 1 Undirected graph used in simulations studies. The edge between nodes *A* and *B* is present in condition 1 and absent in condition 2.

4 Results

We have applied SourseSet procedure implemented in the Sourceset package to investigate its performance in this context. We considered a number of transformations of the original count data generated by the negative binomial graphical model:

1. raw data: untransformed data;
2. square root;
3. logarithm;
4. Anscombe residuals ([2]);
5. deviance residuals;
6. Pearson residuals;
7. random quantile residuals ([5]).

To obtain residuals 4-7, data matrices from two conditions were merged and a null negative binomial model assuming constant gene expression across conditions was estimated for each gene. Residuals from these models are then continuous, and in general follow a more symmetric distribution with respect to the original count data.

In regard to using source set when count data are available, not all transformations of the data achieve the same results. From Table 1, it appears that standard transformations, such as the logarithm or the square root transformation, produce good results, competitive with the best trasformations based on residuals from generalized linear modelling. Pearson residuals always perform poorly. This might be explained by the fact that, for enumerative data, the distribution of the Pearson statistic is often much more nearly chi squared than is that of the deviance (e.g., see [7]). Assuming that our conclusions from this first simulation study are valid, then further work with different distributions and in more general settings would be very useful for practitioner statisticians.

Table 1 Simulation study results: the average number of times (out of 100) the source set $D = \{A, B\}$ was correctly identified (columns 1 and 3) and the average number of times the estimated source set strictly covered the true source set (columns 2 and 4).

	$\lambda_{noise} = 0.1$		$\lambda_{noise} = 0.5$	
	$\hat{D} = D$	$\hat{D} \supset D$	$\hat{D} = D$	$\hat{D} \supset D$
Raw	42.30	0.94	39.26	0.44
Square root	91.84	5.38	92.30	4.10
Log	91.02	7.02	93.16	4.64
Anscombe	86.60	2.88	83.36	1.94
Deviance	91.76	6.30	93.30	4.48
Pearson	42.46	0.72	39.40	0.34
RQR	93.82	3.28	88.82	2.38

References

1. Allen, G.I., Liu Z. (2013) A Local Poisson Graphical Model for inferring networks from sequencing data. IEEE Trans Nanobioscience. **12**(3):189-98.
2. Anscombe, F. J. (1961). Examination of residuals. Technical report, Princeton University, Princeton, United States.
3. Djordjilović V, Chiogna M. (2018). Searching for a source of difference in Gaussian graphical models. arXiv preprint arXiv:1811.02503.
4. Djordjilović V., Chiogna M., Romualdi C., Salviato E. (2020). Searching for the Source of Difference: A Graphical Model Approach. In: Raposo M., Ribeiro P., Sério S., Staiano A., Ciaramella A. (eds) Computational Intelligence Methods for Bioinformatics and Biostatistics. CIBB 2018. Lecture Notes in Computer Science, vol 11925. Springer, Cham.
5. Dunn, P., & Smyth, G. (1996). Randomized Quantile Residuals. J Comput Graph Stat, **5**(3), 236-244.
6. Hue Nguyen, T. K. and Chiogna, M. (2021). Structure learning of undirected graphical models for count data. J Mach Learn Res. In press.
7. Larntz, K. (1978). Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics. J Am Stat Assoc. **73**, 253-263.
8. Salviato E, Djordjilović V, Chiogna M, Romualdi C. (2019). SourceSet: A graphical model approach to identify primary genes in perturbed biological pathways. PLoS Comput Biol. **15**(10):e1007357.

Snipped robust inference in mixed linear models

Inferenza robusta nei modelli lineari con effetti misti mediante snipping

Antonio Lucadamo, Luca Greco, Pietro Amenta, Anna Crisci

Abstract In this contribution, a robust approach to estimation and inference in mixed linear models for longitudinal and repeated measures data is proposed. The method relies on the idea of snipping. The model is fit after discarding some entries of the data matrix leading. The method performs simultaneous estimation and cellwise outlier detection. Standard errors for the fixed effects are obtained by parametric bootstrap. The behavior of the proposed method is investigated by a real data example.

Abstract *In questo lavoro proponiamo un metodo per la stima e l'inferenza robusta nei modelli lineari misti per dati longitudinali o misure ripetute. Il metodo si basa sull'idea dello snipping. La procedura di stima prevede che alcune osservazioni relative alla singola unitá possano essere eliminate. Gli errori standard associati alle stime degli effetti fissi sono calcolati mediante bootstrap parametrico. Il comportamento del metodo é stato verificato applicandolo ad un insieme di dati.*

Key words: Bootstrap, Mixed linear model, Snipping, Stochastic optimization

1 Introduction

Mixed linear models (MLM) [14] provide a successful approach to the analysis of longitudinal and repeated measures (or more in general clustered data). Actually, they allow modeling the linear relationship between the response and a set of covariates while taking into account the heterogeneity across subjects under study by the inclusion of random effects. The MLM state that $y_{it} = x_{it}\beta + Z_{it}u_i + \varepsilon_{it}$ with

Antonio Lucadamo, Luca Greco, Pietro Amenta
Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Italy, Piazza Arechi II, Benevento, e-mail: {luca.greco, antonio.lucadamo, amenta}@unisannio.it, Anna Crisci
University of Naples Federico II, e-mail: anna.crisci@unina.it

$i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots T_i$ (in particular we may have $T_i = T\ \forall i$) and $N = \sum_{i=1}^{n} T_i$ is the overall number of entries. In matrix notation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$. Here $\mathbf{y}$ is the response, $\boldsymbol{\beta}$ is the $p-$vector of fixed effects, $\mathbf{u}$ is the $q-$vector of random effects, $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices for the fixed and random effects, respectively, and $\boldsymbol{\epsilon}$ is the vector of independent error terms. Covariates may be time-dependent but also subject-specific. The main assumption behind this approach is that the responses are conditionally independent, given the random effects. The classical MLM formulation assumes that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I}_n)$ and $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Psi}_{\mathbf{u}})$. It follows that

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_{\varepsilon}^2 \mathbf{I}_n) \qquad \textit{and} \qquad \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_{\varepsilon}^2 \mathbf{I}_n + \mathbf{Z}^T \boldsymbol{\Psi}_{\mathbf{u}} \mathbf{Z})$$

In such a framework, parameter estimation is usually performed by (restricted) maximum likelihood and all inferences are driven by the likelihood function. Nevertheless, the fitted model can be highly influenced by the presence of outliers, unexpected anomalous values departing from model assumptions. Outliers can be of different nature. It may be that all the measurements made on the same unit exhibit anomalous patters with respect to the bulk of the data. Such data anomalies are called structural outliers but they represent a very extreme situation in longitudinal data. On the contrary, it is likely that only some observations on the same subject are separately contaminated. Only few outlying subject specific occasion-wise entries could deteriorate the classical inferential procedures. In the presence of such data inadequacies, one should avoid misleading conclusions. In other words, there is the need to implement some robust inferential procedures that allow the fitted model to be resistant against the occurrence of outliers.

An appealing approach to obtain robust parameter estimates in a MLM framework is given by the employ of bounded influence estimators stemming from weighted versions of the likelihood function. The reader is pointed to the book by [13] for a detailed account on several approaches for robust fitting of MLMs. In particular, the multivariate normal formulation of MLMs has been the starting point in [3] in order to develop robust techniques for MLMs, based on constrained S-estimators. Similarly, [4] proposed MM-estimators for the main effects parameter, whereas [11] suggested some robust proposals based on a constrained MCD estimator.

In this contribution, we would like to provide further insights on the development of robust techniques in the MLM framework. The proposed technique is meant to be robust against entry-wise outliers and is entirely built on a cell-wise contamination model. The main idea is that outlying entries have to be discarded. Here we apply the approach developed along the lines of [2] and developed by [6, 7] in robust multivariate estimation and cluster analysis. Other entrywise robust proposal have been suggested by [1] and [15]. A methodology that allows simultaneous estimation and outliers detection will be described in Section 2. Its behavior in finite samples will be illustrated through a real data example in Section 3.

2 Snipped MLM fitting

Let us assume $\lfloor N\alpha \rfloor$ entries of the data y_{it} are contaminated, therefore obtaining subject specific occasion-wise outliers. Let $\mathbf{w}$ be a binary vector of the same length of $\mathbf{y}$, such that $\sum_{it} w_{it} = \lceil N(1-\alpha) \rceil$. The element w_{it} is an indicator of y_{it} being contaminated or not. When $w_{it} = 0$, then the i^{th} data point has been contaminated in its t^{th} measurement. According to this approach, one observation is snipped when one or more of its dimensions are discarded, but at least one is retained in the analysis. Potentially all observations can be snippe, that is, the rate of genuine observation is not fixed for a given level of contamination α. In other words, snipping treats a fixed fraction α of the data entries as it they were missing (see also [5]).

When the i^{th} observation is contaminated, we assume it has been drawn from an almost arbitrary subject-specific distribution $g_i(y_i)$ in R^{T_i}. The entry-wise contamination model can be expressed as follows

$$f(y_i|u_i) = w_i \phi_{T_i}(y_i; X_i, Z_i, u_i, \beta) + (1 - w_i) g_i(y_i) \tag{1}$$

where $f(y_i|u)$ denotes the density of y_i conditionally on random effects. The reader should refer to [6, 7] and [8] for more details, even if in a different framework. Then, under the entry-wise contamination model (1)

$$y_i \sim N_{d_i}\left(\mu_i(d_i), \Sigma_i(d_i)\right)$$

where $d_i = \sum_t w_{it}$, $\mu_i(d_i) = x_i^S \beta$, $\Sigma_i(d_i) = z_i^S \Psi_u z_i^{'S}$; x_i^S and z_i^S have been obtained after removing those rows corresponding to contaminated entries in x_i and z_i respectively, where x_i is $T_i \times p$ and z_i is $T_i \times q$. Then, the dimension of $\mu_i(d_i)$ and $\Sigma_i(d_i)$ depends on the non null entries in $w_i = (w_{i1}, w_{i2}, \ldots, w_{iT_i})$.

The log-likelihood under the entry-wise contamination model is

$$\ell(\beta, \Psi_u, \sigma_\varepsilon^2) = \sum_{i=1}^{n} \log \phi_{d_i}\left(y_i; \mu_i(d_i), \Sigma_i(d_i)\right) + \sum_{i=1}^{n} \log g_i(y_i) \tag{2}$$

where, now, $g_i(y_i)$ is a density in $R^{T_i - d_i}$. In a similar fashion, one could define a restricted log-likelihood function. Under the separation condition, the log-likelihood (2) can be maximized by ignoring the contributions of those contaminated entries [10]. Some additional constraints [6] are needed in order to be able to maximize (2):

$$\sum_i w_{it} > 0, \qquad \sum_i w_{it} w_{it'} > 0 \tag{3}$$

observation for each time occasion, $\sum_i w_{it} > 0$; the number of genuine entries should allow estimation of variance components, $\sum_i w_{it} w_{it'} > 0$.

Maximization of the loglikelihood function (2) is a complex problem since it is maximized over both W and $\theta = (\beta, \sigma_\varepsilon^2, \Psi_u)$, that is we aim at finding

$$\text{argmax}_W \sup_{\theta} \sum_{i=1}^{n} \log \phi_{d_i}\left(y_i; \mu_i(d_i), \Sigma_i(d_i)\right) \ .$$

Here, the problem is tackled by using a stochastic optimization algorithm involving an acceptance-rejection scheme, set up along the lines of [2] and [6, 7]. The algorithm proceeds as follows (see also [9] for an application to robust Cox regression):

1. Initialize $W = W(0)$ by randomly discarding $\lfloor N\alpha \rfloor$ entries
2. At the s^{th} iteration:
   a. obtain the Maximum Likelihood Estimates (MLE) or the Restricted Maximum Likelihood Estimates (REML) $\hat{\theta}(s)$ based on the current set of non trimmed entries $W(s-1)$;
   b. form a new set $W(s)$ by switching one entry in $W(s-1)$ with a previously discarded entry;
   c. if $W(s)$ satisfies the conditions in (3), then obtain $\hat{\theta}^*(s)$ and accept the candidate set with probability $p(s) = \min[1, A]$ with

$$A = exp\left\{ \frac{log(s)}{D} [\ell(\hat{\theta}^*(s)) - [\ell(\hat{\theta}(s))] \right\};$$

3. stop after k_{max} iterations or when the current maximum is not updated for r_{max} iterations.

We notice that whenever the likelihood evaluated over the candidate set increases, the current parameter estimate is updated with probability one. The acceptance probability decreases with the number of iterations and is proportional to the likelihood ratio when the candidate set gives rise to a lower likelihood. The maximum number of iterations k_{max} should be set large enough. The tuning parameter D allows control of the speed of convergence and acceptance ratio. Here we set $k_{max} = 10000$, $D = 0.1N(1-\alpha)$ and r_{max}=50. The algorithm should been tuned in order to escape local maxima. Furthermore, in order to increase the chance to get the absolute maximum, the algorithm should be run from several different initial sets. Then, the solution leading to the largest likelihood is considered. The fitting process is completed by the estimation of the random effects. They are estimated on the snipped set of observations at convergence by using standard results. In order to evaluate standard errors for the regression coefficients vector β, it is natural to resort to the application of the bootstrap. It is worth noting that conditioning on the selected set on non trimmed entries would lead to underestimate uncertainty. Snipping introduces a further source of uncertainty to be taken into account. Here we propose to resort to bootstrap, new samples are generated according to the fitted model and the proposed algorithm is applied to each replication. By using parametric bootstrap, each replication is obtained as $y^*_{it} = x_{it}\hat{\beta} + z_{it}u^*_i + \varepsilon^*_{it}$ where u^*_i is sampled from a normal distribution with null mean vector and variance covariance matrix $\hat{\Psi}_u$ and ε^*_{it} from a $N(0, \hat{\sigma}^2_\varepsilon)$ distribution. According to a non parametric bootstrap strategy, u^*_i and ε^*_{it} are sampled with replacement from the fitted random effects and the residuals, respectively [16]. The estimates from each bootstrap sample are then used to

obtain standard errors and confidence intervals. The method clearly depends on the snipping level α. When the snipping level is larger than the actual contamination rate, all the outliers are expected to be discarded. The efficiency loss will increase with the number of genuine entries wrongly included in the snipped set. On the contrary, the consequences of setting α too small are more dangerous, since some outliers will still affect inference. In order to improve the results, one should monitor the changes in the fitted model as α varies. One strategy could be to fit the model for different snipping level and evaluate to what extent the likelihood or the coefficients estimates change. Another approach would be to set α in order to minimize the average sum of squares of the differences between the estimates and the estimates evaluated over each bootstrap replication. This approach has been introduced in [9].

3 Real data example

In order to investigate the proposed method, we consider data set from the well-known study on investment theory by Yehuda Grunfeld [12]. This study analyzes the effect of the real value of the firm and of the real capital stock on real gross investments. The study involved 10 U.S. firms (considered as random effects) over 20 years, 1935 - 1954. Before performing the analysis we scaled the three variables, because they are on very different scales. The model can be written as $y_{ij} = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i + \varepsilon_{ij}, \quad i = 1,2,\ldots,10 \quad j = 1,\ldots,20$ where Y are the standardized gross investments, X_1 the standardized real value of the firm and X_2 the standardized real capital stock. The algorithm is designed to optimize a snipped restricted likelihood. The estimates for fixed and random effects, with snipping at 10% and 15% are given in table 1.

Table 1 Grunfeld data: estimates (with s.e.) by snipped REML ($\alpha = 0.10, 0.15$).C.I. by bootstrap

$\alpha = 0.10$				$\alpha = 0.15$			
Parameter	Estimates	St.Errors	95% CI	Parameter	Estimates	St.Errors	95% CI
Fixed effects							
μ	-0.012	0.122	(-0.026, 0.015)	μ	-0.006	0.112	(-0.027, 0.016)
β_1	0.615	0.062	(0.475, 0.768)	β_1	0.631	0.059	(0.478, 0.780)
β_2	0.369	0.026	(0.283, 0.453)	β_2	0.381	0.025	(0.274, 0.459)
Random effects							
σ^2_u	0.147			σ^2_u	0.124		
σ^2_ε	0.052			σ^2_ε	0.047		

Standard errors and percentile 95% confidence intervals obtained by parametric bootstrap (based on 999 replicates) for the fixed effects are introduced. The REMLs of parameters result different from the classical estimates. The intercept value changes from 0.000 to -0.012 and -0.006, with $\alpha = 0.10, 0.15$, respectively. The β_s estimates change from 0.665 to 0.615 and 0.631, and from 0.428 to 0.369 and 0.381, for $\alpha = 0.10, 0.15$, respectively.

4 Conclusions

In this paper a methodology that allows simultaneous estimation and outliers detection is introduced. It is meant to be robust against entry-wise outliers and it is built on a cell-wise contamination model. This robust approach, based on snipping, is an useful method for estimation and inference in mixed linear models for longitudinal data. The application to a real dataset highlights how the procedure detects outliers and estimates the parameters. Furthermore, for inference purpose, parametric bootstrap has been applied. Further simulation analyses and deeper studies are necessary to better evaluate and appreciate the goodness of the proposal.

References

1. Agostinelli, C., Leung, A., Yohai, V.J., Zamar, V.J.: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Test, **24**, 441–461, (2015)
2. Chackraborty, B, Chaudury, P, On an optimization problem in robust statistics, Journal of computational and Graphical Statistics, **17** 683–702, (2008)
3. Copt, S., Victoria-Feser, M.P., High breakdown inference for mixed linear models. Journal of the American Statistical Association, **101**, 292–300. (2006)
4. Copt, S., Heritier, S., Robust alternatives to the F-test in mixed linear models based on MM-estimates, Biometrics, **63**, 1045–1052, (2007)
5. Danilov, M., Yohai, V.J., Zamar, R.H.: Robust estimation of multivariate location and scatter in the presence of missing data. Journal of the American Statistical Association, **107**, 1178–1186, (2012)
6. Farcomeni, A.: Robust constrained clustering in presence of entry-wise outliers. Technometrics, **56**, 102–111, (2014)
7. Farcomeni, A.: Snipping for robust k-means clustering under component-wise contamination. Statistics and Computing, **24**, 909–917, **38**, 963–974, (2014)
8. Farcomeni, A., Greco, L.: Robust Methods for Data Reduction, CRC Press (2015)
9. Farcomeni, A., Viviani, S.: Robust estimation for Cox regression, Biometrical journal, **53**, 956–973, (2011)
10. García-Escudero, L A, Gordaliza, A, Matrán, C, Mayo-Iscar, A., A general trimming approach for robust cluster analysis: The Annals of Statistics, **36**, 1324–1345 (2008)
11. Greco, L., Lunardon N., Ventura L., Pairwise robust estimation of multivariate location and scatter, Proceedings of S.Co. 2011, ISBN: 9788861297531, (2011).
12. Grunfeld, Y.: The Determinants of Corporate Investment, unpublished Ph.D. thesis, Department of Economics, University of Chicago, (1958).
13. Heritier, S., Cantoni, E., Copt, S., Victoria-Feser,M.P., Robust methods in biostatistics, Wiley, New York, (2009)
14. Laird, NM, Ware, JH, Random-effects models for longitudinal data, Biometrics, **38**, 963–974, (1982)
15. Rousseeuw, PJ, Van Den Bossche, W., Detecting Deviating Data Cells, Technometrics, **60**, 135-145 (2018)
16. Shang, J., Cavanaugh, J.E., An assumption for the development of bootstrap variants of the Akaike information criterion in mixed models, Statistics & Probability Letters, **78**, 1422–1429, (2008)

4.6 Advances in time series

A spatio-temporal model for events on road networks: an application to ambulance interventions in Milan

Un modello spazio-temporale per eventi su network stradali: analisi degli interventi delle ambulanze nel comune di Milano

Andrea Gilardi and Riccardo Borgoni and Jorge Mateu

Abstract The algorithms for optimal management and deployment of ambulances within a municipality require a spatio-temporal model to forecast hotspots and minimise the response times. Ambulance interventions represent an example of a point pattern occurring on a linear network, which was created starting from the main streets of Milan. The constrained spatial domain raises particular challenges and unique methodological problems that cannot be ignored for proper model development. Hence, this paper presents a non-separable spatio-temporal model for analysing the emergency interventions that occurred in the street network of Milan from 2015 to 2017. A dynamic latent factor model is adopted for capturing the temporal evolution, while the spatial dynamics are modelled using a network-readaptation of a kernel estimator.

Abstract Gli algoritmi per la gestione delle ambulanze all'interno di un comune necessitano di modelli statistici che possano prevedere l'insorgere di criticità, in maniera tale da poter minimizzare i tempi di intervento. Gli interventi in emergenza delle ambulanze rappresentano un esempio di processo di punto su network stradale, creato partendo dalla rete stradale di Milano. Il supporto spaziale del fenomeno sviluppa diverse problematiche sia da un punto di vista metodologico che applicato, che non possono essere ignorate per la creazione di un modello appropriato. In questo paper analizziamo la distribuzione degli interventi in emergenza delle ambulanze nel comune di Milano tra il 2015 ed il 2017, sviluppando un modello dinamico a fattori latenti per la componente temporale ed uno stimatore kernel non-parametrico per l'intensità spaziale, riadattato nel caso di dati su network.

Key words: ambulance interventions, point pattern on networks, spatial networks, spatio-temporal data

---

Andrea Gilardi; Riccardo Borgoni
Department of Economics, Management and Statistics, University of Milan Bicocca, e-mail: andrea.gilardi@unimib.it; riccardo.borgoni@unimib.it

Jorge Mateu
Department of Mathematics, Universitat Jaume I, Castellón (ES), e-mail: mateu@uji.es

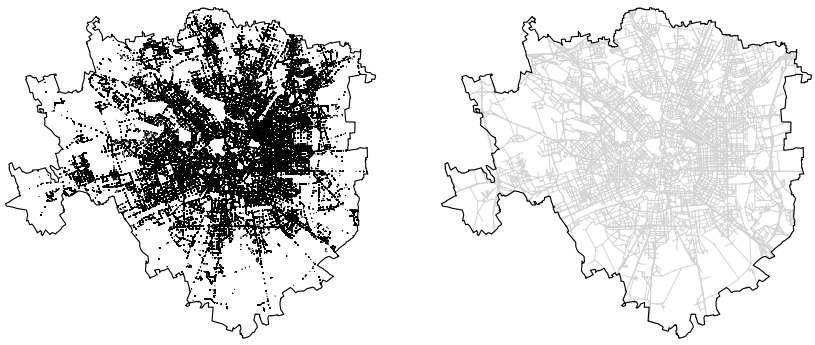
1 Introduction

The algorithms for optimal staff management and ambulances deployment within a municipality require a spatio-temporal model to forecast hotspots and minimise the expected response times. The predictions are required at a fine spatial and temporal resolution, due to intricate spatio-temporal patterns in emergency intervention data, which are particularly relevant for a hectic city like Milan.

Ambulance interventions represent a typical example of a point pattern occurring on a linear network, an increasingly popular type of events presenting several challenges related to the tangled and non-homogeneous nature of their spatial support [1]. Several authors explained the perils of re-adapting classical planar techniques, such as K-function or Kernel Density Estimator (KDE), to network data without considering the network's structure [5, 7]. The recent surge of interest can also be linked with the rapid development of several open-source spatial databases (such as Open Street Map), that provide the starting point for creating a computational representation of a road network.

2 Data: Ambulance Interventions

The data at hand included all emergency calls registered in the municipality of Milan (IT) from 2015-01-01 to 2017-12-31, which required an ambulance intervention and were handled by the regional Emergency Medical System (EMS). We removed all records with missing spatial or temporal coordinates, and we included only the first intervention when multiple ambulances were dispatched for the same (typically



(A) (B)

Fig. 1 *Left*: Locations of ambulance interventions in Milan from 2015 to 2017. *Right*: The most important streets of the road network. In both cases, we can recognise several white areas corresponding to parks (Parco Sempione), pedestrian areas (Citylife), and non-urban places.

life-threatening) event. The final sample included 495,950 interventions, 163,488 occurred in 2015, 165,368 in 2016 and 167,094 in 2017.

The spatial distribution of the EMS calls is reported in Figure 1(A). We note that the events resemble a street network structure, highlighting the city ring road and the most critical arterial thoroughfares. Hence, we argue that a spatio-temporal model of emergency interventions should not ignore their peculiar spatial support. The empty areas in Figure 1(A) correspond to non-urban places, mainly located in the south or the west. We can also clearly distinguish the shapes of several iconic locations of Milan, such as City Life, Parco Sempione or Scalo Farini.

We examined the temporal dimension of EMS interventions and determined the seasonal patterns that govern the total number of emergency calls. More precisely, we noticed that the average number of hourly events during the weekdays follows a particular trend: after a rapid increase in the early morning, the time series reaches its maximum around 10:00, slowly declines until 20:00 and then decreases until the night. The weekends present a similar distribution, with more interventions during the night hours (probably linked with the city's nightlife) and fewer events in the late morning. The time series of ambulance interventions also exhibits a weekly seasonal pattern, and the global minima are registered around August, in conjunction with national holidays. The dynamic latent factor model introduced in Section 3.1 was defined taking into account these seasonal patterns, which are discussed by [6, 4].

A linear network, typically denoted by L, is defined as the union of a finite set of segments, say l_i, lying in a planar region S:

$$l_i = [\mathbf{u}_i, \mathbf{v}_i] = \{\mathbf{s} : \mathbf{s} = t\mathbf{u}_i + (1-t)\mathbf{v}_i;\ 0 \leq t \leq 1\}; \quad \mathbf{u}_i, \mathbf{v}_i \in S \subseteq \mathbb{R}^2.$$

The endpoints of l_i are denoted by $\mathbf{u}_i$ and $\mathbf{v}_i$, and, in this paper, S denotes the polygonal boundary of Milan. The computational structure of the road network was created starting with data downloaded from Open Street Map (OSM) and selecting only the most important[1] street segments.

Spatial networks can also be seen as graph objects, where the edges correspond to the street segments, while the nodes are usually placed at road junctions [2]. We took advantage of the graph representation to simplify Milan's road network, excluding the small groups of isolated road segments. More precisely, we created a binary adjacency matrix between pairs of edges, defining two edges as *connected* if the corresponding road segments share one point at their geographical boundaries. Then, we clustered the segments and removed the isolated groups (typically denoted as *components* in the graph-analysis literature). This procedure creates a fully connected road network, which has relevant consequences on the kernel estimator presented in Section 3.2.

The linear network obtained after applying the pre-processing steps described above is depicted in Figure 1(B). It is composed of approximately 11,000 edges, and it covers more than 1850km, traversing almost every part of the city. We can

[1] We filtered only the street segments that, in the OSM jargon, are classified as *motorways*, *trunks*, *primary roads*, *secondary roads*, *tertiary roads*, and *unclassified roads*. Using the Italian classification, they range from *Autostrada* to *Strada Comunale*.

notice several similarities between Figure 1(A) and 1(B), and, once again, we can recognise several iconic places.

After creating the street network, we decided to exclude all ambulance interventions that occurred farther than 50 metres from the closest street segment, since we assumed that they occurred in other parts of the city network, and we projected the remaining ones into the linear network. We removed approximately 5% of the EMS data. Finally, we explored the spatio-temporal nature of EMS data, observing the presence of space-time interactions in the hourly distributions. More precisely, we noticed that from 08 AM to 08 PM the interventions are concentrated near the city centre, close to the office areas and the main buildings, while, during the night hours, they are scattered all around the municipality. These interactions are captured by the weighted network kernel estimator detailed in Section 3.2.

3 Statistical Methods

Following and extending the approach introduced in [9, 4], we consider a continuous one-dimensional linear network L and a discrete temporal dimension $\mathscr{T}$ divided into intervals of one hour. Let y_t denote the number of emergency calls that were recorded at time $t \in \mathscr{T}$, and let $\mathbf{s}_{i,t}, i = 1, \ldots, y_t$ be the location of ith event. Then, we assume that, independently for each $t \in \mathscr{T}$, the point process $\{\mathbf{s}_{i,t} : i = 1, \ldots, y_t\}$ can be modelled as a *Non-homogeneous Poisson Process* (NHPP) on a linear network with intensity function $\lambda_t(\mathbf{s})$ [3, 1]. Furthermore, we assume that

$$\lambda_t(\mathbf{s}) = \mu_t g_t(\mathbf{s}), \quad \mathbf{s} \in L;\ t \in \mathscr{T}, \tag{1}$$

where μ_t represents the temporal dimension of the EMS counts, while $g_t(\mathbf{s})$ is the spatial component of the process. Even though Equation 1 looks like the classical separability assumption for spatio-temporal point processes, the notation $g_t(\mathbf{s})$ implies that the spatial component depends on the temporal distribution of the data. These space-time interactions are taken into account adding a set of weights into the kernel function used to estimate $g_t(\mathbf{s})$, as detailed in Section 3.2.

In the next sections, we briefly introduce a time series model to capture the evolution of μ_t, and we describe with greater details the procedures for estimating $g_t(\mathbf{s})$ using a re-adaptation of the planar weighted kernel estimator for point pattern data on linear networks.

3.1 Temporal model

Following the approach detailed in [6, 4], we modelled the temporal component μ_t using a dynamic latent factor model. The hourly, daily, and weekly seasonalities were included by imposing a set of constraints on the factors and loadings matri-

ces, while penalised and cyclic cubic regression splines were adopted to impose a smooth evolution on EMS counts.

3.2 Spatial model

As mentioned before, the spatial component of the EMS interventions is modelled using a network-readaptation of Jones-Diggle corrected weighted kernel estimator, which, given a location $\mathbf{s} \in L$ and a time period u, can be written as

$$\hat{g}_u(\mathbf{s}) = \frac{\sum_{t \in \mathscr{T}} \sum_{i=1}^{y_t} w_{\mathbf{s}_i}(t,u) K_N(\mathbf{s}, \mathbf{s}_{i,t})}{\sum_{t \in \mathscr{T}} \sum_{i=1}^{y_t} w_{\mathbf{s}_i}(t,u)}.$$

We assumed that the weight function, hereby denoted as $w_{\mathbf{s}_i}(t,u)$, depends only on the temporal lag between u and the historical data. The weights are used to incorporate a space-time interaction into the KDE, giving more importance to EMS calls that occurred in the temporal proximity of u, and creating a non-separable structure into $\lambda_t(\mathbf{s})$. We refer to [9, 4] for more details on the weights' estimation process.

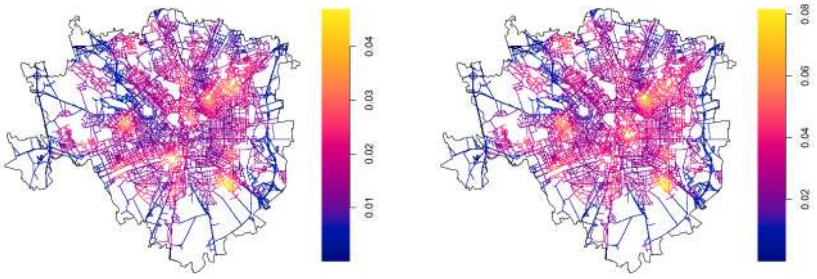
The function $K_N(\mathbf{s}, \mathbf{s}_{i,t})$ denotes the Jones-Diggle corrected network KDE, as introduced by [8]. More precisely, considering a location $\mathbf{s} \in L$ and a time period $t \in \mathscr{T}$, the estimator is defined as

$$K_N(\mathbf{s}, \mathbf{s}_{i,t}) = \frac{K(\mathbf{s} - \mathbf{s}_{i,t})}{c_L(\mathbf{s}_{i,t})}, \tag{2}$$

where K denotes a planar bivariate kernel function, $\mathbf{s}_{i,t}$ is an historical ambulance intervention, and $c_L(\mathbf{s}_{i,t})$ represents the convolution of the kernel K with arc-length measure on the network, defined as $c_L(\mathbf{s}) = \int_L k(\mathbf{v} - \mathbf{s})\, d_1\mathbf{v}$. Equation 2 is analogous to the planar KDE, where the Jones-Diggle correction is replaced using an integral over the network. Despite a slightly suboptimal statistical efficiency, the KDE estimator in Equation 2 can be computed rapidly using the fast Fourier transformation, which is essential considering the size of the network and the volume of EMS calls. The other statistical properties are extensively described in [8], whereas alternative approaches are discussed by [1].

4 Results and Conclusions

We exemplified the algorithm described in Section 3 considering two future temporal occasions: 2018-01-03 at 03:00 (left) and 2018-01-03 at 15:00 (right). The results are reported in Figure 2. The map on the left shows that EMS interventions are spread in several parts of Milan, highlighting nightlife areas such as Porta Genova or San Lorenzo, while the map on the right draws attention to other zones close to Duomo and significant working places. In both cases, the main train station, Pi-



(A) (B)

Fig. 2 Estimates of the spatial intensity function, $\hat{g}_u(\mathbf{s})$, considering two future time periods: 2018-01-03 at 03:00 (left) and 2018-01-03 at 15:00 (right).

azzale Loreto, and several retirement houses (such as Pio Albergo Trivulzio) are highlighted. The two maps are represented using different scales in order to better point out the temporal fluctuation of ambulance intervention intensity.

As further steps, we are developing a methodology for properly assessing the fit of the suggested model. Moreover, we plan to extend the planar spatio-temporal estimators for relative risk to network data to investigate and compare the spatial dynamic of EMS calls having different severity levels.

References

1. Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G. and Davies, T.M., 2020. Analysing point patterns on networks—A review. Spatial Statistics, p.100435.
2. Barthélemy, M., 2011. Spatial networks. Physics Reports, 499(1-3), pp.1-101.
3. Diggle, P.J., 2013. Statistical analysis of spatial and spatio-temporal point patterns. CRC press.
4. Gilardi, A., Borgoni, R., Pagliosa, A. and Bonora, R., 2018, Spatiotemporal Prevision for Emergency Medical System Events in Milan. Book of Short Papers SIS 2018, pp. 1697-1702
5. Lu, Y. and Chen, X., 2007. On the false alarm of planar K-function when analyzing urban crime distributed along streets. Social science research, 36(2), pp.611-632.
6. Matteson, D.S., McLean, M.W., Woodard, D.B. and Henderson, S.G., 2011. Forecasting emergency medical service call arrival rates. Annals of Applied Statistics, 5(2B), pp.1379-1406.
7. Okabe, A. and Sugihara, K., 2012. Spatial analysis along networks: statistical and computational methods. John Wiley & Sons.
8. Rakshit, S., Davies, T., Moradi, M.M., McSwiggan, G., Nair, G., Mateu, J. and Baddeley, A., 2019. Fast Kernel Smoothing of Point Patterns on a Large Network using Two-dimensional Convolution. International Statistical Review, 87(3), pp.531-556.
9. Zhou, Z. and Matteson, D.S., 2015, August. Predicting ambulance demand: A spatio-temporal kernel approach. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2297-2303).

Forecasting electricity demand of individual customers via additive stacking

Previsione della domanda di elettricità di consumatori individuali attraverso modelli additivi per l'aggregazione di esperti

Christian Capezza, Biagio Palumbo, Yannig Goude, Simon N. Wood, and Matteo Fasiolo

Abstract Smart grids rely on renewable energy sources and distributed production, which lead to increased variability in the electricity load. Then, accurate forecasts of individual household electricity demand will be a key element for a cost-effective management of smart grids. However, individual electricity demand forecasting is particularly challenging because of the lower signal-to-noise ratio compared to the aggregate demand. Therefore, we propose a new method for stacking probabilistic forecasts, which borrows information across households while taking into account their individual characteristics. The proposed method is an extension of regression stacking where mixture weights vary with covariates via an additive model structure.
Abstract *Le smart grid sono caratterizzate dalla produzione distribuita e l'uso di fonti rinnovabili, che portano a una maggiore variabilità nella domanda di elettricità. Per una loro gestione efficace, è dunque fondamentale avere previsioni accurate della domanda a livello di consumatori individuali. Tuttavia, il rapporto segnale-rumore più basso rispetto alla domanda aggregata rende questo tipo di previsioni particolarmente difficile. Dunque, proponiamo un nuovo metodo per l'aggregazione di previsioni probabilistiche che utilizza informazioni congiunte di diversi consumatori, ma allo stesso tempo tiene conto delle loro caratteristiche individuali. Il metodo è un'estensione del classico "regression stacking", in cui i pesi sono modellati come funzioni di covariate attraverso un modello additivo.*

Christian Capezza and Biagio Palumbo
Department of Industrial Engineering, University of Naples Federico II, Naples, Italy e-mail: christian.capezza@unina.it; biagio.palumbo@unina.it

Yannig Goude
Électricité de France R&D, Paris, France, e-mail: yannig.goude@edf.fr

Simon N. Wood
School of Mathematics, University of Edinburgh, Edinburgh, UK, e-mail: simon.wood@ed.ac.uk

Matteo Fasiolo
School of Mathematics, University of Bristol, Bristol UK, e-mail: matteo.fasiolo@bristol.ac.uk

Key words: Electricity Demand Forecasting, Ensemble Methods, Generalised Additive Models, Probabilistic Forecast, Regression Stacking.

1 Introduction

The need to reduce carbon emissions is leading to an expansion in the use of weather dependent, renewable energy sources and the electrification of the transportation system. Then, electricity production and storage are increasingly decentralised and this setting is more complex than a centralised system where industrial operators control the production and the limited storage. The main challenge for modern grid management systems is to be able to satisfy a demand that is both larger, due to the electric vehicles, and more uncertain, because of the less flexible production. Therefore, new, smart policies need to be adopted to avoid the expansion of the physical capacity of the electricity network with expensive infrastructural works.

One aim is to reduce the daily demand peak by adopting demand-side tools such as dynamic electricity pricing and remotely controlled consumption. To achieve this, with the increase of the distributed production and storage, electricity demand forecasts at a low level of aggregation will become more important with respect to the aggregate demand, e.g., at regional scale. However, predictive accuracy decreases with the level of granularity. In fact, the daily electricity demand profile is smooth when the demand is averaged across customers, while individual household profiles have a lower signal-to-noise ratio and are less predictable.

In this work, we define a set of experts that provide probabilistic forecasts and are fitted separately to each household to capture the heterogeneous dynamics of the individual customer demand, such as smooth daily demand components or abrupt change-points. To deal with the low signal-to-noise ratio of the individual demand, we "borrow strength" across the different households through a weighted aggregation of the experts. Weights are estimated in a single model, using data from all the households. The main novelty of this work is that we use an additive model for the weights, i.e. they can depend on covariates such as the time of the day, the day of the week, individual household characteristics and so on, using linear combinations of parametric and smooth effects based on spline basis expansions.

Note that the aggregation of several point predictors was firstly introduced by Stone [11], with the aim to improve the predictive performance. Then, Breiman [2] proposed the so called regression stacking. Our work is closely related to Yao et al. [14], who use stacking to average Bayesian predictive distributions, as in our work, but with fixed weights. Other aggregation methods let the experts' weights vary with time [9], while we use a more flexible model where weights are semi-parametric functions of all the covariates. Finally, in Coscrato et al. [6] experts' weights depend non-parametrically on covariates, however the authors only focus on point predictions, not probabilistic forecasts.

2 Additive Stacking

In this section we briefly illustrate the methodology we developed in Capezza et al. [3], to which we refer for further details. Let $p_k(y_i|\boldsymbol{x}_i)$ be the i-th conditional density estimate produced by the k-th expert. Probabilistic stacking is performed by forming a mixture, $\sum_{k=1}^{K} \alpha_k p_k(y_i|\boldsymbol{x}_i)$, where we let the weights α_k vary with the covariates via an additive model structure. The weights are parametrised as $\alpha_{ki} = \exp \eta_{ki} / (\sum_{a=1}^{K} \exp \eta_{ai})$, for $k = 1, \ldots, K$, where η_{ki} is the linear predictor of the k-th expert, evaluated at the i-th observation, and η_1 is fixed to zero for identifiability. We model linear predictors as linear combinations of parametric, random or smooth effects, based on spline basis expansions, so that they linearly depend on unknown regression coefficients $\boldsymbol{\beta}$ which must be estimated. An improper multivariate Gaussian prior, centered at zero and with precision matrix given by $\sum_{g=1}^{G} \lambda_g \mathbf{S}_g$, where the $\mathbf{S}_g$'s are positive semi-definite matrices and $\lambda_1, \ldots, \lambda_G$ are positive smoothing parameters, controls the wiggliness of the smooth effects. Then, regression coefficients can be estimated via maximum a posteriori (MAP) estimation, i.e. by maximising the Bayesian posterior log-density

$$\log p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\lambda}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \alpha_{ki}(\boldsymbol{\beta}) p_k(y_i|\boldsymbol{x}_i) - \frac{1}{2} \sum_{g=1}^{G} \lambda_g \boldsymbol{\beta} \mathbf{S}_g \boldsymbol{\beta}. \tag{1}$$

We select the smoothing parameters by maximising a Laplace approximation to the marginal likelihood (LAML). We are able to adopt the likelihood based fitting methods of Wood et al. [13], aimed at generalised additive models (GAMs [7]), because we perform stacking in a probabilistic, rather than loss based, context. The proposed additive stacking is quite flexible because of the many effect types available under standard GAM models, see e.g. Wood [12], moreover, using a probabilistic Bayesian framework allows us to adopt well-founded and computationally efficient statistical methods for model estimation and inference.

The parametrisation of the proposed stacking model is non-linear in the regression coefficients, then it is difficult to interpret the effects of covariates on the weights of the experts. In this work, we address this problem by using the accumulated local effects (ALE) of Apley and Zhu [1], which allow to visualise the main effect of the covariates on the aggregation weights. Moreover, another novelty in this work is that we also quantify the uncertainty of the ALE effects, without extra computational cost, based on the posterior distribution of the regression coefficients.

3 Disaggregate Electricity Demand Forecasting

We consider the data set from the Commission for Energy Regulation (CER) trial [5], which contains electricity demand y_i^c, for $i = 1, \ldots, N$, measured in kWh and at 30min resolution by smart meters at 2672 Irish households, $c = 1, \ldots, C$. The data set covers the whole of 2010 and contains survey information about each household,

such as the occupation of the chief income earner, the number of white goods, if the customer owns or rents the property, as well as hourly temperatures from the National Centers for Environmental Information (NCEI). The data set consists of more than 30 millions observations.

We consider four experts. Under *LastMonth*, the predictive density is obtained through kernel density estimation based on the most recent 30 available observations for each customer, at the same time of the day; its strength is that the electricity demand distribution can change abruptly with the time of the day. *GaulssInd* is a log-normal generalised additive model for location scale and shape (GAMLSS, [10]). *GaulssInd* is meant to capture smooth components of the daily individual profiles, as well as the temperature, calendar and autoregressive effects, which are typically used to model the aggregated demand. *Dynamic* is a log-normal GAM model, where the mean of the logarithm of the data is modelled by a smooth effect of the time of the day and is fitted only to the data from the last three days, which makes it quicker to adapt in case of sudden changes. *GaulssCommon* is a log-normal GAMLSS model and is the only expert fitted to all customers jointly and uses the effects of household specific survey variables, together with smooth effects of covariates such as the time of the day. Because of the heterogeneity across customers, predictions under *GaulssCommon* are strongly biased, however they are a valuable baseline forecast to predict household demand for anomalous consumption trends.

We combine the four experts through the additive stacking model proposed in Section 2. Covariates include: the mean and standard deviation of the consumption of the individual customer up to the current week, in order to identify customers with rich consumption dynamics, on which we expect that *GaulssInd* performs well; time of the day; time of the year, which allows *GaulssCommon* to provide a starting baseline prediction that becomes less useful during the year, as more data are available; the number of consecutive days a customer has been out of home before the current day, which, when large, is expected to give a large weight to *Dynamic* expert that is the quicker to adapt to sudden changes in the demand; we also consider the relative past predictive performance of each expert, based on the exponentially weighted average forecaster (EWA) [4], so that experts are given a larger weight if they are performing particularly well on a customer in the previous days. Note that households' electricity demand was not standardized for household size. In fact, in the individual models fitted separately for each household, the household size is constant within each model, therefore no standardization is required. In the *GaulssCommon* expert and the additive stacking model, which use all households' data together, we included covariates that take into account the difference among household sizes, such as the mean and standard deviation of the consumption of each customer observed in the previous weeks, or the number of white goods.

To avoid overfitting, the training data set for estimation of the stacking model is built on a rolling basis as follows. We fit experts on the first five weeks, then we store their probabilistic forecasts on week 6, afterwards we fit experts on the first six weeks and store probabilistic forecasts for week 7, and so on. By iterating this, we obtain out-of-sample probabilistic forecasts from all experts, which can be used to fit the stacking model. Then, we use again a rolling basis to build a test data set

	Dynamic	GaulssCommon	GaulssInd	LastMonth	Stacking
Log-loss	-0.059	-0.019	-0.244	28.812	**-0.376**
Square loss	0.317	0.335	0.291	0.298	**0.279**
CRPS	0.216	0.230	0.203	0.204	**0.195**
Pinball 0.5	0.138	0.151	0.134	0.136	**0.130**
Pinball 0.9	0.121	0.123	0.105	0.105	**0.100**
Pinball 0.99	0.042	0.031	0.030	0.028	**0.024**

Table 1 Average predictive losses of each model. The lowest loss in each category is bold.

and compare the stacking predictive performance with that of the experts. We fit the stacking model to the data from weeks 6–9 and use it to produce probabilistic forecasts for week 10, then we use data from weeks 6–10 and predict the demand on week 11, and so on. By iterating this, we obtain probabilistic forecasts of the additive stacking that can be compared with those produced by the experts.

We use several loss functions to evaluate the predictive performance of each model. In particular, we consider the log-loss, i.e., the negative log-likelihood evaluated on the test data, the square loss, the continuous ranked probability score and the pinball loss [8] at quantiles 0.5, 0.9 and 0.99. In Table 1 we report the predictive losses under each model, averaged over the entire data set. Additive stacking outperforms all the experts under each predictive loss. This result is noteworthy, as the additive stacking model is estimated via likelihood-based MAP and LAML methods, which are directly related to the log-loss, but not to the other losses. Moreover, the improvement is achieved at any time of the day and that, relative to the individual experts, the stacking achieves larger improvements in terms of the pinball loss on the highest quantiles. This indicates that stacking is doing a better job at predicting the spikes in the daily demand. *LastMonth* performs particularly poorly on the log-loss, because it is based on a thin tailed mixture of Gaussian densities, which generates large losses on outlying demand observations. However, *LastMonth* is more competitive on the other losses.

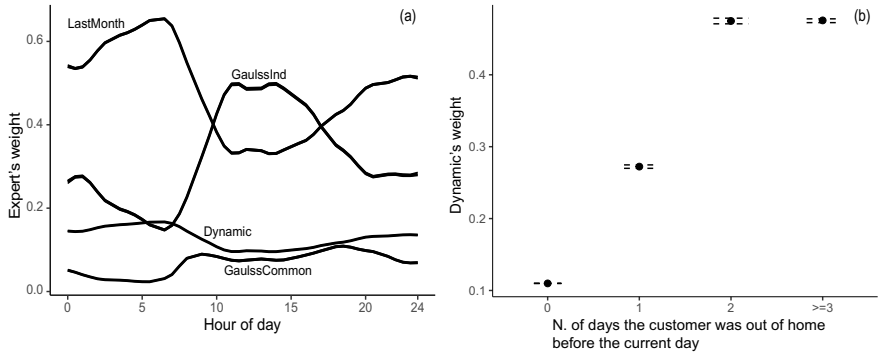


Fig. 1 ALE effects on the stacking weights of some covariates: a) effect of time of day on each expert, b) effect of the number of previous days a customer was out of home on *Dynamic*.

We use ALE plots of Apley and Zhu [1] for the interpretation of the effects of covariates on stacking weights. For brevity we show only a few of these plots. In Figure 1a, *LastMonth* has the largest average weight, but on the key working hours, when demand dynamics are more complex, *GaulssInd* is the dominant model. The weight of *Dynamic* is low on average (Figure 1a), however it strongly increases when a customer was out of home in the previous days (Figure 1b), because it adapts quicker to sudden changes in the customer demand.

4 Conclusion

The proposed additive stacking allows to combine predictive densities in a flexible way and the fast direct MAP and LAML methods for model fitting allow to deal with a large data set. The results are promising because the additive stacking overcomes experts under several loss functions, moreover ALE plots provide interpretability of the covariates effects on the weights. One interesting use of such improved probabilistic forecasts is the optimisation of energy costs under daily-max tariffs by means of home battery scheduling, with the aim to make the daily aggregate demand profile flatter.

References

1. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. J. Roy. Stat. Soc. B Met. **82**(4), 1059–1086 (2020)
2. Breiman, L.: Stacked regressions. Mach. Learn. **24**(1), 49–64 (1996)
3. Capezza, C., Palumbo, B., Goude, Y., Wood, S.N., Fasiolo, M.: Additive stacking for disaggregate electricity demand forecasting. To appear in Ann. Appl. Stat.
4. Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge University Press (2006)
5. Commission for Energy Regulation: CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010 [dataset]. 1st Edition. Irish Social Science Data Archive. SN: 0012-00. www.ucd.ie/issda/CER-electricity (2012)
6. Coscrato, V., de Almeida Inácio, M.H., Izbicki, R.: The NN-Stacking: Feature weighted linear stacking through neural networks. Neurocomputing (2020)
7. Hastie, T.J., Tibshirani, R.: Generalized additive models. CRC press (1990)
8. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1), 33–50 (1978)
9. McAlinn, K., West, M.: Dynamic Bayesian predictive synthesis in time series forecasting. J. Econometrics **210**(1), 155–169 (2019)
10. Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. J. R. Stat. Soc. C-Appl. **54**(3), 507–554 (2005)
11. Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. Roy. Stat. Soc. B Met. **36**(2), 111–133 (1974)
12. Wood, S.N.: Generalized additive models: an introduction with R. CRC press (2017)
13. Wood, S.N., Pya, N., Säfken, B.: Smoothing parameter and model selection for general smooth models. J. Am. Stat. Assoc. **111**(516), 1548–1575 (2016)
14. Yao, Y., Vehtari, A., Simpson, D., Gelman, A.: Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Anal. **13**(3), 917–1007 (2018)

Hierarchical Forecast Reconciliation on Italian Covid-19 data

Riconciliazione gerarchica su time-series: applicazione ai dati Covid-19 italiani

Andrea Marcocchia, Serena Arima and Pierpaolo Brutti

Abstract Due to the spread of the Covid-19 pandemic, many different actions to limit personal freedom have been decided according to the trend of the epidemiological situation. Often these decisions were made at the national level or, in a second phase, at the regional one. In order to ensure the implementation of actions and behaviours more in line with the real epidemiological situation, it would be useful to decide and give specific information at a more granular territorial level (provinces) even if one of the main problem is the quality of the data. This paper presents how the use of forecast reconciliation techniques can help forecasts at a detailed territorial level, while exploiting the more robust information available at the aggregate level.

Abstract *Con la comparsa della pandemia da Covid-19 si sono rese necessarie azioni di limitazione delle libertà personali, decise in funzione dell'andamento della situazione epidemiologica. Tali decisioni sono state prese a livello nazionale o, in una seconda fase, regionale. Per garantire la messa in atto di azioni maggiormente in linea con la reale situazione epidemiologica di un territorio, sarebbe utile agire ad un livello di maggiore granularità territoriale (province). Uno dei principali problemi nel prendere decisioni a un livello più dettagliato risiede nella qualità dei dati. In questo lavoro si presenta come l'uso di tecniche di forecast reconciliation possono aiutare a migliorare le previsioni ad un livello granulare, sfruttando contemporaneamente le, più robuste, informazioni disponibili a livello aggregato.*

Key words: forecasting, time series, forecast reconciliation, hierarchical time series, Covid-19 data

---

Andrea Marcocchia
Sapienza University of Rome, e-mail: andrea.marcocchia@uniroma1.it

Serena Arima
University of Salento, e-mail: serena.arima@unisalento.it

Pierpaolo Brutti
Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction

The main idea behind forecast reconciliation in a hierarchy is that observed demands at each level will always add up to the observed demands at higher levels. It is usually desirable that the same holds true also for forecasts (that is the "forecasting coherence"). If forecasting at the different levels is done independently, we usually have forecast incoherence, meaning that the bottom level forecasts do not add up.
The various components of the hierarchy can interact in a variety of complex ways. A change in one series at one level, can have an impact on other series at the same level, as well as on series at higher and/or lower levels. By modeling the entire hierarchy of time series simultaneously, we obtain better forecasts of each of the component series. *Reconciliation* is the process that fix incoherent forecasts. Another important benefit of forecast reconciliation is that, although usually the research interest lays in the most disaggregated data, these are also noisier than the others; in contrast, the "total" series, although less interesting, is typically way more resilient to noise: forecast reconciliation is capable to extract the most relevant information from both series [1].

2 Models

This section presents the used models: the first part explains the forecast algorithms, while the second subsection adds the description of the reconciliation step.

2.1 Forecast models

Any type of prediction methods can be used in the first phase because the reconciliation models have no limits in this regard. The used forecast methods are:

- **ARIMA with and without external covariates**: considering the difficulty in identifying a single parameterization that guarantees a good performance on all the series of the hierarchy, we proceeded using an automatic selection method of the best set of parameters for each series. In an other case, an external variable has been added as additional information. The considered variable indicates for each date whether it is in a lockdown condition or not. The variable is the same at all levels of the hierarchy, therefore restrictions were not considered at the regional or the provincial level, but only at the national level. It should be borne in mind that in the considered period (up to the beginning of November 2020), the closure measures imposed by the Italian government were mainly national, with few measures at the regional level, as lately more frequently decided.
- **Exponential Smoothing (ETS)**: also with regard to this method, the best model was selected for each series.

- **Segmented regression**: this approach allows to divide the time series into different segments and to use a simple linear regression to estimate the trend of each segment. It was decided to use this approach after noting that there were purely linear trends within the time series. The number of segments that have been created is equal to 3. The reason is that, until November 2020, there was an initial period of severe growth in the number of daily infections, then a decrease during the summer and then a new growth phase after summer.

2.2 Reconciliation models

The reconciliation problem can be formalize as follows:

- Consider a multi-level hierarchy, where level 0 denotes the completely aggregated series and level k contains the most disaggregated time series and assume that the observations are recorded at times $t = 1,2,\ldots,n$ and that we are interested in forecasting each series at each level at times $t = n+1,n+2,\ldots,n+h$.
- The additive structure of the time-series involved can be exemplified in a three level hierarchy: $Y_t = \sum_i Y_{i,t}$ represents the most aggregated series, obtained as sum of all the series that belong to the first level ($Y_{i,t}$), that is $Y_{i,t} = \sum_j Y_{ij,t}$ so that the first level series is obtained starting from the second level. As final stage, $Y_{ij,t} = \sum_z Y_{ijz,t}$ represents the aggregation of the lowest level of the hierarchy (the third level, with $k = 3$). The notation refers to $Y_{ij,t}$ as the values of series ij at time t and to Y_t as the aggregation of all the series at time t.
- More generally, let $\mathbf{Y}_{i,t}$ be the vector that collects the values of all the series for $i = 1,\ldots,k$, observed at a specific level i at time t. Now define $\mathbf{Y}_t = \left[\mathbf{Y}_{1,t},\ldots,\mathbf{Y}_{k,t}\right]'$. According to this notation, we can define $\mathbf{Y}_t = \mathbf{S}\,\mathbf{Y}_{k,t}$, where $\mathbf{S}$ is a summing matrix used to aggregate the lowest level series. If we assume that the number of bottom level time series is m_k and that m is the total number of series (considering all the levels), then $\mathbf{S}$ is an $(m \times m_k)$ matrix. The $\mathbf{S}$ matrix can be partitioned by the levels of the hierarchy: the top row is a unit vector of length m_k and the bottom section is a $m_k \times m_k$ identity matrix. The symbols m and m_i, are the total number of series in the total hierarchy and at level i, so it is always true that $m_i > m_{i-1}$ and $m = m_0 + m_1 + \cdots + m_k$.
- If we compute forecasts for each period $n+1,n+2,\ldots,n+h$ for a generic series X, we can define this forecasted series $\widehat{Y}_{X,n}(h)$. Applying the same logic, $\widehat{Y}_n(h)$ denotes the h-step ahead prediction of the total. The final step is to define $\widehat{\mathbf{Y}}_n(h)$ as the vector consisting of these base forecasts, stacked in the same series order as for $\mathbf{Y}_t$;
- Bearing all this in mind, all existing hierarchical forecasting methods can be written as $\widetilde{\mathbf{Y}}_n(h) = \mathbf{S}\mathbf{P}\widehat{\mathbf{Y}}_n(h)$. The effect of the $\mathbf{P}$ matrix is to extract and combine the relevant elements of the base forecasts $\widehat{\mathbf{Y}}_n(h)$, which are then summed by $\mathbf{S}$ to give the final revised hierarchical forecasts, $\widetilde{\mathbf{Y}}_n(h)$.

The choice of $\mathbf{P}$ is a crucial step: the effect of its choice is to extract and combine the relevant elements of the base forecasts according to: $\mathbf{SP}\widehat{\mathbf{Y}}_n(h)$.
According with this formulation, the following reconciliation approaches have been used in order to reconcile the base forecasts introduced in the previous paragraph. Net of some exceptions, all the reconciliation methods were applied for each prediction method:

- **Top-Down**: this method entails the forecasting of the completely aggregated series (higher level of the hierarchy), and then the disaggregation of the forecasts based on historical proportions [5]. In this approach $\mathbf{P} = \left[\mathbf{p} \mid \mathbf{0}_{m_k \times (m-1)}\right]$, where $\mathbf{p}$ is a vector of proportions that sum to one. Different methods of top-down forecasting lead to different proportionality vectors $\mathbf{p}$. Some possible choices are the *Average historical proportions* that reflects the average of the historical proportions of the bottom-level series, the *Proportions of the historical averages* that captures the average historical value of the bottom-level series $Y_{j,t}$ relative to the average value of the total aggregate Y_t or the *Forecast proportions* that uses proportions based on forecasts rather than historical data.
- **Bottom-Up**: the idea is to forecast each of the disaggregated series at the lowest level of the hierarchy, and then to obtain forecasts at higher levels of the hierarchy by using simple aggregations. In this case $\mathbf{P} = \left[\mathbf{0}_{m_k \times (m-m_k)} \mid \mathbf{I}_{m_k}\right]$, so the $\mathbf{P}$ matrix extracts only bottom level forecasts from $\widehat{\mathbf{Y}}_n(h)$, which are then summed by $\mathbf{S}$ to give the bottom-up forecasts.
- **Mint**: the forecast reconciliation through trace minimization is an approach that incorporates the information coming from the full covariance matrix of forecast errors in order to obtain a set of coherent forecasts. It minimizes the mean squared error of the coherent forecasts across the entire collection of time series under the assumption that they are unbiased [3].
- **Bayesian approach**: as typical of this inferential paradigm, the central point is to explore and exploit the posterior distribution of the reconciled forecast. The distribution of the reconciled forecasts can be obtained by sampling from the posterior predictive distribution using a Gibbs sampler. Using such techniques it is possible to keep into account some prior belief about the component time series, so that during the reconciliation step, more importance is given to the more robust time series by taking into account the in-sample variance of each base forecast [2].
- **Ordinary or Weighted Least Square**: the revised forecast at each node will be a weighted average of the forecasts from all nodes. The weights are obtained using linear regression, where all the independent forecasts (from all nodes) are regressed against a set of dummy variables indicating which of the bottom-level series contribute to each node[4]. The mean squared reconciliation error, computed using the differences between the reconciled and independent forecasts, is as small as possible. In the OLS case the $\mathbf{P}$ matrix is equal to $(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'$.

3 Data

The data used are the Italian Covid-19 incidence (new daily confirmed cases) obtained from the Github repository of the *Protezione Civile* agency. The data are daily updated and the information are released on a province basis. The first considered observation is recorded on 24^{th} February 2020 and the last one on 3^{rd} Novembrer 2020 for a total of 254 observations. The data are divided into two groups: the training set (from the first day until 14^{th} October 2020 for 234 observations) and the test set (from 15^{th} October 2020 until the last day for 20 observations). The dataset is made by 107 bottom-level time series, related to the Italian provinces.
Starting from these bottom level time series, a hierarchical structure has been created looking at the italian administrative organization, so the provinces have been aggregated into regions, the regions into the five macrozones partion and the macrozes summed up to obtain the total. The original data are transformed applying a logarithmic function. Due to the fact that the data are forecasted using models not well performing on count data. In detail, for a single value x_i the following transformation is applied: $x_i = log(x_i + \varepsilon)$.

4 Results

Each forecast and reconciliation method leads to different performances depending on the level of the hierarchy on which we focus. The most interesting level, for the purposes of the research, is the bottom-level series, that is the provinces. In over 70% of the 107 Italian provinces, the prediction adjusted by the reconciliation step is more accurate, in terms of RMSE, than the basic prediction.

Table 1 Count of bottom-level series where a method over-performs the others

Method	Count	Method	Count	Method	Count
Segmented	27	Ets + TD-GSF	4	Arima with cov. + TD-FP	2
Ets + OLS	13	Segmented + WLS	4	Segmented + TD-FP	1
Arima + TD-FP	10	Ets + WLS	3	Arima with cov. + OLS	1
Arima + Mint	9	Arima with cov.	3	Arima + TD-GSA	1
Ets + Mint	8	Arima	3	Arima + OLS	1
Ets + TD-FP	8	Ets + TD-GSA	3	-	-
Arima with cov. + Mint	4	Ets	2	-	-

Analyzing Table 1, it is possible to observe that the prediction method that most often has the best performance is the segmented regression. However, it must be considered that there are some Italian provinces in which the trend of Covid-19 daily cases in the considered period has remained stable and close to 0 even in the test period. Focusing on some of the larger Italian provinces, such as Rome (see Figure 1), the best score obtained without the reconciliation was of an RMSE equal to 0.59

(obtained applying segmented regression), while by inserting the reconciliation step, a value of 0.15 is reached (the ETS method was used for the base forecast). This behaviour is true for most of the bigger Italian provinces.

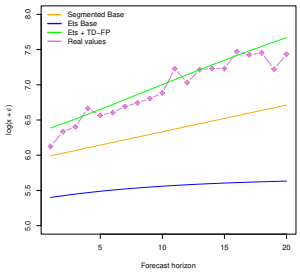


Fig. 1 Forecasted and real values for Covid-19 time series in Rome

It is important to note that among the reconciliation approaches there is no technique that systematically performs better than others. It must be borne in mind that the performance of the various series is profoundly different. Even at a higher level of aggregation (regions, macrozones, ...) there are important differences in the behaviour, which make it difficult to identify a single method valid all over the series.

5 Future works

Given the high variability of methods that over-perform others on the bottom-level time series, the construction of a method that, using a validation frame, is able to identify the best model would be very useful. Reconciliation techniques that make machine learning have recently been presented in the literature, and this could help in solving the problem by providing more accurate predictions.

References

1. G. Athanasopoulos and R. Ahmed and R. J. Hyndman: Hierarchical forecasts for Australian domestic tourism. International Journal of Forecasting **25**, 146-166 (2009)
2. F. Eckert and R. Hyndman and A. Panagiotelis: Forecasting Swiss Exports using Bayesian Forecast Reconciliation
3. S. Wickramasuriya and G. Athanasopoulos and R. Hyndman: Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. Journal of the American Statistical Association - **114**, 1-45 (2018)
4. G. Athanasopoulos and R. Hyndman and R. Ahmed and H.L. Shang: Optimal combination forecasts for hierarchical. Computational Statistics & Data Analysis - **55**, 2579-2589 (2011)
5. Charles W. Gross and Jeffrey E. Sohl: Disaggregation methods to expedite product line forecasting (1990)

Link between Threshold ARMA and tdARMA models

Relazione tra modelli a soglia ARMA e modelli ARMA con parametri dipendenti dal tempo

Guy Mélard and Marcella Niglio

Abstract In the present contribution, we propose a link between Threshold Autoregressive Moving Average (TARMA) and Time-Dependent ARMA (*td*ARMA) models. We show that a proper parametrization allows to include the TARMA model in the large class of *td*ARMA structures. The main advantage that can be obtained from this result is the derivation of the asymptotic properties of the estimators of TARMA parameters that can be obtained under weaker conditions with respect to those in the available literature.

Abstract *Nel presente contributo proponiamo una relazione tra modelli a soglia autoregressivi media mobile (TARMA) e modelli ARMA con parametri dipendenti dal tempo (tdARMA). Diamo evidenza che una opportuna parametrizzazione consente di includere il modello TARMA nell'ampia classe dei modelli tdARMA. Il principale vantaggio che può essere ottenuto da questo risultato, è la derivazione delle proprietà asintotiche degli stimatori dei parametri TARMA che possono essere ottenuti sotto condizioni più deboli di quelle disponibili in letteratura.*

Key words: Threshold model, time-dependent ARMA model

1 Introduction

In time series analysis, the dynamic of data has often been modelled through the introduction of time-dependent coefficient structures. Starting from [8] different proposals have been given in this domain.

Guy Mélard
Université Libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES, Bruxelles, Belgique e-mail: gmelard@ulb.ac.be

Marcella Niglio
Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy e-mail: mniglio@unisa.it

In the present contribution, the attention is focused on some generalizations of the ARMA structure where the dependence of the coefficients to the time is differently modelled. More precisely we will start considering Threshold ARMA (TARMA) models [10]:

$$X_t = \sum_{i=1}^{k} \left[\sum_{j=1}^{p} \phi_j^{(i)} X_{t-j} - \sum_{j=1}^{q} \theta_j^{(i)} \varepsilon_{t-j} \right] \mathscr{I}_{\{Y_{t-d} \in \mathscr{R}_i\}} + \varepsilon_t, \tag{1}$$

where X_t is the variable of interest at time t, k is the number of regimes, the $\phi_j^{(i)}$, $j = 1, ..., p$, and $\theta_j^{(i)}$, $j = 1, ..., q$, are, respectively, the autoregressive and moving average coefficients of the ARMA models for the i-th regime, $i = 1, ..., k$, Y_{t-d} is the threshold variable, d is the threshold delay, $\mathscr{I}_{\{\cdot\}}$ is an indicator function, $\mathscr{R}_i$ is a subset of the real line such that $\mathscr{R} = \bigcup_{i=1}^{k} \mathscr{R}_i$ with $\mathscr{R}_i \cap \mathscr{R}_s = \emptyset$, for $i \neq s$, and $\{\varepsilon_t\}$ a sequence of independent and identically distributed (i.i.d.) random variables with null mean and finite moments of order $4 + \delta$, $\delta > 0$, with ε_t independent from Y_t and Y_t a stationary and ergodic process.

Even if model (1) can be shortly described as "local linear ARMA" because, within each regime, X_t follows an ARMA model, its overall structure is more complex and goes beyond the linear domain. This is the reason why general results for the statistical properties of model (1), such as stationarity and ergodicity, have only been faced for well-defined parametrizations with endogenous threshold variable ($Y_{t-d} = X_{t-d}$): [3], consider a simplified structure with $\theta_j^{(i)} = \theta_j$, for $j = 1, \ldots, q$ where, in other words, the moving average coefficients do not change among regimes; [6] define sufficient conditions for the stationarity of model (1) with $p = 1$ whereas more recently [4] focus the attention on the ergodicity of first-order threshold ARMA processes (with $p = q = 1$).

As clarified before, in all cited literature the examined threshold model is characterized by an endogenous threshold variable. It makes, at the same time, the model less general with respect to model (1) but even more complex, when its dynamic structure needs to be investigated.

A recent contribution in this domain is given in [2], where, differently from model (1) the switching structure at known dates is related to an observed process with values in a finite set.

In the following, we consider a further variant of (1) that allows connecting the TARMA model to the large class of tdARMA models ([1]). We are going to present the model, the main differences with respect to model (1), and how these differences can support the estimation of the model parameters.

2 Threshold ARMA model

In (1), t is assumed to vary in $\mathbf{Z}$, the set of integers.

In the following we consider the process X_t starting at time $t = 1$, such that $X_t = 0$ and $\varepsilon_t = 0$ for $t < 1$.

A tdARMA model is defined by

$$X_t = \sum_{j=1}^{p} \phi_{tj} X_{t-j} - \sum_{j=1}^{q} \theta_{tj} \varepsilon_{t-j} + \varepsilon_t, \tag{2}$$

where the coefficients ϕ_{tj}, $j = 1, \ldots, p$, and θ_{tj}, $j = 1, \ldots, q$, depend on a vector of parameters β and ε_t is like above.

To better understand the relation between the TARMA and the tdARMA models, we introduce the following notation: let $\mathscr{I}_{t-d}^{(i)}$ be a short form for the indicator function $\mathscr{I}_{\{Y_{t-d} \in \mathscr{R}_i\}}$, such that $\mathscr{I}_{t-d}^{(i)} = 1$ if $y_{t-d} \in \mathscr{R}_i$ and $\mathscr{I}_{t-d}^{(i)} = 0$ otherwise, for $i = 1, \ldots, k$, model (1) can be written as (2) where

$$\phi_{tj}(\beta) = \sum_{i=1}^{k} \phi_j^{(i)} \mathscr{I}_{t-d}^{(i)}, \quad j = 1, \ldots, p,$$

$$\theta_{tj}(\beta) = \sum_{i=1}^{k} \theta_j^{(i)} \mathscr{I}_{t-d}^{(i)}, \quad j = 1, \ldots, q,$$

and $\beta = (\phi_1^{(1)}, \cdots, \phi_p^{(k)}, \theta_1^{(1)}, \cdots, \theta_q^{(k)})$, with $\beta \in \mathbf{B}$ an open set of a Euclidean space $\mathbf{R}^{(p+q)k}$ and let β_0 (an interior point of $\mathbf{B}$) be the corresponding vector of the true parameters. Let $e_t(\beta)$ be the residual defined iteratively by

$$X_t = \sum_{i=1}^{k} \left[\sum_{j=1}^{p} \phi_j^{(i)} X_{t-j} - \sum_{j=1}^{q} \theta_j^{(i)} e_{t-j}(\beta) \right] \mathscr{I}_{\{Y_{t-d} \in \mathscr{R}_i\}} + e_t(\beta), \tag{3}$$

for $t = 1, 2, \ldots$. Following [5] and the notation of [7], model (2) can be iteratively given as:

$$\mathbf{X}_t(\beta) = \sum_{r=0}^{t-1} \left[\sum_{s=0}^{r} \left(\prod_{\ell=0}^{r-s-1} \mathbf{J}\mathbf{A}_{t-\ell}(\beta) \right) \mathbf{K} \left(\prod_{j=0}^{s-1} \mathbf{A}_{t-r+s-j}(\beta_0) \right) \right] \mathbf{E}_{t-r}, \tag{4}$$

where:

$$\underset{[(p+q)\times 1]}{\mathbf{X}_t}(\beta) = \begin{bmatrix} X_t \\ \vdots \\ X_{t-p+1} \\ e_t(\beta) \\ \vdots \\ e_{t-q+1}(\beta) \end{bmatrix}, \quad \underset{[(p+q)\times(p+q)]}{\mathbf{A}_t(\beta)} = \begin{bmatrix} \Phi_t & \tilde{\Theta}_t \\ \underset{(q\times p)}{\mathbf{0}} & \mathbf{C} \end{bmatrix}, \quad \underset{[(p+q)\times 1]}{\mathbf{E}_t} = \begin{bmatrix} \varepsilon_t \\ \underset{[(p-1)\times 1]}{\mathbf{0}} \\ \varepsilon_t \\ \underset{[(q-1)\times 1]}{\mathbf{0}} \end{bmatrix},$$

with:

$$\Phi_t = \begin{bmatrix} \phi_{t1} & \phi_{t2} & \dots & \phi_{tp} \\ \mathbf{I} & & & \mathbf{0} \\ {\scriptstyle (p-1)} & & & {\scriptstyle (p-1)\times 1} \end{bmatrix}, \quad \tilde{\Theta}_t = \begin{bmatrix} \theta_{t1} & \theta_{t2} & \dots & \theta_{tq} \\ \mathbf{0} & & & \\ {\scriptstyle (q-1)\times q} & & & \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{0} & 0 \\ {\scriptstyle [1\times(q-1)]} & \\ \mathbf{I} & \mathbf{0} \\ {\scriptstyle (q-1)} & {\scriptstyle [(q-1)\times 1]} \end{bmatrix},$$

whereas $\mathbf{I}$ is the identity matrix, $\mathbf{0}$ a null vector or matrix, $\mathbf{K} = [k_{u,v}]$, for $u, v = 1, \dots, (p+q)$, is a null matrix with two elements replaced with $k_{1,1} = k_{(p+1),1} = 1$ and $\mathbf{J} = [j_{u,v}]$ is an identity matrix with two elements replaced with $j_{1,1} = 0$ and $j_{(p+1),1} = -1$.

Given these results, model (4) can be shortly written as:

$$\mathbf{X}_t(\beta) = \sum_{r=0}^{t-1} \Psi_{tr}(\beta)\mathbf{E}_{t-r}, \tag{5}$$

with $\Psi_{tr}(\beta) = \sum_{s=0}^{r} \left(\prod_{\ell=0}^{r-s-1} \mathbf{J}\mathbf{A}_{t-\ell}(\beta)\right) \mathbf{K} \left(\prod_{j=0}^{s-1} \mathbf{A}_{t-r+s-j}(\beta_0)\right)$.

Finally note that the results given in this section for the TARMA model with exogenous threshold variable can be applied to the case where the threshold variable is endogenous, $Y_{t-d} = X_{t-d}$, and, in our knowledge, it is a novelty that has not been considered in the cited literature. Note also that this section (and the next one) can be written also for a vector TARMA model (VTARMA), see e.g. [9].

3 MA representation of the TARMA model

The notation introduced in Section 2 allows obtaining the MA representation of $e_t(\beta)$. In fact, noting that $e_t(\beta)$ is the $(p+1)$-th element in $\mathbf{X}_t(\beta)$, then:

$$e_t(\beta) = \sum_{r=1}^{t-1} \psi_{tr}(\beta)\varepsilon_{t-r}. \tag{6}$$

with $\psi_{tr}(\beta) = \mathbf{U}'_{p+1}\Psi_{tr}(\beta)\mathbf{U}_1$, where $\mathbf{U}_1$ and $\mathbf{U}_{p+1}$ are two $[(p+q)\times 1]$ null vectors with the first and the $(p+1)$-th elements replaced with 1 respectively and $\mathbf{U}'_{p+1}$ the transpose of $\mathbf{U}_{p+1}$.

It is then possible to obtain the first three derivatives of $e_t(\beta)$ with respect to the elements of β, e.g. for the first-order derivative with respect to β_i, $i = 1, \dots, m$, where $m = (p+q)k$,

$$\frac{\partial e_t(\beta)}{\partial \beta_i} = \sum_{r=1}^{t-1} \psi_{tir}(\beta)\varepsilon_{t-r}, \tag{7}$$

where

$$\psi_{tir}(\beta) = \frac{\partial \psi_{tr}(\beta)}{\partial \beta_i}. \tag{8}$$

Then we let $\psi_{tir} = \psi_{tir}(\beta_0)$. Starting from here, we restrict the TARMA model to the case of a strictly exogenous variable Y_{t-d}. Otherwise, in the case of a TARMA with an endogenous threshold variable, the coefficients ψ_{tir} will be random variables.

Given the first three derivatives of $e_t(\beta)$ like in (7) and using the results in [5] and [7] for homoscedastic tdVARMA models, we can obtain, under the conditions A1-A6 in [7], see the Appendix, quasi maximum likelihood estimators for β whose asymptotic properties are derived.

These conditions can be checked under relatively weak assumptions on the TARMA model except A4. Indeed, A1 and A3 are trivially true, and it can be seen that A2, A5, and A6 are verified if the k ARMA models involved in (1) are stationary and invertible, by proceeding like in [7]. This is because the Frobenius norm of the products in (4) can then be bounded by a power of some constant $\Phi < 1$ by using properties of the eigenvalues of companion matrices.

It is unfortunately not possible to check (A4): it remains a condition. A natural requirement is that the k ARMA models involved in (1) are identifiable (so each of them has no common root for their autoregressive and moving average polynomials) but it is not enough to guarantee the existence and invertibility of the information matrix.

References

1. Azrak R., Mélard G.: The exact quasi-likelihood of time-dependent ARMA models. J Stat Plan Infer, **68**, 31–45 (1998)
2. Boubacar Maïnassara Y., Rabehasaina L.: Estimation of weak ARMA models with regime changes. Stat Infer Stoch Proc, **23**, 1–52 (2020)
3. Brockwell P.J., Liu J., Tweedie R.L.: On the existence of stationary threshold autoregressive moving-average processes. J Time Series Anal, **13**, 95–107 (1992)
4. Chan K.S., Goracci G. (2019): On the ergodicity of first-order threshold autoregressive moving-average processes. J Time Series Anal, **40**, 256–264 (2019)
5. Francq C., Gautier A.: Estimation of time-varying ARMA models with Markovian changes in regime, Stat Probabil Lett, **70**, 243–25 (2004)
6. Liu J., Susko E.: On strict stationarity and ergodicity of a non-linear ARMA model, J Appl Probabil, 29, pp. 363-373 (1992)
7. Mélard G. (2021), An indirect proof for the asymptotic properties of VARMA model estimators, Economet Stat, in press. https://doi-org/10.1016/j.ecosta.2020.12.004
8. Nicholls, D.F. and Quinn, B.G.: Random Coefficient Autoregressive Models: An Introduction, Springer-Verlag, New York (1982)
9. Niglio, M. and Vitale, C.D.: Threshold vector ARMA models, Commun Stat-Theor M, **44**, 2911–2923 (2015)
10. Tong, H.:Threshold Models in Non-linear Time Series Analysis, Springer-Verlag, New York (1983)

Appendix: the assumptions of [7]

.

We use the notations introduced in Sections 2 and 3. We consider a homoscedastic tdVARMA model of dimension ℓ (equal to $p+q$ above) and denote Σ the co-

variance matrix of the innovations E_t supposed to be invertible (this is not the case here).

We suppose

(A1) that the coefficient matrices $\Phi_{ti}(\beta)$ et $\Theta_{tj}(\beta)$ are of class C^3 in β in an open set B which contains the true value β_0;

(A2) denoting $\|.\|_F$ the Frobenius norm of a matrix, that there exist positive constants N_l, $l = 1,...,4$, and $0 < \Phi < 1$ such that, for $v = 1,...,t-1$, and $i = 1,...,m$, $\sum_{r=1}^{t-1} \|\Psi_{tr}\|_F^2 < N_1$, $\sum_{r=1}^{t-1} \|\Psi_{tr}\|_F^4 < N_2$, $\sum_{r=v}^{t-1} \|\Psi_{tir}\|_F^2 < N_3 \Phi^{v-1}$, $\sum_{r=v}^{t-1} \|\Psi_{tik}\|_F^4 < N_4 \Phi^{v-1}$ and three others for second and third order derivatives;

(A3) $\kappa = E\left(\mathrm{vec}(E_t E_t^T)\mathrm{vec}(E_t E_t^T)^T\right) = E\left((E_t E_t^T) \otimes (E_t E_t^T)\right)$ exists and does not depend on t;

(A4) the limits $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E_{\beta_0}\left(\frac{\partial E_t^T(\beta)}{\partial \beta_i} \Sigma^{-1} \frac{\partial E_t(\beta)}{\partial \beta_j}\right) = V_{ij}$ exist for $i, j = 1,...,m$, where the matrix $V = (V_{ij})_{1 \le i,j \le m}$ is strictly positive definite;

(A5) $\frac{1}{n^2} \sum_{d=1}^{n-1} \sum_{t=1}^{n-d} \sum_{r=1}^{t-1} \|\Psi_{tir}\|_F \left\|\Psi_{t+d,j,r+d}\right\|_F = O\left(\frac{1}{n}\right), \quad i, j = 1,...,m;$

(A6) Similarly

$$\frac{1}{n^2} \sum_{d=1}^{n-1} \sum_{t=1}^{n-d} \left[\sum_{r=1}^{t-1} M_{t0rr}^{jiT} \Xi(\Sigma) M_{tdrr}^{ij} + \sum_{r_1=1}^{t-1} \sum_{r_2=1}^{t-1} M_{t0r_2r_1}^{jiT} K_{\ell,\ell}(\Sigma \otimes \Sigma) M_{tdr_1r_2}^{ij} \right.$$
$$\left. + \sum_{r_1=1}^{t-1} \sum_{r_2=1}^{t-1} M_{t0r_2r_1}^{jiT} (\Sigma \otimes \Sigma) M_{tdr_2k_1}^{ij} \right] = O\left(\frac{1}{n}\right),$$

with a commutation matrix $K_{\ell,\ell}$, $\Xi(\Sigma) = \kappa - \mathrm{vec}(\Sigma).\mathrm{vec}(\Sigma)^T - (\Sigma \otimes \Sigma) - K_{\ell,\ell}(\Sigma \otimes \Sigma)$, and, for $r', r'' = r, r_1, r_2$, $M_{tfr'r''}^{ij} = \mathrm{vec}(\Psi_{t+f,i,r'+f}^T \Sigma^{-1} \Psi_{t+f,j,r''+f})$, $f = 0, d, i, j = 1,...,m$.

4.7 Bayesian nonparametrics

Bayesian nonparametric prediction: from species to features

La previsione in Statistica Bayesiana nonparametrica: modelli di specie e generalizzazioni

Lorenzo Masoero, Federico Camerlenghi, Stefano Favaro and Tamara Broderick

Abstract In species sampling models, observations represent the species' labels of distinct animals in a population. Feature sampling models generalize species sampling models by allowing every observation to belong to more than one species, now called features. In the present paper we review some results to face prediction in the species sampling framework via Bayesian nonparametric tools. We then move to introduce the most recent results on prediction in the context of feature models, first discussed in [17] and further developed here. We conclude with a discussion on possible future developments.

Abstract *Nei modelli di specie, le osservazioni rappresentano le specie degli animali all'interno di una popolazione. Un'importante generalizzazione dei modelli di campionamento di specie si ottiene quando ogni osservazione può appartenere a più specie contemporaneamete, che ora prendono il nome di caratteristiche. Nel presente lavoro, faremo una rassegna dei principali risultati sulla previsione nell'ambito dei modelli di campionamento di specie, con approccio di tipo Bayesiano nonparametrico. Successivamente, introdurremo i risultati più recenti sulla previsione per modelli di caratteristiche; tali risultati sono stati dimostrati in [17] e qui vengono ulteriormente sviluppati. Per concludere, delineeremo alcuni importanti problemi aperti.*

Lorenzo Masoero
Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA. e-mail: lom@mit.edu

Federico Camerlenghi
Department of Economics, Management and Statistics, University of Milano - Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy. e-mail: federico.camerlenghi@unimib.it

Stefano Favaro
Department of Economics and Statistics, University of Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy. e-mail: stefano.favaro@unito.it

Tamara Broderick
Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA. e-mail: tbroderick@csail.mit.edu

Key words: Prediction, exchangeability, species models, feature models, Indian Buffet process

1 Introduction

"Science cannot limit itself to theorize about accomplished facts but must foresee". As emphasized by B. de Finetti [6] here, one of the fundamental goals of Statistics consists in predicting the outcome of a certain experiment given n analogous observations. In the present paper, we discuss and review how the problem of prediction is performed in the Bayesian nonparametric literature for two distinct frameworks: species sampling models and feature sampling models. In particular we would like to show the most recent findings in the context of feature models. We also point out some important open problems.

To fix the notation, we consider a sequence of observations $Z_1, Z_2, \ldots$ defined on a common probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and taking values in a Polish space $\mathbb{Z}$, which is assumed to be endowed with a Borel σ-field $\mathscr{Z}$. We further indicate by $\mathsf{P}_{\mathbb{Z}}$ the set of all probability measures on the space $(\mathbb{Z}, \mathscr{Z})$. In species sampling models, each data point represents an individual and Z_i stands for the species' label of the ith individual. There are a lot of contributions to face prediction within the species framework, especially to estimate the number of new species that will appear in future samples. See, e.g., [9, 4, 18] for the frequentist approach, and [15, 20] for a Bayesian nonparametric framework.

Feature allocation models generalize species sampling models by allowing each individual to belong to more than one species, now called features. Each individual displays an unknown finite set of features selected from a countable collection, more precisely Z_i may be conveniently represented as a counting measure $Z_i = \sum_{k\geq 1} a_{k,i}\delta_{x_k}$, where $\{x_k\}_{k\geq 1}$ is the infinite collection of features while $a_{k,i} = 1$ if ith individual belongs to the kth feature, $a_{k,i} = 0$ otherwise. Feature models first appeared in ecology when animals are captured using traps and each observation is an incidence vector collecting the presence or absence of each species in the traps, see, e.g., [2] and [3]. These models found applications in other areas, such as in biosciences (see, e.g., [13]), in the analysis of choice behaviour arising from psychology, marketing and computer science [10], etc. See [1] and references therein. These models found important applications in genetics, where a feature is interpreted as a variant with respect to a reference genome. Researchers have developed a wide range of approaches for predicting the number of new features, often interpreted as amount of new genetic variation, in a follow-up study, when a pilot study is available. See, e.g., [12, 13, 21]. A full Bayesian nonparametric approach to face prediction in the context of feature models has been developed in [17].

In order to perform prediction, both for species and feature models, one has to assume a kind of similarity or analogy across data. In the Bayesian nonparametric literature, *exchangeability* is the simplest form of homogeneity across data. This notion has been introduced by de Finetti [7], and it amounts to assuming that the distri-

bution of the sequence $\{Z_i\}_{i\geq 1}$ is invariant under finite permutations of its elements. Thanks to the de Finetti representation theorem, this is tantamount to assuming that there exists a probability measure γ on $\mathsf{P}_{\mathbb{Z}}$ such that

$$\mathbb{P}(Z_1 \in A_1, \ldots, Z_n \in A_n) = \int_{\mathsf{P}_{\mathbb{Z}}} \prod_{i=1}^{n} p(A_i)\gamma(\mathrm{d}p) \tag{1}$$

for any $A_1, \ldots, A_n$ Borel sets in $\mathscr{Z}$ and for any $n \geq 1$. Equation (1) can also be rephrased as follows: $Z_i \overset{\text{iid}}{\sim} \tilde{p}$ where $\tilde{p}$ is a random probability measure with distribution γ.

The rest of the paper is structured as follows. In Section 2 we review some important results for prediction in species sampling model. Section 3 contains our most recent findings to face prediction for feature models. We conclude with a discussion, pointing out some interesting lines of research we are investigating.

2 Prediction for species sampling models

Lijoi et al. (2007) [15] developed a useful Bayesian nonparametric approach to face prediction in the context of species sampling models. Assume that $Z_1, \ldots, Z_n$ is a sample of size n, they faced the problem of estimating how many new species will be discovered in a future sample of arbitrary size. More precisely, denoting by $Z_{n+1}, \ldots, Z_{n+m}$ an additional sample of size m, in [15], the authors investigated the random quantity $K_m^{(n)}$, which counts the number of distinct species in $Z_{n+1}, \ldots, Z_{n+m}$ that have not been recorded in $Z_1, \ldots, Z_n$. The authors focused on a very large class of priors, called Gibbs-type priors [8]; among them we find the noteworthy Pitman–Yor process [19]. The Pitman-Yor random probability measure $\tilde{p}$ is characterized by two parameters $\sigma \in (0,1)$ and $\theta > 0$, and it can be defined via a stick-breaking procedure. More precisely, it equals $\tilde{p} = \sum_{j\geq 1} \tilde{\pi}_j \delta_{Z_j^*}$, with

$$\tilde{\pi}_1 = V_1, \quad \tilde{\pi}_j = V_j \prod_{i=1}^{j-1}(1 - V_i) \text{ for } j \geq 2,$$

where the $(Z_j^*)_{j\geq 1}$'s are i.i.d. random variables taking values in $(\mathbb{Z}, \mathscr{Z})$, with common distribution P_0, and the V_i's are independent Beta random variables with parameters $(\theta + i\sigma, 1 - \sigma)$. In addition the sequences $(V_i)_{i\geq 1}$ and $(Z_i^*)_{i\geq 1}$ are assumed to be independent. We would like to recall the results of [15] for the Pitman–Yor case.

Proposition 1. *Under a Pitman–Yor model, the posterior distribution of $K_m^{(n)}$ can be expressed as*

$$\mathbb{P}(K_m^{(n)} = k | Z_1, \ldots, Z_n) = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \cdot \frac{\prod_{i=j}^{j+k-1}(\theta + i\sigma)}{\sigma^k} \mathscr{C}(m, k; \sigma, -n + j\sigma),$$

where $k \in \{0,1,\ldots,m\}$, $\mathscr{C}(n,k;\sigma,\gamma) := \frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(-\sigma j-\gamma)_n$ *is the noncentral generalized factorial coefficient, and* j *is the number of distinct species out of the sample* $(Z_1,\ldots,Z_n)$.

Here the posterior distribution of $K_m^{(n)}$ is analytically tractable and available in closed form. Starting from the contribution of Lijoi et al. [15], prediction via Bayesian nonparametric tools has been faced for many other quantities of interest in the context of species sampling problems (see, e.g., [16] and [5]). It seems natural to extend these results to feature models: we will discuss prediction within the features' framework in the next section.

3 Prediction for feature models

Few results on prediction are available for feature sampling models from a Bayesian nonparametric viewpoint. In this case, we assume that the observations $\{Z_i\}_{i\geq 1}$ are modeled as Bernoulli processes, that is to say

$$Z_i := \sum_{k\geq 1} a_{i,k}\delta_{\tilde{x}_k}, \tag{2}$$

and the $a_{i,k}$'s in (2) are independent Bernoulli random variables with parameter $\tilde{\tau}_k$, i.e., $a_{i,k}|\tilde{\mu} \overset{\text{ind}}{\sim} \text{Bern}(\tilde{\tau}_k)$, conditionally on the realization of a measure $\tilde{\mu}$, which works as a prior distribution. The random measure $\tilde{\mu}$ is supposed to be almost surely discrete, i.e., $\tilde{\mu} = \sum_{k\geq 1}\tilde{\tau}_k\delta_{\tilde{x}_k}$. We assume that $\tilde{\mu}$ is a Beta process, i.e. $\tilde{x}_k \overset{\text{iid}}{\sim} \text{Unif}(0,1)$ and the $\tilde{\tau}_k$s are drawn from a Poisson point process with intensity

$$\rho(s) = \alpha\frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}s^{-\sigma-1}(1-s)^{c+\sigma-1}\mathbf{1}_{(0,1)}(s) \tag{3}$$

where $\alpha > 0$, $c > -\sigma$ and $\sigma \in (0,1)$. See [11]. The predictive distribution for this model has been nicely investigated by James (2017) [14] in a more general framework. However James (2017) [14] focused on singular predictive inference, i.e. on the outcome of the next observation Z_{n+1} given the past $Z_1,\ldots,Z_n$. Here and in [17], we are able to investigate the more general situation of m-step ahead prediction, i.e., to predict the outcome of the next m observations $Z_{n+1},\ldots,Z_{n+m}$, given the initial sample $Z_1,\ldots,Z_n$, where $m \geq 1$ is arbitrary. Here we would like to investigate three different statistics:

- $K_m^{(n)}$: the total number of hitherto unseen features that will be selected in the additional sample of arbitrary size m;
- $M_{m,r}^{(n)}$: the number of new features out of $Z_{n+1},\ldots,Z_{n+m}$ that have been observed with frequency r in the additional sample;
- $O_{m,j}^{(n)}$: the number of features out of $Z_{n+1},\ldots,Z_{n+m}$ coinciding with the j-th distinct feature, say x_j, observed in the initial sample $Z_1,\ldots,Z_n$.

It is possible to prove the following.

Theorem 1. *Let $\{Z_i\}_{i\geq 1}$ be a Bernoulli process, whose prior is a Beta process $\tilde{\mu}$. Then we have the following distributional results:*

$$K_m^{(n)}|Z_1,\ldots,Z_n \sim \text{Poiss}\left(\alpha\sum_{h=1}^{m}\frac{(c+\sigma)_{n+h-1}}{(c+1)_{n+h-1}}\right),$$
$$M_{m,r}^{(n)}|Z_1,\ldots,Z_n \sim \text{Poiss}\left(\alpha\binom{m}{r}\frac{(1-\sigma)_{r-1}(c+\sigma)_{n+m-r}}{(c+1)_{n+m-1}}\right),$$
$$O_{m,j}^{(n)}|Z_1,\ldots,Z_n \sim \text{BeB}(m;m_j-\sigma,n-m_j+c+\sigma),$$

where Poiss *denotes the Poisson distribution, whereas* BeB *is the Beta-Binomial distribution, i.e., a Binomial random variable with parameters $(m,\tilde{\pi})$, and $\tilde{\pi}$ is the random probability of success, distributed as* $\text{Beta}(m_j-\sigma,n-m_j+c+\sigma)$.

The previous results concerning $K_m^{(n)}$ and $M_{m,r}^{(n)}$ have been presented in [17] to face prediction of genetic variants, while the distributional result on $O_{m,j}^{(n)}$ is new and it can be proved by exploiting the probability generating function of $O_{m,j}^{(n)}$. It is also possible to determine asymptotic properties for all the statistics discussed there. See [17] for details on $K_m^{(n)}$ and $M_{m,r}^{(n)}$. Moreover, we are able to study the asymptotic behavior of $O_{m,j}^{(n)}$, in particular one can prove that $O_{m,j}^{(n)}/m|Z_1,\ldots,Z_n \xrightarrow{a.s.} \tilde{\pi}$, as $m\to+\infty$, where $\tilde{\pi}\sim\text{Beta}(m_j-\sigma,n-m_j+c+\sigma)$; finally $O_{m,j}^{(n)}$ satisfies the following central limit theorem

$$\frac{O_{m,j}^{(n)}-m\tilde{\pi}}{\sqrt{m\tilde{\pi}(1-\tilde{\pi})}}\Big|Z_1,\ldots,Z_n \xrightarrow{d} \text{N}(0,1) \quad \text{as } m\to+\infty.$$

4 Discussion

In the present paper we discussed how prediction problems are performed via Bayesian nonparametric tools. We briefly reviewed the literature for species sampling models to introduce the new results we have found in the context of feature models (see Theorem 1). The Bayesian nonparametric framework is pretty natural to face m-step ahead prediction, thus solving a fundamental problem of Statistics.
The results presented in Section 3 are particularly useful in many applied context, such as in genetics (see [17]). We finally point out that the distribution of $K_m^{(n)}$, in the context of features, depends on the initial sample $Z_1,\ldots,Z_n$ only via the sample size n, while, in the context of species, $K_m^{(n)}$ is a function of both n and the number of distinct values appearing in the sample, i.e., j. Thus, a question arises: in the context of features, is it possible to enrich the predictive structure, exploiting the whole information contained in the sample to perform prediction? This is a question we

are trying to answer by choosing a different prior for $\tilde{\mu}$. Finally, the problem of prediction in presence of multiple-sample information is still open and merits further investigation. In genetics, e.g., this is useful to deal with observations coming form different tissues of the same organism.

References

1. Ayed, F., Battiston, M., Camerlenghi, F., Favaro, S.: A Good-Turing estimator for feature allocation models. Electron. J. Statist., **13**, 3775-3804 (2019)
2. Colwell, R., Chao, A., Gotelli, N.J., Lin, S., Mao, C.X., Chazdon, R.L., Longino, J.T.: Models and estimators linking individual-based and sample–based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology, **5**, 3–21 (2012)
3. Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., Ellison, A.M.: Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecological Monographs, **84**, 45–67 (2014)
4. Efron, B., Thisted, R.: Estimating the number of unseen species: How many words did Shakespeare know? Biometrika, **63**, 435–447 (1976)
5. Favaro, S., Lijoi, A., Prünster, I.: Conditional formulae for Gibbs-type exchangeable random partitions. Ann. Appl. Probab., **23**, 1721–1754 (2013)
6. de Finetti, B.: Probabilismo. Logos **14**, 163–219 (1931)
7. de Finetti, B.: La prévision: ses lois logiques, ses sources subjectives. Ann. Inst. H. Poincaré, **7**, 1–68 (1937)
8. Gnedin, A., Pitman, J.: Exchangeable Gibbs partitions and Stirling triangles. Zap. Nauchn. Sem. POMI, **325**, 83–102 (2005)
9. Good, I.J.: On the population frequencies of species and the estimation of population parameters. Biometrika, **40**, 237–264 (1953)
10. Görür, D., Jäkel, F., Rasmussen, C.E.: A choice model with infinitely many latent features. 23rd International Conference on Machine Learning (2006)
11. Ghahramani, Z., Griffiths, T.L.: Infinite latent feature models and the Indian buffet process. In: Advances in Neural Information Processing Systems, pp. 475–482 (2006)
12. Gravel, S.: Predicting discovery rates of genomic features. Genetics, **197**, 601–610 (2014)
13. Ionita-Laza, I., Lange, C., Laird, N.M.: Estimating the number of unseen variants in the human genome. Proc. Natl. Acad. Sci., **106**, 5008–5013 (2009)
14. James, L.F.: Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors. Ann. Statist., **45**, 2016–2045 (2017)
15. Lijoi, A., Mena, R.H., Prünster, I.: Bayesian nonparametric estimation of the probability of discovering new species. Biometrika, **94**, 769–786 (2007)
16. Lijoi, A., Prünster, I., Walker, S.G.: Bayesian nonparametric estimators derived from conditional Gibbs structures. Ann. Appl. Probab., **18**, 1519–1547 (2008)
17. Masoero, L., Camerlenghi, F., Favaro, S., Broderick, T.: More for less: Predicting and maximizing genetic variant discovery via Bayesian nonparametrics. Biometrika, doi:10.1093/biomet/asab012 (2021)
18. Orlitsky, A., Suresh, A.T., Wu, Y.: Optimal prediction of the number of unseen species. Proc. Natl. Acad. Sci., **113**, 13283–13288 (2016)
19. Pitman J., Yor M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab., **25**, 855–900 (1997)
20. Rodríguez, A., Quintana, F.A.: On species sampling sequences induced by residual allocation models. J. Stat. Plan. Inference, **157-158**, 108–120 (2015)
21. Zou, J., Valiant, G., Valiant, P., Karczewski, K., Chan, S.O., Samocha, K., Lek, M., Sunyaev, S., Daly, M., MacArthur, D.G.: Quantifying the unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. Nature Communications, **7**, 13293 (2016)

A framework for filtering in hidden Markov models with normalized random measures

Filtraggio di modelli di Markov nascosti in presenza di misure aleatorie normalizzate

Filippo ASCOLANI, Antonio LIJOI, Igor PRÜNSTER and Matteo RUGGIERO

Abstract The vast majority of explicitly available posterior characterizations in Bayesian nonparametrics refer to the exchangeable case, a restrictive assumption for time-dependent phenomena. Alternative formulations that accommodate partial exchangeability include hidden Markov models (HMMs), where the exact derivation of the posterior distribution given data collected at past times (*optimal filtering*) remains a challenging task. Here we outline a general framework based on duality for the analysis of HMMs which feature normalized random measures. The posterior tractability is ensured by combining certain projective properties of the infinite-dimensional distributions involved with the existence of a suitable duality relation between the hidden signal and an appropriate death process. Under these conditions, the filtering distributions are all finite mixtures, paving the way for closed form inferential strategies.

Abstract *I risultati analitici disponibili nel campo della statistica Bayesiana nonparametrica riguardano perlopiù il caso scambiabile, in cui la distribuzione deli dati è invariante rispetto a permutazioni finite. Poichè questa assunzione non è adatta per studiare fenomeni dinamici, sono state proposte molte alternative per i modelli di Markov nascosti. In questo lavoro proponiamo una classe generale che è dotata di grande trattabilità analitica, grazie a una relazione di dualità tra il segnale nascosto e un opportuno processo di Markov.*

Filippo ASCOLANI
Bocconi University and BIDSA, e-mail: filippo.ascolani@phd.unibocconi.it

Antonio LIJOI
Bocconi University and BIDSA, e-mail: antonio.lijoi@unibocconi.it

Igor PRÜNSTER
Bocconi University and BIDSA,e-mail: igor.pruenster@unibocconi.it

Matteo RUGGIERO
University of Torino and Collegio Carlo Alberto, e-mail: matteo.ruggiero@unito.it

Key words: Bayesian nonparametrics, hidden Markov models, duality, optimal filtering, completely random measures, partially observed Markov processes

1 Completely random measures and hidden Markov models

The standard Bayesian nonparametric specification for exchangeable data is

$$Y_i \mid X \overset{\text{iid}}{\sim} X, \quad X \sim \Pi, \tag{1}$$

where observations Y_i live in a Polish space $\mathbb{Y}$ with Borel sigma algebra $\mathscr{Y}$, and Π is a distribution on the space $\mathscr{P}_{\mathbb{Y}}$ of probability measures on $(\mathbb{Y},\mathscr{Y})$ that plays the role of the prior. A general approach to construct such priors considers suitable transformations of completely random measures (CRMs); see [5, 6]. Denote by $(\Omega,\mathscr{F},\mathbb{P})$ a probability space and by $\mathbb{M}_{\mathbb{Y}}$ the space of boundedly finite measures on $(\mathbb{Y},\mathscr{Y})$, with corresponding Borel sigma algebra $\mathscr{M}_{\mathbb{Y}}$.

Definition 1. A measurable function μ from $(\Omega,\mathscr{F},\mathbb{P})$ to $(\mathbb{M}_{\mathbb{Y}},\mathscr{M}_{\mathbb{Y}})$, is a *completely random measure* if, for any disjoint collection $A_1,\ldots,A_n \in \mathscr{Y}$, the random variables $\mu(A_1),\ldots,\mu(A_n)$ are mutually independent.

In this work we focus on CRMs without deterministic drift and fixed points, so that μ can be characterized through the *Lévy-Khintchine* representation of its Laplace transform

$$\mathbb{E}\left[e^{-\lambda\mu(A)}\right] = e^{-P_0(A)\psi(\lambda)}, \quad \psi(\lambda) := \int_{\mathbb{R}_+}\int_{\mathbb{Y}}(1-e^{-\lambda s})\,\nu(\mathrm{d}s,\mathrm{d}y), \quad A \in \mathscr{Y}, \lambda > 0,$$

where ν is a measure on $\mathbb{R}_+ \times \mathbb{Y}$ satisfying $\int_{\mathbb{R}_+}\int_A \min\{1,s\}\,\nu(\mathrm{d}s,\mathrm{d}y) < \infty$, $A \in \mathscr{Y}$, called *Lévy intensity*, and $\psi(\lambda)$ is the *Laplace exponent*. Here we will focus on *homogeneous* intensities $\nu(\mathrm{d}s,\mathrm{d}y) = \rho(\mathrm{d}s)\alpha(\mathrm{d}y)$, where ρ is a measure on $\mathbb{R}_+$ and the intensity of jumps does not depend on the jump location, and we further assume α is a finite non-atomic measure on $\mathbb{Y}$, with normalized version $P_0(\cdot) = \alpha(\cdot)/\alpha(\mathbb{Y})$. Typical examples are the *gamma* process, whereby $\nu(\mathrm{d}s,\mathrm{d}y) = s^{-1}e^{-s}\mathrm{d}s\,\alpha(\mathrm{d}y)$, and the σ-*stable* process, whereby $\nu(\mathrm{d}s,\mathrm{d}y) = (\Gamma(1-\sigma))^{-1}\sigma s^{-1-\sigma}\mathrm{d}s\,\alpha(\mathrm{d}y)$, for $0 < \sigma < 1$.

The following class of models is due to [9].

Definition 2. Let μ be a CRM such that $0 < \mu(\mathbb{Y}) < \infty$ almost surely. Then $p(\cdot) = \mu(\cdot)/\mu(\mathbb{Y})$ is called *normalized random measure with independent increments* (NRMI).

In the following we will further assume $\nu(\mathbb{R}_+ \times \mathbb{Y}) = \infty$ and $\psi(\lambda) < \infty$ for any positive λ, that guarantee almost sure finiteness and positivity of $\mu(\mathbb{Y})$. The class of NRMIs is large and encompasses many priors of interest: for example, the well-known Dirichlet process can be defined as a normalized gamma process. Moreover,

the class of NRMIs enjoys a certain degree of analytical tractability, that makes posterior inference feasible (see [4]). In addition, a NRMI p admits the representation as discrete measure

$$p = \sum_{j \geq 1} W_j \delta_{Z_j}, \quad Z_j \overset{\text{iid}}{\sim} P_0,$$

with the weights $\{W_j\}$ independent from the locations $\{Z_j\}$. A sample $(Y_1, \ldots, Y_n)$ will therefore yield ties with positive probability, and can be represented by the unique observed values $(Y_1^*, \ldots, Y_k^*)$ with associated multiplicities $\mathbf{m} = (m_1, \ldots, m_k) \in \mathbb{Z}_+^k$.

A natural extension of (1), that accommodates a partially exchangeable framework where observations are collected at different times $0 = t_0 < t_1 < \ldots$, is given by

$$Y_{t_n}^i \mid X_{t_n} \overset{\text{iid}}{\sim} X_{t_n}, i = 1, \ldots, n_{t_n}, \quad X \sim Q, \tag{2}$$

where $X = \{X_t : t \geq 0\}$ and Q is the law of a stochastic process indexed by $\mathbb{R}_+$ with state space $\mathscr{P}_{\mathbb{Y}}$. If X_t is a Markov process, then (2) is a *hidden Markov model* (HMM) (see [3]), and we denote by Q_0 the initial distribution of the process and by P_t its transition function. For brevity, let $Y_k := Y_{t_k}$ and $Y_{0:n} := (Y_0, \ldots, Y_n)$. The main objects of interest are the *update* operator $\mathscr{U}_Y$, that returns the posterior distribution given observations Y at a fixed time, and the *prediction* operator, defined as applied to measures ξ by

$$\mathscr{P}_t(\xi)(\mathrm{d}x') = \int \xi(\mathrm{d}x) P_t(x, \mathrm{d}x'). \tag{3}$$

The so-called *filtering distribution* $P_n := \mathscr{L}(X_{t_n} \mid Y_{0:n})$ can then be obtained recursively by considering $P_0 = \mathscr{U}_{Y_0}(Q_0)$ and, for $n \geq 1$, $P_n = \mathscr{U}_{Y_n}\left(\mathscr{P}_{t_n - t_{n-1}}(P_{n-1})\right)$.

2 A general framework for optimal filtering

We provide general requirements on Q that lead to computing explicitly the filtering distributions. Let $X_t^K := (X_t(A_1), \ldots, X_t(A_K))$ be the projection of X_t over an arbitrary measurable partition $A_1, \ldots, A_K$ of $\mathbb{Y}$.

We make the following assumptions:

A1 Q_0 is induced by a NRMI with Lévy intensity ν.

A2 Q is such that $\mathscr{L}\left(X_t^K \mid X_0 = x\right) = \mathscr{L}\left(X_t^K \mid X_0^K = x^K\right)$, for any partition of K elements.

A3 Denoting by π^K the distribution of X_0^K and by P_t^K the transition function induced by Q relative to the partition, we assume π^K is reversible with respect to P_t^K, i.e. $\pi^K(\mathrm{d}x) P_t^K(x, \mathrm{d}x') = \pi^K(\mathrm{d}x') P_t^K(x', \mathrm{d}x)$. Moreover, we assume P_t^K admits a strictly positive transition density.

A4 The distribution of X_t^K given $(Y_t^1, \ldots, Y_t^n)$ has density $h(x^K, Y^*, \mathbf{m}) \pi^K(x^K)$ for a suitable function $h(\cdot)$, that depends on the unique values Y^* and multiplicities $\mathbf{m} \in \mathbb{Z}_+^k$. We assume X_t^K is *dual* to a time-homogeneous death process M_t on

$\mathbb{Z}_+^k$, that is

$$\mathbb{E}\left[h(X_t^K, Y^*, \mathbf{m}) \mid X_0^K = x^K\right] = \mathbb{E}\left[h(x^K, Y^*, M_t) \mid M_0 = \mathbf{m}\right].$$

We denote by $p_{\mathbf{m},\mathbf{n}}(t)$ the transition probabilities of M_t, with $\mathbf{m} \in \mathbb{Z}_+^k$, $\mathbf{n} \in L(\mathbf{m})$ and $L(\mathbf{m}) = \{\mathbf{n} \mid \mathbf{n} \leq \mathbf{m}\}$, where $\mathbf{n} \leq \mathbf{m}$ if $n_j \leq m_j$ for any $j = 1, \ldots, k$.

Example 1. Define the transition $P_t(x, \mathrm{d}x') = e^{-\beta t}\delta_x(\mathrm{d}x') + (1 - e^{-\beta t})Q_0(\mathrm{d}x')$. Then A1–A3 are immediately verified, while A4 reads

$$\mathbb{E}\left[h(X_t^K, Y^*, \mathbf{m}) \mid X_0^K = x^K\right] = e^{-\beta t}h(x^K, Y^*, \mathbf{m}) + 1 - e^{-\beta t},$$

so that $p_{\mathbf{m},\mathbf{m}}(t) = 1 - p_{\mathbf{m},\mathbf{0}}(t) = e^{-\beta t}$.

Example 2. The HMM induced by the Fleming–Viot process (see [8]) satisfies A1–A4 with Q_0 being the law of a Dirichlet process. Indeed, even if its transition function is known up to an infinite series, [7] proved that the projections are dual to a pure death process with rates $(m_j/2)(\alpha(\mathbb{Y}) + |\mathbf{m}| - 1)$, $|\mathbf{m}| = \sum_{j=1}^k m_j$, for jumping from $\mathbf{m}$ to $\mathbf{m} - \mathbf{e}_j$, where $\mathbf{e}_j$ denotes the vector of all zeroes except the j-th element. See [2] for an investigation of the predictive properties of this model.

Notice that, except for A1, the above requirements regard the finite-dimensional projections, typically more tractable especially in terms of transition functions. In this respect, note that NRMIs includes three classes of random measures for which the distribution of the projections is known explicitly (Dirichlet, normalized inverse-Gaussian and normalized stable processes).

3 Main results

In this Section we show how the tractability of NRMIs can be combined with duality in A4 to prove explicit a priori and a posteriori properties.

3.1 Prior properties

The first result shows that the invariance property of the projections extend to the distribution Q_0 itself.

Proposition 1. *Consider* (2) *with Q satisfying A1–A4. Then Q_0 is the invariant measure for the stochastic process with transition P_t.*

Proof. It follows from A3, since $\mathscr{L}(X_t^K)(\mathrm{d}x') = \int \pi^K(\mathrm{d}x)\, P_t^K(\mathrm{d}x, \mathrm{d}x') = \pi^K(\mathrm{d}x')$ and the fact that random measures are characterized by their finite dimensional distributions. □

Hence if $X_0 \sim Q_0$, $X_t \sim Q_0$ as well, and before conditioning on the data the same Bayesian nonparametric model for exchangeable data as in (1) is propagated to each time t.

Since Q is the law of a collection of random probability measures, one is immediately interested in the support properties. The *weak support* is the smallest closed set in the Borel sigma algebra $\mathscr{B}\{\mathscr{P}_{\mathbb{Y}}^{\mathbb{R}_+}\}$, generated by the product topology of weak convergence, and can be seen as a measure of flexibility: indeed, each neighborhood of an element of the support has positive probability under Q. The next proposition shows that our proposal yields a full weak support, relative to the support of P_0.

Proposition 2. *Let $\mathbb{S}$ be the support of P_0. Then the weak support of a model satisfying A1–A4 is given by $\mathscr{P}_{\mathbb{Y}}^{\mathbb{R}_+}(\mathbb{S})$.*

Proof. The marginal law Q_0 has weak support $\mathscr{P}_{\mathbb{Y}}(\mathbb{S})$. Then the result follows as for Proposition 1 in [2] using A2. □

When dealing with temporal data, it is often of interest to quantify the dependence between measures at different times. A simple way consists in using the correlation between the observables, that in this case can be computed exactly, as the next result highlights.

Proposition 3. *Consider* (2) *with Q satisfying A1–A4. Then*

$$Corr\left(Y_t^i, Y_{t+s}^j\right) = -p_{1,1}(s) \int_{\mathbb{R}_+} u \left\{ \frac{d^2}{du^2} \psi(u) \right\} e^{-\psi(u)}\, du.$$

In particular $Corr\left(Y_t^i, Y_{t+s}^j\right) \geq 0$ for any t and s.

Proof. Since X_t is almost surely discrete we have $\mathrm{Corr}(Y_t^i, Y_{t+s}^j) = \mathbb{P}(Y_t^i = Y_{t+s}^j) = p_{1,1}(s)\mathbb{P}(Y_t^i = Y_t^j)$. The latter is recovered from a reasoning similar to Proposition 2 in [4]. □

Considering for instance Example 1, with Q_0 being the law of a Dirichlet process, the formula reduces to $\mathrm{Corr}(Y_t^i, Y_{t+s}^j) = e^{-\beta t}/(\alpha(\mathbb{Y}) + 1)$.

3.2 Posterior properties

As shown in Proposition 1, at each fixed time t we have a sampling model as in (1) with the marginal law Q_0 in place of Π. We denote by $H(\cdot \mid \mathbf{m})$ the associated posterior distribution given data Y, with unique values $Y_1^*, \ldots, Y_k^*$ and multiplicities $\mathbf{m}$. In the notation of Section 1, $\mathscr{U}_Y(Q_0)(\mathrm{d}x) = H(\mathrm{d}x \mid \mathbf{m})$.

The next result shows that the prediction operator yields a finite mixture of such distributions.

Theorem 1. *Consider model* (2) *with Q satisfying A1–A4. Then*

$$\mathscr{P}_t\left(H(dx,\boldsymbol{m})\right) = \sum_{\boldsymbol{n}\in L(\boldsymbol{m})} p_{\boldsymbol{m},\boldsymbol{n}}(t) H(dx,\boldsymbol{n})$$

Proof. Given an arbitrary partition, from A2 and A3 we have $\mathscr{L}(X_t^K \mid \mathbf{m})(\mathrm{d}x') = \int P_t^K(x,\mathrm{d}x')h(\mathrm{d}x,Y^*,\mathbf{m})\pi^K(\mathrm{d}x) = \pi^K(\mathrm{d}x')\mathbb{E}\left[h(X_t^K,Y^*,\mathbf{m}) \mid X_0^K = x'\right]$. The result now follows from A4. □

For instance, in the case of Example 1, it reads $\mathscr{P}_t\left(H(\mathrm{d}x,\mathbf{m})\right) = e^{-\beta t}H(\mathrm{d}x,\mathbf{m}) + (1-e^{-\beta t})Q_0(\mathrm{d}x)$.

Thanks to linearity of the prediction operator, the filtering distributions can be derived explicitly.

Theorem 2. *Consider* (2) *with Q satisfying A1–A4. Given unique values* $Y_1^*,\ldots,Y_k^*$ *it holds*

$$\mathscr{L}(X_0 \mid Y_0) = H(dx \mid \boldsymbol{n}_0),$$

with $\boldsymbol{n}_0$ *multiplicities of* Y_0*. Moreover, there exist* $M_n \subset \mathbb{Z}_+^k$ *and weights* $w_{\boldsymbol{n}}$ *such that*

$$\mathscr{L}\left(X_{t_n} \mid Y_{0:n}\right) = \sum_{\boldsymbol{n}\in M_n} w_{\boldsymbol{n}} H(dx \mid \boldsymbol{n}).$$

Proof. Since prediction operator (3) is linear, we apply the same reasoning of Proposition 2.3 in [7]. □

Since Q_0 is a NRMI, the posterior distribution $H(\cdot \mid \mathbf{n})$ is analytically tractable, at least conditionally to a suitable latent variable (see Theorem 2 in [4]). Thus, thanks to the finiteness of the mixture, devising conditional or marginal algorithms for sampling becomes a feasible operation. The results will be detailed and developed in [1].

References

1. Ascolani, F., Lijoi, A., Prünster, I. and Ruggiero, M.: Optimal filtering for hidden Markov models featuring normalized random measures. Work in progress.
2. Ascolani, F., Lijoi, A. and Ruggiero, M.: Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. Bayesian Anal., in press (2021)
3. Cappé, O., Moulines, E. and Ryden, T.: Inference in Hidden Markov Models. Springer, New York (2005)
4. James, L. F., Lijoi, A. and Prünster, I.: Posterior analysis for normalized random measures with independent increments. Scand. J. Stat. **36**(1), 76–97 (2009)
5. Kingman, J.F.C.: Completely Random Measures. Pacif. J. Math. **21**, 59–78 (1967)
6. Lijoi, A. and Prünster, I.: Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, pp. 80–130, Cambridge Univ. Press, Cambridge (2010)
7. Papaspiliopoulos, O. and Ruggiero, M.: Optimal filtering and the dual process. Bernoulli. **20**(4), 1999–2019 (2014)
8. Papaspiliopoulos, O., Ruggiero, M. and Spanó, D.: Conjugacy properties of time-evolving Dirichlet and gamma random measures. Electron. J. Stat. **10**(2), 3452–3489 (2016)
9. Regazzini, E., Lijoi, A. and Prünster, I.: Distributional results for means of normalized random measures with independent increments. Ann. Stat. **31**(2), 560–585 (2003)

On the convex combination of a Dirichlet process with a diffuse probability measure

Sulla combinazione convessa di un processo di Dirichlet con una misura di probabilità diffusa

Camerlenghi Federico and Corradin Riccardo and Andrea Ongaro

Abstract The Dirichlet process is a discrete nonparametric prior, widely used in Bayesian nonparametrics. As a consequence of the almost sure discreteness, a sample from the Dirichlet process may display ties with positive probability, and each observed value has always positive probability to be observed again. In some real problems, this property may be too restrictive or unrealistic. We propose a convex combination of a Dirichlet process with a diffuse probability measure to overcome this possible limitation. We finally discuss an application to text data.

Abstract *Il processo di Dirichlet è una misura di probabilità aleatoria largamente usata in ambito Bayesiano nonparamterico. Dal momento che esso è un processo quasi certamente discreto, un campione estratto da un processo di Dirichlet può contenere osservazioni uguali con probabilità positiva, inoltre ogni valore del campione ha sempre una probabilità positiva di essere riosservato. In talune applicazioni, quest'ultima proprietà potrebbe essere troppo restrittiva od irrealistica. Di conseguenza, per superare questa limitazione, proponiamo l'uso di una combinazione convessa tra un processo di Dirichlet ed una misura di probabilità diffusa. Per concludere, discuteremo un'applicazione del modello a dati testuali.*

Key words: Dirichlet process, Bayesian nonparametrics, Hapax legomena, Unique values, Convex combination

Camerlenghi Federico
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy. e-mail: federico.camerlenghi@unimib.it

Corradin Riccardo
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy. e-mail: riccardo.corradin@unimib.it

Ongaro Andrea
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy. e-mail: andrea.ongaro@unimib.it

1 Introduction

The Dirichlet Process (DP) is a combinatorial stochastic process, early introduced by Ferguson [3]. Due to its mathematical tractability and flexibility, it is commonly used in a large variety of applications, and nowadays its distribution is a natural choice to model the prior opinion in Bayesian nonparamterics.

Let $\mathbb{X}$ be a Polish space, endowed with its Borel σ-field $\mathscr{X}$. Let $Q_0(\cdot)$ be a diffuse probability measure defined on the measurable space $(\mathbb{X}, \mathscr{X})$. There are several equivalent definitions and representations of the DP, the stick-breaking construction is arguably one of the most popular representations in terms of computational convenience. More precisely a DP is an almost surely discrete random probability measure $\tilde{p}$ which equals

$$\tilde{p}(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{X_j}(\cdot),$$

where $\{p_j\}_{j\geq 1}$ is a sequence of random weights with $\sum_{j=1}^{\infty} p_j = 1$ a.s., and $\{X_j\}_{j\geq 1}$ is a sequence of random elements i.i.d. from $Q_0(\cdot)$, where $\{p_j\}_{j\geq 1}$ is independent of $\{X_j\}_{j\geq 1}$. Moreover the distribution of the sequence of weights $\{p_j\}_{j\geq 1}$ is described by the one-parameter GEM distribution, i.e. $\tilde{p}_1 := V_1$ and $\tilde{p}_j := V_j \prod_{r=1}^{j-1}(1-V_r)$, as $j \geq 2$, where $V_j \overset{iid}{\sim} \text{Beta}(1,\alpha)$ and $\alpha > 0$. In the sequel, we will write $\tilde{p} \sim DP(\alpha, Q_0(\cdot))$, where α is termed mass parameter of the process, and it plays the role of a concentration parameter, and $Q_0(\cdot)$ is the centering measure. For an extensive discussion on the DP and its properties, see e.g. [4].

If we now consider a sample from a Dirichlet process, i.e. $X_1, \ldots, X_n | \tilde{p} \overset{iid}{\sim} \tilde{p}$ and $\tilde{p} \sim DP(\alpha, Q_0(\cdot))$, thanks to the almost sure discreteness of $\tilde{p}$, one may observe ties among the X_is. This is appreciable from the predictive distribution of a DP, which describes a generalized Pólya Urn sampling scheme [1]. In order to have a glimpse on this, we will denote by $X_1^*, \ldots, X_k^*$ the $k \leq n$ distinct values out of the sample $(X_1, \ldots, X_n)$, and by $n_1, \ldots, n_k$ the respective frequencies. The distribution of the X_{n+1} element, given $X_1, \ldots, X_n$ is equal to

$$P(X_{n+1} \in \text{A} \mid X_1, \ldots, X_n) = \frac{\alpha}{\alpha+n} Q_0(A) + \sum_{j=1}^{k} \frac{n_j}{\alpha+n} \delta_{X_j^*}(A), \quad A \in \mathscr{X},$$

from which it is apparent that the jth unique value X_j^* has positive probability, which equals $n_j/(\alpha+n)$, to be sampled again at the $(n+1)$th sampling step. In some specific application, there are observations which appear only once in the sample and cannot be observed again. Thus, in Section 2, we introduce a suitable modification of the Dirichlet process in order to handle these specific situations. We also describe the distribution of the random partition induced by the data. Finally, in Section 3, we discuss how our proposal is useful in the context of text data.

2 Convex combination of the DP with a diffuse measure

We aim to derive an extension of the DP to the case where, out of a sample $X_1,\dots,X_n$, a subset of the realizations have null probability of being sampled again. To this end, we consider a convex combination of the DP with a diffuse probability measure $Q_0(\cdot)$, where $Q_0(\cdot)$ equals the centering measure of the DP. We then obtain a new process $\tilde{q}$ of the form

$$\tilde{q}(\cdot) = \beta\tilde{p}(\cdot) + (1-\beta)Q_0(\cdot). \tag{1}$$

When we consider $X_1,\dots,X_n|\tilde{q} \overset{iid}{\sim} \tilde{q}$, then each X_i can be sampled from $\tilde{p}$, with probability β, and from Q_0 with probability $1-\beta$. In the latter case, X_i is sampled from a diffuse measure and it cannot be selected again, i.e. it is observed only once in a sample of arbitrary size. We now denote by $X_1,\dots,X_k^*$ the unique values in the sample $(X_1,\dots,X_n)$ from $\tilde{q}$. We further indicate by τ_1 the latent number of elements sampled from the diffuse term $Q_0(\cdot)$. We can characterize the *exchangeable partition probability function* (EPPF) of $X_1,\dots,X_n$ and τ_1, which is the probability to observe a partition of the sample into k sets of distinct values with frequencies $n_1,\dots,n_k$. See [8] for further details. We are now able to describe the EPPF of the model in the following proposition.

Proposition 1 *Let $X_1,\dots,X_n$ be a sample form a random probability measure $\tilde{q}(\cdot) = \beta\tilde{p}(\cdot) + (1-\beta)Q_0(\cdot)$, with $\tilde{p} \sim DP(\alpha, Q_0(\cdot))$. The joint distribution of the partition induced by the observations and τ_1 equals*

$$\Pi^{(n)}_{k,\tau_1}(n_1,\dots,n_k) = \binom{k_1}{\tau_1}\beta^{n-\tau_1}(1-\beta)^{\tau_1}\frac{\alpha^{k-\tau_1}}{(\alpha)_{n-\tau_1}}\prod_{j=1}^{k}(n_j-1)!,$$

where $k_1 := \sum_{j=1}^{k}\mathbb{1}_{\{n_j=1\}}$ is the number of distinct observations that are observed only once, and $(a)_b := \Gamma(a+b)/\Gamma(a)$, with $a,b>0$, denotes the Pochhammer symbol.

Proof. Let $J_1,\dots,J_n$ be a set of suitable latent random variables, with $J_i = 1$ if the i-th observations belongs to the discrete component of $\tilde{q}$, $J_i = 0$ otherwise, and $\tau_1 = \sum_{i=1}^{n}\mathbf{1}_{[J_i=0]}$. We further denote by k_1 the number of unique values with frequency 1 out of a sample $X_1,\dots,X_n$. Without loss of generality, we assume the first k_1 observations $X_1,\dots,X_{k_1}$ be the unique elements observed only once. Then we have

$$\begin{aligned}\Pi^{(n)}_{k,\tau_1}(n_1,\dots,n_k) &= \int_{\mathbb{X}^k}\mathbb{E}\left[\prod_{j=1}^{k}\prod_{i=1}^{n_j}\beta^{J_i}\tilde{p}^{J_i}(\mathrm{d}X_i^*)(1-\beta)^{1-J_i}P_0^{1-J_i}(\mathrm{d}X_i^*)\right]\\ &= \int_{\mathbb{X}^k}\mathbb{E}\left[\left(\binom{k_1}{\tau_1}\prod_{j=1}^{\tau_1}(1-\beta)P_0(\mathrm{d}X_i^*)\right)\left(\prod_{j=\tau_1+1}^{k}\prod_{i=1}^{n_j}\beta\tilde{p}(\mathrm{d}X_i^*)\right)\right]\end{aligned}$$

$$= \binom{k_1}{\tau_1} \beta^{n-\tau_1} (1-\beta)^{\tau_1} \frac{\alpha^{k-\tau_1}}{(\alpha)_{n-\tau_1}} \prod_{j=1}^{k} (n_j - 1)!,$$

where the last equality holds thanks to $\tilde{p}$ distributed as a Dirichelt process.

Thanks to the previous proposition, we can describe the probability distribution of a sample $(X_1, \ldots, X_n)$ from a convex combination of a DP with a diffuse measure, including the number of unique elements arising from the diffuse part. We can further exploit the EPPF of Proposition 1 to perform inference on the main parameter of the model α, β and τ_1.

3 Application to text data

We describe a possible application of the model proposed in Section 2 in the context of text data. Random probability measures as the Dirichlet process are commonly used to model several data types, and their usage in natural language processing analysis has grown severally over the past decade. On the one hand the DP is sufficiently flexible to capture the structure of the data, but, on the other hand, it does not take into account observations that cannot be recorded again in the sample.

It is well known that text data may show words observed only once. These unique words, commonly termed as *hapax legomena*, have been recognized as peculiar usage of words by the authors, and they represent an interesting problem to study from a statistical perspective. See e.g. [6] for further details on word frequency distributions. Hereby we aim to use the results of Section 2 to model text data, allowing for the presence of unique words.

Let $\pi(\alpha)$ and $\pi(\beta)$ denote the prior distributions for the parameters α and β respectively. We further assume that $\pi(\alpha)$ is a $Gamma(a,b)$ distribution and $\pi(\beta)$ is a $Beta(c,d)$ distribution. Given an observed sample, we can perform inference on the parameters characterizing the distribution of the data, and the number of words associated with the diffuse measure, using the EPPF to build a MCMC algorithm and to update iteratively from

$$\begin{aligned}
\pi(\alpha \mid \tau_1, \beta, X_1, \ldots, X_n) &\propto \alpha^{a-1} e^{-b\alpha} \frac{\alpha^{k-\tau_1}}{(\alpha)_{n-\tau_1}}, \\
\pi(\beta \mid \alpha, \tau_1, X_1, \ldots, X_n) &\propto \beta^{c+n-\tau_1-1} (1-\beta)^{d+\tau_1-1}, \\
P(\tau_1 = t \mid \alpha, \beta, X_1, \ldots, X_n) &\propto \binom{k_1}{t} \beta^{n-t} (1-\beta)^{t} \frac{\alpha^{k-t}}{(\alpha)_{n-t}},
\end{aligned}$$

where, thanks to conjugacy, the posterior distribution of β is a $Beta(c_n, d_n)$ random variable, with $c_n = c + n - \tau_1$ and $d_n = d + \tau_1$. In order to sample realizations from $\pi(\alpha \mid \tau_1, \beta)$, we resort to a Metropolis-Hastings [7, 5] step with Gaussian proposal, on a logarithmic scale. Finally we note that the posterior distribution of τ_1 is a discrete random variable with values in $\{0, \ldots, k_1\}$.

As a text data set, we consider the story of *"Alice's Adventures in Wonderland"* [2], available on the Project Gutenberg (www.gutenberg.org). We are interested in both modeling the distribution of the words, and the probability of having unique words which cannot be observed again. We pre-processed the text data set by discarding the symbols and the main stopwords. The data set is composed by a total number of $n = 12\,715$ distinct words, $1\,220$ of which are used just once in the story. We split the story in two parts, with the first $6\,358$ words as data set used to estimate the model. Figure 1 shows the frequency spectrum on log-log scale of the

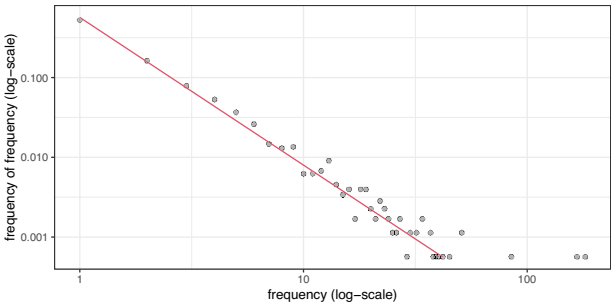


Fig. 1 Frequency spectrum on a log-log scale. The red line denotes the ZIPF distribution evaluated with the maximum likelihood estimate parameter.

words used in the first part of *"Alice's Adventures in Wonderland"* story. We can appreciate the presence of a large amount of elements with frequency one, hence our proposal is particularly suited for this kind of data.

We model the frequencies' distribution with the EPPF function described in Proposition 1. We initialize the prior parameter by setting $\alpha \sim Gamma(1,1)$ and $\beta \sim Beta(1,1)$, i.e. assign to the parameter β an uniform distributed prior over $(0,1)$. We estimate the model with $20\,000$ iterations, which include a burn-in period of $10\,000$ iterations. Figure 2 shows the traceplots for the sample of the main

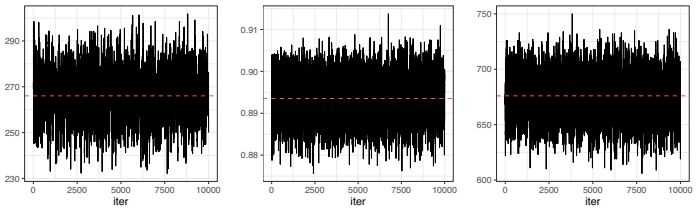


Fig. 2 Left to right: traceplots of α, β and τ_1.

parameters of the model from their posterior distributions. There is no visual suggestion against the convergence of the chain. Furthermore, the Geweke diagnostic detected are -0.999, 0.808 and 1.009 for α, β and τ_1 respectively, and the accep-

tance rate of the Metropolis-Hastings steps to sample from the posterior distribution of α is equal to 0.387: these results suggest that the chain is stable.

	posterior mean	0.95 posterior CI	True value	Geweke diagnostic
α	266.030	(239.326–295.005)	-	-0.999
β	0.894	(0.881–0.906)	0.904	-0.808
τ_1	676.053	(638.975–712.000)	657	1.009

Table 1 Posterior summaries for the model estimated with the first part of the book.

Table 1 shows the posterior summaries of the model, for the *"Alice's Adventures in Wonderland"* story data. The model identifies averagely $\hat{\tau}_1 = 676.053$ words as unique words, and also the posterior estimate of the probability of being an unique term is equal to $\hat{\beta} = 0.894$, where $(1-\hat{\beta}) = 0.106$ corresponds to the point estimate of the probability for a word of being a hapax legomena.

The parameter τ_1 represents the number of hapaxes, i.e. words which are assigned to the contaminant component of the model Q_0. We can easily check the accuracy of our estimation by looking at how many unique words with frequency 1 in the first part of the story remain with frequency 1 also in the second part, which we refer to as the true value of τ_1. The point estimate 676.053 is close to the true value 657 respectively. Furthermore the 0.95 posterior credible interval for τ_1 is covering the true number of hapaxes. Although with the Dirichlet process there is a positive probability that a word observed once in a sample remains unique in a larger sample, it is not possible to model directly uniqueness in the corresponding population. For comparison, we estimated the number of words with frequency one in the first part of the sample not observed in the second part of the sample, using a Dirichlet process. We obtained as 0.95 posterior confidence interval $(481.223 - 482.955)$, far from the true value 657.

References

1. Blackwell, D., MacQueen, J. B.: Ferguson Distributions Via Polya Urn Schemes. Ann. Statist. **1**, 2, 353–355 (1973)
2. Carroll, L.: Alice's Adventures in Wonderland. Urbana, Illinois: Project Gutenberg
3. Ferguson, T. S.: A Bayesian analysis of some nonparametric problems. Ann. Statist. **1**,2, 209–230 (1973)
4. Ghosal, S., Van Der Vaart, A.: Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press, Cambridge (2017)
5. Hastings, W.: Monte Carlo sampling methods using Markov chains and their application. Biometrika. **57**: 97–109 (1970)
6. Harald Baayen, R.: Word Frequency Distributions. Springer Netherlands (2001)
7. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**(6): 1087–1092 (1953)
8. J. Pitman.: Exchangeable and partially exchangeable random partitions. Probability Theory and Related Fields **102**, 145–158 (1995)

Detection of neural activity in calcium imaging data via Bayesian mixture models

Rilevamento dell'attività cerebrale in dati di calcium imaging tramite modelli mistura bayesiani

Laura D'Angelo, Antonio Canale, Zhaoxia Yu, and Michele Guindani

Abstract The recent technological advancements in the field of miniature microscopy have made it possible to study the brain activity in alive animals at a neuronal level. In particular, the technique of calcium imaging has allowed to record the neurons' activity over time in response to external stimulation, thus enabling the study of how neurons encode information. However, this technique has several limitations, and statistical tools are necessary to extract valuable information from the noisy time series. In this paper we analyze a neuron's response to a range of visual stimuli: exploiting a Bayesian mixture model we detect the neuron's spiking activity and study the distribution of the activations in relation to the type of stimulus.

Abstract *Il progresso tecnologico nel campo della microscopia miniaturizzata ha recentemente reso possibile studiare l'attività cerebrale in animali vivi al livello dei singoli neuroni. In particolare, la misurazione della concentrazione intracellulare di ioni di calcio ha reso possibile rilevare l'attività dei neuroni in risposta a stimoli esterni e, di conseguenza, studiare come questi codifichino l'informazione. Tuttavia, questa tecnica ha varie limitazioni, e sono necessari nuovi strumenti statistici per estrarre l'informazione contenuta nei dati. In questo contributo presentiamo l'analisi della risposta di un neurone a una serie di stimoli visivi: tramite un modello mistura bayesiano riusciamo ad identificare le attivazioni, e a studiare come queste varino in risposta a diversi stimoli.*

Key words: Bayesian nonparametrics, Dirichlet process, Mixtures of Finite Mixtures, Model-based clustering, Nested Dirichlet process.

Laura D'Angelo & Antonio Canale
Department of Statistical Sciences, University of Padova, Padova, Italy,
e-mail: laura.dangelo.1@phd.unipd.it, e-mail: antonio.canale@unipd.it

Zhaoxia Yu & Michele Guindani
Department of Statistics, University of California, Irvine, Irvine, U.S.A.
e-mail: zhaoxia@ics.uci.edu, e-mail: mguindan@uci.edu

1 Introduction

The understanding of the brain functioning is still an important open question in neuroscience. In particular, in the past few years, the study of how neurons react to stimuli and encode information has become a central topic of research [10]. This was made possible by technological advances in miniaturized microscopes which enabled the analysis of the neuronal activity through the technique of calcium imaging. This technique is based on the observation of the intra-cellular calcium concentration through fluorescent molecules that bind to the calcium ions. Fluctuations of such concentration can be used as a proxy of the neuronal activity: because of a physiological process, when a neuron fires, calcium floods the cell, and a transient spike in the intra-cellular calcium level is produced [7].

On one side, the ability to observe the neurons' behavior opened the way to the understanding of how neurons react to external stimuli, however, it also posed new challenges on how to extract valuable information from the observed noisy fluorescent calcium traces. These traces are in fact not directly available for the analysis of a neuron's spiking activity, as the real object of study is not the calcium itself, but the underlying spike trains, i.e. the exact firing times and the related intensities. Extracting the spike trains from the calcium trace is not straightforward, as the technique itself has some limitations as, for example, the presence of measurement error and the slow decay of the fluorescence trace compared to the underlying neuronal activity. Several statistical methods have been proposed to detect the spikes: [5] proposed an efficient on-line algorithm based on a lasso penalty, while [8, 9] proposed to replace the L_1 with an L_0 penalization. Previous methods provide high accuracy in spike detection but lack of uncertainty quantification. In this sense, a Bayesian approach would be preferable. In this paper we focus on the Bayesian approach recently proposed in [3], a specification that explicitly takes into account the presence of varying external stimuli, and allows to study whether and how the neuronal activity is affected by them.

2 Allen brain observatory data

We focus our analysis on a publicly available data set of calcium imaging data provided by the Allen Institute for brain science [1] and we analyze the activity of a neuron located in the mouse visual cortex. During this experiment, a mouse is placed in front of a screen, and different types of visual stimuli are presented, while its neuronal activity is recorded. Here, it is of interest not only to detect the spikes, but also to understand how the neuron's activity varies in relation to the type of visual stimulus. Specifically, we aim to understand if different types of visual stimuli induce a similar response in the neuron's activity, and if there is any pattern in the response, i.e., recurring amplitudes of the spikes. We consider two experiments for the same neuron, for a total of 5 different visual stimuli (experimental conditions). The fluorescent calcium traces for the two experiments are shown in Fig. 1 and Fig. 2

with a black line, while the colored backgrounds indicate the type of visual stimulus presented in that moment. Each experiment has a duration of about an hour: since the traces are recorded at a frequency of 30 Hz, for both experiments we have over 100,000 time points.

3 Model for Spikes' Detection and Analysis

The observed fluorescent trace is considered to be a noisy realization of the unobserved (latent) true calcium concentration. To model the underlying calcium dynamic, it is often assumed an autoregressive process with jumps at the neuron's activations. (see, e.g., [12, 5, 8, 9]). Denoting with y_t the observed trace at time t, and with c_t the underlying true calcium concentration, for $t = 1, \ldots, T$, the model can be written as

$$y_t = b + c_t + \varepsilon_t \tag{1}$$

$$c_t = \gamma c_{t-1} + A_t + \omega_t \tag{2}$$

where $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$ and $\omega_t \sim \mathrm{N}(0, \tau^2)$ are independent random variables describing the measurement error. The parameter b represents the observed baseline level, so that, in absence of activity, the concentration is centered around zero; while γ is the autoregressive parameter. Finally, the parameters A_t describe the activity at each time t: with this model specification, each A_t will either be equal to zero, if no spike occurs at time t, or will assume a non-zero positive value corresponding to the spike amplitude (jump) at a neuron's activation. Specifying an adequate and flexible prior for the parameters A_t is extremely important, as it will determine the accuracy of the estimated spike trains.

Figures 1 and 2 clearly show that the neuron's activity is affected by the type of stimulus and that, possibly, a clustering structure of its response exists between experimental conditions. The nested Dirichlet process [11] is a flexible nonparametric prior that was proposed to deal with this kind of settings, where it is of interest to obtain a cluster at both the unit level (here, the spikes) and the distribution level (the stimuli). More recently, [4] proposed the common atom model (CAM), a modification of the original nested Dirichlet process that does not suffer from the degeneracy issue pointed out by [2]. Starting from the CAM, [3] combined its nested structure with the generalized mixtures of finite mixtures of [6] in place of the two levels of Dirichlet process. We adopt the latter model specification for the prior on the neuronal activity A_t and, following [3], we place a spike-and-slab base measure on the parameters A_t. This is a mixture base measure, with a spike at zero, modeling the absence of activity, and a Gamma "slab" component, modeling the amplitude of the neuron's firing activity.

4 Data Analysis

We applied the model of [3] to the fluorescence traces of a neuron located in the mouse visual cortex of the Allen brain dataset described in Section 2. Specifically, in the considered experiments, the mouse is subjected to five types of visual stimuli: *static grating, drifting grating, natural scene,* and two different natural movies, namely *natural movie one* and *natural movie three*. The stimuli go from simple synthetic and static images, to more complex movies. Fig. 1 and Fig. 2 show the estimated spike trains (yellow lines) for the two experiments: the lines represent the inferred neuronal activities, after removing of the measurement error and of the slow decay of the calcium. At each time point it is shown the actual presence or absence of a spike, and its amplitude when present. It is evident that the neuron's activity can be very different between experiments as well as between experimental conditions. In general, the neuron has a much more intense activity during the first experiment, with a larger number of detected spikes and higher amplitudes. The posterior mean of the firing rate (number of detected spikes per second) and related 95% credible interval are equal to $0.75\ (0.72, 0.79)$ for the first experiment, and $0.19\ (0.16, 0.21)$ for the second experiment. In the following paragraphs we will analyze in detail the neuron's activity in each of the two experiments.

Within each experiment, it is of interest to understand which stimuli led to a similar response in the neuron's activity, which corresponds to studying the clustering of the stimulus-specific distributions. In the first experiment, *natural scene* and *natural movie one* are clustered together, while *static grating* is put into a separate distributional cluster. The firing rate can be used to describe the intensity of the neuron's activity in response to a stimulus, with higher rates associated to more activity. The posterior expectations (together with the 95% credible intervals) of these rates for the three stimuli are equal to $0.235\ (0.206, 0.264)$ for *static grating*, $1.498\ (1.434, 1.566)$ for *natural scene*, and $0.884\ (0.797, 0.973)$ for *natural movie one*. The more complex stimuli led to an increased activity.

We now analyze the clustering of the amplitudes of the spikes: first, we obtained a posterior point estimate of the partition by minimizing the variation of information loss [13]; then, we computed a representative value for each cluster by averaging through the group-specific parameters A_l (keeping the partition fixed). The amplitudes were grouped into 5 clusters, with parameters equal to 0.20, 0.44, 0.60, 0.77 and 0.99. Figure 3 (left) shows how these clusters are distributed within each experimental condition: each bar shows the relative frequency of each cluster during the stimulus. The first cluster (associated with the smallest amplitude) is largely the most frequent, and accounts for over 75% of the spikes in all the conditions. On the contrary, the largest spikes (corresponding to an amplitude equal to 0.99) are only found during *natural scene*, which is also the stimulus with the largest firing rate. Hence, during *natural scene*, the neuron had the most intense activity, with more frequent spikes and larger spikes' amplitudes.

Turning now to the second experiment, each stimulus was allocated to a different distributional cluster, except for *drifting grating*, which was clustered along with the absence of stimuli condition. The firing rates (posterior mean and 95% credible

intervals) were estimated equal to 0.0379 $(0.0254, 0.0535)$ for *drifting grating*, to 0.164 $(0.123, 0.203)$ for *natural movie one*, and to 0.471 $(0.417, 0.523)$ for *natural movie three*. If compared to the previous experiment, we notice a strong reduction of the overall activity, with *natural movie three* being the only stimulus leading to a comparable level of activation. The analysis of the clustering of the spikes' amplitudes shows the existence of only three clusters, with amplitudes 0.16, 0.38 and 0.75. Figure 3 (right) shows how these clusters are distributed within each experimental condition: again, most spikes are associated with the smallest amplitude. Also in this case the largest cluster is found only in one experimental condition (*natural movie three*), i.e., in correspondence with the most intense activity.

These findings highlight how the behavior of a neuron can be very complex and not trivial to understand, as in different situations its response can vary quite heavily. In particular, we noticed how its activity can be very different not only between experiments, but also between stimuli that are similar. These insights confirm that the study of neurons' behavior is an open field of research, and that further inves-

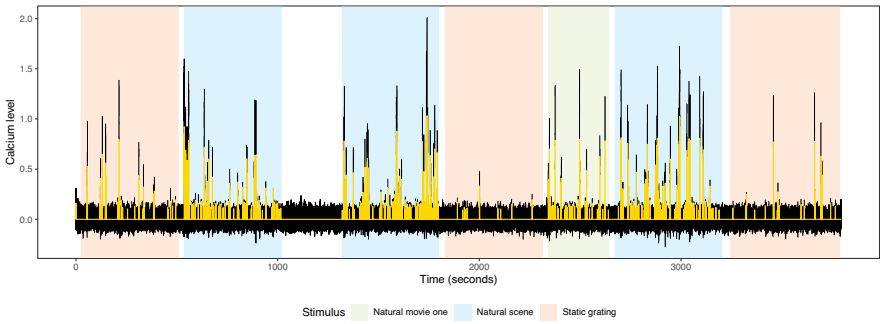


Fig. 1 Observed fluorescence trace (black line) for Experiment 1 and estimated spike train (yellow line). The background colors correspond to the type of visual stimulus.

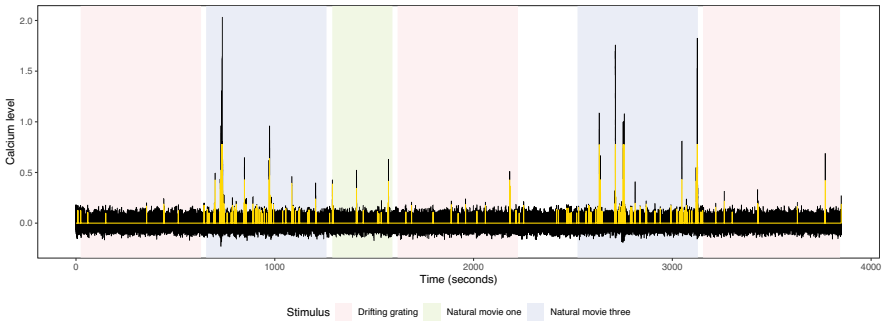


Fig. 2 Observed fluorescence trace (black line) for Experiment 2 and estimated spike train (yellow line). The background colors correspond to the type of visual stimulus.

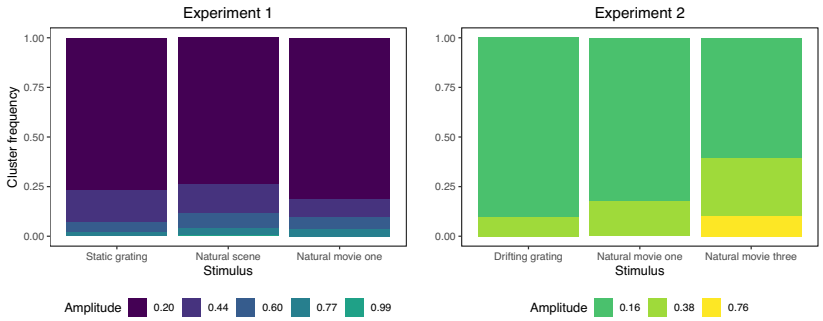


Fig. 3 Posterior estimate of the clustering of the amplitudes of the positive spikes, for the two experiments. Each bar shows the relative frequency of the observational clusters within each experimental condition.

tigation is needed in order to fully understand how single neurons and neuronal networks respond to stimulation and encode information.

References

1. Allen Institute for Brain Science. Allen brain observatory. http://observatory.brain-map.org/visualcoding (2016)
2. Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., Rodríguez, A.: Latent nested nonparametric priors (with discussion). Bayesian Analysis **14**, 1303—1356 (2019).
3. D'Angelo, L., Canale, A., Yu, Z., Guindani, M.: Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. arXiv:2102.09403 (2021).
4. Denti, F., Camerlenghi, F., Guindani, M., Mira, A.: A common atom model for the Bayesian non-parametric analysis of nested data. arXiv:2008.07077 (2020).
5. Friedrich, J., Zhou, P., Paninski, L.: Fast online deconvolution of calcium imaging data. PLOS Computational Biology **13**, 1–26 (2017)
6. Frühwirth-Schnatter, S., Malsiner-Walli, G., Grün, B.: Generalized mixtures of finite mixtures and telescoping sampling. arXiv:2005.09918 (2020).
7. Grienberger, C., Konnerth, A.: Imaging calcium in neurons. Neuron **73**(5), 862–885 (2012)
8. Jewell, S., Witten, D.: Exact spike train inference via L0 optimization. The Annals of Applied Statistics **12**, 2457–2482 (2018).
9. Jewell, S. W., Hocking, T. D., Fearnhead, P., and Witten, D. M.: Fast nonconvex deconvolution of calcium imaging data. Biostatistics **21**, 709-–726 (2019).
10. Nakajima, M., Schmitt, L. I.: Understanding the circuit basis of cognitive functions using mouse models. Neuroscience Research **152**, 44–58 (2020)
11. Rodríguez, A., Dunson, D. B., and Gelfand, A. E.: The nested Dirichlet process. Journal of the American Statistical Association **103**, 1131-–1154 (2008).
12. Vogelstein, J. T., Packer, A. M., Machado, T. A., Sippy, T., Babadi, B., Yuste, R., Paninski, L.: Fast nonnegative deconvolution for spike train inference from population calcium imaging. Journal of Neurophysiology **104**, 3691-–3704 (2010).
13. Wade, S. and Ghahramani, Z.: Bayesian cluster analysis: point estimation and credible balls (with discussion). Bayesian Analysis **13**, 559-–626 (2018).

4.8 Clustering for complex data

Clustering categorical data via Hamming distance

Raggruppamento di dati categoriali attraverso la distanza di Hamming

Edoardo Filippi-Mazzola and Raffaele Argiento and Lucia Paci

Abstract Clustering methods have typically found their application when dealing with continuous data. However, in many modern applications data consist of multiple categorical variables with no natural ordering. In the heuristic framework the problem of clustering these data is tackled by introducing suitable distances. In this work, we develop a model-based approach for clustering categorical data with nominal scale. Our approach is based on a mixture of distributions defined via the Hamming distance between categorical vectors. Maximum likelihood inference is delivered through an expectation-maximization algorithm. A simulation study is carried out to illustrate the proposed approach.

Abstract *Le tecniche di clustering trovano normalmente la loro applicazione su variabili continue. Tuttavia, in molti contesti applicativi, i dati sono categorici senza un ordine naturale. All'interno del framework euristico, la clusterizzazione di questi dati avviene grazie all'utilizzo di metriche adeguate. In questo lavoro, proponiamo un approccio probabilistico per la clusterizzazione di dati categorici nominali. Il nostro approccio si basa su una mistura di distribuzioni derivate dal concetto di distanza di Hamming. Proponiamo l'utilzzo di un algoritmo EM per la stima di massima verosimiglianza dei parameteri del modello. L'approccio è validato su dataset simulati.*

Key words: Expectation-Maximization algorithm, Hamming distribution, mixture modeling, nominal data

---

Edoardo Filippi-Mazzola
Institute of Computational Science, Università della Svizzera italiana, Lugano e-mail: edoardo.filippi-mazzola@usi.ch

Raffaele Argiento
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan e-mail: raffaele.argiento@unicatt.it

Lucia Paci
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan e-mail: lucia.paci@unicatt

1 Introduction

Clustering is a widely used data analysis technique with application in many research fields such as life, environmental and social sciences. The main goal of cluster analysis is to investigate underlying structures in the data and identify subsets of observations such that data belonging to the same group are close to each other in a measurable space.

The most popular technique for clustering analysis is the *K-means* algorithm (MacQueen, 1967). The main limitation of this procedure is that it can be used to cluster only continuous data. Indeed, most of the clustering algorithms in the literature have been proposed to deal with continuous data and very few methods have been developed for clustering categorical data. The main challenge in dealing with nominal categorical data is the lack of a metric space in which data points are positioned with measurable coordinates.

Ralambondrainy (1995) proposed a numerical coding scheme for categorical attributes, i.e., convert multiple category variables into binary attributes and treated them as numeric using the K-means algorithm. However, converting the categorical data into numeric values does not necessarily produce meaningful results, especially when data do not have a natural ordering. Moreover, the arbitrary choice of encoding scheme can lead to different results. Huang (1998) introduced the *K-modes* algorithm that modifies the K-means algorithm by replacing means with modes. However, the number of clusters must be known to run the K-modes algorithm. As an alternative, Zhang et al (2006) developed the *UH-Vector* algorithm to identify underlying clusters based on the Hamming distance. The algorithm uses a chi-squared type statistic to determine the cluster center, the cluster radius and determines the cluster's assignment. One advantage of this procedure is that the number of clusters is not needed in advance, rather it is determined by the algorithm.

Celeux and Govaert (2015) discussed latent class models, i.e, mixture models for clustering categorical and mixed-type data. In this framework, our contribution is to propose a model based approach for clustering categorical data with nominal scale. Building upon the work Zhang et al (2006), we define a new distribution based on the concept of the Hamming distance. From a probabilistic point of view, clustering is usually addressed via mixture modeling (Frühwirth-Schnatter et al (2019), Argiento and Iorio (2019)). Hence, a mixture of Hamming distributions is proposed to model non-ordinal categorical data. A hierarchical formulation of the mixture model is exploited to facilitate the computation via the Expectation-Maximization (EM) algorithm. A simulation study is carried out to assess the performance of the proposed algorithm.

The paper is organized as follows. Section 2 presents the categorical sample space and the Hamming distribution. Section 3 introduces the mixture model and the adapted EM algorithm. In Section 4 we discuss the results of a simulation study. We conclude with a brief summary in Section 5.

2 The Hamming distribution

Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ be a vector of categorical variables collected on a set of observations $i = 1, \ldots, n$. Each variable j, for $j = 1, ..., p$, can assume m_j possible levels called *categories* defining the finite set A_j; namely, $x_{ij} \in A_j$, where $A_j = \{a_{j1}, a_{j2}, \ldots, a_{jm_j}\}$. Hence, in our setting, we consider data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ collected over the discrete sample space $\Omega_p = A_1 \times A_2 \times \cdots \times A_p$, i.e.,

$$\Omega_p = \left\{ \mathbf{x} = (x_1, \ldots, x_p)^\top | x_1 \in A_1, \ldots, x_p \in A_p \right\}$$

The set Ω_p is a discrete sample space of size $N = \prod_{j=1}^{p} m_j$.

To determine the distance between two elements of Ω_p, we can use the *Hamming distance* as an indicator of mismatch that counts how many attributes differ between the two. Analytically, the Hamming distance between two points in Ω_p is

$$d(\mathbf{x}_i, \mathbf{x}_h) = \sum_{j=1}^{p} 1 - \delta_{x_{ij}}\left(x_{hj}\right),$$

where δ_x denotes the Dirac measure on x, i.e.,

$$1 - \delta_{x_{ij}}\left(x_{hj}\right) = \begin{cases} 0 \text{ if } x_{ij} = x_{hj} \\ 1 \text{ if } x_{ij} \neq x_{hj}. \end{cases}$$

The Hamming distance defines a proper metric space in Ω_p and the distance that separates two points in Ω_p can only assume a finite number of integer values that range from 0 to p.

Consider a center parameter $\mathbf{c} = (c_1, ..., c_p)^\top \in \Omega_p$ and a scale parameter $\sigma > 0$. A random variable $\mathbf{X} = (X_1, \ldots, X_p)$ with support Ω_p follows an *Hamming distribution* with center $\mathbf{c}$ and scale σ if its p.m.f. for each $\mathbf{x} \in \Omega_p$ is the following

$$p(\mathbf{x}|\mathbf{c}, \sigma) = \prod_{j=1}^{p} \left(1 + \frac{m_j - 1}{\exp\{1/\sigma\}}\right)^{-1} \exp\left\{ - \frac{1 - \delta_{c_j}(\mathbf{x}_j)}{\sigma} \right\}, \tag{1}$$

and we write $\mathbf{X} \sim \text{Hamming}(\mathbf{c}, \sigma)$. It is possible to show that $\sum_{x \in \Omega_p} p(\mathbf{x}|\mathbf{c}, \sigma) = 1$, namely that $p(\mathbf{x}|\mathbf{c}, \sigma)$ defined in (1) is a proper p.m.f.

2.1 Maximum likelihood estimation

For a given sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and a scale parameter $\sigma > 0$, we denote with $l(\mathbf{c}, \sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)$ the log-likelihood corresponding to (1). The maximum likelihood estimate of the two parameters is based on a two-step procedure. The first step involves the optimization with the center parameter $\mathbf{c}$, that is

$$\arg\max_{\hat{\mathbf{c}}\in\Omega_p} l(\hat{\mathbf{c}} \mid \mathbf{x}_1,\ldots,\mathbf{x}_n,\sigma) = \arg\min_{\hat{\mathbf{c}}\in\Omega_p}\left\{\sum_{i=1}^{n} d(\mathbf{x}_i,\hat{\mathbf{c}})\right\}. \tag{2}$$

According to (2), $\hat{\mathbf{c}}$ does not depend on σ and turns out to be the Fréchet mean with respect to the square root of the Hamming distance. In practice, we look for the set of points whose sum of distances from point $\hat{\mathbf{c}} \in \Omega_p$ is the lowest. To accomplish that, we follow the *HD-vector* algorithm (Zhang et al, 2006), that is we define O_p the set of augmented data consisting of all points in Ω_p at Hamming distance smaller or equal of 1 from the observed data and then solve (2) over O_p.

The second parameter that has to be estimated is the scale parameter σ. Given the center $\hat{\mathbf{c}}$, the maximum likelihood estimator for the scale parameter σ has no closed formulation and it is numerically approximated.

3 Mixture modeling

We address model-based clustering assuming that each observation $\mathbf{x}_i$, $i = 1,\ldots,n$ comes from a linear combination of K different Hamming distributions, that is

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_i|\mathbf{c}_k,\sigma_k), \tag{3}$$

which we refer to as a *mixture of Hamming distributions*. Each $p(\mathbf{x}_i|\mathbf{c}_k,\sigma_k)$ is a component of the mixture with center $\mathbf{c}_k$ and scale parameter σ_k. Moreover, the parameter $\pi = (\pi_1,\ldots,\pi_K)$, is referred to as the vector of component weights. For each $k = 1,\ldots,K$, π_k is the probability that a generic observation i belongs to component k. The latter means that the observation $\mathbf{x}_i$ has been generated from an Hamming distribution with parameters $\mathbf{c}_k$ and σ_k. Finally, $0 \le \pi_k \le 1$ for each k with the constrain $\sum_{k=1}^{K} \pi_k = 1$.

Inference under mixture model is computationally simplified when expressing model (3) in a hierarchical fashion by introducing a latent set of variables $z_i \in \{1,\ldots,K\}$, $i = 1,\ldots,n$, called cluster allocators (Bishop, 2006). In particular $z_i = k$ denotes the event that data i belongs to cluster k, and $P(z_i = k) = \pi_k$ independently for $i = 1,\ldots,n$. It is not difficult to realize that if $p(\mathbf{x}_i \mid z_i) = p(\mathbf{x}_i \mid \mathbf{c}_{z_i},\sigma_{z_i})$, then the joint p.m.f. $p(\mathbf{x}_i,z_i) = p(\mathbf{x}_i \mid z_i)P(z_i = k)$ induced as marginal for $\mathbf{x}_i$ in (3).

Working with the joint distribution $p(\mathbf{x}_i,z_i)$ simplifies the computation. In particular, once the number of components K is fixed, the implementation of a EM algorithm is as follows. Given the initial values of the parameters $\mathbf{c}_k$, σ_k and π_k, $k = 1,\ldots,K$, the algorithm iterates between the following steps:

- **E-step**: using the current parameters $\mathbf{c}_k$, σ_k and π_k, evaluate the posterior probability for observation i to belong to cluster l (also called *responsibilities*):

$$p(z_i = l \mid \mathbf{x}_i) = \frac{\pi_l p(\mathbf{x}_i \mid \mathbf{c}_l,\sigma_l)}{\sum_{k=1}^{K} \pi_k p(\mathbf{x}_i \mid \mathbf{c}_k,\sigma_k)}, \quad l = 1,\ldots,K$$

- **M-step**: using the current responsibilities $p(z_i = l \mid \mathbf{x}_i)$, estimate the parameters $\mathbf{c}_k$, σ_k and π_k, for $k = 1, \ldots, K$, by maximizing the expected complete-data log-likelihood; this is achieved by adapting the procedure in Section 2.1.

The estimation procedure assumes that K is fixed. In practice, when the number of clusters is relatively small, a simple way to estimate K is by comparing the values of model selection criteria calculated for various mixture models with fixed number of groups. Then, the number of components is chosen according to the optimal value of model selection criterion. Finally, the clustering estimation is obtained by assigning each observation to the k-th component with the highest responsibility.

4 Simulation results

We carried out a simulation study to evaluate the ability of the proposed approach to correctly allocate the observations within the clusters. We simulate categorical data from the mixture of Hamming distributions in (3), with different values of: the number of components K, the sample size n, the cluster size n_k and the scale parameter σ, see Table 1. In the first two experiments the number of attributes is $p = 2$, while in the other scenarios the number of attributes is $p = 10$. Performances are evaluated using the Adjusted Rand-Index (RI) (Rand, 1971) and compared with the results obtained by employing the HD-vector algorithm of Zhang et al (2006).

Table 1 displays the results that we obtained in the different settings. Here, we select the number of clusters according to both the AIC and the BIC indexes; the selected K is equal to the true number of clusters in all experiments. The first two experiments consider data sampled from a mixture with just one component, i.e., $K = 1$. In other words, data come from the Hamming distribution in (1). The results show that the maximum likelihood estimator described in Section 2.1 provides an accurate estimate of the center of the distribution and a good approximation of the scale parameter.

In the last three experiments, data are simulated from a Hamming mixture distribution with an increasing number of components, i.e., $K = \{3,4,5\}$. The results show the capability of the EM algorithm to recover the true model parameters. Moreover, our approach outperforms the HD-vector algorithm in recovering the true clustering. As we expected, the number of incorrect clustering assignment increases, as the number of components and the sparsity of the data increases, such as in the last scenario. However, our method shows improved results over the HD-vector algorithm also with increased complexity of the data.

5 Summary and future work

In this work we have proposed a model-based approach to cluster categorical data with nominal scale based on the Hamming distance. We first introduced a valid

Table 1 Results on simulated scenarios.

Scenario	K	n	cluster size n_l	**True** σ_l	**Estimated** σ_l	RI EM	RI HD
1	1	100	-	{ 0.5}	{ 0.457}	-	-
2	1	100	-	{ 1}	{ 1.02}	-	-
3	3	150	{ 50, 50, 50}	{ .4, .7, .9}	{ .39, .68, .93}	93.3%	31.3%
4	4	200	{ 70, 30, 40, 60}	{ .4, .7, .9, .8}	{ .39, .72, .84, .75}	88.5%	63.4%
5	5	220	{ 70, 30, 40, 60, 20}	{ .4, .7, .9, .8, 1}	{ .38, .8, .81, .8, .85}	83.2%	58.3%

probability mass function based on the Hamming distance; links to heterogeneity measures such as the Gini and entropy indeces can be illustrated (not shown here for brevity). Then, a mixture of Hamming distributions has been specified to cluster categorical data. The hierarchical formulation of the mixture model has been employed to facilitate the computation via a EM algorithm. The main drawback is that the number of clusters is fixed. However, in practice, the number of clusters is usually unknown and needs to be estimated. Future works will find us to develop a fully Bayesian approach with unknown number of clusters. Finally, the simulation study shown the capability of our approach to recover the underlying clusters. Real-data applications of the proposed method are ongoing.

References

Argiento R, Iorio MD (2019) Is infinity that far? a Bayesian nonparametric perspective of finite mixture models. arXiv: 190409733

Bishop CM (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg

Celeux G, Govaert G (2015) Latent class models for categorical data. In: C H, Melia M, Murtagh F, Rocci R (eds) Handbook of cluster analysis, Chapman & Hall/CRC

Frühwirth-Schnatter S, Celeux G, Robert C (2019) Handbook of Mixture Analysis. CRC Press, Taylor & Francis Group

Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 2(3):283–304

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp 281–297

Ralambondrainy H (1995) A conceptual version of the k-means algorithm. Pattern Recognition Letters 16(11):1147 – 1157

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336):846–850

Zhang P, Wang X, Song PXK (2006) Clustering categorical data based on distance vectors. Journal of the American Statistical Association 101(473):355–367

Penalized model-based clustering for three-way data structures

Clustering penalizzato basato su un modello per dati con struttura tridimensionale

Andrea Cappozzo, Alessandro Casa, and Michael Fop

Abstract Recently, there has been an increasing interest in developing tools able to find groups in matrix-valued data. To this extent, matrix Gaussian mixture models (MGMM) represent an extension of the popular model-based clustering based on normal mixtures. Unfortunately overparametrization issues, already affecting the vector-variate framework, are exacerbated in the MGMM one, where the number of parameters grows quadratically with both row and column dimensions. To overcome this limitation, we introduce a sparse model-based clustering approach for three-way data structures. The proposed penalized estimation scheme, shrinking the estimates towards zero, achieves more stable and parsimonious clustering in high-dimensional scenarios. An application to satellite images underlines the benefits of the proposal.
Abstract *Il crescente interesse verso metodi in grado di identificare gruppi in dati matriciali ha portato allo sviluppo di modelli di mistura matriciali Gaussiani (MGMM) che rappresentano una naturale estensione del clustering basato su misture di normali. L'eccessiva parametrizzazione, che già interessa il contesto vettoriale, è particolarmente evidente nei MGMM dove il numero di parametri cresce all'aumentare sia del numero di righe che di colonne. Al fine di superare questa limitazione, in questo lavoro si introduce un approccio di clustering basato su modelli sparsi per dati matriciali. La procedura di stima penalizzata adottata permette di ottenere un clustering più stabile e parsimonioso in scenari ad alta dimensione. Un'applicazione a immagini satellitari evidenzia i vantaggi del metodo proposto.*

Key words: Model based clustering, Matrix distribution, EM-algorithm, Penalized likelihood, Sparse matrix estimation

---

Andrea Cappozzo,
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: andrea.cappozzo@unimib.it

Alessandro Casa, Michael Fop
School of Mathematics & Statistics, University College Dublin e-mail: alessandro.casa@ucd.ie, michael.fop@ucd.ie

1 Introduction and motivation

Model-based clustering is a probabilistic-based approach to account for heterogeneity in a population, useful for discovering subgroups in data [2]. This framework assumes that each cluster corresponds to a different component of a finite mixture, with the Gaussian distribution being the standard choice when dealing with continuous data [4]. Nonetheless, the ever-increasing complexity of real-world datasets is jeopardizing the usage of standard Gaussian Mixture Models (GMM), as they tend to be over-parametrized in high-dimensional spaces [1]. To this extent, parameters regularization by means of penalized estimation has been proven useful in performing model-based clustering and variable selection in such scenarios [9].

The aforementioned problem complicates even further when dealing with three-way data structures, where, for each statistical unit, multiple variables are recorded simultaneously on various occasions and dimensions. These increasingly common situations lead to a complex statistical framework, for which the observations are assumed to be realizations of some matrix-variate distribution. In details, for a given sample of n standardized matrices $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, with $\mathbf{X}_i \in \mathbb{R}^{p \times q}$, the GMM extension to the three-way data context is provided by the matrix normal mixture model (MGMM [7]), in which the marginal density of each $\mathbf{X}_i$ reads:

$$f(\mathbf{X}_i; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \tau_k \phi_{p \times q}(\mathbf{X}_i; \mathbf{M}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Gamma}_k). \tag{1}$$

The number of mixture components is denoted by K, τ_k's are the mixing proportions with $\tau_k > 0, \forall k = 1, \ldots, K$, $\sum_{k=1}^{K} \tau_k = 1$ and $\phi_{p \times q}(\cdot; \mathbf{M}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Gamma}_k)$ is the k-th component density of a $p \times q$ matrix normal distribution:

$$\begin{aligned}\phi_{p \times q}(\mathbf{X}_i; \mathbf{M}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Gamma}_k) = (2\pi)^{-\frac{pq}{2}} |\boldsymbol{\Omega}_k|^{\frac{q}{2}} |\boldsymbol{\Gamma}_k|^{\frac{p}{2}} \\ \exp\left\{ -\frac{1}{2} \mathrm{tr}(\boldsymbol{\Omega}_k (\mathbf{X}_i - \mathbf{M}_k) \boldsymbol{\Gamma}_k (\mathbf{X}_i - \mathbf{M}_k)') \right\},\end{aligned}$$

with $\mathbf{M}_k$ representing the mean matrix and $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$ are the rows and columns precision matrices with dimensions $p \times p$ and $q \times q$, respectively. The previously mentioned over-parametrization issue deeply affects the model in (1), since the number of parameters $\boldsymbol{\Theta} = \{\tau_k, \mathbf{M}_k, \boldsymbol{\Psi}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ scales quadratically with both dimensions p and q. Motivated by this, the present manuscript proposes a matrix-variate extension to the two-way penalized model-based clustering framework introduced in [9], assuming that $\mathbf{M}_k$, $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$, $k = 1, \ldots, K$, possess some degree of sparsity. The resulting model flexibly accounts for cluster-wise conditional independence patterns, providing a unified way for jointly reducing the number of estimated parameters and eliminating redundant variables, combining advantages of state-of-the-art procedures for matrix-variate clustering [5, 8].

The remainder of the paper proceeds as follows: in Section 2 we introduce our new proposal and we discuss its main methodological aspects. An application to

soil classification by means of satellite images is reported in Section 3. Section 4 summarizes the novel contributions and highlights future research directions.

2 Penalized matrix-variate mixture model

A penalized likelihood approach is introduced for parameter estimation. The resulting objective function to be maximized with respect to $\boldsymbol{\Theta}$ is:

$$\ell(\boldsymbol{\Theta};\mathbf{X}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i;\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k) \right\} - p_{\lambda_1,\lambda_2,\lambda_3}(\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k), \quad (2)$$

with the penalization term $p_{\lambda_1,\lambda_2,\lambda_3}(\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k)$ being equal to

$$p_{\lambda_1,\lambda_2,\lambda_3}(\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k) = \sum_{k=1}^{K} \lambda_1 ||\mathbf{P}_1 * \mathbf{M}_k||_1 + \sum_{k=1}^{K} \lambda_2 ||\mathbf{P}_2 * \boldsymbol{\Omega}_k||_1 + \sum_{k=1}^{K} \lambda_3 ||\mathbf{P}_3 * \boldsymbol{\Gamma}_k||_1.$$

With $*$ we denote element-wise product, $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ are matrices with non-negative entries, λ_1, λ_2 and λ_3 are penalty coefficients and $||\mathbf{A}||_1 = \sum_{jh} |A_{jh}|$. A dedicated EM-algorithm is devised for inference by firstly defining a suitable *penalized complete log-likelihood* for model (2):

$$\begin{aligned}\ell_C(\boldsymbol{\Theta};\mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \Big[&\log \tau_k - \frac{pq}{2}\log 2\pi + \frac{q}{2}\log|\boldsymbol{\Omega}_k| + \frac{p}{2}\log|\boldsymbol{\Gamma}_k| + \\ &-\frac{1}{2}\operatorname{tr}\left\{ \boldsymbol{\Omega}_k (\mathbf{X}_i - \mathbf{M}_k)\boldsymbol{\Gamma}_k (\mathbf{X}_i - \mathbf{M}_k)' \right\} \Big] - p_{\lambda_1,\lambda_2,\lambda_3}(\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k) \quad (3)\end{aligned}$$

where as usual for mixture models $z_{ik} = 1$ if observation $\mathbf{X}_i$ belongs to the k-th component, and 0 otherwise. The E-step at the t-th iteration requires computing the estimated a posteriori probabilities of class membership $\hat{z}_{ik}^{(t)}$, achieved via the standard updating formula. On the other hand, the M-step involves a partial optimization strategy. Let us denote with m_{lsk}, x_{lsi}, ω_{lsk}, γ_{lsk} and p_{ls1} the element in the l-th row and s-th column of matrices $\mathbf{M}_k$, $\mathbf{X}_i$, $\boldsymbol{\Omega}_k$, $\boldsymbol{\Gamma}_k$ and $\mathbf{P}_1$. The sparse estimation of $\mathbf{M}_k$ is achieved via a cell-wise coordinate ascent algorithm, where $\hat{m}_{lsk}^{(t)} = 0$ if

$$\begin{aligned}\left| \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[\sum_{\substack{r=1 \\ r\neq l}}^{p} \hat{\omega}_{lrk}^{(t-1)} \left(\sum_{c=1}^{q} \left(x_{rci} - \hat{m}_{rck}^{(t)} \right) \hat{\gamma}_{csk}^{(t-1)} \right) + \right. \right. \\ \left. \left. + \; \hat{\omega}_{llk}^{(t-1)} \left(\sum_{\substack{c=1 \\ c\neq s}}^{q} \left(x_{lci} - \hat{m}_{lck}^{(t)} \right) \hat{\gamma}_{csk}^{(t-1)} \right) + \hat{\omega}_{llk}^{(t-1)} x_{lsi} \hat{\gamma}_{ssk}^{(t-1)} \right] \right| \leq \lambda_1 p_{ls1}, \quad (4)\end{aligned}$$

otherwise, $\hat{m}_{lsk}^{(t)}$ is obtained by solving

$$\hat{n}_k^{(t)}\hat{\omega}_{llk}^{(t-1)}\hat{m}_{lsk}^{(t)}\hat{\gamma}_{ssk}^{(t-1)} + \lambda_1 p_{ls1}\,\text{sign}\left(\hat{m}_{lsk}^{(t)}\right) = \sum_{i=1}^{n}\hat{z}_{ik}^{(t)}\sum_{r=1}^{p}\sum_{c=1}^{q}\hat{\omega}_{lrk}^{(t-1)}x_{rci}\hat{\gamma}_{csk}^{(t-1)} + \\ - \hat{n}_k^{(t)}\left(\sum_{\substack{r=1\\ r\neq l}}^{p}\sum_{\substack{c=1\\ c\neq s}}^{q}\hat{\omega}_{lrk}^{(t-1)}\hat{m}_{rck}^{(t)}\hat{\gamma}_{csk}^{(t-1)}\right) \quad (5)$$

with respect to $\hat{m}_{lsk}^{(t)}$, where $\hat{n}_k^{(t)} = \sum_{i=1}^{n}\hat{z}_{ik}^{(t)}$. Lastly, expressions for estimating sparse precision matrices $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$ rely on dedicated modifications of the the coordinate descent graphical LASSO [3].

3 Application to Satellite Data

The data encompass 397, 211 and 237 satellite images of respectively grey soil, damp grey soil and soil with vegetation stubble. Each scene (represented by $q = 9$ pixels) is recorded $p = 4$ times with different spectral bands, resulting in $n = 845$ samples of 4×9 matrices. The methodology described in Section 2 is employed to perform clustering on this three-way dataset, mimicking the analyses performed in [5, 7]. Table 1 reports the classification error rate, Adjusted Rand Index and number of estimated parameters for our method and two competing procedures, namely constrained MGMM [5] and standard GMM applied to the two-way representation of the data, obtained by unfolding the original three-way data into a 845×36 matrix. Our model not only succeeds in better retrieving the true underlying data partition, but it is also the most parsimonious, displaying the lowest number of non-zero estimated parameters. The resulting sparse structures retrieved by our proposal are

Table 1 Misclassification errors, Adjusted Rand Index and number of free estimated parameters for three clustering procedures, Satellite Data. *Sparsemixmat* denotes the proposal introduced in the present paper.

	Sparsemixmat	Sarkar et al. [5]	Mclust [6]
Misclassification error	0.0793	0.0828	0.3053
Adjusted Rand Index	0.7883	0.7772	0.3841
# of free parameters	218	275	850

showcased in Figure 1, where estimated parameters for the three soil types are displayed. Matrix entries that are shrunk to 0 by the penalized estimator are highlighted with an $\times$ symbol in the plots. As expected, the clustering is mainly driven by the different patterns in the mean matrices, while the column-precision matrices $\hat{\boldsymbol{\Gamma}}_k$, $k = 1, \ldots, 3$ possess the highest level of sparsity.

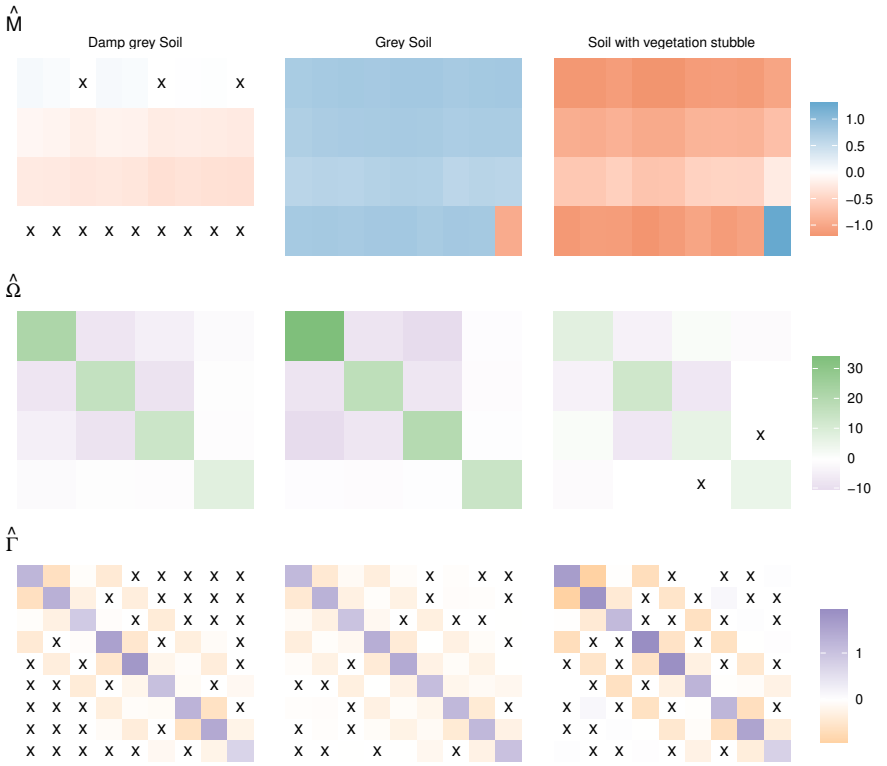


Fig. 1 Estimated sparse mean matrices (upper plots), row-precision matrices (middle plots) and column-precision matrices (lower plots) for the three clusters, satellite data. Matrix entries that are shrunk to 0 by the penalized estimator are highlighted with an $\times$.

4 Conclusion

The present work has introduced a novel penalized matrix-variate mixture model, able to capture heterogeneity and redundancy in three-way data structures. By means of sparse estimation, we are able to overcome the over-parametrization issue occuring in MGMM when either the row or the column dimensions increase.

Future research directions aim at deriving an efficient procedure for performing model selection. Jointly determining the best values for the penalty coefficients, as well as the number of mixture components define a challenging computational problem: feasible solutions are currently being investigated.

References

[1] C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, 71:52–78, 2014.

[2] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-Based Clustering and Classification for Data Science*, volume 50. Cambridge University Press, jul 2019.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, jul 2008.

[4] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2004.

[5] S. Sarkar, X. Zhu, V. Melnykov, and S. Ingrassia. On parsimonious models for modeling matrix data. *Computational Statistics and Data Analysis*, 142:106822, 2020.

[6] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1):289–317, 2016.

[7] C. Viroli. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21(4):511–522, 2011.

[8] Y. Wang and V. Melnykov. On variable selection in matrix mixture modelling. *Stat*, 9(1):1–11, dec 2020.

[9] H. Zhou, W. Pan, and X. Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3:1473–1496, 2009.

Does Milan have a smart mobility? A clustering analysis approach

Milano ha una mobilità veramente smart? Un approccio basato sulla cluster analysis

Nicola Cornali, Matteo Seminati, Paolo Maranzano and Paola M. Chiodini

Abstract This paper aims at evaluating the level of mobility services and infrastructures in the neighbourhoods of Milan to establish which areas are better equipped concerning the citizens' needs. The main objective is to measure the overall degree of mobility in the city by ranking the 88 administrative neighbourhoods (NILs) according to their transportation services. We propose a two-steps statistical approach. In the first step, we build several aggregative indicators based on various mobility variables. In the second stage, the above indices are used to rank the NILs and cluster them in homogenous groups with similar mobility levels. The cluster analyses successfully group the neighbourhoods in a suitable way, distinguishing key areas of the city, the city centre, workplaces and suburbs.

Abstract *Il presente lavoro valuta il livello dei servizi e delle infrastrutture di mobilità negli 88 quartieri amministrativi di Milano (NIL) per stabilire quali aree sono meglio attrezzate per soddisfare le esigenze dei cittadini. L'obiettivo principale è quello di misurare il grado complessivo di mobilità della città, raggruppando i NIL in base ai loro servizi di trasporto. Proponiamo un approccio a due stadi in cui prima costruiamo diversi indicatori compositi sulla base di alcune variabili di mobilità e poi usiamo tali indici per raggruppare i NIL in gruppi con livelli di mobilità simili. Gli algoritmi di clustering riescono a raggruppare i quartieri in modo adeguato, distinguendo le aree chiave della città, il centro città, i luoghi di lavoro e le periferie.*

Key words: Clustering; Aggregative indices; AMPI index; Milano NILs; Smart mobility;

Nicola Cornali, e-mail: cornali.nicola@gmail.com
Matteo Seminati, PlusValue, e-mail: matteo.seminati@plusvalue.org
Paolo Maranzano, University of Bergamo, e-mail: paolo.maranzano@unibg.it
Paola M. Chiodini, University of Milano-Bicocca, e-mail: paola.chiodini@unimib.it

1 Introduction

This paper aims at evaluating the levels of mobility services and infrastructures in the different neighbourhoods of Milan to establish which areas are better equipped concerning the citizens' needs. The research's main objective is to measure the overall degree of mobility in the city by ranking the 88 administrative neighbourhoods, or NILs, according to their transportation services. We then propose a two-steps statistical approach. In the first step, we build several composite indicators aiming at quantifying the mobility level for each neighbourhood of the city based on various mobility variables. In the second stage, the above indices are used to rank the NILs and cluster them in homogenous groups characterised by similar mobility levels.

2 Available data and two-stage statistical modelling

The dataset used for the analysis is obtained merging different open data provided by the municipality of Milan and concerning the 88 administrative neighbourhoods of the city. We considered a set of variables that characterize smart urban mobility along three relevant dimensions: public transport means, sharing mobility means, and mobility infrastructures. All the considered variables refer to 2018 and have been weigthed by the resident population. Table 1 reports the list of twelve elementary indicators considered in our application. Note that the number of stops and the number of rides for each means of transport has to be considered separately, hence we considered four variables for the stops and four for rides.

Variable	Description
Metro/tram/bus/trolleybus stops	Number of metro/tram/bus/trolleybus stops in each NIL
Metro/tram/bus/trolleybus rides	Average weekly metro/tram/bus/trolleybus rides in each NIL
Parking spaces GuidaMi	Number of all GuidaMi parking spaces in each NIL
BikeMi slots	Number of all BikeMi spaces in each NIL
Recharging columns	Number of columns for electric recharging in each NIL
Bike slots	Number of all bike spaces in each NIL

Table 1: Description of variables considered

The first step is the creation of multiple mobility measures using a composite approach to rank the NILs. We firstly choose and normalize a group of elementary indicators, computed as the ratio between the selected variables and the resident population in 2018. Subsequently, we aggregate them using four types of composite indicators: the geometric mean, the 0-1 mean, the Adjusted Mazziotta-Pareto Index (AMPI) and the Static Jevons Index (JJI) [3].

Let $j = 1,...,m = 12$, be the subscript of the elementary indicators and let $i = 1,...,n = 88$ be the subscript for each NIL. Recall that the AMPI index grounds on a min-max transformation of the original values such that the normalized values (y_{ij}) are constrained in the range 60 to 70. Denoting with M_{yi} the mean, S_{yi} the standard deviation and cv_{yi} the coefficient of variation of the normalized values y_{ij} for unit i, the composite index is given by $AMPI_i^{+/-} = M_{y_i} \pm S_{y_i} cv_{y_i}$ where the sign ± depends on the kind of the phenomenon to be measured. Therefore, the AMPI decomposes the score of each unit into two parts: M_{zi} that represents the mean level and $S_{zi} cv_{zi}$, i.e. the penalty. The penalty is a function of the indicators' variability concerning the mean value, and it is used to penalize the units [3]. The JJI for unit i represents the geometric mean of all the elementary indicators and is defined as $JJI_i^t = \prod_{j=1}^{m} ((x_{ij}^t)/(x_{rj}^t) \cdot 100)^{1/m}$ where x_{rj}^t is the reference value, for instance the average, and x_{ij}^t is the value of indicator j for unit i, at time t $\forall x_{ij}^t > 0$ (j=1,. . . ,m; i=1,. . . ,n; t=t_0, t_1).

We then implemented a leave-one-out influence analysis to identify the optimal index. Following the methodology proposed by [3], we compared the indices according to several descriptive statistics computed on the distances among the values of the indicators estimated by including or eliminating each of the underlying elementary variables. Priority is given to those indicators with low variability and small deviations when varying the sub-indices (robustness). Let r_{ij} be the rank of the i_{th} NIL computed without the j_{th} indicator and let r_i be the rank of the i_{th} NIL computed using all the elementary indicators $j = 1,...,m = 12$. The algorithm consists of dropping each j_{th} elementary indicator from the list of m base indicators and then re-calculating the index using the remaining $m-1$ indicators. At each iteration j, the algorithm computes for each i_{th} NIL the absolute difference (or shift) between its rank position in the full-index ranking and its position in the leave-one-out ranking, i.e. $d_{ij} = |r_{ij}\text{-}r_i|$ $\forall i = 1,...,n = 88$. Having $m = 12$ elementary indicators, for each aggregative indicator, the algorithm returns $m = 12$ vectors of shifts. The sensitivity of each composite indicator is then evaluated by computing the sample mean of the shifts and the sample standard deviation for all the vectors of shifts. Therefore, we have obtained $m = 12$ averages ($\bar{X}$) and standard deviations (S_X). Finally, these values are summarised by computing the corresponding sample averages, sample standard deviations and sample variability coefficients (i.e. the ratio between the standard deviation and the mean). Thus, for every aggregative indicator, we obtain the global average shift ($\mu_{\bar{X}}$), the standard deviation of the average shifts ($\sigma_{\bar{X}}$), the variability coefficient of the average shifts ($VC_{\bar{X}}$), the average standard deviations of the shifts (μ_{S_X}), their standard deviations (σ_{S_X}), and their variability coefficients (VC_{S_X}).

Once the optimal aggregative indicators has been identified, it is used to group the city's neighbourhoods. The aim is to find out clusters of NILs

that present similar mobility degree values but are distinguished from each other as much as possible. We implemented a cluster analysis using the $k-means$ algorithm with several parameters settings. We estimated the algorithm using from two to six groups and trying multiple random starting points. The definitive number of groups has been selected by combining the GAP statistic [4], the silhouettes, and the majority rule-of-thumb for several indicators ([1, 2]).

3 Results

Indices	$\mu_{\bar{X}}$	$\sigma_{\bar{X}}$	$VC_{\bar{X}}$	μ_{S_X}	σ_{S_X}	VC_{S_X}
Mean 0-1	1.309	1.140	0.871	1.973	1.398	0.709
Static Jevons	2.298	0.907	0.395	2.317	0.536	0.231
AMPI	1.335	1.106	0.828	2.095	1.501	0.716
Geometric mean	1.468	1.249	0.851	2.091	1.474	0.705

Table 2: Sensitivity analysis results

The descriptive statistics of each composite indicator are reported in Table 2. Regarding the indices' variability, the static JJI performs very well, as it presents the lowest values both for the mean and the standard deviation. However, both the global average shift and the average variability are the largest among those estimated. This lack of robustness of the JJI may derive directly from its construction method, which only considers index numbers and ignores the sample variability. The AMPI performs well in terms of the average shift but has slightly higher variability values. The two results appear to be somewhat complementary and suggest to consider both the indices within the cluster analysis.

Whether JJI or AMPI is considered, the criteria for selecting the optimal number of groups are consistent. The majority rule-of-thumb suggests considering three or four groups, while both the silhouette and the GAP statistic suggest exactly four clusters. Therefore, we proceed to cluster the NILs setting the number of potential clusters to four. We considered outlier seven NILs as they present a null value of mobility or show a shallow level of resident population (≤ 50 inhabitants). We then decided to directly consider them as a separated cluster not comparable with the others.

We performed and compared three different K-means algorithms specifications: the first one uses as clustering variable just the Jevons index (JJI clustering), the second uses only the AMPI (AMPI clustering), whereas the third combine both indices (JJI-AMPI clustering). The cluster analysis results are represented in Figure 1.

The four clusters can be characterized by ascending order of mobility:

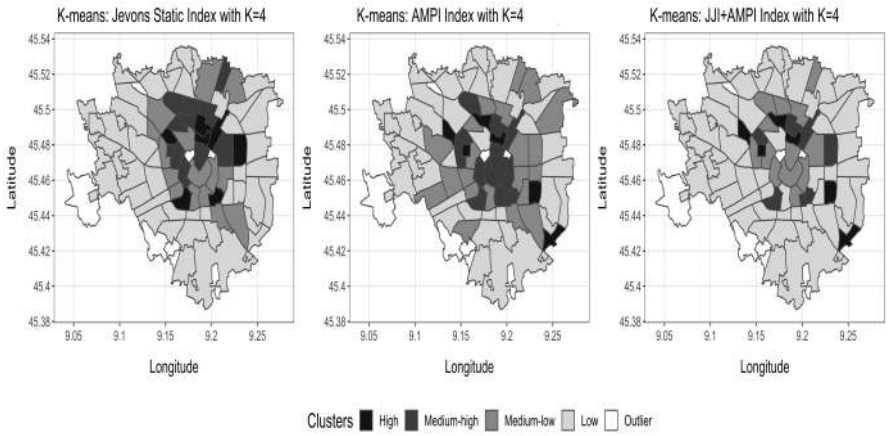


Fig. 1: Clustering maps

- **Low**: suburbs and peripheral areas, characterized by deficient levels of mobility. They are located in the southern and western part of Milan or at the eastern border;
- **Medium-low**: NILs with a high resident population level but with low mobility service availability. These are areas where people transit or commute and live. They host a high number of offices and companies;
- **Medium-high**: medium-high mobility areas composed by university poles, former industrial areas, and strongly inhabited and wealthy-class neighbourhoods;
- **High**: high mobility areas where citizens and commuters transit to reach offices or specific NILs characterized by a strong presence of a single mobility service.

The JJI clustering can well separate peripheral areas and suburbs by the historical city centre, the transit sites, and nightlife areas. The high-mobility group includes the main commuting points, the nightlife areas, as well as the Politechnical university area at East. The medium-low cluster mainly coincides with the city centre and some further inhabited areas. The suburbs form a unique cluster surrounding the centre.

The AMPI-based clustering can separate those areas not well-discriminated by the JJI, such as the Linate airport area (at south East) or the recent and highly-coveted neighbour of City Life (at north-west). The algorithm groups them within the high-mobility cluster. Moreover, the AMPI identifies the city centre more clearly than JJI, which now coincides with the limited traffic zone (Area C) extended to the commuting points and nightlife neighbours. The cluster associated with a medium-low mobility degree here includes all the university campuses and several inhabited areas. As in the JJI-based

case, the AMPI well separates the suburbs, i.e. low mobility degree, and extends the middle-mobility NILs, i.e. medium-low group a larger area.

The JJI-AMPI cluster inherits the useful properties of both the single-index approaches. However, the result appears to be strongly dependent on AMPI's behaviour, as it holds the high values of AMPI and Jevons' low values. The algorithm well-separate the city centre by the suburbs, covering the entire peripheral areas of Milan as in the JJI-based case. The city centre is classified as medium-low as in the JJI output. The medium-low cluster also includes some northern NILs, which are highly inhabited areas hosting several university campus and train stations. The JJI-AMPI algorithm separates into two different clusters those NILs that were individually classified as high-mobility areas. In this scenario, the AMPI high-mobility clusters remain in the high-mobility group, while JJI high-mobility NILs are now classified as medium-high mobility areas. Moreover, the medium-high group includes some other highly mobile neighbourhoods and commuting points, whereas the high mobility cluster includes transit-sites-NILs.

4 Conclusions

In this paper, we investigated smart mobility's actual state in Milan using a two-stage statistical approach. In the first stage, we built a set of aggregative indicators to measure the mobility degree at each neighbourhood of the city. In the second stage, the aggregative indices are used to rank and cluster the NILs in homogenous groups characterised by similar mobility degrees. Whether by using each indicator individually or by combining them, the cluster analyses are successful in grouping neighbourhoods distinguishing critical areas of the city, such as interchange hubs and university zones. The results highlight substantial differences in terms of mobility among the old town, which is characterised by very high mobility levels, and the suburbs, which have fewer mobility services.

References

1. T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.
2. W. J. Krzanowski and Y. Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34, 1988.
3. M. Mazziotta and A. Pareto. *Synthesis of indicators: The composite indicators approach*, pages 159–191. Springer, 2017.
4. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63(2):411–423, 2001.

A Fuzzy clustering approach for textual data

Un approccio di clustering di tipo fuzzy per dati testuali

Irene Cozzolino, Maria Brigida Ferraro and Peter Winker

Abstract Document clustering is a process of partitioning a corpus of documents into distinctive clusters based on the content similarity. Traditional (hard or fuzzy) document clustering algorithms are usually relying on the vector representation of documents based on the bag-of-words (BOW) approach, leading to very high dimensions in the vector representation of the corpus. In recent years, spectral clustering has been extensively applied in the field of text classification with support vector machines (SVMs) in combination with string kernels, but little has been done in the field of fuzzy document clustering with kernel-based methods. This work proposes a novel approach to text clustering, by grouping documents into clusters based on a new version of fuzzy spectral clustering with string kernels.

Abstract *Le procedure di clustering applicate a dati testuali hanno come obiettivo quello di identificare gruppi di documenti caratterizzati da contenuti simili. Gli algoritmi di clustering tradizionali (hard o fuzzy) utilizzano il modello bag-of-words (BOW), rappresentando ciascun documento come un vettore sparso contenente il numero di occorrenze delle parole in esso contenuto. Negli ultimi anni il clustering spettrale è stato ampiamente utilizzato nell'ambito della classificazione dei documenti tramite le support vector machines (SVMs) insieme con gli string kernels; tuttavia molto poco è stato fatto nel campo del clustering di tipo fuzzy in combinazione con i metodi kernel-based. In questo lavoro proponiamo una procedura di clustering spettrale di tipo fuzzy da applicare a dati testuali che impiega l'utilizzo di string kernels.*

Key words: Document clustering, Fuzzy approach, Kernel-based clustering, String kernels

---

Irene Cozzolino
Sapienza University of Rome, e-mail: irene.cozzolino@uniroma1.it

Maria Brigida Ferraro
Sapienza University of Rome, e-mail: mariabrigida.ferraro@uniroma1.it

Peter Winker
Justus Liebig University of Giessen, e-mail: Peter.Winker@wirtschaft.uni-giessen.de

1 Introduction

In modern applications, statistical models involve working with large collection of textual documents that should be properly analysed in order to extract the significant information contained in them.

Clustering techniques can be very useful in the text domain, where the objects to classify can be of different granularities such as documents, paragraphs, sentences or terms. Indeed, document clustering algorithms are commonly used to automatically organize, understand, search and summarize large electronic archives.

Over the past decades, different clustering algorithms, such as K-means [17], have been widely applied to textual data.

However depending on the type of documents analyzed, it might be often the case that they do not contain only information related to a single topic. Furthermore, there might be an overlap of contents characterizing different knowledge domains. Hence, documents may contain information that is relevant to different areas of interest to some degree. With soft clustering methods documents are attributed to several clusters simultaneously and thus, useful relationships between domains may be uncovered, which would otherwise be neglected by hard clustering methods.

This is the idea behind the work in [19], where after representing documents as term frequency vectors, the authors have modified the fuzzy K-means algorithm for clustering text documents based on the cosine similarity coefficient rather than on the Euclidean distance, leading to the Hyper-spherical Fuzzy K-Means algorithm (H-FKM). Further studies [3], [10], [18] have proposed other approaches for using fuzzy clustering algorithms in the document clustering process.

2 A fuzzy version of spectral clustering with string kernels

2.1 Basic concepts of spectral clustering

Spectral clustering methods arise from concepts in spectral graph theory [20]. The basic idea is to construct a weighted graph from the initial data set where each node represents a pattern and each weighted edge simply takes into account the similarity between two patterns. In this framework the clustering problem can be seen as a graph cut problem.

The key to spectral clustering is to select an adequate similarity measure, which can well describe the intrinsic structure of data points. Data in the same groups should have high similarity. In the document clustering-domain, string kernel functions are usually adopted as similarity measures [16]. The most commonly used string kernel function is usually referred to as *spectrum*. The idea behind it is to compare two documents by means of the substrings of exactly same length they contain, where a substring is defined as a sequence of n characters occurring in the text though not necessarily contiguously. The more substrings of length n the docu-

ments have in common, the more similar they are. The generic form of string kernels is given by equation [13]:

$$k(x,x') = \sum_{s \in A^\star} num_s(x) num_s(x') \lambda_s \tag{1}$$

where $A^\star$ represents the set of all non empty strings, x and x' are two sequences of characters and λ_s is a decay factor that can be used to weight the presence of a certain feature in a text. In the *spectrum* string kernel the decay factor is maintained fixed for all the matching substrings. In our work we study the effects of spectral clustering by keeping the value of the decay parameter λ_s fixed to 1.1 (default value) and varying the length parameter, n, which controls the length of the substrings.

The general approach to spectral clustering is to use a standard clustering method, such as K-means, on the first K eigenvectors of the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{S}$, built from the similarity matrix $\mathbf{S}$ and the degree matrix $\mathbf{D}$. This last one is a diagonal matrix with main diagonal elements equal to the degrees of the nodes. The Laplacian matrix is usually normalized. For a more detailed explanation on spectral clustering see, e.g., [6], [9] and [11].

2.2 Fuzzy version

Fuzzy spectral clustering algorithms can be developed as a straightforward extension. We adapted the fuzzy version as reported in [7] to string kernels. It is sufficient to replace the K-means algorithm with the fuzzy K-means [1] when analyzing the normalized eigenvectors of the Laplacian matrix.

The fuzzy K-means minimizes the functional:

$$J(\mathbf{U},\mathbf{H}) = \sum_{i=1}^{N} \sum_{g=1}^{K} u_{ig}^m ||\mathbf{x}_i - \mathbf{h}_g||^2 \tag{2}$$

with respect to the membership matrix $\mathbf{U}$ and the prototype matrix $\mathbf{H}$ with the constraint $\sum_g u_{ig} = 1$. The parameter m controls the fuzziness of the memberships: for high values of m the algorithm tends to set all the memberships equals, while for m tending to one we obtain the K-means algorithm where the partition is crisp.

3 Results

In this section we report the results of the application of the fuzzy version of spectral clustering with string kernels on Reuters-21578 dataset [15], which contains stories for the Reuters news agency and it is publicly available. We use the classes “trade” vs “ship” (with, respectively, 326 and 144 documents) and “crude” vs “money-fx”

(with 374 and 293). In the pre-processing phase, we removed the punctuation signs and the numbers; the remaining words have been lower-cased and we applied the Porter's Stemmer algorithm [21].

We considered string kernel by using the function `stringdot` in the package `kernlab` [14], specifying the option `spectrum` which considers only matching substring of exactly length n. In order to learn more about the influence of the length parameter n on clustering results, we run the algorithm over a range of values for n, from $n = 3$ to $n = 8$.

Concerning fuzzy K-means, we run the function `FKM` of the package `fclust` [8] on the normalized eigenvectors of the Laplacian matrix. For each value of n we let m vary from 1.1 to 2. Moreover, three random starts were used in order to limit the risk of hitting local optima.

We evaluate the performance of the clustering algorithm by using the average fuzzy silhouette index [5] and the fuzzy adjusted Rand index (ARI) [4]. The former is an internal validation measure relying only on information in the data; it evaluates the goodness of the clustering structure without respect to external information. The latter is an external validation measure which is used when the "true" cluster labels are known in advance.

The cluster validity indexes returned in both cases $n = 5$ and $m = 2$ as the optimal hyper-parameters, whose values are reported in Table 1. Table 2 and Table 3 are agreement tables between the known partition and the partition determined by the clustering algorithm. Note that the fuzzy version of spectral clustering correctly classifies the majority of observations, but misclassifies 14 observations in the first example and 12 in the second one.

Table 1 Cluster validity indexes: fuzzy silhouette and fuzzy adjusted Rand index.

Classification	Fuzzy silhouette	Fuzzy ARI
"trade" vs "ship"	0.8688571	0.9794686
"crude" vs "money-fx"	0.8961476	0.9872041

Table 2 Classification of "trade" vs "ship" by fuzzy spectral clustering with string kernel.

Classes	Cluster 1	Cluster 2	Cases
ship	143	1	144
trade	13	313	326
Total	156	314	470

The performance of the fuzzy spectral clustering algorithm is consistent with the corresponding hard version but, whilst in the hard case the obtained membership

Table 3 Classification of "crude" vs "money-fx" by fuzzy spectral clustering with string kernel.

Classes	Cluster 1	Cluster 2	Cases
crude	363	11	374
money-fx	1	292	293
Total	364	303	667

degrees were either 1 or 0, highlighting a clear assignment of the objects to the clusters, in the fuzzy approach the misclassified objects for both problems are assigned to the wrong clusters with membership degrees in the interval $[0.5, 0.7]$ while the correctly classified objects present higher membership degrees, ranging in the interval $[0.7, 1]$.

Adjacency matrix and Laplacian matrix are commonly used representations for weighed graph [12]. In Figure 1 we represent the Laplacian graphs for both examples, highlighting the assignment of objects to clusters.

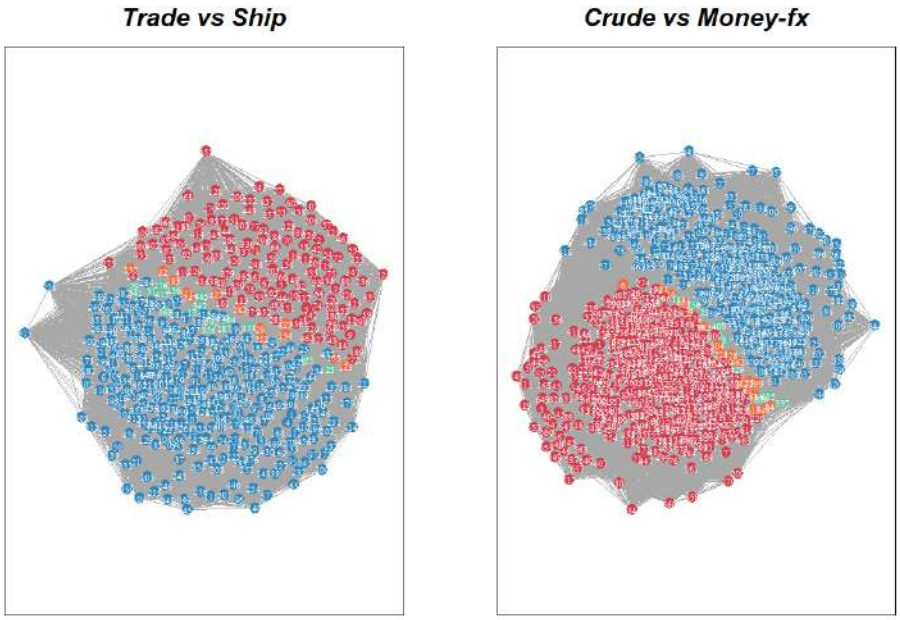


Fig. 1 The fuzzy spectral clustering with string kernel solution with $n = 5$ and $m = 2$; red and blue points denote objects assignments to Cluster 1 and Cluster 2, respectively, with membership degrees higher than 0.70. In particular, for each cluster, light colours (orange and jade green) denote membership degrees in the intervals $[0.50, 0.70)$.

This research work attains the goal of discovering ways to improve document clustering algorithms, since the fuzzy approach provides additional insights going beyond those obtained by crisp clustering. Future works include to test the proposed clustering framework against other fuzzy approaches, such as the hyper-spherical fuzzy K-means.

References

1. J.C. Bezdek: Pattern Recognition with Fuzzy Objective Function Algorithm. Plenum Press, New York (1981)
2. D. M. Blei, A. Y. Ng and M. I. Jordan: Latent Dirichlet Allocation. Journal of Machine Learning Research, **3**, 993–1022 (2003)
3. C. Borgelt, A. Nurnberger: Fast Fuzzy Clustering of Web Page Collections. PKDD Workshop on Statistical Approaches for Web Mining (2004)
4. R. J. Campello. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters, **28**, 833–841 (2007)
5. R. J. Campello, E. R. Hruschka. A fuzzy extension of the silhouette width criterion for cluster analysis. Pattern Recognition Letters, **157**, 2858–2875 (2006)
6. W. E. Donath, A. J. Hoffman. Lower bounds for the partitioning of graph. IBM Journal of Research & Development, **17**, 420–425 (1973)
7. M. B. Ferraro, P. Giordani. A review and proposal of (fuzzy) clustering for nonlinearly separable data. International Journal of Approximate Reasoning, **115**, 13–31 (2019)
8. M.B. Ferraro, P. Giordani, A. Serafini. fclust: an R package for fuzzy clustering. R J, **9** (2019), available via https://doi .org /10 .32614 /RJ -2019 -017
9. M. Fiedler. Algebraic connectivity of graphs. Czechoslovak Mathematical Journal, **23**, 298–305 (1973)
10. M. Friedman et al. A Fuzzy-Based Algorithm for Web Document Clustering. IEEE Annual Meeting of the Fuzzy Information, **2**, 524–527 (2004)
11. L. Hagen, A. B. Kahng. New spectral methods for radio cut partitioning and clustering. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, **11**, 1074–1085 (1992)
12. R. Janani, S. Vijayarani. Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization. Expert Systems With
13. A. Karatzoglou, I. Feinerer: Text Clustering with String Kernels in R. In: Advances in Data Analysis, 91–98. Springer (2006)
14. A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis. kernlab – An S4 Package for Kernel Methods in R. Journal of Statistical Software, **11**, 1–20 (2004)
15. D. Lewis: Reuters-21578 Text Categorization Test Collection. (1997)
16. H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins. Text Classification using String Kernels. Journal of Machine Learning. **2**, 419–444 (2002)
17. J.B. MacQueen. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathe-matical Statistics and Probability, University of California Press. 281–297 (1967)
18. M.E.S. Mendes Rodrigues, L. Sacks. Evaluating fuzzy clus- tering for relevance-based information access. Proceedings of the 12th IEEE International Conference on Fuzzy Systems (2003)
19. M.E.S. Mendes Rodrigues, L. Sacks. A Scalable Hierarchical Fuzzy Clustering Algorithm for Text Mining. Proceedings of the 5th International Conference on Recent Advances in Soft Computing (2004)
20. A. Ng, M. Jordan, Y. Weiss. On spectral clustering: analysis and an algorithm. Advances in Neural Information Processing Systems. **14**, 849–856 (2002)
21. M. F. Porter: An algorithm for suffix stripping. Program. **14** 130–137 (1980)

Valid Double-Dipping via Permutation-Based Closed Testing

Inferenza per Analisi Esplorative Tramite Metodi di Permutazione e Closed Testing

Anna Vesely, Livio Finos, Jelle J. Goeman and Angela Andreella

Abstract Functional Magnetic Resonance Imaging (fMRI) cluster analysis is widely popular for finding neural activation associated with some stimulus. However, it suffers from the spatial specificity paradox, and making follow-up inference inside clusters is not allowed. Valid double-dipping can be performed by closed testing, which determines lower confidence bounds for the number of active voxels, simultaneously over all regions. Moreover, a permutation framework adapts to the unknown joint distribution of the data. In the fMRI context, we evaluate two methods that rely on closed testing and permutations: permutation-based true discovery guarantee by sum tests, and permutation-based All-Resolutions Inference.

Abstract *L'analisi dei cluster di risonanza magnetica funzionale (fMRI) permette di individuare l'attivazione neurale associata a qualche stimolo. Tuttavia, il metodo soffre del paradosso della specificità spaziale, e non permette inferenza di follow-up all'interno dei cluster. La procedura di closed testing consente il double-dipping, fornendo i limiti inferiori di confidenza per il numero di voxel attivi, simultaneamente su tutte le regioni. Inoltre, i test di permutazione permettono di adattare il metodo alla distribuzione congiunta non nota dei dati. Si esaminano tramite analisi di dati fMRI due metodi basati sul closed-testing via permutazioni: garanzia di true discovery tramite test di somma, e All-Resolutions Inference.*

Anna Vesely
Department of Statistical Sciences, University of Padua, Italy,
e-mail: anna.vesely@phd.unipd.it

Angela Andreella
Department of Statistical Sciences, University of Padua, Italy,
e-mail: angela.andreella@phd.unipd.it

Livio Finos
Department of Developmental Psychology and Socialization, University of Padua, Italy,
e-mail: livio.finos@unipd.it

Jelle J. Goeman
Biomedical Data Sciences, Leiden University Medical Center, Netherlands,
e-mail: j.j.goeman@lumc.nl

Key words: True Discovery Proportion, Permutation Test, Multiple Testing, Selective Inference, fMRI Cluster Analysis

1 Introduction

In functional Magnetic Resonance Imaging (fMRI) analysis, brain activation is measured as the correlation between the sequence of cognitive stimuli and blood oxygenation levels. The brain image comprises approximately 300,000 volume units (voxels), each of which may be tested for significant neural activity. This results in a multiple testing problem with about 300,000 statistical tests.

Controlling Type I error at the voxel level is conservative. Cluster-extent based thresholding, which exploits the signal's spatial nature to analyze data at the level of clusters of contiguous voxels, is less conservative than voxel-wise inference. However, it suffers from the spatial specificity paradox [14]. As the method tests the null hypothesis that there is no active voxel in the cluster, rejecting such null hypothesis does not provide any information on the proportion of active voxels and their spatial location. Moreover, follow-up inference inside the cluster ("drilling down") leads to a double-dipping problem and inflated Type I error rate [8].

Closed testing allows to solve this problem by constructing lower confidence bounds for the proportion of active voxels, simultaneously over all possible brain regions [4]. Simultaneity ensures that the bounds remain valid even when the region of interest is selected post hoc, and when drilling down within clusters. An example of method using closed testing in this way is All-Resolutions Inference (ARI) [9].

We compare the performance of two methods that rely on closed testing via permutations: true discovery guarantee by sum tests [11], and ARI [2]. Both employ permutation testing in order to adapt to the unknown joint distribution of the data. We focus on the permutation framework since it often offers an improvement in power over the parametric approach, especially with multiple hypotheses [6].

The paper is organized as follows. In Section 2, we illustrate the methods under analysis. In Section 3, we evaluate their performance on real fMRI data.

2 Permutation-based closed testing for true discovery proportion

Let B denote the brain, composed of $|B| = m$ voxels, and $A \subseteq B$ denote the unknown subset of truly active voxels. When studying a brain region $S \subseteq B$ with significance level α, we are interested in making inference on the number of true discoveries $a(S) = |A \cap S|$, i.e. the number of truly active voxels within the region. The following methods allow to construct simultaneous lower $(1-\alpha)$-confidence bounds $\bar{a}(S)$, i.e.,

$$P\left(a(S) \geq \bar{a}(S)\ \forall S \subseteq B\right) \geq 1-\alpha.$$

Simultaneous lower confidence bounds for the True Discovery Proportion (TDP), $\pi(S) = a(S)/|S|$, can be immediately derived as $\bar{\pi}(S) = \bar{a}(S)/|S|$.

The methods are based on permutation testing. Let T_i be a generic test statistic for the null hypothesis that there is no activation in voxel i ($i \in \{1,\ldots,m\}$). Let p_i be the corresponding p-value. Subsequently, suppose to randomly draw ω elements from a group of data transformations that preserve the distribution of the test statistics under the null. Denote by $T_1^j,\ldots,T_m^j$ the test statistics corresponding to the j-th transformation. An α-level permutation test with random transformations is defined in [5]. Under some condition on the group of transformations, which is usually satisfied for continuous data, the test is exact when ω is a multiple of $1/\alpha$. Finally, the power increases with ω. For the usual choices of the significance level (e.g., $\alpha = 0.05$), a value $\omega \geq 200$ is generally suitable.

2.1 sumSome: permutation-based true discovery guarantee by sum tests

The method, introduced in [11] and implemented in [12], is a general procedure for sum-based tests. This means that the test statistic T_S for the null hypothesis that there is no activation in the region S may be written as $T_S = \sum_{i\in S} T_i$, where T_i is a generic statistic. The procedure is defined as an iterative shortcut for closed testing, based on the branch and bound algorithm, which converges to the full closed testing results, often after few iterations. Even if it is stopped early, it defines valid lower confidence bounds for the number of true discoveries.

At each iteration $n \in \mathbb{N}$, the procedure constructs a collection of functions $U^n_{S,z}$: $\{z,\ldots,m\} \to \mathbb{R}$, with $S \subseteq B$ and $z \in \{0,\ldots,|S|\}$. They are increasing in n, in the sense that $U^n_{S,z}(v) \leq U^{n+1}_{S,z}(v)$. Moreover, they are such that $P(a(S) > |S| - z) \geq 1-\alpha$ if $\max_v\{U^n_{S,z}(v)\} < 0$. Lower confidence bounds for the number of true discoveries are determined by studying the sign of the maximum of each function.

Theorem 1. *For each n,* $\bar{a}^n(S) = |S| - \max\left\{z \in \{0,\ldots,|S|\} : \max_v\{U^n_{S,z}(v)\} \geq 0\right\}$ *is a lower* $(1-\alpha)$*-confidence bound for* $a(S)$*, simultaneously for all* $S \subseteq B$*. The bounds are increasing in n, as* $\bar{a}^n(S) \leq \bar{a}^{n+1}(S)$.

After a finite number of iterations, the confidence bounds $a^n(S)$ converge to those obtained from full closed testing, as defined in [4].

2.2 pARI: permutation-based All-Resolutions Inference

The method proposed in [2] merges the strengths of ARI [9] with the permutation-based method of [6]. Let define $l_1,\ldots,l_m$ a critical vector if and only if

$$\Pr(\cap_{i=1}^{|N|}\{q_{(i)} \geq l_i\}) \geq 1-\alpha \tag{1}$$

where $N = B \setminus A$ is the set of inactive voxels, and $q_{(i)}$ $(1 \leq i \leq |N|)$ are their sorted p-values. *pARI* computes l_i using the p-values null distribution p_i^j in the computation of a calibration parameter called $\lambda_\alpha \in \Lambda \subseteq \mathbb{R}$, i.e.:

$$\lambda_\alpha = \sup\{\lambda \in \Lambda : w^{-1}|\{1 \leq j \leq w : p_i^j \geq l_i(\lambda)\ \forall i\}| \geq 1-\alpha\}.$$

The λ_α-calibration permits to incorporate the unknown dependence structure of the data into the critical vector's choice. The following Theorem then computes the lower bounds for the number of true discoveries.

Theorem 2. *Let $l_1, \ldots, l_m$ satisfy* (1)*. Then for every $\emptyset \neq S \subseteq B$,*

$$\bar{a}(S) = \max_{1 \leq u \leq |S|} 1 - u + |\{i \in S : p_i \leq l_u\}|$$

is a lower $(1-\alpha)$ confidence bound of $a(S)$, simultaneously for all $S \subseteq B$.

In this paper, we consider two families of candidate vectors, i.e., $\mathscr{F} = \{l(\lambda_\alpha) : \lambda_\alpha \in \Lambda\}$. The first one is inspired by Simes' probability inequality [10], i.e., $l_i(\lambda_\alpha) = \frac{(i-\delta)\lambda_\alpha}{m-\delta}$, while the second one is derived from the asymptotically optimal rejection curves (AORC) [3], i.e., $l_i(\lambda_\alpha) = \frac{(i-\delta)\lambda_\alpha}{(m-\delta)-(i-\delta)(1-\lambda_\alpha)}$. The shift parameter $\delta \in \{0, \ldots, m-1\}$ determines how sensitive $l(\lambda_\alpha)$ will be to the smallest p-values.

For further details about *pARI*, please refer to [2]. The method is implemented in [1].

3 Application on Rhyme Data

We analyzed the `Rhyme` data collected by [15], where $K = 13$ subjects were presented with pairs of either words or pseudowords, and asked to make rhyming judgments for each pair.

Brain activation was measured as correlation between the sequence of cognitive stimuli and Blood-Oxygen-Level-Dependent (BOLD) response, detected in fMRI. Data were analyzed by means of a mixed model. For each voxel $i \in \{1, \ldots, m\}$ and each subject $k \in \{1, \ldots, K\}$, let β_{ik} represent a parameter for the contrast describing neural activation under the word stimulus. We determined an estimate $\hat{\beta}_{ik}$. Pre-processing and this first-level data analysis were performed using FSL [7].

For each voxel $i \in \{1, \ldots, m\}$, denote by μ_i the between-subject mean activation. In order to test the null hypothesis $H_{0i} : \mu_i = 0$ against the alternative $H_{1i} : \mu_i \neq 0$, we defined the one-sample t-statistic $T_i = \hat{\mu}_i / \sqrt{\hat{\sigma}_i^2 / K}$. Here $\hat{\mu}_i = \sum_{k=1}^K \hat{\beta}_{ik}/K$ and $\hat{\sigma}_i^2 = \sum_{k=1}^K (\hat{\beta}_{ik} - \hat{\mu}_i)^2/(K-1)$.

The t-statistics were computed for $\omega = 200$ transformations of the data. The first was the identity, and the remaining were randomly drawn from the group of sign-flipping transformations, as in [13]. These statistics were used as input for the analysis with the two methods of interest.

We analyzed clusters computed from Random Field Theory, considering $|T| > 3.2$ and $|T| > 4$. We applied *sumSome* by using as group statistics the sum of the t-statistics and Cauchy p-value combination. Then we applied *pARI* with Simes and AORC families of candidate vectors using $\delta = 27$ to account for signal spreading out in clusters with size at least equals 27 voxels.

3.1 Results

Results are displayed in Table 1. Figure 1 shows the lower confidence bounds for the TDP of clusters obtained from *sumSome* with the sum of t-statistics. In concordance with earlier studies, we found activation on Paracingulate Gyrus (PG), Lateral Occipital Cortex (LOC), Superior Frontal Gyrus (STG), Frontal Operculum Cortex (FOC), Putamen (P), Inferior Frontal Gyrus (IFG), Lingual Gyrus (LG), Occipital Fusiform Gyrus (OFG), Insular Cortex (IC), Cingulate Gyrus (CG), Superior Parietal Lobe (SPL), Post Central Gyrus (PCG).

Table 1 Cluster-forming threshold, size, lower confidence bound for the TDP, and coordinates of the maximum t-statistic. Clusters with $\bar{\pi}(S) = 0$ are not shown.

Cluster	Threshold	Size	% active				Voxel Coordinates		
S	t	$\lvert S\rvert$	$\bar{\pi}(S)$				x	y	z
			sumSome		pARI				
			t-statistics	Cauchy	Simes	AORC			
LOC/LG/OFG/PG/SFG FOC/P/IFG/IC/CG	3.2	34115	90.25%	80.63%	88.6%	89.2%	4	12	48
LOC/LG/OFG	4	11045	79.28%	84.69%	90.82%	92.06%	−6	−56	−12
FOC/P/IFG/IC	4	6930	67.63%	75.73%	85.38%	87.37%	−42	14	−6
PG/SFG/CG	4	2100	13.43%	29.43%	56.95%	60.53%	4	12	48
Left SPL/PCG	3.2	1546	0%	0%	2.32%	2.78%	−24	−62	44

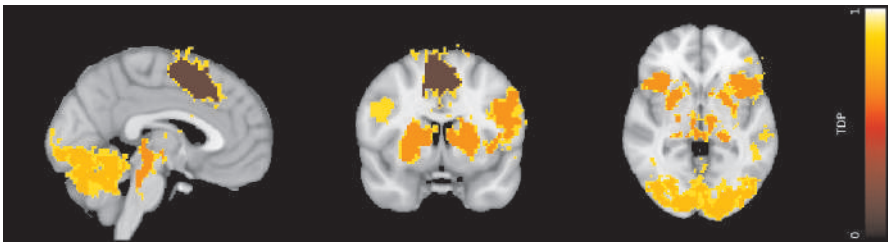


Fig. 1 Map of the lower confidence bounds for the TDP using *sumSome* with the sum of t-statistics.

The approaches have different power properties. *sumSome* appears to be more powerful for large clusters, in particular when using the sum of t-statistics. On the

contrary, *pARI* gains power in the case of smaller clusters. The performance of the AORC family moderately overtakes the Simes family's results.

4 Discussion

In fMRI cluster analysis, we compared two methods based on closed testing via permutations, *sumSome* and *pARI*. They compute lower confidence bounds for the proportion of active voxels within clusters, which are robust against post-hoc selection. As a result, they allow for double-dipping, avoiding the well-known spatial specificity paradox. When employing these approaches in the analysis of task-related fMRI data, we obtained appreciable activation in clusters defined by Random Field Theory.

References

1. Andreella, A. (2020). pari: Permutation-based All-Resolutions Inference. http://doi.org/10.5281/zenodo.4275924.
2. Andreella, A., Hemerik, J., Wouter, W., Finos, L., and Goeman, J. (2020). Permutation-based True Discovery Proportions for fMRI cluster analysis. arXiv:2012.00368v1f.
3. Finner, H., Dickhaus, T., and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *The Annals of Statistics*, 37(2):596–618.
4. Goeman, J. J., and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597.
5. Hemerik, J., and Goeman, J. J. (2018). Exact testing with random permutations. *TEST*, 27(4):811–825.
6. Hemerik, J., Solari, A., and Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106(3):635–649.
7. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *NeuroImage*, 62(2):782–790.
8. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience – the dangers of double dipping. *Nature Neuroscience*, 12:535–540.
9. Rosenblatt, J. D., Finos, L., Wouter, D. W., Solari, A., and Goeman, J. J. (2018). All-Resolutions Inference for brain imaging. *NeuroImage*, 181:786–796.
10. Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.
11. Vesely, A., Finos, L., and Goeman, J. J. (2020). Permutation-based true discovery guarantee by sum tests. arXiv:2102.11759.
12. Vesely, A. (2020). sumSome: Permutation true discovery guarantee by sum-based tests. http://doi.org/10.5281/zenodo.4531437.
13. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, 92:381–397.
14. Woo, C. W., Krishnan, A., and Wager, T. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, 91:412–419.
15. Xue, G., and Poldrack, R. A. (2020) Rhyme judgment. *OpenNeuro*, 10.18112/openneuro.ds000003.v1.0.0.

4.9 Data science for complex data

Text mining on large corpora using Taltac4: An explorative analysis of the USPTO patents database

Text mining su corpora di grandi dimensioni utilizzando Taltac4: Un'analisi esplorativa del database dei brevetti USPTO

Pasquale Pavone, Arianna Martinelli and Federico Tamagni[1]

Abstract. This paper aims to make a brief presentation of the main features and potential of the Taltac4 freeware software through an exploratory analysis of a large corpus (more than 600 million of occurrences) which includes all the abstracts of the USPTO patent documents. Patents have been extensively used as a source of information on innovative activity but the textual content of patent documents has not been fully exploited in existing research. Our preliminary results are promising and suggest that text analysis of patent abstracts can help developing new classification of innovative activities, overcoming the shortcomings of existing classifications of technologies.

Abstract. *In questo lavoro vengono brevemente illustrate le principali caratteristiche e potenzialità del software freeware Taltac4 attraverso un'analisi esplorativa di un corpus di grandi dimensioni (più di 600 milioni di occorrenze) che include tutti gli abstract dei documenti dei brevetti USPTO. I brevetti sono stati ampiamente utilizzati come fonte di informazioni sull'attività innovativa, ma il loro contenuto testuale non è stato pienamente sfruttato nella ricerca esistente. I nostri risultati preliminari sono promettenti e suggeriscono che l'analisi testuale degli abstract dei brevetti può aiutare a sviluppare una nuova classificazione delle attività innovative, superando le carenze delle classificazioni esistenti delle tecnologie.*

Key words: Text mining, large corpus, textual Big Data, patents

Pasquale Pavone, Sant'Anna School of Advanced Studies; pasquale.pavone@santannapisa.it:
Arianna Martinelli, Sant'Anna School of Advanced Studies; arianna.martinelli@santannapisa.it
Federico Tamagni, Sant'Anna School of Advanced Studies; federico.tamagni@santannapisa.it

1 Introduction

The enormous availability of textual data produced by the mass digitization of documents has generated considerable empirical data for scientific investigation. In the social sciences, the use of text mining tools and statistical methods to analyze textual Big-Data has become unavoidable. In this context, TaLTtaC4 represents an open-access tool offering great potential to analyze large collection of textual data.

TaLTaC is the acronym for "Trattamento automatico Lessico-Testuale per l'analisi del Contenuto" (lexical-textual automatic treatment for content analysis). It has been under development since 1999 within the research group coordinated by Prof. Bolasco, and has been designed for automatic text analysis in the dual logic of Text Analysis and Text Mining [4]. The previously released freeware version of TaLTaC, named TaLTaC2.11.3, has a limit on the size of the corpus it can analyze; in particular, it can analyze Corpora in text file format, with a maximum size of 150GB and 100,000 documents. The newly released TaLTtaC4 represents a substantial step forward, as it does not face limits on the size of the corpora's size to be analyzed, other than the storing limits those imposed by the machine on which TaLTtaC4 is working.

Technically, TaLTaC4 (T4) represents a multi-platform software that maximizes the exploitation of the hardware's computational capabilities. T4's architecture is divided between Graphical User Interface and computing core, communicating with each other via HTTP protocol. The computing core is capable of processing textual data in multi-process mode and thus exploits the host machine multi-core capabilities [3].

The aim of this paper is to provide a presentation of the T4 potential, analysing the large corpus of the United States Patent and Trademark (USPTO) patent documents. In the economic and innovation literatures, patent data are widely employed to measure innovative activities [6] and, over the years, scholars have been very active in exploiting information in patent documents to develop indicators highlighting patent intensity as well as different characteristics of the inventions disclosed in patents. For instance, as patent documents do not have any direct indication of the value of the inventions, some 'indirect' measures such as patent citations [14], patent renewals, patent families [8] and patent scope have been developed and validated to know more about the characteristics and qualities of inventive outputs. The number of patent claims (i.e. the list of the subject-matters protected by the patent) or the number technological domains covered in the patents have been used to measure the scope of the patents both from the technological [11] and legal point of views [9,7].

All these indicators exploit information easily retrieved from a limited section of the patent document, which is the first page. However, patents are granted over a complete disclosure of the protected invention which is described in detail in the abstract and in the remainder of the document. Increasingly easy access to the entire patent text (e.g., via the EPO-PATSTAT Database, Google patent database, web-scraping), together with advances in text mining techniques, brought about new research attempts, exploiting various parts of the text to unfold a number of

invention's characteristics such as patent similarity [1], patent novelty, or the degree of basicness of a patent (i.e. relation to basic vs. applied science).

New techniques and tools as T4 allowing to better exploit the information content of the patent documents can provide the basis for further and more sophisticated analysis of the innovative process at different levels (e.g. firm, region, country).

The paper is organized as follows. In Sec. 2 we provide an outline of the logic work in T4, while in Sec. 3 we present the preliminary results that only concern an exploratory analysis applied to the Corpus of abstracts of the USPTO patent. Finally, conclusions are drawn in Sec. 4.

2 Methodology and logic of work on T4

Through T4, textual information - unstructured by nature - is structured in two main databases, defining the two analysis domains: the Vocabulary DB, for the lexical analysis and the Fragments DB for textual analysis.

In lexical analysis, the study object is the lexicon, and the single word represents the elementary unit of analysis. However, depending on the corpus characteristics and the research question, multi-word expressions, lemmas or word stems can be considered units of lexical analysis, instead of words. In natural processing languages (NLP), particular attention is devoted to recognizing the nominal multiword expressions in a corpus [13]. These expressions represent the specialized terminology of a sector and their recognition makes it possible to work with semantic unambiguous lexical units.

In the Vocabulary DB, each unit of analysis can be associated with annotations of grammatical, semantic and statistical nature. Each of these properties constitutes an example of meta-information attributed to the lexical units, which can be retrieved by querying the Vocabulary database's corresponding fields in which this information is stored. Additionally, T4 produces several vocabularies for a multi-level lexical analysis, in which every layer corresponds to a vocabulary with the different lexical units defined. The extraction/selection of the vocabulary parts serve to "tell" the lexical characteristics of the corpus by highlighting the significant elements of each "part of speech", or to "illustrate" certain subsets of units and the relations existing between them.

In text analysis, the object of study is the corpus, and the unit of analysis is the context unit, i.e., a fragment of text, whether it is a sentence, a section of a document, an entire document, or a group of documents. In analogy to Lexical Analysis, each context unit constitutes an entry in the Fragments database to which are associated both the modes of the a priori coded variables and the textual annotations (categorizations) resulting from Textual Analysis. These annotations can be of various kinds: i) syntactic, obtained through the categorization of documents in which certain syntactic structures or groups of variable elements are present; ii) semantic, concerning automatic categorizations on the basis of certain lexicons, and iii) quantitative. Strings of text, which can be the occurrences of lexical units, both of their classes and relationships between classes or between individual units and classes,

are searched through Regular Expressions (RE). The result of such elaboration is to recover the fragments that verify the textual query; to inventory the list of the extracted strings; to annotate eventually the fragments.

Based on different lexicon fragments obtained through the analysis, both lexical matrices (words x categories) and textual matrices (fragments x words) can be extracted. These matrixes can be used to represent the extracted information using infographic tools or can be further analyzed using other statistical tools.

3 Explorative analysis of USPTO Corpus

The corpus under analysis includes 5,573,936 abstracts of patents granted by the USPTO between 1980 and 2015. Each patent is assigned to at least one IPC (International Patent Classification) code, indicating the subject to which the invention relates. The IPC classification is a hierarchical classification system consisting in 5 levels of different granularity to which correspond a different number of digits[2].

After the first parsing of texts, the Vocabulary (lexical DB) includes 1,469,138 different words referred to 641,666,177 total occurrences. Through grammatical tagging of vocabulary entries, it was possible to define word lemmas as lexical analysis units. Based on grammatical annotations, we apply a hybrid multiword expression (MWEs) recognition system [5,12], based on string search according to syntactic structures. Through this technique, we identify 3,570 MWEs with at least 3,000 occurrences in the corpus[3].

In order to explore patents' content, all lexical units classified as adjectives and nouns (lemmas and MWEs) were selected to build a series of matrices for the graphical representation. To observe and study the general relationships between the elements of the matrices, we use the correspondence analysis [2]. As our first objective of the study is the analysis of the temporal evolution of innovative activities, we construct yearly matrices of the type <Lexicon x Year>. The correspondence analysis (CA) on the yearly matrices highlight similarity and differences of lexical profiles over time.

Figure 2 shows the distribution of the years and the selected lexical units., on the bi-dimensional plane spanned by the first two factors from the CA. The figure unfolds

[2] An IPC class has the form of H04J 1/10. The first letter represents the "section", combined with a two digits' number, it represents the "class" (H04), and the final letter indicates the "subclass". The digits after the subclass indicates the "group" and after the oblique stroke the least two digits indicates the "main group". A three-digit IPC class is at the level of subclass. For a complete overview see: https://www.wipo.int/classifications/ipc/en/

[3] The 20 most recurring MWEs we have found are: *mobile device, control device, control system, computer program, control circuit, computer system, virtual machine, mobile terminal, network device, light guide, management system, optical system, medical device, gas turbine, processing system, video data, film transistor, data processing, bit line, solar cell.*

a chronological development of the patent content and through cluster analysis we identify four temporal clusters represented with the different colours.

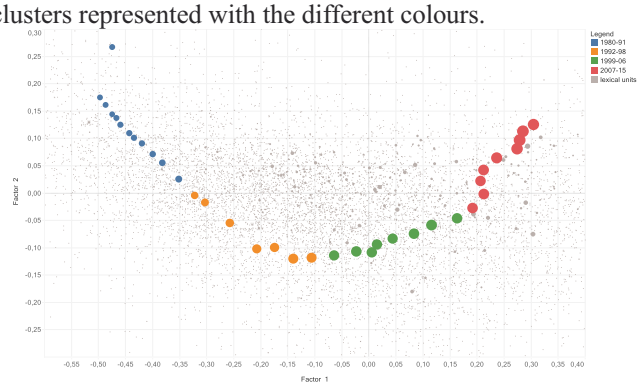


Figure 2: Distribution on the factorial plane *f1f2* of the lexicon and the years grouped in four clusters.

While this result is interesting, it would be compelling to observe in detail the elements characterizing these different temporal moments. We undertake this further step of the analysis using a different matrix of the type <Lexicon x IPC_3_DIGIT>, where IPC_3_DIGIT indicates the three-digit level of the primary IPC classification of each USPTO patents in our corpus. The CA on this matrix highlights the similarity of patent groups defined through their three-digit IPC codes. In this case the cluster analysis applied on the CA results, allows us to identify groups of patents with high technological semantic similarity.

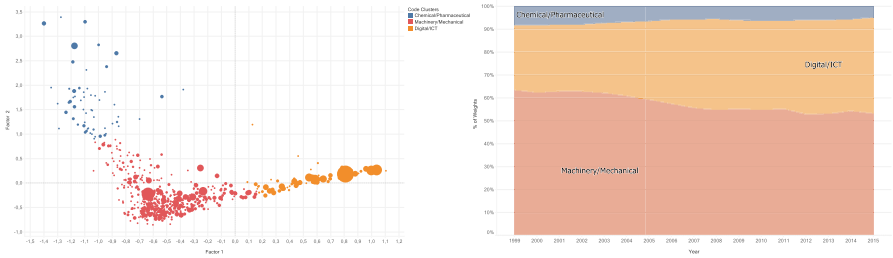


Figure 3: Distribution on the factorial plane of the codes (left) and of the percentage weight of each detected Industry over time (right)

This conceptual homogeneity emerges from the prevailing theme or semantic trait in each group, read through their characteristic dictionaries highlighted by test-values [10]. This procedure allowed us to define three clusters of Patent codes for each year, which single out specific industrial activities. Specifically, the following three industries were recognized: Chemical/Pharmaceutical; Machinery/Mechanical; Digital/ICT. Figure 3 shows the distribution on the factorial plane of the IPC_3_DIGIT (left) and the weight of each Industry detected over time (right). It is possible to clearly observe how Digital/ICT has gradually occupied a greater space in the world of patents, going from 28% in 1999 to 42% in 2015.

4 Conclusion

As working with textual Big Data is becoming increasingly common and relevant for research, in general, and for social sciences research in particular, T4 represents a freeware essential tool for text mining of very large corpora. Its initial application to USPTO patent data, as shown here, was particularly successful in identifying text patterns, in turn mapping into meaningful and well recognizable industry classes. This initial result indicates that text analysis can provide a viable way to overcome some shortcomings of the existing classification of innovation activities based on IPC codes. The further step of our research will exactly move in the direction to build a new taxonomy based on a fuzzy categorization of patents' membership within a system of industrial categories defined through text analysis. A key ingredient to this aim will be the integration in the software of the most recent text analysis tools, in particular those aimed at identifying the universe of topics in a corpus.

References

1. Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strateg. Manag. J. 39, 62–84.
2. Benzécri, J.-P., 1992. Correspondence analysis handbook, Statistics, textbooks and monographs. Marcel Dekker, New York.
3. Bolasco, S., De Gasperis, G., 2017. TaLTaC 3.0. A Multi-level Web Platform for Textual Big Data in the Social Sciences, in: Data Science and Social Research. Springer, pp. 97–103.
4. Bolasco, S., Morrone, A., Baiocchi, F., 1999. A paradigmatic path for statistical content analysis using an integrated package of textual data treatment, in: Classification and Data Analysis. Springer, pp. 237–246.
5. Bolasco, S., Pavone, P., 2010. Automatic dictionary-and rule-based systems for extracting information from text, in: Data Analysis and Classification. Springer, pp. 189–198.
6. Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey. Journal of economic literature, 28(4), 1661-1707. Retrieved from www.jstor.org/stable/2727442
7. Kuhn, J. M., & Thompson, N. (2017). The Ways We've Been Measuring Patent Scope are Wrong: How to Measure and Draw Causal Inferences with Patent Scope. Available at SSRN 2977273.
8. Lanjouw, J. O., Pakes, A., & Putnam, J. (1998). How to count patents and value intellectual property: The uses of patent renewal and application data. The journal of industrial economics, 46(4), 405-432.
9. Lanjouw, J. O., & Schankerman, M. (2001). Characteristics of patent litigation: a window on competition. RAND Journal of economics, 129-151.
10. Lebart, L., Salem, A., Berry, L., 1998. Exploring Textual Data, Text, Speech and Language Technology. Springer Netherlands. https://doi.org/10.1007/978-94-017-1525-6
11. Lerner, J., 1994. The importance of patent scope: an empirical analysis. RAND Journal of Economics, 319-333.
12. Pavone, P., 2018. Automatic Multiword Identification in a Specialist Corpus, in: Tuzzi, A. (Ed.), Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. Springer International Publishing, Cham, pp. 151–166. https://doi.org/10.1007/978-3-319-97064-6_8
13. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D., 2002. Multiword expressions: A pain in the neck for NLP, in: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, pp. 1–15.
14. Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. The RAND Journal of Economics, 172-187.

Emotion pattern detection on facial videos using functional statistics

Riconoscimento di pattern emozionali in video di volti attraverso la statistica funzionale

Rongjiao Ji, Alessandra Micheletti, Natasa Krklec Jerinkic, Zoranka Desnica

Abstract There is an increasing scientific interest in automatically analysing and understanding human behavior, with particular reference to the evolution of facial expressions and the recognition of the corresponding emotions. In this paper we propose a technique based on Functional ANOVA to extract significant patterns of face muscles movements, in order to identify the emotions expressed by actors in recorded videos. We determine if there are time-related differences on expressions among emotional groups by using a functional F-test. Such results are the first step towards the construction of a reliable automatic emotion recognition system [1]

Abstract *C'è un crescente interesse scientifico nell'analizzare e intepretare automaticamente il comportamento umano, soprattutto rispetto all'evoluzione delle espressioni del volto e al riconoscimento delle corrispondenti emozioni espresse. In questo lavoro proponiamo una tecnica, basata sull'ANOVA Funzionale per estrarre pattern significativi dei movimenti dei muscoli facciali, al fine di identificare le emozioni espresse da alcuni attori in video registrati. In particolare determiniamo se, in istanti specifici, ci siano differenze nell'evoluzione delle espressioni fra diversi gruppi di emozioni, applicando un F-test funzionale. Questi risultati sono il primo passo verso la costruzione di un sistema affidabile per il riconoscimento automatico delle emozioni.*

Key words: functional ANOVA, emotion, expression evolution, action units

---

Rongjiao Ji, Alessandra Micheletti
Universitá degli Studi di Milano, e-mail: rongjiao.ji@unimi.it, alessandra.micheletti@unimi.it

Natasa Krklec Jerinkic
University of Novi Sad, e-mail: natasa.krklec@dmi.uns.ac.rs

Zoranka Desnica
3Lateral DOO, e-mail: zoranka.desnica@3lateral.com

[1] This work was funded by European Unions Horizon 2020 research and innovation programme under the Marie Skodowska Curie grant agreement No 812912 for the project BIGMATH.

1 Introduction

The study of human facial expressions and emotions never stops in our daily life while we communicate with others. Following the increased interest in automatic facial behavior analysis and understanding, the need of a semantic interpretation of the evolution of facial expressions and of human emotions has become of interest in recent years [4]. In this paper, based on a work cooperated with the Serbian company 3Lateral, which has special expertise on building visual styles and designs in animation movies, we want to explore functional statistical instruments to identify the emotions while analyzing the expressions through recorded videos of human faces. The final aim of this research is to use this information to better and more realistically establish virtual digital characters, able to interact autonomously with real humans.

The data that we consider are multivariate longitudinal data, showing the evolution in time of different face muscles contraction. Functional Data Analysis (FDA) offers the possibility to analyze the entire expression evolution process over time and to gain detailed and in-depth insight into the analysis of emotion patterns. The basic idea in functional data analysis is that the measured data are noisy observations coming from a smooth function. Ramsay and Silverman [6] describe the main features of FDA, that can be used to perform exploratory, confirmatory or predictive data analysis. Ullah and Finch [7] published a systematic review on the applications of functional data analysis, where they included all areas where FDA was applied.

In our application, Functional ANOVA can be used to determine if there are time-related differences between emotion groups by using a functional F-test [2]. Functional ANOVA yields the possibility to determine if a functional response can be described by scalar or functional variables.

The structure of this paper goes as follows. In Section 2 we briefly describe the RAVDESS dataset from where the expression data of interest is extracted. Section 3 includes some methods of functional data analysis that we implemented in our application, and in Section 4 our results are presented.

2 The RAVDESS Dataset

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [5] fits our needs for studying the human expression evolution and emotion identification, as it contains 24 professional actors (12 female, 12 male) to offer the performance with good quality and natural behavior under the emotions: calm, happy, sad, angry, fearful, disgusted and surprised. Also a neutral performance is available for each actor. The actors are vocalizing one lexically-matched statement in a neutral North American accent (Kids are talking by the door).

To avoid being lost in the difference of individual facial appearances, when analyzing the expressions and emotions, researchers mostly focus on the movements of individual facial muscles which are encoded by the Facial Action Coding System

(FACS) [3]. FACS is a common standard to systematically categorize the physical expression of emotions, extracting the geometrical features of the faces and then producing temporal profiles of each facial movement. Such movements, corresponding to contraction of specific muscles of the face, are called *action unit (AU)*. As action units are independent of any interpretation, they can be used for any higher-order decision-making process including recognition of basic emotions. Following the FACS rules, OpenFace [1], an open-source software, is capable of recognizing and extracting facial action unit from facial images or videos. We applied OpenFace to extract the engagement degrees of action units for the videos in RAVDESS. The extracted action units include 17 functions for each video, taking values in $[0,5]$, sampled in about 110 time points (which is also the number of frames in each video).

3 Functional Statistical Methods

We will represent the action units evolution recorded on each video as a multivariate time series $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_d(t), \ldots, Y_D(t)), t \in [0,T]$ containing a set of D univariate longitudinal functions ($D = 17$ in our case), each defined on the finite interval $[0,T], 0 < T < +\infty$. The observation of $\mathbf{Y}$ on our sample of videos provides the set $\mathbf{Y_1}, \ldots, \mathbf{Y_n}$ of multivariate curves, that we represent as multivariate functional data.

It is essential to align the action units functions into a common registered internal timeline that follows the same pronunciation speed, to control the influence of the specific pronounced sentence and to detangle it from the influence of the emotions. Therefore, we need to isolate the phase variability of the action units curves, but keeping, at the same time, the amplitude-phase unchanged to maintain the information of the intensity level of the action units.

The phase variation is normally represented by a random change of time scale, which is mostly a non-linear transformation. We use the warping functions $T_i : [0,T] \to [0,T], i = 1, \ldots, n$, assuming that they are increasing functions independent of amplitude variation. They map unregistered chronological time t_i^* to registered internal time t so that $T_i^{-1}(t_i^*) = t$, with $E[T_i(t)] = t$. The observed time-warped curves, represented through a Karhunen-Loeve expansion based on a functional basis $\mathbf{f}_{j,d}$, are

$$\tilde{Y}_{i,d}(t) = Y_{i,d}(T_i^{-1}(t_i^*)) = \mu_d(T_i^{-1}(t_i^*)) + \sum_{j \geq 1} C_j \mathbf{f}_{j,d}(T_i^{-1}(t_i^*)),$$

We used a spline basis and followed the principal components based registration method with a generative process [9], whose codes are available in the R package "registr" [8].

Using the registered curves representing the AUs evolution in each video, we then investigated if there exist patterns which could discriminate the different emotions, using a Functional ANOVA model.

Let $y_{k,g}(t)$ be the evolution of one specific action unit in the video $k \in \{1,\dots,K\}$ (in our case $K = 48$) for emotion $g \in \{1,\dots,7\}$. We can assume that

$$y_{k,g}(t) = \mu_0(t) + \alpha_g(t) + \varepsilon_{k,g}(t), \tag{1}$$

where $\mu_0(t)$ is the grand mean function due to the pronounced sentence and to the actor, independent from all emotions. The term $\alpha_g(t)$ is the specific effect on the considered action unit of emotion g, while $\varepsilon_{k,g}(t)$ represents the unexplained zero mean variation, specific of the k-th video within emotion group g. To be able to identify them uniquely, we require that they satisfy the constraint $\sum_{g=1}^{7} \alpha_g(t) = 0, \forall t$.

By grouping the videos representing the same emotion, we can define a $8K \times 8$ design matrix $\mathbf{Z}$ for this model, with suitable 0 and 1 entries, as described in [6, Section 9.2], and rewrite Equation 1 in matrix form: $\mathbf{y} = \mathbf{Z}\beta + \varepsilon$, where $\beta = [\mu_0(t), \alpha_1(t), \dots, \alpha_7(t)]^T$.

To estimate the parameters we use the functional least squares fitting criterion

$$\hat{\beta}(t) = \arg\min_{\beta} \sum_{g=1}^{8} \sum_{k=1}^{K} \int_0^T [y_{k,g}(t) - \langle z_{k,g}, \beta(t) \rangle]^2 dt, \tag{2}$$

subject to the constraint $0 = \sum_{i=1}^{7} \alpha_i(t) = \sum_{j=2}^{8} \beta_j(t), \ \ \forall t$.

In order to investigate which emotions are significantly influencing the change of the action units patterns, for each emotion $\tilde{g}$ and for each action unit we test the null hypothesis $H_0 : \alpha_{\tilde{g}}(t) = 0$.

Similarly to the classical univariate ANOVA model, the statistics used to test H_0 is

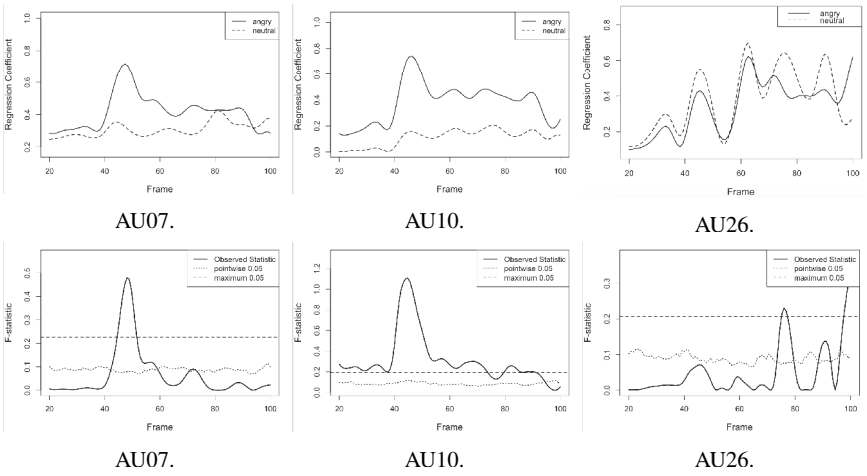
$$FRATIO(t) = \frac{MSR(t)}{MSE(t)}$$

whose distribution under H_0 is estimated through a permutation test.

4 Results

As mentioned before, we first aligned the curves by separating the amplitude and phase variability. We choose to align the curves by AU25, which represents the lip movement, and then we adjusted the time frames of the other AUs according to this rescaling.

We then applied the F-test described in the previous section to detect, for each emotion, which AUs have a mean behaviour significantly different from the neutral performance and in which time period during the videos. In Figure 2 we illustrate the results for emotion angry, as an example.



AU07. AU10. AU26.

AU07. AU10. AU26.

Fig. 2: The functional coefficients of action units 07 (Lid Tightener), 10 (Upper Lip Raiser) and 26 (Jaw Drop) under neutral and angry emotion and the corresponding F-test results

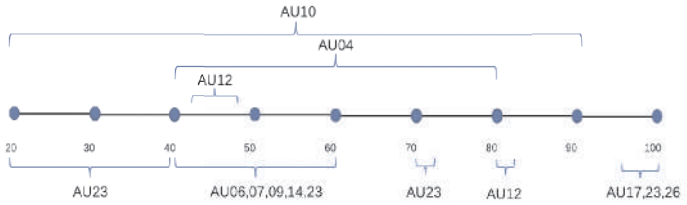


Fig. 3: Which and where AU values are affected significantly by angry emotion

The first row of Figure 2 illustrates the estimated mean $\mu_0(t)$ (neutral emotion) and the angry emotion effects for three action units. The second row displays the observed F-statistics curves together with the pointwise and maximum 95% quantile for the F-distribution in the dashed and horizontal dotted lines respectively. Thus when the observed F-statistics is higher than the critical level lines, the emotion has a significant effect on the AU's pattern. We found in general three main situations of influence of one emotion on expression evolution: 1. locally strengthening (Figure 1d: AU07 in frame range 45 to 55) 2. locally inhibiting (Figure 1f: AU26 in frame range 70 to 90) 3. globally strengthening (Figure 1e: AU10 in almost the whole time). Further, we pointed out the time zones of significant effects of the angry emotion on the action units in Figure 3, which is beneficial to understand and detect dynamically when and how the facial muscles contractions differ from the baseline.

Table 1 summarizes for each emotion of interest the related action units that show significant changes from the neutral case for our videos dataset. Similarly to the example of angry, we found that for happy and disgust emotions more action units

Table 1: Emotions with corresponding significant action units

Emotions	Related Action Units
Calm	06, 07, 10, 12, 14, 23
Happy	01, 06, 07, 10, 12, 14, 17, 23, 25, 26
Sad	04, 06, 10, 14, 17, 20, 23, 25
Angry	04, 06, 07, 09, 10, 12, 14, 17, 23, 26
Fearful	04, 09, 10, 12, 14, 15, 17, 23, 25, 26
Disgust	04, 06, 07, 09, 10, 12, 14, 17, 23, 25, 26
Surprised	06, 09, 10, 12, 14, 15, 17, 23, 25, 26, 45

have the globally strengthening effect on a large time range. Sad emotion sometimes affects the action units to be more constant than in neutral case. Emotion Fearful has more influence on upper half face (brows, eye lids and nose), while emotion calm is more related with the center of the face (Cheek Raiser, Lid Tightener and Lip Corner Puller). Surprised emotion is the only emotion where AU45 is significantly influenced.

As a conclusion, our results can be joined in a multivariate setting and exploited to build a classifier able to automatically recognize the emotions. This task is left to subsequent works.

References

1. B. Amos, L. Bartosz, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016. https://cmusatyalab.github.io/openface/.
2. J. Dannenmaier, C. Kaltenbach, T Kölle, and G. Krischak. Application of functional data analysis to explore movements: walking, running and jumping-a systematic review. *Gait & postureh*, pages 182–189, 2020.
3. R. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
4. A.J. Fridlund. Human facial expression: An evolutionary view. *Academic Press*, 2014.
5. S.R. Livingstone and F.A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. https://smartlaboratory.org/ravdess/.
6. J. Ramsay and B.W. Silverman. *Functional data analysis*. Springer, 1997.
7. S. Ullah and F. F. Caroline. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 2013.
8. J. Wrobel. Register: Registration for exponential family functional data. *Journal of Open Source Software*, 3(22):557, 2018.
9. J. Wrobel, V. Zipunnikov, J. Schrack, and J. Goldsmith. Registration for exponential family functional data. *Biometrics*, 75(1):48–57, 2019.

The spread of contagion on Twitter: identification of communities analysing data from the first wave of the COVID-19 epidemic

La diffusione del contagio su Twitter: identificazione delle comunità di utilizzatori attraverso l'analisi dei dati della prima ondata dell'epidemia COVID-19

Gianni Andreozzi, Salvatore Pirri, Giuseppe Turchetti, Valentina Lorenzoni

Abstract The coronavirus (COVID-19) pandemic suddenly spread, and still is, on the web and on social media leading to rapid diffusion of discussion and fake news about the pandemic. Given the influence social media can exert on users and possibly on the general population, it is important to understand the dynamic of the discussion on public health related issues. Using tailored methods, the present study identified communities of Twitter users that emerged in the preliminary phase of the first wave of the COVID-19 pandemic, providing insight into the dynamic that characterize the discussion about COVID-19 and offering inputs to plan and drive future communications on health-related issues on social media.

Abstract *La pandemia di coronavirus (COVID-19) si è improvvisamente diffusa - e lo ancora lo è - sul web e sui social portando a una rapida diffusione di discussioni sulla pandemia, con conseguente diffusione anche di fake news. Data l'influenza che i social possono esercitare sugli utenti, e verosimilmente anche sulla popolazione generale, è importante comprendere la dinamica attorno alle discussioni. Utilizzando appropriate metodologie, il presente studio ha identificato le comunità di utenti di Twitter che sono emerse nella fase preliminare della prima ondata della pandemia, fornendo informazioni sulla relatiae dinamica offrendo input per pianificare la comunicazione in tema di salute pubblica sui social.*

Key words: COVID-19, twitter, influencer, communities

[1] Gianni Adreozzi, Institute of Management, Scuola Superiore Sant'Anna, Pisa, Italy; email: gianni.andreozzi@santannapisa.it

[2] Salvatore Pirri, Institute of Management, Scuola Superiore Sant'Anna, Pisa, Italy; email: salvatore.pirri@santannapisa.it

[3] Giuseppe Turchetti, Institute of Management, Scuola Superiore Sant'Anna, Pisa, Italy; email: Giuseppe.Turchetti@santannapisa.it

[4] Valentina Lorenzoni, Institute of Management, Scuola Superiore Sant'Anna, Pisa, Italy; email: valentine.lorenzoni@santannapisa.it

Introduction

1.1 Background

The pandemic originated from Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), also named coronavirus (COVID-19) pandemic, suddenly had, and still has, a great resonance on the web, particularly on social media. The widespread diffusion of social media as well as the interruption of social relation experienced in many countries because of restrictive measures set to fight the pandemic led the epidemic to goes viral also on the web [2][6][7] Moreover many institutions used to web and social media to spread information about the disease, updates and recommendations for safe behaviors. Anyway the uncontrolled flux of information that pervaded social media also caused the rapid diffusion of misleading information [1][4].Given the importance of public health information circulating on social media, particularly during health emergencies - as the one we are experiencing - it has become more and more important to understand and monitor the flow of information on social media about public health issues and to understand the dynamic of the network to proper and timely monitor facts that could affect behaviours among the general population but also to understand how the discussion on the web could be effectively driven.

1.2 Objective

Accordingly, the present study aims at identifying and characterize communities of users emerged around the discussion about COVID-19 on Twitter in the preliminary phase of the first wave of the COVID-19 pandemic in Italy.

1.3 Data

The database used in the analysis comprised a total of 75,766 tweets collected using the Twitter API and the *rtweet* package for R. Tweets were included in the analysis if they contained at least one of the selected hashtags, which were updated daily by looking at Twitter's trending tab (i.e., #COVID-19, #coronavirus). Retweets were included in the selection and for the purpose of the present analysis we use data referred to tweet shared from the 15th of February 2020 (about one week before the first confirmed case in Italy) to the 26th of March 2020.

Statistical analysis

Given the low volume of tweets, a graph was constructed based on the retweets to depict the network of users involved in the discussion. The graph had a total of 31403 nodes.
Communities were analysed using Louvain method [5] with a resolution of 2.5. The algorithm starts by assigning to each node a different community, Then, for each node *i* it considers the neighbours *j* of *i* and it evaluates the gain of modularity that would take place by removing *i* from its community and by placing it in the community of *j*. The node *i* is then placed in the community for which this gain is maximum (in case of a tie we use a breaking rule), but only if this gain is positive. If no positive gain is possible, *i* stays in its original community. This process is applied repeatedly and sequentially for all nodes until no further improvement can be achieved. The second phase of the algorithm consists in building a new network whose nodes are now the communities found during the first phase. The steps previously described are then applied to this new network. A "pass" is defined as a combination of these two phases. By construction, the number of meta-communities decreases at each pass, consequently most of the computing time is used in the first pass. The passes are iterated until there are no more changes and a maximum of modularity is attained. Four communities encompassing more than 89% of all nodes were identified.
The language of the communities was explored by looking at word collocations.

Results

The four communities identified were quite clearly distinguishable from one another: one contained accounts from institutional sources such as the Health Ministry, and news outlets; a second one included right leaning politicians and the node with the highest degree overall, representing the account of the leader of one of the biggest Italian parties; a third one contained generally younger users and the fourth one consisted of neither institutional accounts nor famous people but generally not verified regular users.

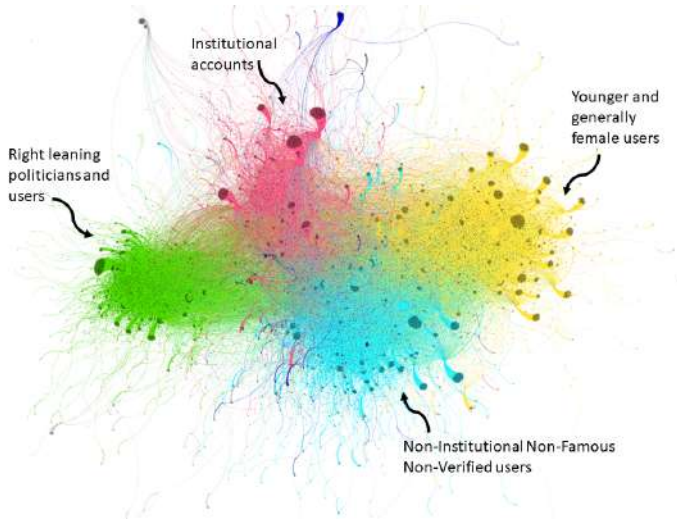


Figure 1: Graph highlighting the communities identified using retweets in the period from the 15th of February to the 26th of February

Figure 2 shows the timelines related to the number of tweets shared all over the study period. The number of tweet and retweets related to COVID-19 rapidly increased in correspondence of the detection of the first autochthone case.

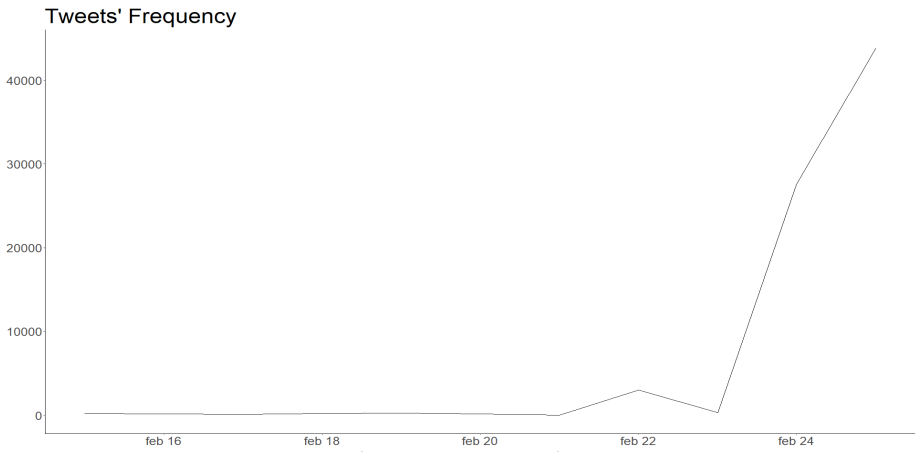


Figure 2: Tweets' frequency from the 15th of February to the 25th of March

Figure 3: Most frequent word collocations for each community

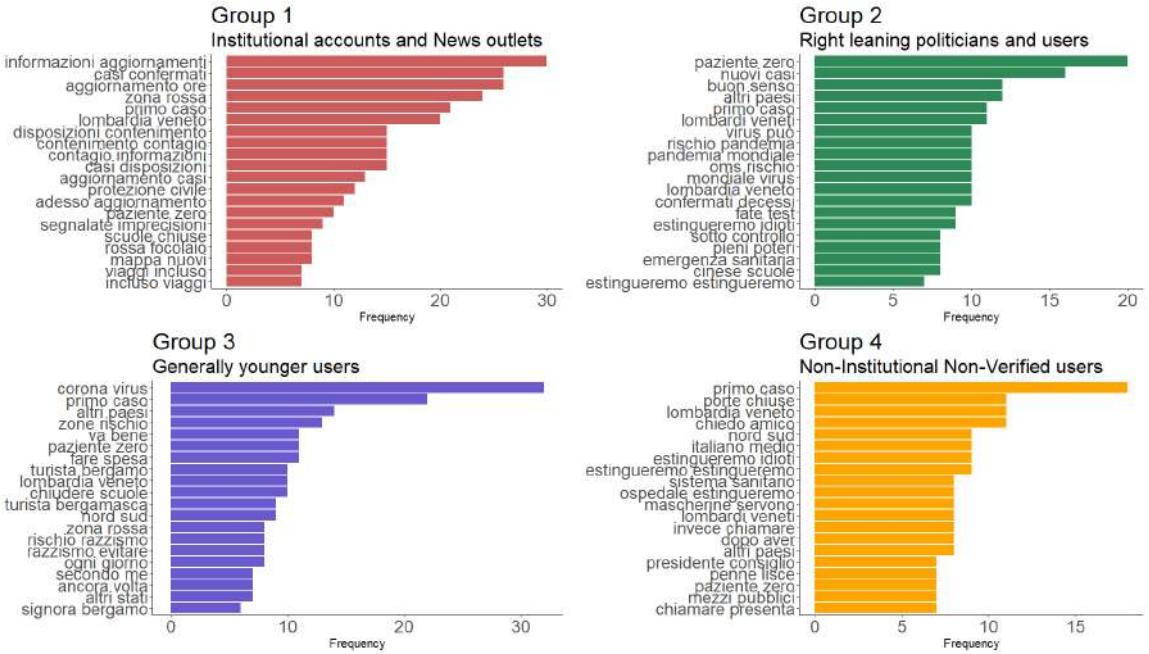


Table 1: Most retweeted tweet in each community

Community	***Text***
1	*#Covid_19 In corso accertamenti in almeno 2 province per falsi infermieri che tentano l'accesso in casa fingendo di dover fare tamponi per #coronavirus. Nessun prelievo porta a porta viene disposto dalle autorità sanitarie senza preventiva comunicazione. #COVID19italia*
2	*#Salvini su #COVID19italia: A Prato c'è enorme comunità cinese e cittadini sono preoccupati, ma il governatore della Toscana, Rossi, ha addirittura accusato di essere fascioleghisti perfino i medici che chiedevano i controlli. Mi auguro che chieda scusa agli italiani e si dimetta.*
3	*Vorrei tornare a 2 settimane fa, a quando il nostro unico problema era sapere dov'era Bugo. #coronavirusitalia*
4	*27enne scappa da Codogno per tornare in Irpinia. Turista di Bergamo se ne va in gita a Palermo. Imprenditore fiorentino accusa i sintomi e, invece di chiamare il 112, si presenta all'ospedale. Non ci estingueremo per il #coronavirusitalia, ci estingueremo perché siamo idioti.*

Figure 3 depicts the most frequent word collocations in each one of the communities. Even with a simple exploratory analysis some of the different themes characterizing each community are clearly visible: information regarding the

containment of the virus in group 1 such as “disposizioni contenimento” and schools in the younger users of group 3 (“chiudere scuole”), public transport, food supplies and face masks (“mascherine servono”) in group 4.

Conclusions

Nowadays, as the experience of the COVID-19 pandemic demonstrated, the web and in particular social media represents one of the main channels through which communication about public health issues could be spread [3]. In details, differently from other social networks, Twitter posts are not meant to be uniquely directed to another specific user or group of users, rather they are available to all the people who are willing to listen to what a user has to say. That characteristic determines rapid and uncontrolled propagation of discussion, moreover in Twitter the dynamic around a topic is typically lead by some accounts, influencers, that exert great influence on other users creating communities around influencers.
Applying robust methods for the detection of communities the present study tries to depict groups of users and to explore how the discussion during the preliminary phase of the first wave of the COVID-19 pandemic evolved in different communities. Results offer inputs to understand how the public space of Twitter divides itself into smaller groups of homogeneous individuals and explores the themes shared in the different communities aiming at helping plan future communication on public health issues by understanding users and the theme that resonate most within the different groups to drive proper health behaviours.

References

1. Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M.B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E., Baddour, K.: Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. Cureus. (2020). https://doi.org/10.7759/cureus.7255
2. Merchant, R.M., Lurie, N.: Social Media and Emergency Preparedness in Response to Novel Coronavirus. J. Am. Med. Assoc. (2020) https://doi.org/10.1001/jama.2020.4469
3. Ni, M.Y., Yang, L., Leung, C.M.C., Li, N., Yao, X.I., Wang, Y., Leung, G.M., Cowling, B.J., Liao, Q.:Mental Health, Risk Factors, and Social Media Use During the COVID-19 Epidemic and Cordon Sanitaire Among the Community and Health Professionals in Wuhan, China: Cross-Sectional Survey. JMIR Ment. Heal. (2020). https://doi.org/10.2196/19009
4. Zarocostas, J.: How to fight an infodemic. Lancet (2020). https://doi.org/10.1016/S0140-6736(20)30461-X
5. Blondel V., Guillaume J., Lambiotte R., Lefebvre E., Fast unfolding of communities in large networks. Journal of Statistical Mechanics Theory and Experiment (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
6. Zhao, Y., Cheng, S., Yu, X., Xu, H.: Chinese public’s attention to the COVID-19 epidemic on social media: Observational descriptive study. J. Med. Internet Res. (2020). https://doi.org/10.2196/18825
7. Zou, C., Wang, X., Xie, Z., Li, D.: Public Reactions towards the COVID-19 Pandemic on Twitter in the United Kingdom and the United States. medRxiv 2020.07.25.20162024. https://doi.org/10.1101/2020.07.25.20162024

Composition-on-Function Regression Model for the Remote Analysis of Near-Earth Asteroids

Modello di Regressione Composizione-su-Funzione per l'Analisi da Remoto di Asteroidi Near-Earth

Mara S. Bernardi, Matteo Fontana, Alessandra Menafoglio, Alessandro Pisello, Massimiliano Porreca, Diego Perugini, Simone Vantini.

Abstract We propose a regression model with compositional response and functional predictor. The motivating application concerns the problem of retrieving the chemical composition of a silicate glass based on its spectral response. This problem is particularly interesting in the context of planetary investigation, where spectroscopy data can be remotely collected.

Abstract *Proponiamo un modello di regressione con risposta composizionale e predittore funzionale. L'applicazione che motiva questo lavoro riguarda il problema di ottenere la composizione chimica di un vetro silicato sulla base della sua risposta spettrale. Questo problema è particolarmente interessante nel contesto delle investigazioni planetarie, dove i dati spettroscopici posso essere raccolti da remoto.*

Key words: functional data, compositional data, regression

1 Introduction

The study of the spectral response of silicate glasses, which are widely present in volcanic rocks, is of great importance in planetary investigations of near-earth asteroids, since remotely sensed spectra can provide information about the constituents of terrains [5, 3, 2]. The motivating application for this work, described in [1], is the problem of retrieving the chemical compositions from reflectance spectra. The dataset analyzed, described in [1], concerns the spectral response of different sili-

Mara S. Bernardi
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy
e-mail: marasabina.bernardi@polimi.it

Matteo Fontana, Alessandra Menafoglio, Simone Vantini
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy

Alessandro Pisello, Massimiliano Porreca, Diego Perugini
Department of Physics and Geology, University of Perugia, Italy

cate glasses presenting a range of chemical compositions that covers the majority of rocks on planet Earth.

This problem poses challenges from a modeling perspective as it involves heterogeneous and complex data whose structure and properties should be appropriately accounted for. During the talk, we will introduce a regression model using the Functional Data Analysis [6] and the Compositional Data [4] frameworks.

The proposed model is described in Section 2. The analysis of the data is described in Section 3. Section 4 draws the conclusions and outlines possible directions of future research.

2 Model

We consider a random sample $\{(\mathbf{y}_i, x_i(t))\}_{i=1,\dots,n}$ where $\mathbf{y}_i$ are vectors of D-parts compositions in the simplex S^D and $x_i(t)$ are curves in the space $L^2(T)$ defined on a compact domain $T \in \mathbb{R}$. To map the compositions to an Euclidean vector space, we apply an *ilr* transformation of the compositional data: $\tilde{\mathbf{y}}_i = ilr(\mathbf{y}_i)$. We model the conditional distribution of $\tilde{\mathbf{y}}_i$ given the functional covariate $x_i(t)$ via the following regression model:

$$\tilde{\mathbf{y}}_i = \boldsymbol{\beta}_0 + \int_T x_i(t)\boldsymbol{\beta}(t)dt + \boldsymbol{\varepsilon}_i, \qquad i = 1,\dots,n,$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^{D-1}$, $\boldsymbol{\beta}(t)$ is a $(D-1)$-dimensional vector whose components are functions in the space $L^2(T)$. The additive error terms $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{D-1}$ are assumed to be i.i.d. (independent and identically distributed) random variables with mean $\mathbf{0}$ and diagonal variance.

3 Application

For the considered application, it is of interest to retrieve the silica and alkaline parts of the chemical compositions since they are linked to evolution of magma and to geodynamic setting respectively. We therefore consider the 3-parts composition: SiO_2, $K_2O + Na_2O$, *Other*. The data are represented in the left panel of Figure 1. We apply an *ilr* transformation using as sign matrix a sequential binary partition that, in the first instance, contrasts the first two components against *Other* and, in the second instance, contrasts SiO_2 against $K_2O + Na_2O$.

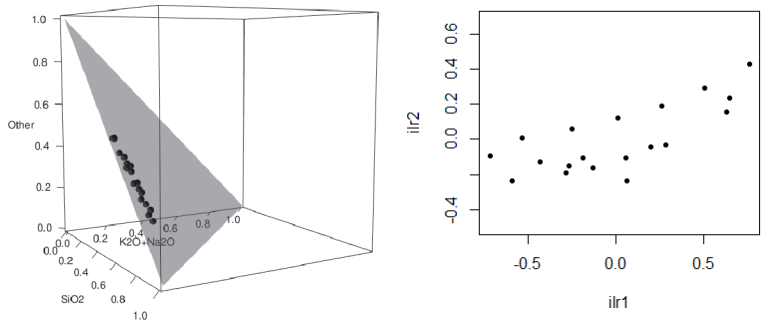


Fig. 1 Plot of the compositions on the simplex (left panel) and of ilr-transformed data (right panel).

We consider the spectral responses as discrete noisy samplings of underlying smooth functions. We suppose that the noise is caused by sampling error and spurious oscillations that are not of interest for the analysis of the phenomenon. Therefore, we apply spline smoothing to the raw data represented in the left panel of Figure 2 to obtain smooth curves used for the analysis and represented in the right panel of Figure 2. Spline smoothing is performed using the R package fda [7].

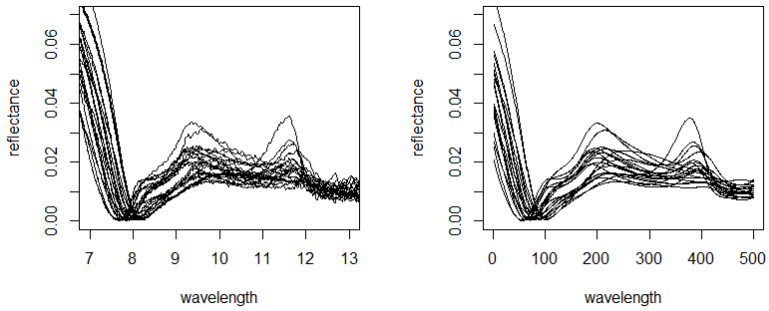


Fig. 2 Plot of the original data (left panel) and of the smoothed data (right panel).

During the talk, we will show the application of the model described in Section 2 to the obtained data using a ridge penalization to estimate the parameters.

Then, we will compare the proposed model with a simple regression model using as covariate the horizontal shift of the spectra that aligns the Christiansen feature. We will show that the proposed method provides better results, highlighting the advantage of the functional approach with respect to a scalar one.

4 Conclusions and future work

The analysis presented during the talk will show the usefulness of the functional data analysis approach for the study of spectral data and the relevance of the information contained in the whole spectrum in retrieving the chemical composition of the material.

Directions of future research concern the integration of the alignment approach proposed in [1] in the model. In this extension of the model, the predictor would be a bivariate functional data composed by the warping function (describing the phase variability) and the aligned data (describing the amplitude variability).

References

1. Bernardi, M. S., Fontana, M., Menafoglio, A., Perugini, D., Pisello, A., Ferrari, M., ... & Vantini, S. (2020). Functional Data Analysis for Spectroscopy Data. Book of Short Papers SIS 2020.
2. De Sanctis, M. C., Altieri, F., Ammannito, E., Biondi, D., De Angelis, S., Meini, M., Pirrotta, S., Vago, J.L., Mugnuolo, R.: Ma_MISS on ExoMars: mineralogical characterization of the martian subsurface. Astrobiology, 17(6-7), 612-620 (2017)
3. Maturilli, A., Helbert, J., Moroz, L.: The Berlin emissivity database (BED). Planetary and Space Science, 56(3-4), 420-425 (2008)
4. Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data. John Wiley & Sons.
5. Pisello, A., Vetere, F. P., Bisolfati, M., Maturilli, A., Morgavi, D., Pauselli, C., Iezzi, G, Lustrino, M., Perugini, D.: Retrieving magma composition from TIR spectra: implications for terrestrial planets investigations. Scientific reports, 9(1), 1-13 (2019)
6. Ramsay, J. O., Silverman, B. W. (2005). Functional Data Analysis. 2nd edn Springer. New York
7. Ramsay, J. O., Wickham, H., Graves, S., Hooker, G.: fda: Functional Data Analysis. R package version 2.4.8.1. (2018)

Determinants of football coach dismissal in Italian League Serie A

Determinanti degli esoneri degli allenatori di calcio nella Serie A italiana

Francesco Porro, Marialuisa Restaino, Juan Eloy Ruiz-Castro, Mariangela Zenga

Abstract The aim of this work is to study the impact of a set of personal and team characteristics on the dismissal of managers (head coaches) in the top division of the Italian Football League during the seasons 2010-2019. We examine the probability of coaches' dismissals by employing survival methods to explore the effects of covariates on coach tenure length. We capture the variation across the seasons and we assess the association between team/coach characteristics and coach dismissals. The set of coach characteristics includes both performance-related and non-performance-related variables. Performance variables include characteristics of teams able to analyze and measure coach or/and team typical performance. Non-performance variables are a group of demographic characteristics related to managers and clubs.

Abstract *Lo scopo del presente lavoro è quello di studiare l'impatto di caratteristiche personali e non personali sulla carriera degli allenatori italiani di calcio nella Serie A nelle stagioni 2010-2019. In particolare, verrà valutata la probabilità che gli allenatori vengano esonerati attraverso l'utilizzo di metodi di sopravvivenza e saranno esplorati gli effetti delle covariate sulla durata del mandato dell'allenatore prima dell'esonero. L'insieme delle covariate comprende variabili legate all'allenatore, alla prestazione della squadra allenata ed al club.*

Key words: Sport statistics, Survival analysis, Cox PH model

Francesco Porro
Università degli Studi di Sassari, Sassari (Italy) e-mail: fporro@uniss.it

Marialuisa Restaino
Università degli Studi di Salerno, San Fisciano (Italy) e-mail: mlrestaino@unisa.it

Juan Eloy Ruiz-Castro
University of Granada, Granada (Spain) e-mail: jeloy@ugr.es

Mariangela Zenga
Università Milano-Bicocca, Milano (Italy) e-mail: mariangela.zenga@unimib.it

1 The role of the head coach in labour marker

In the last years, the role of footbal head coach became comparable to a company manager. Even if the role of a footbal head coach varies across countries and within country depending on club owners' preferences, the required skills of this figure are several. The head coach plans and directs training and recommends acquisition or trade of players for professional athletic team; he/she assesses player's skills, assigns team positions and evaluates own and opposition team capabilities to determine game strategy. Moreover he/she coaches or directs coach, professional athletes to instruct players in techniques of game, participates in discussions with other clubs to sell or trade players and may participate on team managed and be designated coach-player or player-manager. Whereas football players can only be traded at particular times during the football season, head coaches can be laid off or hired throughout the season, as well as in the closed season between May and August. In this paper, we analysed the length of the career for head coach in professional football by using survival analysis [1, 4].

2 Data

We have a data set composed of 225 head coaches who were in charge of football league games played by the 20 teams in the top of professional football in Italy, covering the seasons 2010–2011 to 2019–2020. This period covers more than 6,000 games. The data are a flow sample in that we observe the start date for all coaches' initial employment spells, including those that overlap the start of the initial football season in our data. Each spell ends with the head coach leaving due to dismissal by the club or a voluntary quit, or else the coach remains in post. The data were obtained principally on transfertmarkt website[1]. The average coaching spells lasted for 270 days (min=7 days) before an exit for dismissal. The 60% of the observed spells ends with a dismissal, and nearly all coaches in the sample contribute at least one exit from a club over the sample period. For seasons 10/11, 11/12 and 15/16 the percentage of dismissal is higher than the other seasons. Moreover the clubs in the lower position of the rank at the end of the season show greater number of dismissals. A large number of coaches exits occurs during the season (67%) while the 27% of the coaches exits occurs at the end of the season. In general when coaches wish to leave the club, the closed season is when their contracts will expire so some departures may just reflect the non-renewal of fixed term contracts. The 6% of dismissal occurs before the first match. A large spike in dismissal occurs during the first round of season.

[1] https://www.transfermarkt.it/

3 Methodology: survival analysis to examine the Head Coach dismissals

The survival analysis is used to examine head coach dismissal. Let T be the time of the duration of the coaching from the beginning season. The period of observation is every season and the censorship is given at the end of every season. The event is the dismissal, while the censor occurs if the exit is a voluntary quit or if at the end of observation period the coach remains in post. The survival function is defined as $S(t) = Pr(T > t) = 1 - F(t)$, while the hazard function defines the instantaneous risk of failure at time t given that failure has not yet occurred

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{1}$$

The Cox proportional hazard model [2]

$$h(t) = h_0(t) \exp(\mathbf{X}\beta) \tag{2}$$

is used to test the impact of the covariates on the time to exit.

4 Results

The Figure 1 reports the survival curves before dismissal, respect to the season. The survival curves for seasons 10/11, 11/12, 15/16 and 18/19 have similar shapes. In particular at the end of the seasons 11/12 and 15/16 the European competition was present and this could affect the survival curves for the coaches. Moreover, the season 19/20 was particularly long due to the postponement due to SARS-Cov-2 pandemic.

Table 1 reports the results for the Cox proportional hazard model respect to the variables on the performance of the teams. As suggested in the literature [3], the career duration of head coaches in football is negatively affected by their lost percentage in the match before the dismissal. In fact, coaches are constantly monitored not only by the management of their clubs but also by football fans. Since head coaches are responsible for the teams' performance, more successful coaches are likely to survive longer. Our results confirm the theory. In fact the variables significantly affected the duration before dismissal are the round of the season (in the second round of season is less probable a dismissal) and the percentage of the proportion of lost match before dismissal (the increasing of loss proportion increases the dismissal probability).

Table 2 reports the results of the Cox proportional hazard model considering the variables related to the market value of the team. The literature [3] proposes that the probability for the coaches to be dismissal is determined by the relative value of their teams. In fact, coaches working with expensive teams, i.e. those with relatively high

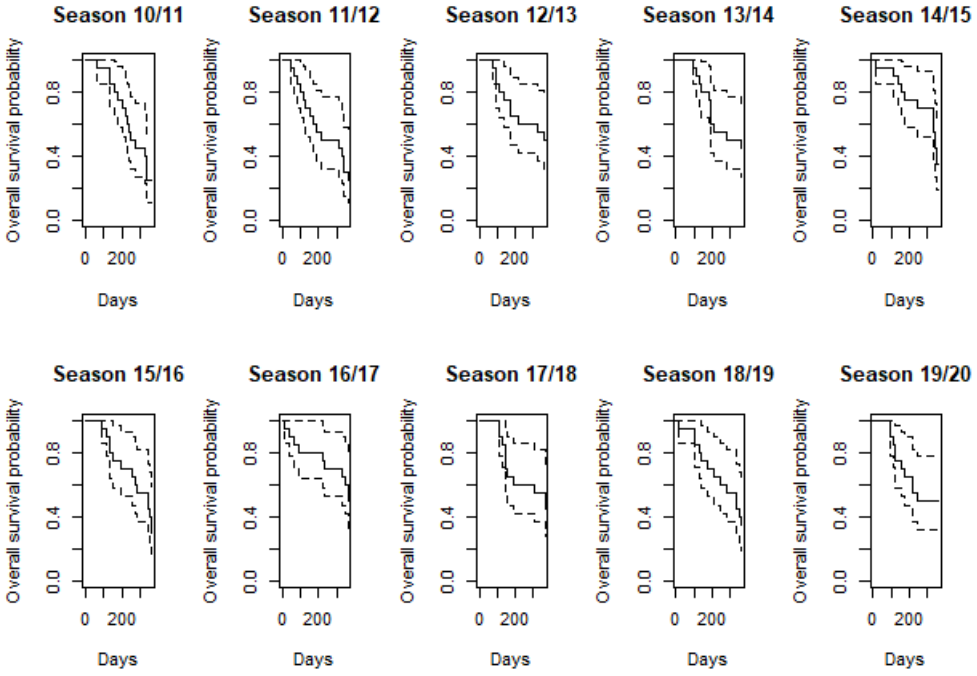


Fig. 1 Survival curves per season

Variable	Coef	exp(Coef)	se	z	p-value
Rank	0.028	1.028	0.046	0.605	0.545
Second round (Yes=1)	-5.494	0.004	1.039	-5.288	< 0.0001
Total won matches	-0.227	0.797	0.321	-0.706	0.48
Total drawn matches	-0.037	0.963	0.119	-0.312	0.755
Scored Goals	0.07	1.072	0.107	0.651	0.515
Conceded Goals	-0.066	0.936	0.106	-0.625	0.532
Scores	0.073	1.075	0.105	0.691	0.489
Proportion of lost match before the exit	1.166	3.209	0.435	2.682	0.007

Table 1 Results of the Cox proportional hazard model for time to dismissal with information on the performance of the coached teams.

Variable	Coef	exp(Coef)	se	z	p-value
Squad of players	-0.013	0.987	0.017	-0.786	0.568
Age	-0.056	0.945	0.094	-0.595	0.448
Foreign players	0.011	1.011	0.017	0.649	0.484
Values Squad of players	0.002	1.002	0.003	0.562	0.426
Market value of the team	-0.139	0.87	0.129	-1.076	0.718

Table 2 Results of the Cox proportional hazard model for time to dismissal with the relative value of their teams.

wage bills, are more likely to be fired when performance lags behind expectations. In our case it seems that this group of covariates does not significantly affect the probability to be dismissal.

Table 3 reports the variables related to the previous career of the coach. The theory [3] states that the coaching experience decreases the probability of being dismissed. In fact, the longer a coach has been working in professional football, the more human capital he has accumulated and this, in turn, reduces the probability of getting fired. In this case, even if the covariates are not significant in the analysis, the sign of the covariates are in line with the theory.

Variable	Coef	exp(Coef)	se	z	p-value
Youth team coach	0.254	1.289	0.191	1.332	0.183
Previous football player	-0.277	0.758	0.299	-0.927	0.354
Previous football player in the same team	0.283	1.326	0.236	1.196	0.232
Coach abroad	0.111	1.117	0.186	0.596	0.551

Table 3 Results of the Cox proportional hazard model for time to dismissal with previous career information of the coach

5 Final remarks

In this paper we analised the carreer of the head coaches for the Italian football clubs in Serie A. We explained the results in the light of the considerations that the head coaches are considered as managers. Our results confirm that the career duration seems to be affected by the performance of the teams, but not by the economic value of their teams or previous career of the coach. As a possible future work, the career of the football head coaches could be analysed by some appropriate stochastic processes.

References

1. Bryson, A., Buraimo, B., Farnell, A. et al.: Time To Go? Head Coach Quits and Dismissals in Professional Football. De Economist **169**, 81–105 (2021).
2. Cox, D.R.: Regression models and life-tables (with discussion). J Roy Stat Soc B, **34**,187–220 (1972).
3. Frick, B., Pestana Barros, C., Prinz, J.: Analysing head coach dismissals in the German "Bundesliga" with a mixed logit approach. European Journal of Operational Research **200**(1), 151–159 (2010).
4. Tozetto, A. B., Carvalho, H. M., Rosa, R. S., et al.: Coach Turnover in Top Professional Brazilian Football Championship: A Multilevel Survival Analysis. Frontiers in Psychology **10**, 1246 (2019).

4.10 Data science for unstructured data

Identification and modeling of stop activities at the destination from GPS tracking data

Identificazione e analisi delle soste a destinazione desunte da dati di tracciamento GPS

Nicoletta D'Angelo, Giada Adelfio, Antonino Abbruzzo and Mauro Ferrante

Abstract This paper aims at analysing tourist behaviour at destination by focusing on the main determinants of their stop activities. A density-based cluster algorithm identifies the stops from GPS tracking data on cruise passengers starting from data on individual trajectories. A Poisson regression model analyses the effects of socio-demographic, and itinerary characteristics on the number of stops made. The results are of interest both from a methodological perspective, related to the analysis and synthesis of GPS tracking data and from an applied perspective concerning tourists' knowledge of spatial behaviour and its implications for destination management.

Abstract *Il presente articolo ha lo scopo di analizzare il comportamento turistico a destinazione, con un focus specifico sulle soste effettuate dai turisti nella destinazione. Vengono analizzati dati desunti da dispositivi GPS raccolti su un campione di crocieristi, a partire dai quali è possibile individuare le soste a destinazione attraverso l'impiego di un opportuno algoritmo. L'effetto delle caratteristiche socio-demografiche e legate all'itinerario intrapreso sul numero di soste effettuate viene studiato attraverso l'impiego di modelli di regressione di Poisson. I risultati sono di interesse sia da un punto di vista metodologico, legato all'analisi e sintesi di dati GPS, che dal punto di vista applicato, per quanto attiene alla conoscenza del comportamento spaziale dei turisti e delle relative implicazioni per il management della destinazione.*

Key words: GPS data analysis; Stop identification algorithm; Tourist behaviour

Nicoletta D'Angelo, Giada Adelfio and Antonino Abbruzzo
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy,
e-mail: nicoletta.dangelo@unipa.it; giada.adelfio@unipa.it; antonino.abbruzzo@unipa.it

Mauro Ferrante
Department of Culture and Society, University of Palermo, Italy,
e-mail: mauro.ferrante@unipa.it

1 Introduction

Collecting data on tourist mobility is of paramount importance for the study tourists' behaviour within a destination [10]. Traditional methods are generally based on post-visit questionnaires or trip diaries, which rely on the accurate recall of the visited places and activities. Moreover, they may introduce a bias on participants' behaviour, since they know to be observed [2]. Nowadays, GPS technology allows collecting information on human mobility at a very high temporal and spatial detail, with any effort required in recalling the visited places. Since the book of Shoval and Isaacson [9], many studies in the tourism field have been carried out by the GPS technology (see [8] for a review of the first decade).

This paper aims to analyse cruise passengers' stop activities derived from data collected on cruise passengers' experience at their destination by integrating a questionnaire-based survey and GPS tracking data. First, a density-based cluster algorithm (DBSCAN) is used to identify stop locations from GPS tracking data. In a second step, a Poisson regression model evaluates the relationships among individual-related variables (collected through the questionnaire-based survey) and itinerary-related characteristics, derived from GPS tracking data on the whole itinerary.

2 Cruise Passengers' Data

In this analysis, we select the cruise tourism segment considering the single exit/entry point and the relatively brief visiting time, which characterises cruise passengers' experience at their destination. These features make the use of GPS technology particularly suitable for analysing their experience at the destination [7]. The data have been collected in Spring 2014 in the city of Palermo (Italy), integrating a questionnaire-based survey and the GPS tracking data on cruise passengers' itinerary at the destination (see [4] for details on data collection procedures). The GPS tracking data consist of coordinate points collected at every 10 seconds. A set of pre-processing operations have been implemented, on the raw GPS data, aimed mainly at removing the outlier observations and impute the missing data. See [1] for details on pre-processing operations.

3 Density-based Algorithm for GPS Data

Identifying stop locations is an essential step for summarising the information in tracking data since they may indicate areas of interest for the individuals. Gong et al. [5] review the main research results on stop location identification and propose a classification of methods according to five groups, namely: centroid-based methods, speed-based methods, duration-based methods, density-based methods and hybrid

methods. Nonetheless, in evaluating any technique for identifying stop locations, the characteristics of the used dataset, the research aims, the study context and the related assumptions, must be considered. In tourism, stops identification may reveal popular places, such as tourist attractions, restaurant locations or shopping centres.

The approach presented in [1] is implemented to identify stop locations from the individual GPS trajectories. According to this approach, starting from a generic trajectory for unit i, $D^{(i)}$, a cluster is defined as a minimum set of spatio-temporal points, which are sufficiently close to each other. The DBSCAN is a density-based algorithm [3], which is designed to identify arbitrary-shaped clusters, where the clusters are sets of spatial points which fall within a certain distance. Concurrently, the algorithm can identify the *noise* points, which are spatial points not belonging to any cluster.

Let $p = (x, y)$ be a point in the trajectory of a generic unit. The ε-*neighbourhood* of a point p is defined by $ne_\varepsilon(p) = \{q \in D : ||p,q|| \leq \varepsilon \in \mathbb{R}^+\}$, where $||p,q||$ is a distance function. If the cardinality of an ε-neighbourhood of a point p, i.e. $|ne_\varepsilon(p)|$, is at least greater than a minimum number of time points ($minpts \in \mathbb{R}^+$) then p is a *core point*.

A point p is *directly density-reachable* from the object q with respect to ε and $minpts$ if $p \in ne_\varepsilon(q)$ and $|ne_\varepsilon(q)| \geq minpts$. A point p is *density-reachable* from the object q with respect to ε and $minpts$ if there is a chain $p_1, \ldots, p_l$, $p_1 = q$, $p_l = p$ such that p_{i+1} is directly density-reachable from p_i. An object p is *density-connected* to object q, with respect to ε and $minpts$, if there is an object o such that both p and q are density-reachable from o with respect to ε and $minpts$.

Definition 1. A *cluster* C is a non-empty subset of D satisfying the following requirements:

- $\forall$ p,q: if $p \in C$ and q is density-reachable from p with respect to ε and $minpts$, then $q \in C$;
- $\forall$ $p,q \in C$: p is density-connected to q with respect to ε and $minpts$.

Let $C_1, \ldots, C_k$ be the clusters of D with respect to ε and $minpts$, then $p \in D$ is a *noise* point if it does not belong to any cluster C_i. The algorithm starts with the first point p in the database D, and it retrieves all the neighbours of a point p with respect to ε and $minpts$. If p is a core point, this procedure will yield a cluster concerning ε and $minpts$. If p is not a core point, no points will be density-reachable from p, and the DBSCAN algorithm will proceed to consider the next point of the database. The DBSCAN has been deployed in the DBSCAN R package [6]. For the purpose of the present study, a value of 30 (5 minutes) for the $minpts$ parameter and of 40 (meters) for the value of distance ε were used.

4 Poisson GLM for the determinants of the stop activities

A generalized linear model (GLM), with the logarithm as link function and family Poisson, is proposed to model the number of stops y_i as a count variable. We include

in our analysis some covariates related to: a) visit duration, b) tourists' synthetic information on the itinerary undertaken and other socio-demographic characteristics and c) indicating whether the cruise passenger has visited a specific touristic attraction.

The covariate related to a) is the *average duration of the stops* (`Avg`).

The covariates related to b) are:

- `Dist`: *maximum distance from the port*, dichotomized in < 3.5 and ≥ 3.5 km.
- `Time`: *total time of the tour*, dichotomized in $<3,5$ and $\geq 3,5$ hours;
- `Lenght`: *total length of the tour*, dichotomized in <11 and ≥ 11 km;
- `Edu`: *education level*, dichotomized in `low` (High school diploma or Bachelor degree) and `high` (Master or PhD);
- `Visit`: *first visit*, indicating whether the cruise passenger is visiting the city for the first time (`yes`) or not (`no`);
- `Inc`: *yearly income*, dichotomized in <40000 and $\geq$ `40000` euro.

Finally, the covariates related to c) are selected touristic attractions in Palermo, among which: `Cathedral`, `Politeama`, `Ballarò`, `Capo`, `Vucciria`.

After exploratory data analysis, the proposed Poisson model is:

$$\begin{aligned}\log(\mathtt{y}_i) = &\alpha + \beta_1 \mathtt{Avg}_i + \beta_2 \mathtt{Dist}_i + \beta_3 \mathtt{Time}_i + \beta_4 \mathtt{Edu}_i + \beta_5 \mathtt{Inc}_i + \beta_6 \mathtt{Cathedral}_i \\ &+ \beta_7 \mathtt{Politeama}_i + \beta_8 \mathtt{Ballarò}_i + \beta_9 \mathtt{Capo}_i + \beta_{10} \mathtt{Vucciria}_i. \end{aligned} \tag{1}$$

5 Results of the analysis

The implementation of the DBSCAN algorithm identifies 1350 stops made by the 218 cruise passengers considered for the analysis, with an average number of stops per tourist equal to 6, with a minimum of 1 and a maximum of 16. The duration of stops ranges from 5 to 130 minutes with a mean of 16.15 and a median of 10.33 minutes. In Figure 1, bivariate plots of the number of stops according to the set of considered categorical covariates are reported.

As for the exploratory analysis, by looking at the plots in Figure 1, the number of stops generally appears higher as the mobility behaviour synthetic characteristics increase. Namely, the higher is the distance from the port, the higher is the number of the stops. Similar considerations hold for the total duration of the tour and the total length of the tour. That is, those who tend to explore more the destination tend to stop more. As for socio-demographic characteristics, the degree of association is less clear. Nonetheless, the median value of the number of stops is slightly higher for those with a higher education level and with higher levels of income. On the other hand, this value is slightly lower for repeated visitors, compared to those who visit the destination for the first time.

For describing the joint effect of the considered variables on the number of stops, the results of the estimated model (1) are reported in Table 1. Among the considered

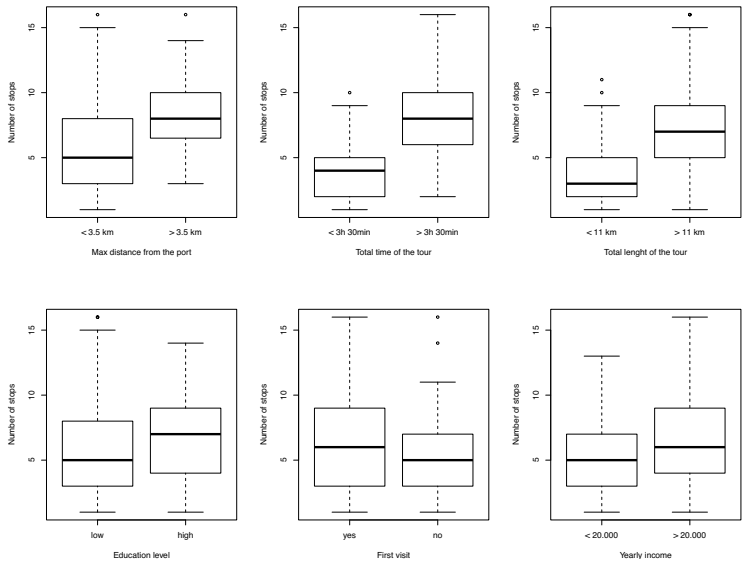


Fig. 1 Relationship between the response variable and the categorical covariates.

itinerary characteristics variables, both the maximum distance from the port and the total tour duration are positively associated with the number of stops. Whereas, the average duration of stops is negatively associated with the total stop number. This result is rather reasonable: people with longer stops tend to perform fewer stops. Regarding socio-demographic characteristics, both education and income are slightly significantly associated with the total stops number. More in detail, the association with the income may be explained with the potential expenditure associated with the stop activities, as well as education may be associated with the visit to museums or other types of attractions. In terms of the visited places, only the Cathedral visit is slightly associated with the total stops number, whereas the other considered places of interest are not significantly associated with the total stops number.

Table 1 Parameters' estimates of the proposed model

Variable	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
`Intercept`	3.5811	0.5562	6.44	0.0000
`Avg`	-0.0921	0.0230	-4.01	0.0001
`Dist`	1.8676	0.4685	3.99	0.0001
`Time`	3.6051	0.3873	9.31	0.0000
`Edu`	0.8033	0.3940	2.04	0.0427
`Inc`	0.7932	0.3359	2.36	0.0191
`Cathedral`	0.8231	0.3619	2.27	0.0240
`Politeama`	0.6747	0.3627	1.86	0.0643
`Ballarò`	0.7094	0.4750	1.49	0.1369
`Capo`	1.2788	0.7728	1.65	0.0995
`Vucciria`	0.9169	0.6461	1.42	0.1574

6 Conclusions

This paper proposed an approach to derive meaningful information on a tourist visit at the destination, starting from GPS tracking data and questionnaire-based information. The complex structure of GPS data requires methods able to synthesize the vast amount of data collected in order to extract relevant information on visitors' behaviour. Among the various types of information which can be derived from GPS tracking data, in this contribution, we focused our attention on stop activities as an important element on tourist visit, since stops are likely to indicate relevant locations at the destination. To identify cruise passengers' stop activities, a spatial clustering algorithm, the DBSCAN, has been applied to the GPS tracking data collected at an individual level. Moreover, thanks to the integration between stop activities, socio-demographic characteristics and other itinerary-related information, it was possible to identify some of the potential determinants of stop activities at the destination. The determination of the number of stops and the analysis of their main determinants is fundamental for service management since the stop locations may identify places where most of the expenditure is concentrated. The results are of interest both from a methodological perspective, related to the analysis and synthesis of GPS tracking data and from an applied perspective concerning tourists' knowledge of spatial behaviour and its implications for destination management.

References

1. Abbruzzo, A., Ferrante, M., De Cantis, S.: A pre-processing and network analysis of GPS tracking data. Spat. Econ. Anal., 1–24 (2020) doi: 10.1080/17421772.2020.1769170
2. East, D., Osborne, P., Kemp, S., Woodfine, T.: Combining GPS & survey data improves understanding of visitor behaviour. Tour. Man., **61**, pp. 307–320 (2017)
3. Ester, M., Kriegel, H.-P., Jorg, S., Xu, X.: A density-based clustering algorithms for discovering clusters. In E. Simoudis, J. Han, and U. Fayyad (Eds.), KDD-96: Proceedings of the second international conference on knowledge discovery and data mining, **96**(34), Portland, Oregon, pp. 226–231 (1996). Retrieved Jan 2021, from https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf
4. Ferrante, M., De Cantis, S., Shoval, N.: A general framework for collecting and analysing the tracking data of cruise passengers at the destination. Curr. Issues Tour., **21**(12), 1426–1451 (2018)
5. Gong, L., Sato, H., Yamamoto, T., Miwa, T., Morikawa, T.: Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. J. Mod. Transp., **23**(3), 202–213 (2015)
6. Hahsler, M., Piekenbrock, M., Arya, S., Mount, D.: Package 'dbscan': Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. R package version (2018) https://github.com/mhahsler/dbscan.
7. Shoval, N.: Tracking technologies and urban analysis. Cities, **25**(1), 21–28 (2008)
8. Shoval, N., Ahas, R.: The use of tracking technologies in tourism research: the first decade. Tour. Geog., **18**(5), 587–606 (2016)
9. Shoval, N., Isaacson, M.: Tourist mobility and advanced tracking technologies. Routledge, London (2009)
10. Stopher, P.: Collecting, managing, and assessing data using sample surveys. Cambridge University Press, Cambridge (2012)

A generalization of derangement
Sulla generalizzazione delle dismutazioni

Maurizio Maravalle and Ciro Marziliano

Abstract As a natural extension of the concept of *derangement*, we define as *derangement-3*, *derangement-4*,..., *derangement-K*, the triplet, quadruplets,..., *K*-plets of permutations that have no common elements in the same place. In this paper we propose a theoretical conjecture for the asymptotic behaviour of higher order derangements and validate it by computer simulation for significant values of *K*.

Abstract *Come naturale estensione del concetto di dismutazione definiremo dismutazioni di ordine tre, quattro,..K rispettivamente le terne, quaterne e le K-ple di permutazioni che non hanno elementi in comune nello stesso posto. Nellarticolo viene presentata una congettura per valutarne le probabilità nel comportamento asintotico di queste dismutazioni di ordine superiore, al variare del numero di elementi, confortato con simulazioni per valori significativi di K.*

Key words: Derangement, Married Couples Problem, Montmort's matching problem, Permutation, Problème des Rencontres, Subfactorial.

1 Introduction

The problem of derangement, hereafter referred to as *derangement-2* was formulated and solved long time ago by Pierre M. de Montmort respectively in 1708 and 1713. Nicholas Bernoulli also solved the problem using the inclusion-exclusion principle. The results are summarised for convenience in Section 2, where the generalization to higher order derangements is also discussed. In Section 3 the conjecture is presented and verified experimentally, via simulation, in Section 4. The problem

Maravalle
University of L'Aquila - Department of Information Engineering, Computer Science and Mathematics. e-mail: maurizio.maravalle@univaq.it

Marziliano
University of L'Aquila - Statistical office. e-mail: ciro.marziliano@univaq.it

remains of finding if possible an exact analytical formulations for the more general *derangement-K*[1].

2 Derangement and generalization

Let S_n be the symmetric group of all permutations of n elements $(1,2,3,\ldots,n)$ whose cardinality is $|S_n|$=n!. A derangement is a permutation in which none of the elements appears in its original position. The number of derangements is indicated by $!n$, called a subfactorial of n and is given by:

$$!n = n! \sum_{i=0,1,2,\ldots,n} \frac{(-1)^i}{i!}. \tag{1}$$

The probability that two elements of S_n do not have some element in same place is:

$$P_2[n] = \frac{!n}{n!} = \sum_{i=0,1,2,\ldots,n} \frac{(-1)^i}{i!} = \sum_{i=2,\ldots,n} \frac{(-1)^i}{i!} \qquad \forall\, n \geq 2.$$

The proof for equation (1) is based on the well known inclusion-exclusion principle. As n increases, the probability converges very rapidly $(n \geq 4)$ to $1/e$. Figure 1 shows a graphical representation of this probability as a function of n. Let us now consider what is the probability $P_3[n]$ that three elements of S_n have no element in the same place. More generally we define as $P_K[n]$ with $n \geq K$, the probability that K permutations of S_n have no common elements in the same place. At present the exact solution to this problem is not known. By combinatorial analysis, it is possible to calculate some $P_K[n]$ values for small values of n.
For example in case of $K = 3$ we have:

$$P_3[3] = \frac{1}{18}; \quad P_3[4] = \frac{1}{24} \ldots$$

for $K = 4$:

$$P_4[4] = \frac{1}{24^2}; \ldots.$$

Table 1 shows the exact results for *derangement-3*, for different values of n, together with the results from simulation discussed in Section 4.

[1] We have introduced the notation derangement-K to differentiate it from K-derangement, which has been used by other authors with a completely different meaning [3, 2, 1].

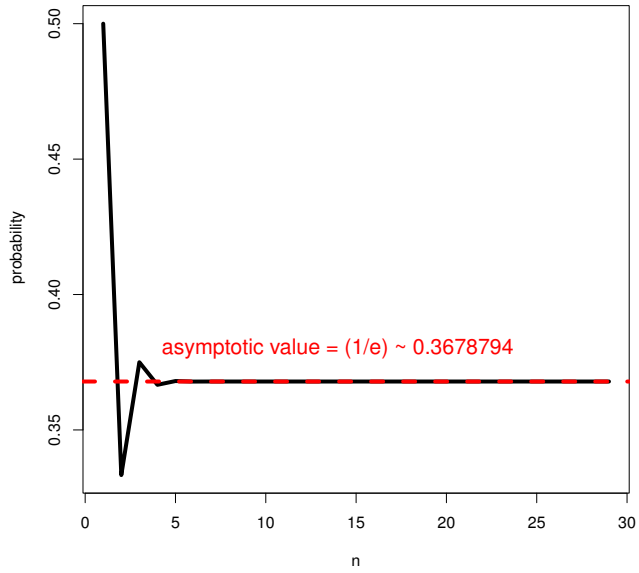


Fig. 1 Derangement

n	N. derangement-3	Probability
3	2	$\frac{2}{3!^2} = \frac{1}{18}$
4	24	$\frac{24}{4!^2} = \frac{1}{24}$
5	552	$\frac{552}{5!^2} = \frac{23}{600}$
6	21280	$\frac{21280}{6!^2} = \frac{133}{3240}$
7	1073760	$\frac{1073760}{7!^2} = \frac{2237}{52920}$
8	70299264	$\frac{70299264}{8!^2} = \frac{26153}{604800}$
9	5792853248	$\frac{5792853248}{9!^2} = \frac{3232619}{73483200}$

Table 1 Number of derangement-3

3 An asymptotic conjecture

Given the difficulty to calculate exactly the probabilities in different cases and by starting from a consideration on *derangement-3*, we attempt to generalize the result previously obtained for *derangement-2*. Taken as X, Y and Z three elements of S_n,

the probability that at two by two don't have elements in the same place is $P_2[n]$. For *derangement-3* we seek the probability of event

$$\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$$

having indicated with $\mathcal{A}$ the event that X and Y do not have common elements, with $\mathcal{B}$ that X and Z do not have elements in common and with $\mathcal{C}$ that Z and Y have no elements in common. For n high enough and assuming that the events are independent, the probability of $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$ should be equal $(1/e)^3$. Same consideration for *derangement-4*, but in this case the pairs that have to be independent are $\binom{4}{2}$ and so for *derangement-5* the pairs will be $\binom{5}{2}$. The asymptotic behaviour, as confirmed by simulation (see Section 4), appears to be correct, except, as expected for the initial values of n. This implies that the events are pairwise independent but not three by three. So, by generalizing, we expect they to be not independent for four by four for *derangement-4*, and so on. A simple check to verify this prediction is the case of $P_3[3]$; if they were independent events, one should have as probability $P_3[3] = (P_2[3])^3 = 1/27$ instead, by contrast $P_3[3] = 1/18 \neq 1/27$. In some sense events are only asymptotically independent, i.e. only for $n \to +\infty$. In general it can be conjectured, however, that for *derangement-K* the limit value of the probability is

$$\lim_{n \to +\infty} P_K[n] = \left(\frac{1}{e}\right)^{\binom{K}{2}} \tag{2}$$

This also means that to have practical relevance, it is necessary to have very mall values of K, because these asymptotic probabilities become extremely small, when K increases, as shown in Table 2.

4 Simulation

For each value of K, the fact that the asymptotic value, beyond the first few values of n, is greater than those estimated via simulation, suggests that, if there is an analytic relationship, this might consists of a fixed component plus a variable one with alternating signs. Furthermore the latter should vanish asymptotically, thereby leaving the constant component only. Another consideration emerging from these simulations, in agreement with calculation of Section 2, is that the probability $P_K[K]$ decreases in the next step $P_K[K+1]$ and then tends to the asymptotic value but in an increasingly slow way for increasing K. On the basis of the simulations, it should be noted that the asymptotic trend is reached, for derangement-2 with very small values of $n \geq 4$. Always through simulation it is recognized that even for $K = 5$ the probability is reached asymptotically for $n \approx 500 \div 800$. In Figure 2, it is

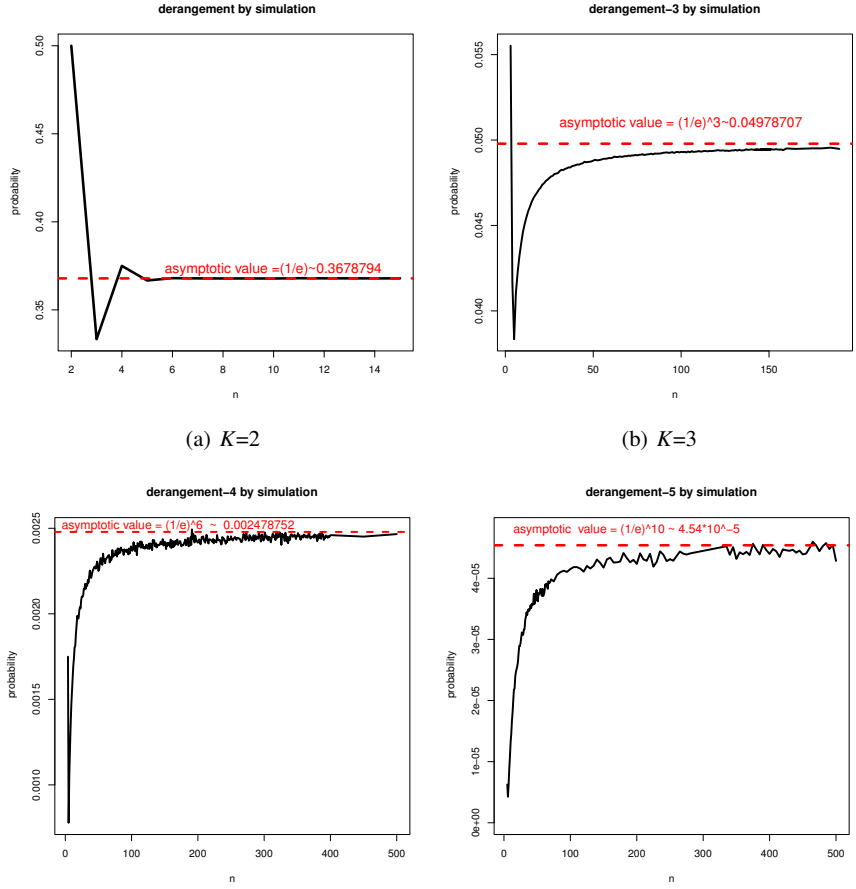
K	asymptotic value
2	0.3678794
3	0.04978707
4	0.002478752
5	4.539993e-05
6	3.059023e-07
7	7.58256e-10
8	6.9144e-13
9	2.319523e-16
10	2.862519e-20
11	1.299581e-24
12	2.170522e-29
13	1.333615e-34
14	3.014409e-40
15	2.506567e-46

Table 2 Asymptotic probabilities values

reported graphically the simulation results for $K = 2,...,5$. Note how the Figure 1 corresponds perfectly to Figure 2(a), case of derangement-2.
All simulation are made using software R.

References

1. Feinsilver, P., McSorley, J.: Zeons, Permanents, the Johnson Scheme, and Generalized Derangements. International Journal of Combinatorics **Volume 2011**, 29 pages (2011). Doi:10.1155/2011/539030
2. Fraticelli, A.: Generalized derangments. http://people.missouristate.edu/lesreid/reu/2009/PPT/tony.pptx (2009)
3. Hassani, M.: Derangements and applications. Journal of Integer Sequences **6**(1) (2003)



(a) K=2 (b) K=3

(c) K=4 (d) K=5

Fig. 2 Simulations graphics

Analysis of clickstream data with mixture hidden markov models

Analisi dei clickstream data tramite i mixture hidden markov model

F. Urso, A. Abbruzzo, and M.F. Cracolici

Abstract Clickstream data is an important source of information for businesses, however it is not easy to manage this data and also to convert the information coming out from it in competitive advantage is not a trivial task. This study considers the application of mixture hidden Markov models to clickstream data extracted from a travel services company's e-commerce portal. We find clusters related to web users' browsing behaviour and geographical position that provide essential indications for developing new business strategies.

Abstract *I clickstream data sono un'importante fonte di informazioni per l'e-commerce, sebbene non siano semplici da gestire e convertire queste informazioni in un reale vantaggio competitivo non è un compito banale. In questo articolo, consideriamo l'applicazione dei mixture hidden Markov model a dati relativi al flusso di clickstream estratti dal portale e-commerce di un'azienda di servizi turistici. Sono stati individuati cluster relativi al comportamento di navigazione degli utenti e alla loro posizione geografica che forniscono indicazioni importanti per lo sviluppo di nuove strategie di business.*

Key words: Clickstream Data, Online browsing behaviour, Mixture hidden Markov models, Tourism 2.0, Web mining

---

F. Urso
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy, e-mail: furio.urso@unipa.com

A. Abbruzzo
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy, e-mail: antonino.abbruzzo@unipa.com

M.F. Cracolici
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy, e-mail: mariafrancesca.cracolici@unipa.com

1 Introduction

Analyzing users' browsing behaviour when exploring e-commerce portals allows companies to gain significant advantages, such as the ability to classify potential customers, tailor their offers accordingly, or identify new opportunities that lead to changes in business strategies. Unfortunately, although the analysis of clickstream data provides essential information on the users' movements exploring a website (5; 3; 9; 6), it does not explain the underlying reasons behind their navigation choices. Furthermore, to develop effective marketing strategies, it would be desirable to identify users' subpopulations based on browsing behaviour. For this reason, statistical models with hidden variables such as mixture hidden Markov models (MHMMs) are a suitable tool for the analysis of clickstreams data. They allow to take into account two levels of uncertainty, a latent process whose evolution explains users' motivations to move from one page to another (13; 8; 7), and a hidden variable related to the presence of clusters representing browsing "profiles" (11; 15; 10). Here, we apply the MHMMs to data collected from the e-commerce portal of the PalermoTravel, a company operating in the hospitality sector, to analyze the differences in user behaviour by identifying the navigation profiles.[1] The paper is structured as follows: Section 2 illustrates the mixture of hidden Markov models. In Section 3, we have applied the model to identify browsing behaviour profiles taking into account user information such as geographic location obtained from IP addresses, access devices and access period.

2 Mixture Hidden Markov models

Hidden Markov models (2; 4; 14) allow analyzing time series whose evolution is supposed to depend on a latent Markov process. The mixture hidden Markov models enable to relax the hypothesis of a single population through latent variables that take into account the different longitudinal patterns in the sequences (15), identifying groups (clusters) of sequences assigned with specific probabilities derived from the data (11). This paper focuses on discrete MHMMs where the latent and the response are assumed discrete random variables.

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})$ be the generic i-th sequence of length T with card$|Y_i| = R$, $U_i = (U_{i1}, U_{i2}, \ldots, U_{iT})$ the i-th hidden random vector with card$|U_i| = S$ and assume n independent sequences. Let $M = \{M^1, M^2, \ldots, M^K\}$ be a set of HMMs, where $\Theta^k = \{\pi^k, A^k, B^k\}$ is the set of parameters for each sub-models M^k, related on each sub-population $k = 1, \ldots, K$. For each sequence Y_i, we define the prior cluster probabilities that the model parameters are the ones related to the k-th sub-model M^k as $P(M^k) = w_k$. The log-likelihood is computed as

[1] PalermoTravel is a pseudonym.

$$\ell(\Theta;Y) = \sum_{i=1}^{n} \log P(Y_i|\Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_{ik} \sum_{u} \pi_{u_1}^k b_{u_1}^k(y_{i1}) \prod_{t=2}^{T} a_{u_{t-1},u_t}^k b_{u_t}^k(y_{it}) \right), \quad (1)$$

where the hidden state sequences $u = (u_1, u_2, \ldots, u_T)$ take all possible combinations of values in the hidden state space S and where y_{it} are the observations of subject i at time t, $\pi_{u_1}^k = P(u_1 = s|\Theta^k)$ with $s \in \{1, \ldots, S^k\}$ is the initial probability of the hidden state at time $t = 1$ in sequence u for cluster k; $a_{u_{t-1},u_t}^k = P(u_t = j|u_{t-1} = i, \Theta^k)$ with $i, j \in \{1, \ldots, S\}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t in cluster k; and $b_{u_t}^k(y_{it}) = P(y_{it} = r|u_t = s, \Theta^k)$ with $s \in \{1, \ldots, S\}$ and $r \in \{1, \ldots, R\}$ is the probability that the hidden state of subject i at time t emits the observed state at t in cluster k. MHMM can be generalized to include time-constant covariates (15) that can be used to estimate cluster memberships w_{ik} of each sequence according to the following multinomial logistic model

$$w_{ik} = P(M^k|X_i) = \frac{e^{X_i\gamma_k}}{1 + \Sigma_{j=2}^{K} e^{X_i\gamma_j}}, \quad (2)$$

where γ_k is the set of coefficients associated with the vector of covariates X_i for observation i and the k-th class, and $\sum_{k=1}^{K} w_{ik} = 1$. The cluster posterior probabilities $P(M^k|Y_i, x_i)$ are obtained as

$$P(M^k|Y_i, X_i) = \frac{P(Y_i|M^k, X_i)P(M^k|X_i)}{P(Y_i|\Theta, X_i)}, \quad (3)$$

where $P(Y_i|\Theta, X_i)$ is the likelihood of the complete MHMM for subject i. In order to obtain the parameters estimates, the forward-backward algorithm (1; 12) can be used in MHMM context as illustrated by Vermunt et al. (15).

3 Results

MHMM has been used to analyze the difference in browsing behaviour among users of the PalermoTravel website. The data (log files) were collected in 2017 from September to December. They consist of 10,252 user sessions of maximum length T=20. These sessions are the sequences of pages viewed from the same IP address in a fixed period. Specifically, we do not consider the page names but their page category, corresponding to the thematic areas of the site: *Homepage*, *Attraction*, *Accommodation*, *Event*, *Experience*, *Service* and *Info* about the company. We consider three time-constant covariates collected from the website log files: IP address geographic area (i.e., Africa, Asia, East Europe, Italy, Latin America, Middle East, North America, North Europe, Oceania, Russia, South Europe), access device distinguishing PC and mobile and access month. These covariates are used to estimate prior cluster probabilities through the multinomial logistic model as in equation 2. Finally, using a selection procedure based on a combination of AIC and entropy, we

have selected the MHMM consists of 4 clusters and different hidden states in each cluster (i.e., 4,5,5,4) by applying a model selection procedure.

As an example, we show in Figure 1 a directed graph representing the path followed by users in cluster 1. Each pie graph represents a hidden state and edges are the transitions between states. Transition probabilities are displayed on the edges. The different colours and sizes of the pie slices represent emission probabilities of observed states (the pages' thematic area). The identified clusters are the following.

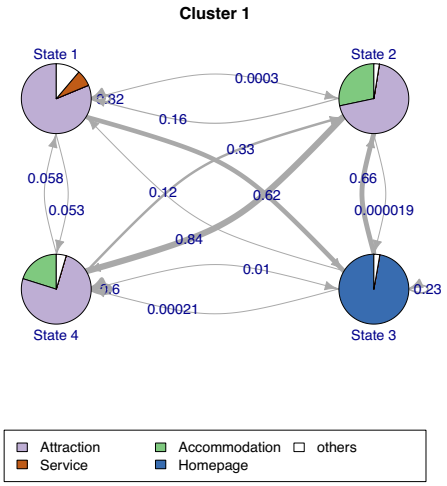


Fig. 1 Cluster 1 hidden Markov process structure. Vertexes represent hidden states, edges show the transition probabilities and the slices' color and size show emission probabilities. Emission probabilities lower than 0.05 are classified as "others".

Cluster 1 includes 15% of the sessions and has a high percentage of Northern Europeans who logged in the website from both PC and mobile, especially in October. Users start their session from state 3 (with probability 0.89). This state emits the observed state *Homepage* with probability 0.97. They move to state 2 with probability 0.66, then to state 4 with probability 0.84 and they stay in this state with probability 0.60. These states emit *Attraction* with probability 0.69 and 0.75 respectively. So, users appear to have a particular interest in general information about the region and less interested in the company's products, which is why the cluster was named **Information seeker**.

Cluster 2 (31.3% of the sessions) largely includes Italian, North American and Northern European users who explored the site using mostly their PC in November and December. Users start their session from state 3 with probability of 0.42 and stay in this state with probability 0.92. State 3 emits *Attraction* with probability 0.91. If users start from state 5 (with probability 0.23) they move to state 2 with probability 0.56 and stay in this state with probability 0.75. These two states emit *Event* with probabilities 0.72 and 0.62 respectively. If users start from state 1 with probability

0.19, they stay in this state with probability 0.96. State 3 emits *Accommodation* with probability 0.59. It seems that users primary interest remains viewing tourist attractions. Still, we also note that navigation can view seasonal events (a sign of the desire to select a period of visit) and apartments. This cluster was named **Potential tourist**.

Cluster 3 (41.3% of the sessions) includes the majority of Italians, with accesses mainly from a PC and in September. Users start from state 2 with probability 0.56. This state emits *Homepage* with probability 0.98. Then they move to other states with probabilities not too different. If they start from state 4 with probability 0.25, they will stay in this state with probability 0.95. This state emits *Accommodation* with probability 0.93. Users are interested in viewing and comparing tourism products, with less interest in the information pages presumably having prior knowledge of the Sicily region, so, this cluster was named **Expert tourist**.

Cluster 4 (12.4% of the sessions) includes most Asians and East Europeans and is characterized by the lowest percentage of access via mobile. Users start from state 3 with probability 0.82 and stay in this state with probability 0.89. State 3 emits *Homepage* with probability 0.99. So, they focused only on viewing the home page and move to different areas of the site. This cluster contains an interesting subgroup of users who would stay there if it reached state 2, this state emits *Info* with probability 0.99. This interest in information relating to the tourism company could be attributed to companies interested in partnership relationships. In light of these considerations, this cluster was named **Casual explorer or Potential partner**.

In summary, focusing on the first three profiles as a representation of the user's interest in purchasing a holiday package, we note that most users are categorized in profiles relating to medium and high interest: cluster two (Potential tourist) and cluster three (Expert tourist). Italian users are mostly present in group 3 and scarcely present in the information seeker profile. Users viewing both tourist information and products (cluster two) are mainly North Americans. In contrast, most North European countries are distributed in all first three profiles, making up most profile 1 of information seeker related to a lack of interest in the company's products. Regarding cluster 4, the users are from countries with a greater "cultural distance" from the Italian such as Slavic or Asian countries. Although scarcely present in the sample, these users are all in this profile showing a superficial interest (rarely accessing areas of the site other than the Homepage) or not purchasing oriented (e.g. potential partners). These results highlight that two users' target exploring the website exist, which come out out two different business models i.e. the business-to-consumer (already adopted by the analyzed firm), and the business-to-business.

In light of the above results, the company should consider making the website more attractive to potential customers from non-Western countries and consider selling products and services to other companies.

References

[1] Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state markov chains. The annals of mathematical statistics **37**(6), 1554–1563 (1966)

[2] Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The annals of mathematical statistics **41**(1), 164–171 (1970)

[3] Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S.: Visualization of navigation patterns on a web site using model-based clustering. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 280–284 (2000)

[4] Collins, L.M., Wugalter, S.E.: Latent class models for stage-sequential dynamic latent variables. Multivariate Behavioral Research **27**(1), 131–157 (1992)

[5] Cooley, R.W., Srivastava, J.: Web usage mining: discovery and application of interesting patterns from web data. Citeseer (2000)

[6] Das, R., Turkoglu, I.: Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. Expert Systems with Applications **36**(3), 6635–6644 (2009)

[7] De Angelis, L., Dias, J.G.: Mining categorical sequences from data using a hybrid clustering method. European Journal of Operational Research **234**(3), 720–730 (2014)

[8] Dias, J.G., Vermunt, J.K.: Latent class modeling of website users' search patterns: Implications for online market segmentation. Journal of Retailing and Consumer Services **14**(6), 359–368 (2007)

[9] Eirinaki, M., Vazirgiannis, M., Kapogiannis, D.: Web path recommendations based on page ranking and markov models. In: Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 2–9 (2005)

[10] Helske, S., Helske, J.: Mixture hidden markov models for sequence data: The seqhmm package in r. arXiv preprint arXiv:1704.00543 (2017)

[11] Van de Pol, F., Langeheine, R.: Mixed markov latent class models. Sociological methodology pp. 213–247 (1990)

[12] Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)

[13] Scott, S.L., Hann, I.H.: A nested hidden markov model for internet browsing behavior. Marshall School of Business pp. 1–26 (2006)

[14] Vermunt, J.K., Langeheine, R., Bockenholt, U.: Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. Journal of Educational and Behavioral Statistics **24**(2), 179–207 (1999)

[15] Vermunt, J.K., Tran, B., Magidson, J.: Latent class models in longitudinal research. Handbook of longitudinal research: Design, measurement, and analysis pp. 373–385 (2008)

Using Google Scholar to measure the credibility of preprints in the COVID-19 Open Research Dataset (CORD-19)

Utilizzo di Google Scholar per misurare la credibilità dei preprints del COVID-19 Open Research Dataset (CORD-19)

Manlio Migliorati, Maurizio Carpita, Eugenio Brentari

Abstract COVID-19 crisis highlighted the difficulty of selecting and accessing credible scientific information as soon as they are produced. Typical peer review process normally takes several months before a scientific paper is published. Open-access preprints repositories (as Arxiv, MedrXiv and BiorXiv) enable fast posting, but without offering any guarantee. In this paper we propose a procedure to attribute a "credibility index" to preprints collected in the COVID-19 Open Research Dataset (CORD-19). Credibility Index is built using Google Scholar, and is based on the higher authors *h*-index corrected to compare researchers from different scientific fields and with different lengths of scientific careers. First results are encouraging, showing how both preprints and archives can be evaluated in this credibility perspective.

Abstract *La crisi determinata dal COVID-19 ha evidenziato la difficoltà di selezionare e accedere a materiale scientifico credibile in tempi rapidi. Il tipico processo di referaggio richiede molti mesi prima che un lavoro sia pubblicato, mentre repository come Arxiv, MedRxiv e BioRxiv consentono una pubblicazione veloce, senza offrire però garanzie di credibilità. In questo articolo proponiamo una procedura che attribuisce un "indice di credibilità" ai preprints raccolti nel COVID-19 Open Research Dataset (CORD-19). L'indice di credibilità è costruito utilizzando Google Scholar, ed è basato sul più alto h-index degli autori, corretto per confrontare ricercatori di settori scientifici diversi e con carriere di diversa durata. I primi risultati sono incoraggianti, e mostrano come tanto i preprint quanto gli archivi possano essere valutati in quest'ottica.*

Key words: Google Scholar, CORD-19, *h*-index, scientific research

University of Brescia, Department of Economics and Management, Contrada S. Chiara 50
e-mail: manlio.migliorati@unibs.it

1 Introduction

COVID-19 crisis highlighted, among others, the difficulty of finding reliable scientific information as soon as they are produced, selecting them in the huge amount of available material [7]. Typical peer review process, adopted by scientific editors, normally takes several months before a paper is published, and this long period is not acceptable in crisis situations, when relevant information must be made accessible as soon as possible. From the other side, open-access preprints repository (as Arxiv, Medrxiv, Biorxiv) enable posting of a scientific material in a very short time, but without offering any guarantee about the reliability of published contents. In this paper we propose a solution to this problem, by assigning a credibility measure, named h^*-index, to preprints posted in repositories, driving users in material selection.
The basic idea is to automatically access Google Scholar for a preprint, retrieving and sinthesizing data about each author (h-index, number of co-authors and lengths of scientific career) in a measure, and assign to the preprint the higher value among authors. First results, derived from applying the procedure to a sample (100 random selected preprints) for each of the 3 repositories are really encouraging, showing how not only preprints, but also archives themselves can be evaluated in this credibility perspective.
Our experience showed us how Google Scholar is a free, wide-ranging bibliographic resource relatively simple to be used and providing several useful information. Sometimes, anywhere, it is not so precise in returning attended results, due both to unregistered authors (registration is on a voluntary base) and some well known issues in authors naming (consistency, homonimy, diacritics) [3].
The paper is structured as follows: section 2 describes the dataset we used, section 3 illustrate our procedure in terms of credibility index definition and implementation of Google Scholar access, section 4 summarize the results we obtained and section 5 presents some conclusions and future directions.

2 The CORD-19 dataset

CORD-19 dataset [9] is a "free resource of tens of thousands of scholarly articles about COVID-19, SARS-CoV-2, and related coronaviruses for use by the global research community". In the Kaggle site [6], where it is possible to find this dataset, too, is explained that "In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19)".
This dataset contains COVID-19 and coronavirus-related research (e.g. SARS, MERS, etc.) from several sources (e.g. PubMed's PMC, Microsoft Academic, World Health Organization, Arxiv, BioRxiv, MedXriv), and offers the possibility of downloading the *metadata.csv* file where it is possible to find in a normalized way several features (and in particular title, abstract, authors, source) to be used for more refined

investigation. We worked on dataset dated July 2020, counting 192,509 references. For our purposes the file was filtered on the base of the source (we are interested only in preprints repositories, i.e. Arxiv, BioRxiv, MedXriv), furtherly restricted to contains some COVID-19 related keywords or in the title either in the abstract[1], to arrive to a set of 8,186 preprints.
For each of the 3 archives we selected a random sample of 100 preprints, constituting the starting point for our analysis.

3 The procedure

3.1 The credibility h^*-index

The credibility h^*-index is based on the h-index [4], modified to compare researchers that work in different scientific fields and with different lengths of scientific careers. A huge number of h-index variations has been proposed (see [1] for a review), and our approach integrate some of these variations.
A simple procedure to calculate the credibility of a scientific preprint based on co-author's h-index was defined in the following way:

1. for each co-author 3 data must be considered:
   h: the classical h-index
   t: the total number of co-authors of the h-papers, i.e. the papers considered in h-index definition, including the author under investigation
   a: the difference of years between the preprint under evaluation and the oldest h-paper plus 1
2. for each co-author the h^*-index is calculated as follows:

$$h^* = \frac{h}{m \times a} = \frac{h^2}{t \times a} \tag{1}$$

   where

$$m = t/h \tag{2}$$

   is the average number of co-authors considering all the h-papers (the idea of dividing h by m can be found in [2], the idea of dividing h by a can be found in [4]). The h^*-index ranges between 0 (h-index=0 for the author) and h (the author wrote all its h-papers alone, in the same year of the preprint under investigation).
3. the credibility index of the preprint is the highest h^* among co-authors as in [5]:

$$h^*_{preprint} = max_{co-authors}(h^*) \tag{3}$$

[1] We used the same query as CORD-19, an OR condition on terms *COVID-19*, *Coronavirus*, *Corona virus*, *2019-nCoV*, *SARS-CoV*,*MERS-CoV*, *Severe Acute Respiratory Syndrome*, *Middle East Respiratory Syndrome.*

3.2 The use of Google Scholar

Starting from preprint title and authors list contained in the dataset, we developed an R procedure that access Google Scholar to retrieve all data needed for calculating h^*-index. Please, note how Google Scholar can be accessed only via Web (no API available), so the only way is:

- building the address of the Google Scholar query page involving preprint title and authors names under investigation;
- executing it via a GET HTTP;
- parsing HTML results pages to find data we are looking for.

Google Scholar policy suggests of avoiding BOT accesses: IP users addresses are banned if they do too many or too fast requests. Consequently, random delays between two and four minutes were inserted between two consecutives accesses, to comply with that access policy.
Apart from details, the algorithm is based on the three main steps, repeated for all preprints in the sample:

- starting from preprint title, grab Google Scholar pages to find authors Google Scholar identifiers (strings of 12 characters). They are the key for querying the system;
- grab Google Scholar pages using title and author's ID to retrieve all data needed for calculating h^*-index. In this phase the R package [8] was used;
- calculate the maximum h^* among authors, and attributes it to the preprint.

4 First results of the analysis

We applied the procedure to a random sample of 100 preprints for each archive, obtaining results in Table 1.

Using h^*-index it is possible also to verify the level of credibility of a whole archive, as summarized in Figure 1. From these first sample results, Arxiv seems to contain "more credible" preprints with respect to the other two archives, showing an higher h^*-index mean, but a h^*-index distribution with higher variability and asymmetry too.

Table 1 Some statistics for the preprint samples of the three archives in CORD-19.

Statistics	Arxiv	Biorxiv	Medrxiv
num sample preprints	100	100	100
num too-many-authors preprints	1	14	9
num no-preprint	1	3	2
num evaluable preprints	98	83	89
num authors	388	723	621
num unknown authors	158	445	422
% unknown authors	40.72	61.55	67.95
mean of authors per preprint	3.96	8.71	6.98
mean of unknown authors per preprint	1.61	5.36	4.74
num preprints without h^*-index	9	4	22
num preprints with h^*-index	89	79	67
total h^*-index	46.48	30.69	31.56
mean h^*-index	0.52	0.39	0.47
st dev h^*-index	0.28	0.18	0.24

Note. We didn't analyse preprints with more than 20 authors (too-many-authors in the table) for avoiding overstress Google Scholar. Moreover, due to the CORD-19 dataset under analysis, it can happens that a preprint is not more available (no-preprint in the table). Authors not found in Google Scholar (because no registered, or because of missed correspondence between CORD-19 and Google Scholar authors names) are reported as unknown. Archive total h^*-index is the sum of h^*-index for archive sample preprints with a h^*-index. At last, the mean of authors per preprint is calculated as (num authors / (num preprints with h^*-index).

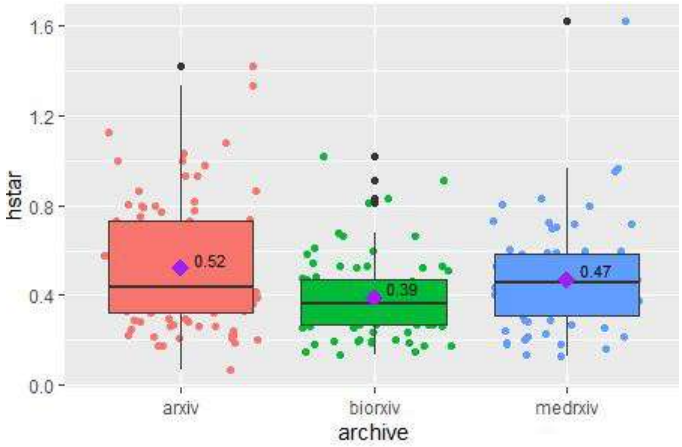


Fig. 1 The h^*-index box-plot for the preprint samples of the three archives in CORD-19.

Described approach is strongly based on the availability of data about authors in Google Scholar. If an (important) author is not registered, the credibility index will suffer for this lackness, eventually producing a lower h^*-index.
In our tests, we verified how often Google Scholar unknown authors in effect are not so impactful in terms of index calculation, but we observed also some cases where important authors, with high number of citations, were not registered.

5 Conclusions and futures directions

In this paper we described the procedure developed with the goal of attributing a "credibility" measure to preprints published in open access repositories as Arxiv, MedRxiv, BioRxiv. This measure can be important in crisis situation as COVID-19, when it is necessary to select and access scientific papers as soon as possible, without delays due to peer review, but with a certain degree of credibility.
We described the h^*-index, based on the classical h-index and assigned to a preprint on the base of higher value among co-authors. Preliminary results concerning credibility of preprints (a sample of 100 papers for each archive) are reported, and it is shown how also archives can be compared with respect the credibility classification of their preprints. From these first results Arxiv seems to be preferable with respect to other archives.
Future directions will address different definitions of the credibility index, studying its properties and comparing it with other indexes. Moreover, the dataset will be updated and the samples dimension increased. At last, the coverage offered by Google Scholar will be further investigated.

References

1. Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F.:h-index: A Review Focused in its Variants, Computation and Standardization for Different Scientific Fields. Journal of Informetrics 3:4, 273-289, 2009
2. Batista, P., Campiteli, M., Kinouchi O.: Is it possible to compare researchers with different scientific interests? Scientometrics, 68, 179-189, 2006
3. Demetrescu, C., Ribichini, A., Schaerf, M.: Accuracy of author names in bibliographic data sources: an Italian case study. Scientometrics 117, 1777–1791, 2018
4. Hirsch, J.E.: An index to quantify an individual's scientific research output, Proceedings of the National Academy of Sciences 102(46):16569-16572, 2005
5. Hirsch, J.E.: halpha: An index to quantify an individual's scientific leadership. Scientometrics 118, 673–686, 2019
6. www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge, accessed 25/02/2021
7. https://blogs.lse.ac.uk/impactofsocialsciences/2020/09/23/are-preprints-a-problem-5-ways-to-improve-the-quality-and-credibility-of-preprints/, accessed 25/02/2021
8. Keirstead, J.: scholar: analyse citation data from Google Scholar. R package version 0.2.0, https://cran.r-project.org/web/packages/scholar/index.html, 2021
9. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R.M., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B.B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D.M., Weld, D.S., Etzioni, O., Kohlmeier, S.: CORD-19: The Covid-19 Open Research Dataset. ArXiv, 2020

Mobile phone use while driving: a Structural Equation Model to analyze the Behavior behind the wheel

Telefono cellulare alla guida: un Modello a Equazioni Stutturali per analizzare il comportanto al volante

Carlo Cavicchia and Pasquale Sarnacchiaro

Abstract The use of mobile phones while driving is one of the main causes of road accidents and it is an ever-growing phenomenon. The key aim of this study is to simultaneously analyze individual Knowledge, Attitudes, and Behaviors toward the use of mobile phones while driving in one of the largest and most populous metropolitan areas of Italy, Naples. The data acquired from 774 questionnaires - administered to subjects evenly divided by gender and with an average age of 39 years - revealed that 69% of the participants had used their mobile phone while driving at least once in their lifetime. A Structural Equation Model shows how the relationship between Knowledge and Behavior passes through the Attitude. According to the collected data and statistical analysis, it is possible to identify factors that can greatly affect the use of mobile phone while driving.

Abstract *L'uso del telefono cellulare alla guida è un fenomeno in forte crescita e rappresenta una della maggiori cause di incidenti stradali. Lo scopo della ricerca è l'analisi simultanea delle conoscenze (Knowledge), delle attitudini (Attitudes) e dei comportamenti (Behaviors) riguardanti l'uso dei telefoni cellulari alla guida, prendendo in esame Napoli, una delle aree metropolitane più grandi e popolosa d'Italia. I dati raccolti da* 774 *questionari - somministrati a patentati, sia uomini che donne, con un'età media di* 39 *anni - rivelano che il* 69% *dei partecipanti ha usato il telefono cellulare alla guida almeno una volta nel corso della propria vita. Un modello a Equazioni Strutturali mostra come la relazione tra comportamenti e conoscenze sia veicolata dalle attitudini. Questo studio, attraverso i dati raccolti e le analisi svolte, permette di identificare i fattori che maggiormente influenzano l'uso del cellulare alla guida.*

Carlo Cavicchia
Econometric Institute, Erasmus University Rotterdam, Rotterdam The Netherlands
e-mail: cavicchia@ese.eur.nl

Pasquale Sarnacchiaro
Department of Economics Management and Institution, University of Naples Federico II, Naples, Italy
e-mail: sarnacch@unina.it

Key words: Knowledge, Attitudes, Behaviors, Cross-Sectional Survey, Measurement model

1 Introduction

Every year around 1.35 million people pass away because of road traffic crashes, while between 20 and 50 million more people suffer non-fatal injuries [2]. There are several factors that increase both the road traffic crashes risk and their resulting risk of injury or death worldwide. Speeding and driving under the influence of alcohol or other psychoactive substances are two of the most important determinants of road accidents and they present significant risk factors for road traffic injuries. However, other risk factors can be identified: non-use of safety devices such as motorcycle helmets, seat-belts, and child restraints, or distraction while driving, including the use of mobile phones [3]. Distracted driving is therefore considered as a major cause of these remarkable numbers. Specifically, the WHO Global Status Report on Road Safety 2018 [2] underlines that people which use mobile phones while driving are approximately 4 times more likely to be involved in a crash than drivers not using a mobile phone. In detail, the use of mobile phones while driving slows reaction times, and makes it difficult to keep in the correct lane, or to keep the correct following distances [2]. In Italy, this tendency seems confirmed; indeed, distraction is presumed to be the primary cause (16.3%) of road crashes, against speeding (10.2%), alcohol-related DUI (3.9%) and drug-related DUI (3.2%) [1], and one of the most important causes of distraction while driving appears to be the use of a mobile phone [6].

In this paper, we analyze the Behaviors enacted by Italian drivers regarding mobile phone use while driving, as well as the level of mobile phone involvement and its frequency of use. The key aim of this study is to simultaneously analyze through a Structural Equation Model (SEM) Knowledge, Attitudes, and Behaviors towards the use of mobile phones while driving in one of the largest and most populous metropolitan areas of Italy, Naples. Analysis of Knowledge, Attitudes and Behaviors about the risks of mobile phone usage while driving can lead us to identify its determinants in order to obtain the means to sensitize public opinion and improve people's awareness regarding the correct Behavior to adopt while driving.

This paper is structured in three sections: the next section deals with a brief description of methods used in the study. In the third section, the results of the SEM are pointed out. A brief conclusion ends the paper.

2 Research Methodology

The SEM is a statistical method for testing and estimating at once causal relationships among multiple independent and dependent latent (LVs) and manifest variables (MVs). SEM entails different sub-models. The structural model comprises the rela-

tionships among the LVs which have to be developed from theoretical considerations. The independent LVs are also referred to as exogenous LVs and the dependent LVs as endogenous LVs. For each of the LVs within the SEM a measurement model has to be defined. These models embody the relationships between the MVs and the LVs, and they can be either reflective or formative. In SEM related literature, two different types of techniques are established: covariance-based ones, as represented by LInear Structural RELations (LISREL,[4]), and variance-based ones, of which the Partial Least Squares (PLS) path modelling [7] is the most prominent representative.

In this paper we used the PLS, performed by SmartPLS (Version 3), because of its less stringent distributional assumptions for the variables and error terms and its ability to work with both reflective and formative measurement models. PLS-SEM is widely used for group comparison, investigating the possible presence of a group-effect in the definition of the LVs. The analysis of the invariance of the measures across different groups is necessary when using PLS-SEM for group comparison. SmartPLS provides permutation-based confidence intervals that allow determining if the correlation between the composite scores of the two groups is significantly lower than one (null hypothesis, H_0: $c = 1$). If the null hypothesis is not rejected, the composite does not differ much in both groups and, therefore, there is compositional invariance. In the next step, permutation-based confidence intervals for the mean values and the variances allow assessing if the composites' mean values and variances differ across groups.

3 Results

3.1 Sample and data characteristics

We analyzed 774 anonymous self-report surveys in the entirety of the metropolitan city of Naples. The questionnaire was anonymous and consisted of demographic information about the participant and three pools of queries focusing on Knowledge, Attitudes and Behaviors concerning the habit and frequency of mobile phone use while driving, for a total of 46 questions. Knowledge and Attitudes were assessed on a three-point Likert scale with options for "agree", "neither agree nor disagree", and "disagree", while inquiries regarding Behavior were presented in a four-answer format of "never", "sometimes", "often", and "always". Important general characteristics of the study sample can be reported: the mean age of the study sample is 39.27 years and most of participants are high school graduates or have a post graduate degree. 89% of the sample had been driving for more than 5 years and 54% of the sample drove a car; only the 27.6% of the interviewed drove both a car and a motorcycle. More than 75% of the sample was aware about the risks about using mobile phone while driving, but 28% of them was unaware that this practice was forbidden by law. Most of the participants thought that mobile phone usage was essential and more than 50% thought that it was necessary for business. Moreover, 24% of the sample admitted

to reading text messages, while only 16% admitted to writing them. Respondents had mainly sought general information about the risks concerning mobile phone use while driving, but only 30% kept themselves up to date on laws that regulated its use while driving. For the purpose of the study, PLS multigroup analyses based on the participants' socio-demographic profile (e.g., gender, age, education level, type of vehicle driven, level of driving experience) were performed.

3.2 Structural Equation Model

PLS-SEM was performed to formalize a scheme for the interpretation of driving Behavior and to detect its drivers. Starting from the considerations elaborated in the previous sections, we hypothesized that Knowledge and Attitude were exogenous LVs, while Behavior was an endogenous LV. Following the criteria summarized in [5], we supposed that Knowledge and Attitude were formative LVs and Behavior was reflective LV. The PLS estimations showed that the relationships between Behavior and Attitude and Attitude and Knowledge were statistically significant (1).

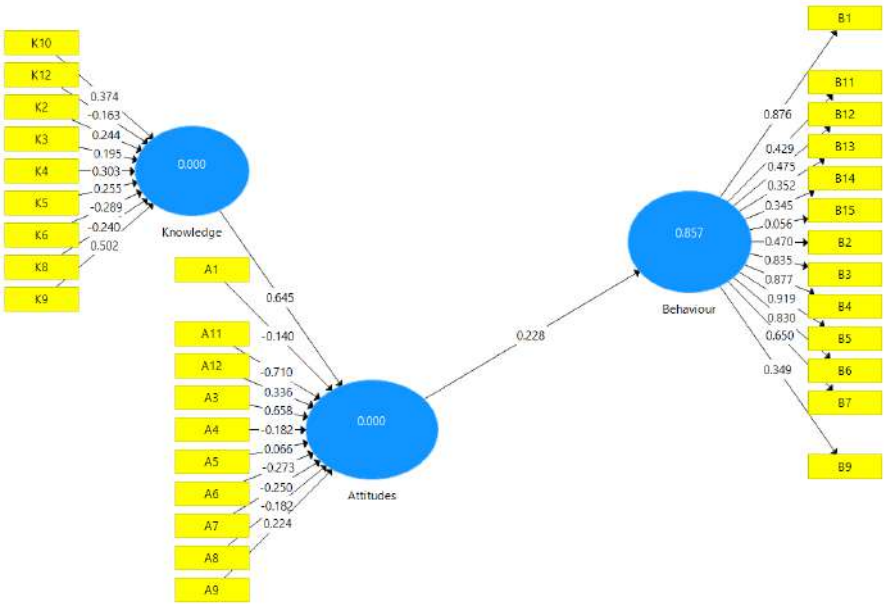


Fig. 1: Structural Equation Model – Path analysis

The goodness of the model was ultimately very strong ($R^2 = 0.856$). With regards to the path coefficients, we observed that the impact of the Knowledge on the Attitude

was considerably greater (0.645) than the impact of the Attitude on the Behavior (0.228). Also the indirect impact of the Knowledge on the Behavior resulted being important (0.147). It is noteworthy that these three impacts and all the outer loadings for latent variables were statistically significant. The direct effect of Knowledge on Behavior was tested but this effect eventually resulted not statistically significant.

4 Conclusions

The results of the present research supported the conclusion that the model well represented the collected data according to the result of the goodness-of-fit test. Similarly to earlier studies, this paper confirmed the goodness of the general structural model in helping to understand and explain how Knowledge, Attitude and Behavior are related. The analyzed population showed a good Knowledge on the subject together with positive Attitudes, and there was a general agreement that using a mobile phone while driving is considered unacceptable, even though the employed Behaviors are knowingly inappropriate according to Italian laws. Through our research we discovered that the relatively elevated education level of the sample and the greater driving experience (measured in years of driving license) of the participants were proven as inversely associated with the Behaviors examined; this means that while the experimental results of this survey can be used for the creation of targeted educational programs, community-based interventions and legal regulations, it might be fundamental to act more firmly in order to directly improve people's overall Behavior while driving. All these measures alone, in fact, may not be sufficient to reduce a phenomenon that is so deeply rooted in the population. This ever-growing phenomenon closely follows the technological evolution of our society and it results in an important indicator of how indispensable mobile phones have become in our daily life, a factor being in turn itself dependent on the increasing functions that can be performed through these devices. Considering that - as previously stated - this phenomenon has a strong impact on the increase in road accidents, on the economy and on public health, another solution might be to promote more restrictive regulations establishing a greater number of controls, using not only qualified personnel, but also innovative technologies possibly suitable for detecting real-time hands-on use of the mobile phone while driving.

References

1. International transport forum. road safety annual report 2019: Italy. Tech. rep. URL https://www.itf-oecd.org/sites/default/files/italy-road-safety.pdf
2. Who global status report on road safety 2018. Tech. rep. URL https://www.who.int/publications-detail/global-status-report-on-road-safety-2018
3. Who road traffic injuries report 2020. Tech. rep. URL https://www.who.int/health-topics/road-safety#tab=tab_2

4. Jöreskog, K.: A general method for estimating a linear structural equation system. ETS Research Report Series (2) (1970)
5. Sarnacchiaro, P., Boccia, F.: Some remarks on measurement models in the structural equation model: an application for socially responsible food consumption. Journal of Applied Statistics **45**(7), 1193–1208 (2018)
6. Trivedi, N., Haynie, D., Bible, J., Liu, D., Simons-Morton, B.: Cell phone use while driving: Prospective association with emerging adult use. Accid. Anal. Prev **16**, 450–455 (2017)
7. Wold, H.: Path models with latent variables: The nipals approach. In: H. Blalock, A. Aganbegian, F. Borodkin, R. Boudon, V. Capecchi (eds.) Quantitative Sociology, International Perspectives on Mathematical and Statistical Modeling, pp. 307–357. Academic Press (1975)

4.11 Demographic analysis

Life expectancy in the districts of Taranto

L'aspettativa di vita nei quartieri di Taranto

S. Cervellera, C. Cusatelli and M. Giacalone

Abstract The "Taranto case" has aroused the scientific interest of demographers, statisticians, epidemiologists and doctors, to understand what is happening today to the health of citizens, finding excesses of mortality and morbidity in certain pathologies strictly connected to pollution, in particular in the Ionian capital and in some neighboring municipalities. We have developed abbreviated mortality tables, using a methodology essentially different from that of Istat due to the fact that we do not rework the probabilities of death at all, which remain real. This solution is valid as Taranto is a large municipality, with almost 200,000 inhabitants.

Abstract *Il "caso Taranto" ha suscitato l'interesse scientifico di demografi, statistici, epidemiologi e medici, per capire cosa accade oggi alla salute dei cittadini, riscontrandosi eccessi di mortalità e morbosità in determinate patologie strettamente connesse all'inquinamento, in particolare nel capoluogo jonico ed in alcuni comuni limitrofi. Abbiamo elaborato tavole di mortalità abbreviate, utilizzando una metodologia essenzialmente differente da quella Istat per il fatto che non rielaboriamo affatto le probabilità di decesso, che rimangono quelle reali. Tale soluzione risulta valida in quanto Taranto è un comune con quasi 200.000 abitanti.*

Key words: mortality tables, biometric functions, Taranto

1 Introduction

In most of the scientific papers on mortality, and also on morbidity, the data coming from Istat (which provides official information through the survey on deaths and causes of death), on the number of the population residing in the municipalities and

[1] S. Cervellera, Municipality of Taranto; email: s.cervellera@comune.taranto.it
C. Cusatelli, University of Bari "Aldo Moro"; email: carlo.cusatelli@uniba.it
M. Giacalone, University of Naples "Federico II"; email: massimiliano.giacalone@unina.it

on the its movement and natural balance, as well as the mortality tables at the minimum level of provincial aggregation and, only exceptionally, for some large Italian municipalities [4].

These surveys, especially in the data collection phase, take place in a complex system falling within SISTAN which, requiring coordination and interaction between multiple subjects, has generated not only delays in the dissemination of information, but also misalignments between official and real data, that is, between the Istat indicators and the official municipal registers [2]. These misalignments generate errors throughout the information use chain, so every analysis is certainly infected. Istat does not provide official data on the extent of the error, so it is considered acceptable in its physiological extent. Even on the historical data relating to the size of the population, there is a misalignment between Istat information and that of the municipal registry, so delays and errors in the recording of the transmitted data can be assumed. For this reason, it is considered more valid to work with the number of deaths originally registered in the registry, rather than with the data subsequently recorded by Istat.

2 The adopted methodology

The mortality tables for contemporaries are generally used in Demography for the study of the population: it would not be possible, here and for our purpose, to follow a generation of individuals until the last one dies. Mainly, the mortality tables tool is used to determine some important biometric functions of the population of age x such as: the probability of death q_x, the survivors l_x, the deaths d_x, the years lived L_x by the entire contingent, the prospective probability of survival p_x, up to life expectancy $\dot{e}_x$. The mortality tables are therefore a very useful tool in the comparative spatial or temporal analysis of the phenomenon in question.

We have formed groupings of age classes $[x, x+s[$ five years, therefore s=5 years, while for the last class $[\text{int}(\omega_t/s)\cdot s, \omega_t]$ the amplitude is variable and equal to $\omega_t - \text{int}(\omega_t/s)\cdot s$, where ω_t is the highest age at death recorded in year t. This methodology is extremely important as it allows you to close the mortality table at age ω_t and determine the complete values of the last row of each table. It differs from that used by Istat which, based on a theoretical estimate built using a Kannisto's model, closes all the tables at 125 years.

For the calculation of the mortality ratios by age group, as the denominator we took into account the number P of the population of each specific class at mid-year t, so that the quotient of the age group $[x, x+s[$ was thus defined:

$$Q_{M\,t}^{[x,x+s[} = \frac{d_{[x,x+s[\,t}}{\frac{1}{2}\left(P_{1.1.t}^{[x,x+s[} + P_{31.12.t}^{[x,x+s[}\right)} \qquad [1]$$

with s=5 except for the last class $[\text{int}(\omega_t/s)\cdot s, \omega_t]$ of amplitude $\omega_t - \text{int}(\omega_t/s)\cdot s$. It should also be noted that the life expectancy of the class [0.5 [is strongly

characterized by the probability of death within the first year of life, generally higher than in subsequent years.

Once the mortality quotients have been obtained, in all age groups and for all t years (from 2010 to 2020), the probability of death can be determined, with a procedure similar to that of the mortality quotients: it was therefore considered to use different calculation methods for the age group [0, 5[, compared to the following ones: the probability of death within the 5th year, therefore:

$$q_{[0,5[\,t} = \frac{2 \cdot 4Q_{M\,t}^{[0,5[}}{2 + 4Q_{M\,t}^{[0,5[}} \qquad [2]$$

while, using the method of Merrel and Reed[1], for the other classes we have:

$$q_{[x,x+5[\,t} = 1 - e^{-Q_{M\,t}^{[x,x+5[} \cdot \left(5 + Q_{M\,t}^{[x,x+5[}\right)} \qquad [3]$$

and in particular for the last age group:

$$q_{[\mathrm{int}(\omega_t/s)s,\omega_t]\,t} = 1 - e^{-Q_{M\,t}^{[\mathrm{int}(\omega_t/s)s,\omega_t]} \cdot \left\{[\omega_t - \mathrm{int}(\omega_t/s)s] + Q_{M\,t}^{[\mathrm{int}(\omega_t/s)s,\omega_t]}\right\}}. \qquad [4]$$

With regard to the probability of death in old age, such as classes from [90, 95[onwards, we therefore use the calculated probabilities directly, unlike Istat which uses theoretical probabilities estimated on the basis of the Kannisto model: with this methodology used by Istat, the tables are closed assuming that the maximum age of survival is 125 years, the same for males and females. In our study, however, we use the maximum age ω_t of deaths by year and by sex: this allows us to know the actual life expectancy for all classes, even the last [int(ω_t/s)·s, ω_t].

Having obtained the various probabilities of death in the age groups, the following biometric functions are determined.

Survivors are linked to the odds of death by the following relationship:

$$l_{[x,x+s[\,t} = l_{[\tilde{x},\tilde{x}+s[\,t}\left(1 - q_{[\tilde{x},\tilde{x}+s[\,t}\right) \qquad [5]$$

where $[\tilde{x},\tilde{x}+s[$ it indicates the previous class, in the case in question with amplitude s=5 except for the final class [int(ω_t/s)·s, ω_t] of ω_t–int(ω_t/s)·s for which we calculate: $l_{[\mathrm{int}(\omega_t/s)s,\omega_t[\,t}$.

Still within the aforementioned classes (and their respective sizes), deaths can therefore be recalculated:

$$d_{[x,x+s[\,t} = l_{[\tilde{x},\tilde{x}+s[\,t} - l_{[x,x+s[\,t} \qquad [6]$$

and similarly, $d_{[\mathrm{int}(\omega_t/s)s,\omega_t[\,t}$, as well as the years lived

$$L_{[x,x+s[\,t} = \frac{s}{2}\left(l_{[\tilde{x},\tilde{x}+s[\,t} + l_{[x,x+s[\,t}\right) \qquad [7]$$

and in particular $L_{[\mathrm{int}(\omega_t/s)s,\omega_t[\,t}$.

[1] The estimate of q_x proposed by Merrel and Reed is considered very efficient and is widely used in the international context, also used by Istat since 1992, so this use makes the results of the tables of the city of Taranto even more compatible and comparable with all the Istat mortality tables.

Finally, life expectancy for the various classes was calculated in general as

$$\dot{e}_{[x,x+s[\,t} = \left(\sum_{i=1}^{\text{int}(\omega_t/s)-1} L_{[is,(i+1)s[\,t} + L_{[\text{int}(\omega_t/s)s,\omega_t]\,t} \right) \Big/ l_{[x,x+s[\,t} = T_{[x,x+s[\,t} \Big/ l_{[x,x+s[\,t} \qquad [8]$$

closing with

$$\dot{e}_{[\text{int}(\omega_t/s)s,\omega_t]\,t} = T_{[\text{int}(\omega_t/s)s,\omega_t]\,t} \Big/ l_{[\text{int}(\omega_t/s)s,\omega_t[\,t} \qquad [9]$$

also written in terms of retro-cumulative functions T of the years lived.

3 Result analysis

The scientific world and institutions have paid a great deal of attention to health conditions and to the analysis of the morbidity of diseases related to pollution and its related mortality in the Ionian area and, in particular, in the municipality of Taranto [1,3,5]. With regard to this local context, even sub-municipal, the mortality tables we have calculated show.

The mortality tables built on sub-municipal areas such as constituencies allow us to determine indicators, called biometric variables, very representative of the quality of life, health and mortality structure of citizens, and life expectancy is one of the most important.

Life expectancy at birth $e_{[0,\,1[}$ in the analysis by five-year classes, with the strong empirical concentration of mortality between $[0,\,1[$ compared to $[1,\,5[$ is well represented by the proxy $e_{[0,\,5[}$ which we indicate with $e°$, that the official statistics of the national institutes normally aggregate the mortality tables and biometric variables in supra-municipal areas such as the provinces, losing important elements of detail (Tab.1).

The detailed analysis carried out in the city of Taranto, which is structured in six administrative districts, presents differences (Tab.2): spatial, gender and historical. For example, the general range goes from the absolute minimum of 74.7 years for males in 2015 and the maximum of 86.5 for females in 2018, and the variability shows a strong gender difference, which sees the Standard deviation of males always lower than females.

The general trends confirm a growth in the 2010-2020 period (Fig.1), which at the gender level shows a further difference, with a regression line coefficient of women ($e°$=78,1739+0,000139247*year) approximately 1/3 lower than that of men ($e°$=61,1669+0,000437763*year).

Unique, in contrast, the trend of life expectancy and $e°$ of the district of Paolo VI, which with an even negative inclination for men ($e°$=82,814-0,000105353*year), raises serious reflections on the quality of life and health in the district, which historically has grown demographically, as a residential area for workers of the steel plant, which in the eighties had over 30,000 workers.

4 Final remarks

In conclusion, the analysis of life expectancy and biometric variables in sub-municipal areas can be very useful in describing the territorial inequalities of citizens' health: coming from various factors, in the City of Taranto it appears to be very influenced by the environmental factor and by pollution since, precisely in Paolo VI, Tamburi and Borgo, very close to the Ionian industrial area and the huge mineral deposit of the metallurgical industry, the lowest life expectancy results, and in Paolo VI even a negative trend is detected.

Table 1: *e° in Districts by gender*

Females	***2010***	***2011***	***2012***	***2013***	***2014***	***2015***	***2016***	***2017***	***2018***	***2019***	***2020***
Borgo	83.27	83.16	82.58	83.86	83.67	84.48	82.73	83.33	83.77	83.80	84.09
Montegran.-Salinella	83.55	85.68	84.90	85.78	84.73	85.60	85.03	86.00	84.26	84.54	85.04
Paolo VI	81.36	82.44	83.75	83.80	82.74	80.73	81.81	83.07	82.31	83.99	82.93
Talsano-Lama-S.Vito	84.62	83.43	85.11	85.32	83.62	85.05	85.63	84.68	86.47	84.94	84.71
Tamburi	83.27	82.28	83.91	83.37	83.28	83.80	84.84	81.26	84.18	84.12	83.36
Tre Carrare-Solito	84.42	84.26	84.98	85.09	85.04	86.14	83.69	84.28	83.33	85.27	85.00
Males											
Borgo	77.59	76.79	76.83	78.87	78.26	74.73	77.18	78.54	79.24	79.97	78.75
Montegran.-Salinella	80.88	79.84	80.06	80.47	80.66	81.73	81.42	81.70	81.36	82.60	81.41
Paolo VI	77.57	79.05	77.48	79.48	78.50	79.84	78.02	79.27	78.60	74.86	79.58
Talsano-Lama-S.Vito	81.04	79.44	80.57	80.97	81.46	81.22	81.05	80.72	80.89	83.33	82.63
Tamburi	75.20	76.71	77.90	78.75	77.69	77.18	78.36	79.60	77.24	77.25	77.39
Tre Carrare-Solito	81.29	79.85	79.69	80.65	80.20	81.35	80.04	82.07	83.30	82.04	82.41

Table 2: *Descriptive statistics*

Females	***Borgo***	***Montegranaro-Salinella***	***Paolo VI***	***Talsano-Lama-S. Vito***	***Tamburi***	***Tre Carrare-Solito***
Mean	83.522	85.010	82.630	84.871	83.425	84.681
Median	83.673	85.031	82.738	84.938	83.371	84.979
St. dev.	0.571	0.734	1.040	0.849	0.975	0.785
Kurtosis	-0.412	-0.002	-0.548	0.685	1.593	0.215
Skewness	-0.179	-0.540	-0.417	-0.059	-1.035	-0.028
Range	1.902	2.456	3.264	3.049	3.585	2.812
Minimum	82.576	83.546	80.729	83.426	81.260	83.330
Maximum	84.478	86.002	83.993	86.475	84.845	86.142
Males						
Mean	77.887	81.102	78.389	81.212	77.570	81.171
Median	78.256	81.364	78.604	81.040	77.389	81.292
St. dev.	1.458	0.810	1.416	1.027	1.143	1.189
Kurtosis	0.924	-0.236	3.314	1.467	1.397	-0.976
Skewness	-0.821	0.091	-1.637	0.723	-0.289	0.355
Range	5.230	2.763	4.975	3.887	4.407	3.616
Minimum	74.735	79.836	74.865	79.444	75.197	79.689
Maximum	79.965	82.599	79.839	83.331	79.604	83.305

Figure 1: City of Taranto - Life expectancy at birth e°

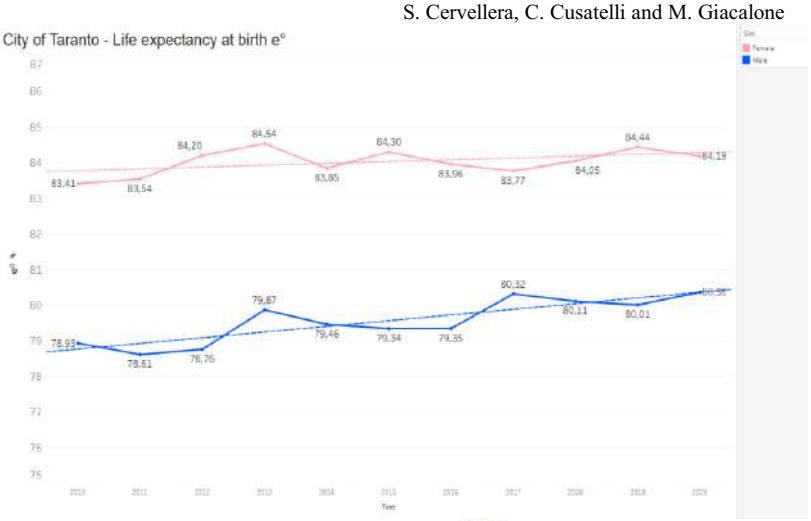
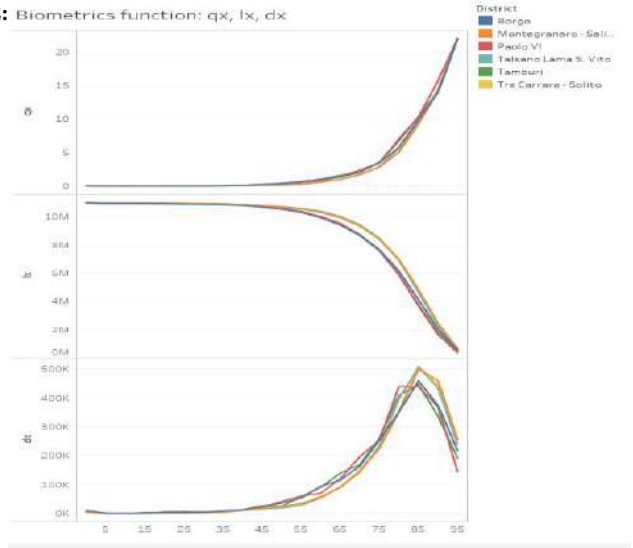


Figure 2: Biometrics function: qx, lx, dx



References

1. G.C. Blangiardo, S. Rimoldi, Vivere (e morire) a Taranto, Statistica & Società, SIS, n.3, 2013
2. S. Cervellera, C. Cusatelli, U. Salinas, Costruzione ed analisi delle tavole di mortalità a Taranto dal 2003 al 2013 e comparazione tra fonti di dati, in: Collana del Dipartimento Jonico, Università degli Studi di Bari "Aldo Moro", vol. 12, Cacucci, Bari, 2014, ISBN: 978-88-6611-387-4
3. P. Comba et al., Environment and Health in Taranto, Southern Italy: epidemiological studies and public health recommendations, Epidemiologia e Prevenzione, n.6, 2012
4. Istat, Tavole di mortalità della popolazione italiana Regioni, province e grandi comuni, Roma, 2013
5. P. Michelozzi, Disastro ambientale a Taranto: il ruolo dell'epidemiologia, Epidemiologia e Prevenzione, n.5, 2012

Family size and Human Capital in Italy: a micro-territorial analysis

Numerosità della famiglia e Capitale Umano in Italia: un'analisi a livello micro-territoriale

Gabriele Ruiu, Marco Breschi, Alessio Fornasin

Abstract: This work analysed the relationship between human capital (measured through formal education) and average family size at the sub-municipal level using data from the 2011 Census of the Italian Population. The results of the estimated error spatial model indicate that, controlling for a vast set of socio-economic characteristics of the population, education is negatively related to average family size.

Abstract *Il lavoro analizza la relazione tra dimensione media della famiglia e capitale umano (misurato attraverso l'istruzione formale) usando dati sulle località italiane provenienti dal Censimento della Popolazione del 2011. I risultati dell'analisi spaziale condotta indicano che, controllando per un ampio set di carattestiche socio-economiche delle località, l'istruzione è negativamente correlata alla dimensione media familiare.*

Key words: family size, education, human capital, spatial regression model

Introduction

The formulation of a microeconomic theory about fertility is due to the seminal contribution of Gary Becker. The most recent formulation of the Beckerian theory (Becker et al. 1990) assumes that families must choose between investing their economic resources on the number of children or their quality (for example, their

[1] Gabriele Ruiu, University of Sassari, email: gruiu@uniss.it

Marco Breschi, University of Sassari, email: breschi@uniss.it

Alessio Fornasin, University of Udine, email: alessio.fornasin@uniud.it

level of human capital), the so-called "quantity-quality trade-off". When human capital is abundant in society, investing in it produces increasing returns, largely surpassing those offered by the strategy based simply on the number of children, while the opposite is true for economies where the level of human capital is low. The presence of a qualified workforce can also induce greater efforts in R&D also by the firms as they know that an increasingly qualified workforce will be able to use increasingly advanced technologies. The investments in technology, in turn, strengthen those in human capital, thus triggering a virtuous circle that pushes individuals to invest more and more in the education of their children. This mechanism leads to the existence of two possible equilibria. The first is the Malthusian one characterised by families with many members and a low level of human capital, the second represented by small families and increasing accumulation of human capital. The prediction of the Beckerian theory is thus clear: a negative relationship exists between the stock of human capital and family size.

This work aims to analyse the relationship between human capital (measured through formal education) and average family size at the sub-municipal level using data from the 2011 Census of the Italian Population.

Data and Method

Our analysis is focused on Italian localities. A locality is a sub-municipal area defined by Istat as: "a more or less vast area of territory, usually known by a proper name, on which one or more grouped or scattered houses are located". They can be inhabited, and in this case, they are divided into three types: *centro abitato* (if public services are present and the locality is to a certain degree autonomous from the rest of the municipality), *nucleo abitato* (houses are contiguous but public services are not present) and *case sparse* (scattered and isolated houses in the territory of a municipality, not belonging to previous types of localities).[2] In the 2011 Population Census the localities were 68,537 (of which 65,045 with at least one resident), divided into 21,657 *centro abitato* (the 91% of the Italian population lived in these localities in 2011), 35,644 *nucleo abitato* (representing the 3% of the Italian population), 7,754 *case sparse* (representing the 6% of the Italian population). 155 different characteristics were reported for each locality, as: the structure by gender and age of the population, the number of employed individuals, the number of residents that are born in other countries, etc. Among the information reported, the most relevant for this work is that the population could be broken down by the following educational levels: university degree, upper secondary school diploma, lower middle school and elementary school certificate. In addition, it is possible to

[2] Note that a *centro abitato* can include several districts (*quartieri*) of a city. For example, all the districts included inside the circumference designed by the Grande Raccordo Anulare (GRA) of Rome constitute a *centro abitato*. Lido di Ostia is a locality outside GRA but has autonomous basic public services with respect to the rest of the city is another *centro abitato* of Rome. At the same time Borgo Lotti is an agglomeration of houses (about 1,000 inhabitants) located outside GRA but it is considered a *nucleo abitato* because of the lack of public services.

calculate the average family size in each locality by using the information about the total number of families and the total number of family components.

Using this rich data source, the current work provides an exploratory analysis at the locality level, in which an indicator of tertiary education and other socio-demographic characteristics of the territory are related to average family size.[3] To the best of our knowledge, this work is the first to propose an analysis at a such fine territorial level. In order to calculate the above mentioned indicator, the expected number of tertiary-educated individuals was obtained for each locality i, applying the following formula:

$$(1)\quad E(L_i) = \sum_{h}^{H} pop_{hi} * f_{hItaly}$$

Where h represents a five-year age group (15-19, 20-24, ..., H = 75 and more), pop indicates the amount of the population in the age class h for the locality i, while f_{hItaly} is the ratio of graduates (bringing together males and females) and population in class h for Italy as a whole. $E(L_i)$ is then rounded to the nearest integer (or to 1 when $E(L_i) < 1$). Subsequently, the IL_i indicator is obtained by dividing the observed number of graduates (L_i) by the expected number. Therefore, the interpretation of the indicator is straightforward, if in a locality i, the IL is greater/lower than 1 it indicates that in i the number of tertiary graduates is higher/lower than expected. This indicator was preferred to a simple ratio of graduated people to resident population, because the latter is influenced by population age structure (university level education is less frequent among the elderly). Furthermore, the IL index may capture those communities where the virtuous circle described by Becker and colleagues has led to a level of formal education that goes beyond the expected.[4]

We have then estimated the following spatial error model:

$$(2)\quad Av\_fam\_size_i = \beta_1 IL_i + x_i\beta' + \varepsilon_i + \rho w_{ij}\xi_i$$

Where Av_fam_size stands for average family size. x_i are a series of regressors such as the population size, the type of locality (*centro abitato*, *nucleo abitato*, *case sparse*) the geographic macro-area, the incidence of empty houses on the total number of houses, the ratio between the houses classified as in excellent conditions

[3] It may be argued that family size is not coincident with fertility, and therefore, we cannot investigate the quantity-quality trade-off. For instance, we have that a family of four members may be either a family with two parents and two children or a couple living with the parents of one of them. Family average size represents thus a case of a dependent variable measured with error. However, when the measurement error is not related to explanatory variables, regression coefficients are still unbiased and consistent even though the power of statistical tests is reduced (Wooldridge 2002); thus we tend to not reject the null too often. In our specific case, we may have that different family structures could derive from different cultural norms among Italian macro-areas. However, we have that among the family with only one nucleus (co-residing people linked together by a couple's relationship or by a parent-child relationship if the child is single), the percentage of families that host members that not belong to the nucleus in 2011 was 3.4, 4.3, 5.4, 4.8, 3.4 respectively in North Western Italy (NW), North Eastern ltaly NE), Central Italy (C), Southern Italy (S), Islands (I). At the same time, the percentage of families with two or more nuclei was 0.9, 1.4, 1.9, 1.7, 1.2, respectively in NW, NE, C, S, I. Hence, it seems that we have not huge differences among macro-areas.

[4] An alternative could be to calculate a graduation rate using direct standardisation to account for the age structure of each locality. Unfortunately, data about the number of graduates for each age class are not available at the locality level.

and the total number of houses, the percentage of the foreign population out of the total resident population, the incidence of divorced persons out of the total resident population, the employment to population ratio.[5,6] The overall error is composed by two components, namely ε, a spatially uncorrelated error term that satisfied the normal regression assumption, and ξ, which is a term indicating the spatial component of error term. The parameter ρ indicates the extent to which the spatial component of the errors are correlated with one another for nearby observations, as given by $w_{ij.}$ The latter are elements of a contiguity matrix W, created adopting queen contiguity criteria (we use the minmax normalisation of the weights, see Drukker et al. 2013 for details). We believe that unobserved local policies or shared cultural values are possible drivers of spatial dependency between neighbouring localities. For this reason, we preferred to run a spatial error regression model rather than an auto-regressive spatial model where instead, it is assumed that the value of the dependent variable for the unit i is directly influenced by the value assumed by the same variable in the nearby unit j.

The authors are aware of the existence of a scientific debate on the actual necessity to conduct inference when working on census data. Jones et al. (2015) recently advanced that, when the analyses is focused at a particularly fine territorial detail (such as, for example, the localities), the inference is necessary in order to take into account the uncertainty due to the low number of observations within each spatial unit. Following Jones et al., we will report both the estimated coefficients, their standard errors and the associated level of statistical significance.

Finally, it must be considered that the number of child and child quality are decided simultaneously in Beckerian model (i.e. we observe a low number of children because parents chose to invest in education and at the same time we observe high investment in education because parents decided to have fewer children). Therefore the recursive relationship between education and fertility implies that the coefficient associated with education in an equation explaining fertility is downward biased (Becker et al. 2010). This means that the estimated β_1 in equation 2 could be more negative than it actually is.

Results and discussion

Figure 1 shows the spatial distribution of the IL_i indicator by means of a thematic map. The colours are attributed by dividing the distribution into five ordered categories containing each the 20% of observations. The indicator tends to assume high values around big cities as Milan, Rome, Turin, Florence and Naples, while it is

[5] The incidence of divorced individuals out of the resident population is included in the model to proxy the level of secularization in each community. See De La Croix et al. (2020).

[6] It could be argued that an older population are also characterised by a lower family size, since the offspring tends to be not present in the household. Hence, the use of the percentage of graduated individuals over total resident population in equation (2) may lead to a spurious correlation with our dependent variable driven by the age structure.

particularly low in the mountain area of Alto Adige (North-Eastern Italy), in South-Eastern Sardinia, in the Northern part of Calabria and Basilicata.

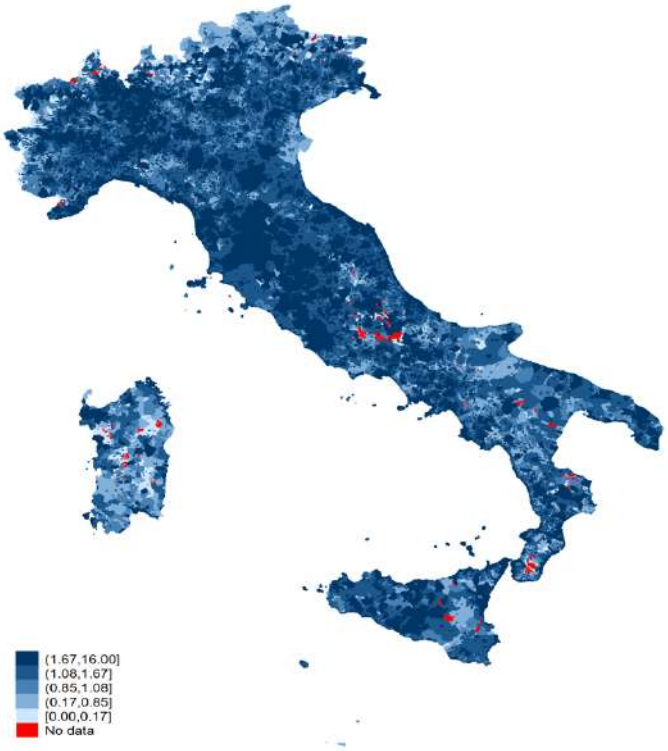


Figure 1: Spatial distribution of the IL_i indicator, 2011 Italian localities.

Table 1 reports the results of the estimation of equation 2. In column 2, we replicate the analysis reported in column 1 but using OLS technique with clustered standard errors at the municipal level. The ρ parameter was not reported to save space however it was positive and strongly significant from a statistical point of view. At the end of the table we reported the estimated Average Total Impact (ATI) associated with the IL index, that is, the sum of the direct effect and of the spill-over effects generated by a one point increase in the indicator.
As expected the relationship between the IL index and family size is negative and strongly significant but not very strong in term of magnitude. For a one point increase in this indicator we have a decrease in the average family size of 0.0084 units. However, note that IL assumes high value when the locality has a number of graduates that is higher than what we should expect. In order words, we are capturing the trade-off between being better than expected in terms of formal education and family size.

Among other controls, it is interesting to note that secularisation as captured by the ratio between divorced individuals and resident population reasonably exerts a negative effect on family size. Note also that, although localities classified as *case sparse* do not offer public services, as schools or kindergartens, larger families tend

to be more concentrated in this type of locality. This may indicate that it still exists a different family structure between urban and rural context.

Table 1: Family size and education, results from a spatial error model.

	(1) Spatial Error Model	(2) OLS
Centri abitati	-0.083*** (0.004)	-0.085*** (0.005)
Case sparse	0.067*** (0.007)	0.073*** (0.006)
Nuclei abitati	REF	REF
IL	-0.008*** (0.002)	-0.011*** (0.003)
Divorced/pop	-2.346*** (0.046)	-2.396*** (0.098)
More than 100,000 inhab.	-0.066 (0.088)	-0.025 (0.039)
From 10,000 to 100,000 inhab.	0.034*** (0.011)	0.052*** (0.006)
From 1,000 to 10,000 inhab.	0.020*** (0.007)	0.033*** (0.004)
Less than 1,000 inhab.	REF	REF
Province capital main locality	-0.162*** (0.052)	-0.178*** (0.019)
North-East	0.118*** (0.005)	0.117*** (0.009)
Center	0.113*** (0.006)	0.113*** (0.009)
South	0.258*** (0.006)	0.259*** (0.010)
Islands	0.138*** (0.009)	0.136*** (0.014)
North-West	REF	REF
Employed/Pop	0.340*** (0.016)	0.353*** (0.033)
Foreign Born/Pop	0.297*** (0.020)	0.299*** (0.041)
Inc. Empty Houses	-0.738*** (0.008)	-0.755*** (0.014)
Excellent Houses	0.123*** (0.007)	0.123*** (0.011)
ATI- for IL	- 0.0084	
N	63331	63331

Standard errors reported in parentheses. Note that N is not equal to 65,045 because of missing values on explanatory variables; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

References

Becker, G.S., Murphy, K. M., Tamura, R.: Human Capital, Fertility and Economic Growth. The Journal of Political Economy, 98 (5), S12–37 (1990).

Becker, S.O., Cinnirella, F., Woessmann, L.: The trade-off between fertility and education: evidence from before the demographic transition. Journal of Economic Growth, 15, 177-204 (2010).

De la Croix, D., Mariani, F., Mercier, M.: Driven by institutions, shaped by culture: Human capital and the secularisation of marriage in Italy. CEPR Discussion Paper No. DP14696 (2020).

Drukker, D.M., Prucha, I.R., Raciborski, R.: A command for estimating spatial-autoregressive models with spatial-autoregressive disturbances and additional endogenous variables. The Stata Journal, 13(2), 287-301, (2013).

Jones K., Johnston R., Manley D., Owen D., Charlton C.: Ethnic residential segregation: A multilevel, multigroup, multiscale approach exemplified by London in 2011. Demography, 52, 1995-2019, (2015).

Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge (2002).

Estimate age-specific fertility rates from summary demographic measures. An Indirect Model Levering on Deep Neural Network.

Stima dei tassi di fecondità specifici per età da misure demografiche di sintesi. Un modello di stima indiretta basato su reti neurali profonde

Andrea Nigri

Abstract The aim of this study is to develop an "indirect" methodology, formulating a model leveraging on deep learning algorithms based on neural networks to derive age-specific fertility profiles from observed or predicted mean age at childbearing.
Abstract Lo scopo di questo studio è sviluppare una metodologia "indiretta", formulando un modello che fa leva su algoritmi di deep learning per derivare i profili di fecondità specifici per età, utilizzando come input l'età media al parto, osservata o prevista

Key words: Fertility, Vital Rates, Deep Neural Network.

1 Introduction

Reliable predictions of age-specific vital rates are crucial in demographic studies, several drawbacks are common, however. Indeed the lack of reliable data or stochastic variation in population counts at a high disaggregation level, made summary demographic measures appealing compared to age specifics one, to model and predict using multiple approaches. In the study of fertility dynamics, as in mortality forecasting, a key advantage of modeling summary measures like mean age at childbearing ($MAB(t)$) is that the predictive model deals only with a single indicator that summarizes the overall level of a demographic index over time, instead of modeling a single time series of rates for each age simultaneously (e.g. Lee-Carter [3],[2]). Although the use of summary measures as an indicator to forecast is appealing, estimating age-specific vital rates is needed to analyze demographic patterns at different ages. This reconstruction is not straightforward.

Demography has a long-standing tradition of developing formal demographic methods and using statistical approaches to indirectly estimating indicators([4],[5]).

Andrea Nigri
Univeristy of Foggia.e-mail: andrea.nigri@unifg.it

According to the UN manual [7], the term "indirect" qualifies the demographic estimation technique that origins in the fact that such technique produces estimates of certain parameters on the basis of information that is only indirectly related to its value.

The aim of this study is to formulate a model leveraging on deep learning algorithms based on neural networks to derive age-specific fertility rates ($fr(a,t)$) from the observed or predicted level of mean age at childbearing. Therefore, the fertility pattern by age and time is indirectly identified by the network. Resulting estimates would be useful for guiding public health interventions, informing about age-specific fertility dynamics in contexts with deficient data collection.

2 Method

In this section, I will describe the DNN model used to forecast the country-specific fertility rates by age and time. To be applied, my model requires that data be arranged in a $age \times year$ matrix Y and vector X. The X vector is composed of input data, that in the present study is the time series of MAB. The Y matrix contains the target output i.e. the fertility rates by age and year. The model learns the hidden pattern in the input data in a given year and gives back the output in the same year. The procedure will be illustrated in Fig. 1.

Aiming at creating a bridge between Deep Neural Network (DNN) and demography, I will describe the steps to obtain the target output. eq. 1 describes the specific NN structure providing the fertility surface $fr(a,t)$ with $a \in \{15,16,...,50\}$ vector of ages and $t \in \{t_1,t_2,...,t_n\}$ vector of years:

$$fr(a,t) = f^{(k)}\left(\begin{bmatrix} w_{1,1}^{(k)} & w_{1,2}^{(k)} & w_{1,3}^{(k)} & \cdots & w_{1,n}^{(k)} \\ w_{2,1}^{(k)} & w_{2,2}^{(k)} & w_{2,3}^{(k)} & \cdots & w_{2,n}^{(k)} \\ w_{3,1}^{(k)} & w_{3,2}^{(k)} & w_{3,3}^{(k)} & \cdots & w_{3,n}^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n,1}^{(k)} & w_{n,2}^{(k)} & w_{n,3}^{(k)} & \cdots & w_{n,n}^{(k)} \end{bmatrix} \begin{bmatrix} H_1^{(k-1)} \\ H_2^{(k-1)} \\ H_3^{(k-1)} \\ \vdots \\ H_n^{(k-1)} \end{bmatrix} + \begin{bmatrix} b_1^k \\ b_2^k \\ b_3^k \\ \vdots \\ b_n^k \end{bmatrix}\right) \tag{1}$$

where, for a generic layer k, $f^{(k)}$ is the activation function, $W^{(k)}$ the weights matrix, $H^{(k)}$ the hidden layers, and $b^{(k)}$ the bias, used to control the triggering value of the activation function. A graphical representation of the DNN model related to eq. 1 is given in Fig. 1.

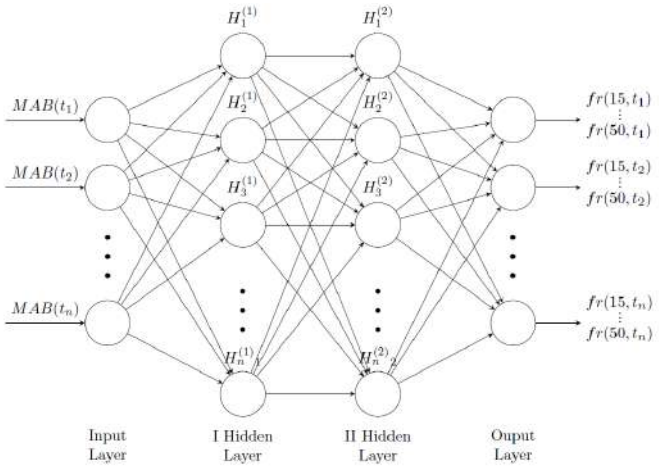


Fig. 1: Graphical representation of the DNN model. Circles represent neurons, and lines synapses. Synapses take the input and multiply it by a weight (the "strength" of the input in determining the output). Neurons add the outputs from all synapses and apply an activation function.

Let $\{MAB(t)\}_{t=t_0}^{t_s}$, for $t_0 < t_s$, be the country-specific observed time series of MAB. Then, each series is split into a train-validation set and a test set, where the first one is used for fitting the model's parameters, while the second one to test the model's prediction and calculate the error. Specifically, the time frames 1965-1995 and 1975-2005 are used as train-validation chunk respectively for the first, second, and third time window, according to common splitting rules 80%-20%. The best hyper-parameter combination obtained in the training phase is used to obtain predictions in the test phase which takes place on the time frames 1996-2005, and 2006-2015 depending on the selected time window.

Hence, let t_τ, with $t_0 < t_\tau < t_s$, be the calendar year corresponding to the last realization in the training-validation set. The values of $MAB(t)$ over the period (t_0, t_τ), $\{e_{0,t}\}_{t=t_0}^{t_\tau}$, represent the input for train-validation, while the corresponding output is $\left\{\hat{fr}(a,t)\right\}_{t=t_0}^{t_\tau}$. The values of MAB over a subsequent period, $\{MAB(t)\}_{t=t_\tau+1}^{t_s}$, represent the input for test, while the corresponding output is $\left\{\hat{fr}(a,t)\right\}_{t=t_\tau+1}^{t_s}$. Thereby, denoting ψ_{nn} as a composition of functions defined on the basis of the NN architecture, the model can be described by:

$$\left\{\hat{fr}(a,t)\right\}_{t=t_\tau+1}^{t_s} = \psi_{nn}\left(\{MAB(t)\}_{t=t_\tau+1}^{t_s} \middle| \hat{W}\right) \tag{2}$$

where $\left\{\hat{fr}(a,t)\right\}_{t=t_\tau+1}^{t_s}$ is the matrix of fertility rates in the test set obtained by ψ_{nn}, that involves the NN weights $\hat{W}$ estimated during the network training. The resulting DNN estimate is a point estimation, not providing any information on the uncertainty given by $\hat{W}$. This is one of the main limitations of deep learning algorithms, where the estimation of prediction intervals is still considered a big challenge.

Nevertheless, the proposed model leverages the use of future target MAB, which can be derived from an extrapolative model (e.g. Lee Carter), official statistics projections, or time-series forecasting.

3 Results

I consider historical fertility data collected by the [1] for Italy, Japan, and USA. Aiming to assess the model robustness and consistency toward the historical data, I carry out an out-of-sample test by considering two time windows of information: 1965-2005, 1975-2015. Each period is split into the train-validation and test set. I use 30 years for training-validation for each time window (respectively, 1965-1995, and 1975-2005). The remaining years are then used for model forecasting / test (respectively, 1996-2005, and 2006-2015), I smooth the DNN estimation of fertility rates by age using P-splines. The RMSE and MAE improvements after the smoothing are on average 1.55% and 1.57%, respectively. This step can be skipped for larger time windows or if the graduation is not of interest.

I now validate the DNN model forecasting performance. Using illustrative applications I study the ability to generate reliable forecasts of fertility rates for ages and periods. These illustrative examples are dedicated to investigating whether the approaches can capture (a) regular and irregular trends over time (b) dynamics of age-specific fertility improvements. In order to show how robust or sensitive my findings are to the reference period, I look at 10 years of fertility forecasts referring to different time windows: 1965–1995, and 1975–2005.

The analysis includes numerical and graphical representation of the goodness of fit. To assess the models' accuracy, I calculate the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) on the validation period, which in the present analysis corresponds to 1996-2005, and 2006-2015 depending on the selected time window.

$$\text{MAE}: \quad \sum^{n} \frac{\mid fr(a,t) - \hat{fr}(a,t) \mid}{n}, \tag{1}$$

$$\text{RMSE}: \quad \sqrt{\frac{\sum^{n} (fr(a,t) - \hat{fr}(a,t))^2}{n}}. \tag{2}$$

Results are provided for three countries (Italy, Japan, and the USA), table 1 shows MAE and RMSE values for estimation periods and each country. Overall, the DNN provides remarkably accuracy.

Table 1: Out-of-sample test: MAE and RMSE for DNN, by country. Training periods: 1965-1996, 1975-2006.

Country	1965-2005		1975-2015	
	RMSE	MAE	RMSE	MAE
Italy	0.003943	0.002869	0.002983	0.002288
Japan	0.006665	0.004515	0.00488	0.003895
USA	0.005466	0.00384	0.004971	0.003631

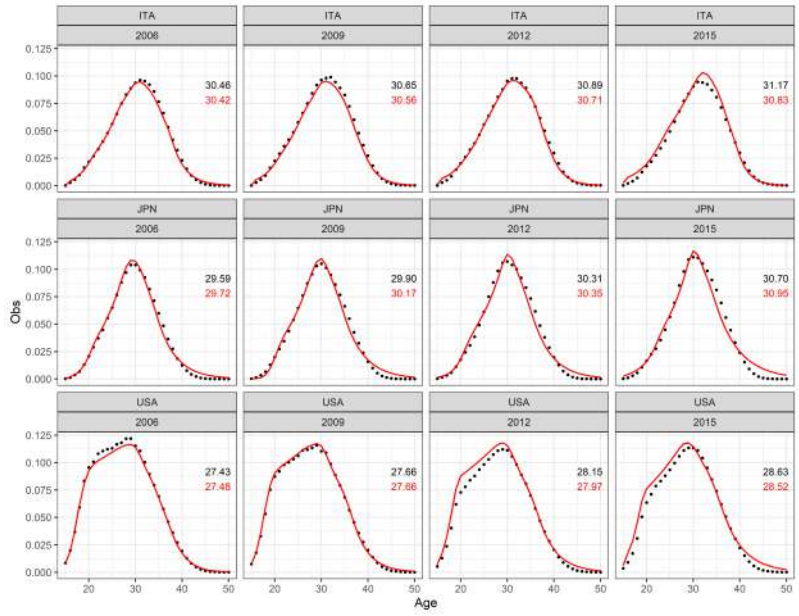


Fig. 1: Estimated age-specific fertility rates $fr(a,t)$, by country for 2006, 2009, 2012, and 2015 based on training period 1975-2006. Black dots are the observed rates. Levels of observed (black) and reconstructed (red) MAB, are reported.

4 Discussion

This manuscript contributes to the current literature on the demographic methods for indirect estimations. I propose a Deep Neural Network framework to indirectly estimate fertility rates from a summary demographic measure, namely, mean age at childbearing. This approach represents an advance among fertility modeling, to be adopted to reconstruct demographic scenarios which are conventionally based on

summary measures. While I apply the methodology to country-specific scenarios, the model could be used to indirectly estimate vital rates for regions or subpopulations with similar fertility profiles. This characteristic makes the proposed model appealing for countries where present information is lacking but past data are available or from surrounding countries or populations. Similarly, this method can be used to derive age-specific fertility for a projected or forecasted value of *MAB*. I acknowledge some drawbacks, however. Despite the small error in the backtesting estimation, the proposed model, in some cases, seems not to be suitable for a coherent reconstruction of *MAB*. This is because even minimal deviations in fertility rates estimation may imply large differences in the number of birth (and thus on a measure of population dynamics). From a methodological perspective, future work will be proposed developing a multi-population extension, relying on a more complex Neural Network architecture.

References

1. Human Fertility Database (2018). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded on 01/12/2020. urlhttps://www.humanfertility.org.
2. Thompson, P. A., W. R. Bell, J. F. Long, and R. B. Miller (1989). Multivariate time series projections of parameterized age-speci
   c fertility rates. Journal of the American Statistical Association 84 (407), 689
3. Lee, R. and Carter, L. (1992). Modeling and forecasting us mortality. Journal of the American Statistical Association, 87:659–671.
4. Pascariu, M. D., Basellini, U., Aburto, J. M., and Canudas-Romo, V. (2017). The linear link: Deriving age-specific death rates from life expectancy.
5. Ševcíková, H., Li, N., Kantorová, V., Gerland, P., and Raftery, A. E. (2016). Age-specific mortality and fertility rates for probabilistic population projections, volume 39. Springer Series on Demographic Methods and Population Analysis.
6. Lee, R. D. (1993). Modeling and forecasting the time series of us fertility: Age distribution, range, and ultimate level. International Journal of Forecasting 9 (2), 187202.
7. United Nations (1955). Age and Sex Patterns of Mortality: Model Life Tables for Under-Developed Countries. Population Studies, No. 22. Department of Social Affairs.

Patterns in the relation between causes of death and gross domestic product

Andamento della relazione tra cause di morte e prodotto interno lordo

Andrea Nigri and Federico Crescenzi

Abstract In this paper, we investigate the relationship between socioeconomic levels and cause-specific mortality. To gain such insight, we offer a novel framework based on a Bayesian hierarchical model for a Dirichlet distribution able to handle competing risks among causes. As a consequence, we are able to investigate the impact of improvements in cause-specific mortality by socioeconomic circumstances that might shed light on untracked economic and demographic dynamics.
Abstract In questo articolo, indaghiamo la relazione tra i livelli socioeconomici e la mortalità causa-specifica. In tal modo, offriamo un nuovo framework basato sul modello gerarchico bayesiano di Dirichlet in grado di gestire i rischi competitivi tra le cause. Siamo quindi in grado di studiare l'impatto dei miglioramenti nella mortalità per specifica per causa in base ai livelli socioeconomici che potrebbero far luce su dinamiche demografiche non ancora tracciate.

Key words: Longevity, Health, CoD.

1 Introduction

Over the last two centuries, the expected number of years that human beings are likely to live has been steadily increasing. Regardless of whether life expectancy improvement has been occurring at a single or at different speeds ([10],[5] ,[6]), its rise has been occurring with a remarkable degree of regularity. However, some authors have found time points at which countries started to diverge from some general trend. Most of these breaks can be explained by the different composition of the causes of death. Indeed, innovations and changes in nearly every branch of

Andrea Nigri
Univeristy of Foggia.e-mail: andrea.nigri@unifg.it

Federico Crescenzi
University of Florence. e-mail: federico.crescenzi@unifi.it

life are responsible for the marked increase in the average lifespan. The factors that contributed to the rise of longevity include better nutrition, improvements in public health, vaccination, and the long-term effects of improvements in early-life conditions. Advances in education, welfare, and infrastructure are other potential determinants of the increase in the average lifespan ([11]). Researchers have also found correlations between life expectancy and a number of other development indicators, such as gross domestic product ([1]; [2]). It is unclear, however, which of these determinants have been and currently are the most important. Socioeconomic groups may be exposed to varying levels of causes of death mortality; this is certainly the case in the USA that are experiencing a recent life expectancy stall. A study of cause-specific mortality may provide rich insight into this phenomenon, therefore, we investigate the relationship between socioeconomic circumstances and cause-specific mortality using a unique dataset obtained from the Human Cause-of-Death Database (HCD). Leveraging on a Bayesian hierarchical model we are able to incorporate socioeconomic circumstances. Furthermore, the framework is able to handle the intrinsic dependence amongst the competing causes. As a consequence, we are able to investigate the impact of improvements in cause-specific mortality by socioeconomic circumstances that might shed light on untracked economic and demographic dynamics.

The remainder of the article is organized as follows. In Section 2 we introduce the dataset and its characteristics regarding the causes of death. In the same section, we properly describe the proposed model. In Section 4, we look at model fit results.

2 Data and Methodology

2.1 Data

The empirical study concerns the USA male mortality for specific years 1999 up to 2013. The cause-of-death data have been taken from the newly developed Human Cause-of-Death Database (HCD)[1], which provides high-quality data on cause-specific mortality. It is coded by using the international classification of diseases (ICD), providing different aggregation levels: full list, intermediate list, and shortlist. Each classification has been developed using the same criteria for all countries, ensuring homogeneity and comparability. Using these data, we obtain a universal and standardized methodology to redistribute deaths between 104 disease categories in five-year age groups. Also, it allows to avoid issues regarding the ICD revisions and ensuring cross-country comparability to different coding practices.
We truncate the cause-of-death analysis at age 80 because of classification quality and the presence of comorbidities [4].
We start from the shortlist and further clustering the death classification. For USA,

[1] The HCD[9] database can be found at: `www.causesofdeath.org`.

we consider the following three major causes of death with the indication of the ICD codes:

- (1) Cancer (C00-D48),
- (2) Circulatory (I00-I52, G45, I60-I69, I70-I99),
- (3) Others (residual class).

We choose circulatory and cancer because they belong to the main causes of death in developed countries. Referring to the study period (1999-2003); we model the proportion of causes of death to levels of Gross Domestic Product (GDP), in log scale, derived from the World Bank ([13]).

2.2 Model

The Gross Domestic Product has been widely used among demographers, and widely recognized as a crucial indicator in order to explain longevity evolution and transitions ([12], [1]). Indeed, GDP is not merely a time-trend index, but rather a "latent factor" incorporating different unobserved latent variables. It implicitly encompasses economic fluctuations, affecting medical innovation and many other variables that directly (or indirectly) influenced the mortality trend. As a result, it exhibits high correlation levels with other social-economic, and health indicators, such as life expectancy at birth, and Hospital beds (per 1,000 people). This poses a relevant restriction referring to the number and type of covariates considered into the model that might jeopardize the reliability of the CoD estimates.
In order to treat competitive risks, we implement a statistical model for proportions where a natural choice is the Dirichlet distribution that allows redistributing the components of a population in categories around a total. Among demographers, several studies address the advantage provided by assuming causes-of-death to follow a Dirichlet distribution (i.e. [7],[8])). The Dirichlet is charaterized by the following density distribution:

$$\mathrm{Dir}(\theta \mid a) = \frac{1}{\mathrm{Beta}(a)} \prod_{i=1}^{K} \theta_i^{a_I - 1}, \text{ where } \mathrm{Beta}(a) = \frac{\prod_{i=1}^{K} \Gamma(a_i)}{\Gamma\left(\sum_{i=1}^{K} a_i\right)}, \text{ and } a = (a_1, \ldots, a_K) \tag{1}$$

In the Bayesian context, an additional advantage is that the Dirichlet distribution is the conjugate for the multinomial distribution, previously used to model causes of death. In this study, a Dirichlet regression is used to provide a coherent estimation of the relationship between mortality rates specific for age and causes, and GDP on logarithmic scale. The model is estimated considering single ages. Further developments may consider to estimate every age class simultaneously, including a random intercept for each age group. The equations below describe the hierarchical structure of the Dirichlet model with parameters a_k. A flat prior is put on the regression coefficients in order to avoid favours to any causes group. Let us suppose to have observed C mutually exclusive causes of death, for X different ages over T calendar

years. The central death rate for cause $c \in C$ is defined as: $m_{ct}(x) = \frac{D_{ct}(x)}{E_t(x)}$ Where $D_{ct}(x)$ is the number of deaths aged x in year t due to cause c, and $E_t(x)$ are the exposures-to-risk aged x in year t. For each age $x \in X$, the aim of our model is to explore the relationship between cause-specific death rates and GDP on a logarithmic scale. Formally, the specification of the model is

$$m_{ct}(x) \sim \text{ Dirichlet } (a_{ij}) \tag{2}$$

$$\log(a_{ij}) = \beta_{0j} + \beta_{1j} log(GDP) \tag{3}$$

$$\beta_{0j}; \beta_{1j} \sim \mathcal{N}(0, \sigma_1) \tag{4}$$

3 Results

We obtained posterior estimates for the proposed model for people aged 40, 60, and 80y. Overall, we found a strong correlation between CoD proportion and log(GDP), with an exception of mortality due to cardiocirculatory disease among age class 40. We pose evidence of homogeneous behavior for Other causes among all ages, which was expected since it embraces the highest share of mortality. Peculiar information has been exhibit by Neoplasms where the negative correlation has been replaced by a positive one among ages 80. After the first decade of the new millennium, the USA show a remarkable life expectancy variation, attributable to the relatively high infant mortality and the high mortality from violence among young adults, as well as a stagnating decline in cardiovascular disease mortality([3]). This might be enough to explain the flat correlation among 40y. Furthermore, this class of diseases is strongly related to social conditions, of which GDP is widely recognized as a good proxy.

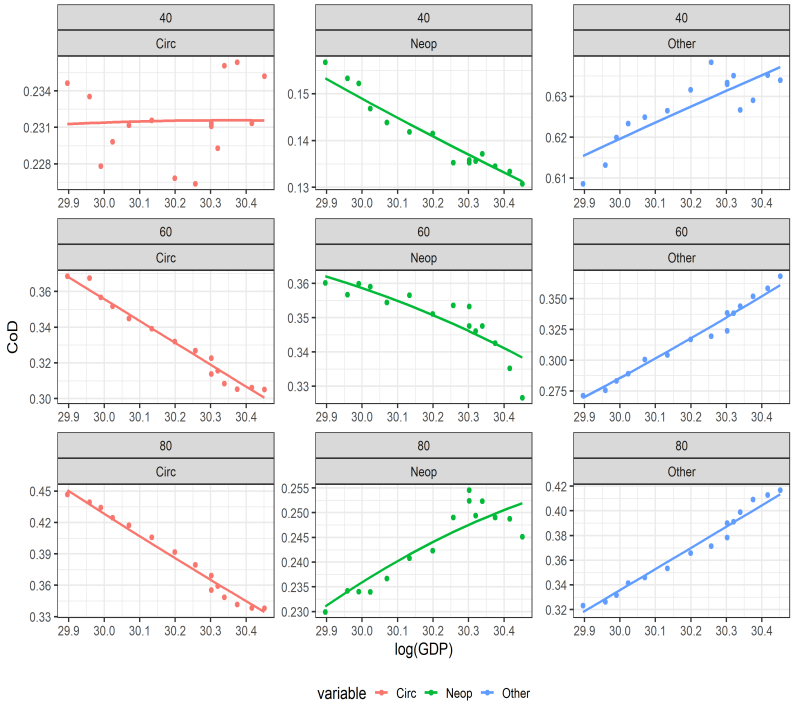


Fig. 1: Relation between CoD proportion and log*GDP*. Panels contain data for specific ages (40,60,80) for years 1999 up to 2013.

4 Conclusions

In this article, we provide a tool to assist public policies in defining their health inequality strategy. Indeed, the estimation of mortality dynamics in lockstep with the evolution of economic indicators can provide fundamental answers to the main questions related to health and socio-economic sustainability, among others. To gain such insight in this paper we have offered a novel framework for estimating the relationship between GDP and causes-specific death rates based on Dirichlet Bayesian hierarchical model. The proposed procedure differs from canonical frameworks, offering a twofold insight. First, the relationship between specific causes of death mortality and economic growth has been described. In doing so a novel framework able to handle competing risks among causes has been proposed. Properly extending the proposed model to multiple countries and social-economic scenarios, researchers would be able to analyze the impact on longevity, by socioeconomic levels, of a hypothetical cause of death mortality evolution. Since different countries are affected