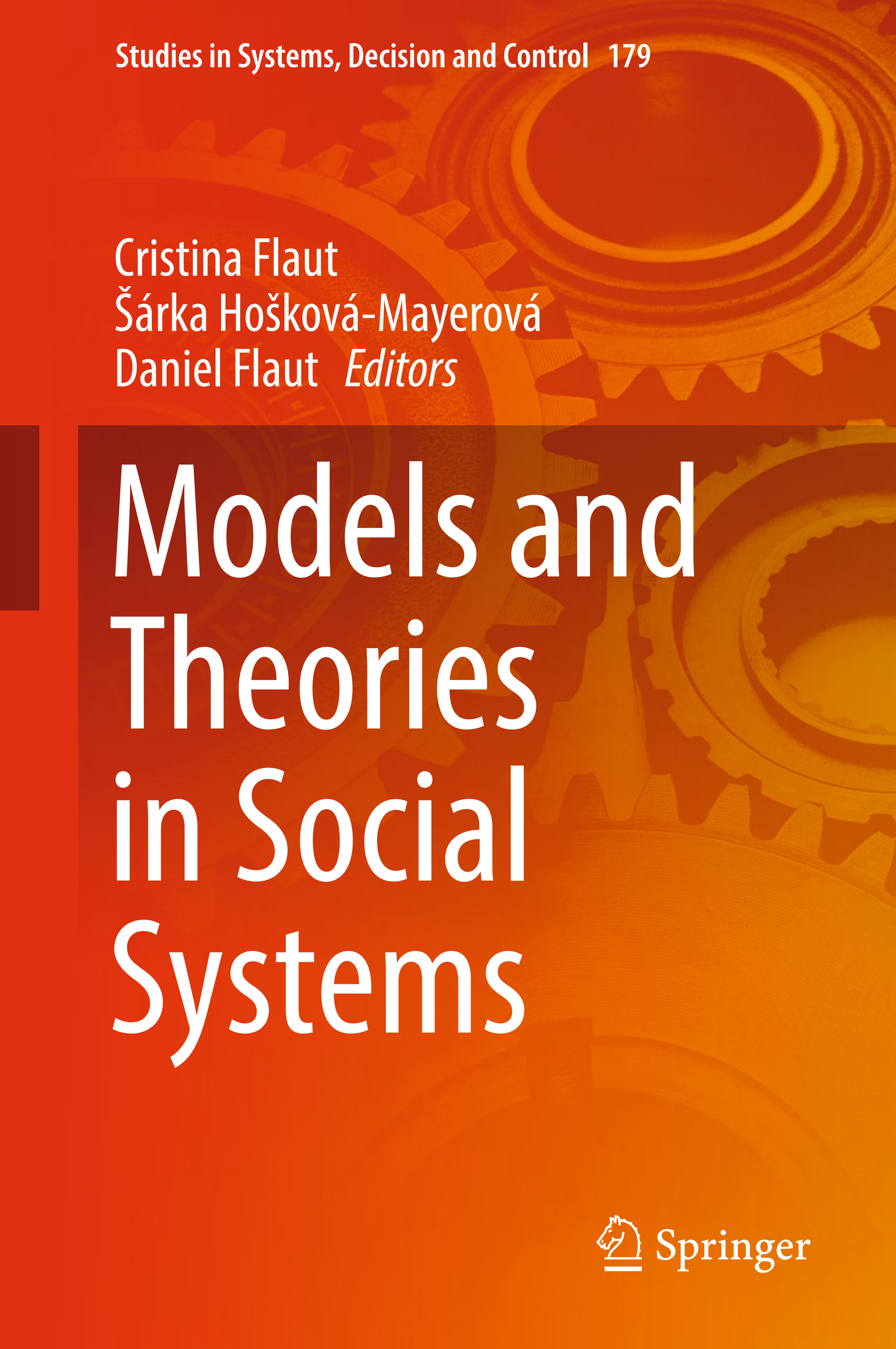
Studies in Systems, Decision and Control 179

Cristina Flaut
Šárka Hošková-Mayerová
Daniel Flaut *Editors*

Models and Theories in Social Systems



Springer

Studies in Systems, Decision and Control

Volume 179

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series "Studies in Systems, Decision and Control" (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control–quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/13304

Cristina Flaut · Šárka Hošková-Mayerová
Daniel Flaut
Editors

Models and Theories in Social Systems

Springer

Editors

Cristina Flaut
Faculty of Mathematics and Computer Science
Ovidius University of Constanţa
Constanţa, Romania

Daniel Flaut
Faculty of History and Political Science
Ovidius University of Constanţa
Constanţa, Romania

Šárka Hošková-Mayerová
Department of Mathematics and Physics, Faculty of Military Technology
University of Defence
Brno, Czech Republic

ISSN 2198-4182 ISSN 2198-4190 (electronic)
Studies in Systems, Decision and Control
ISBN 978-3-030-00083-7 ISBN 978-3-030-00084-4 (eBook)
https://doi.org/10.1007/978-3-030-00084-4

Library of Congress Control Number: 2018953292

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The book *Models and Theories in Social Systems* is part of the series *Studies in Systems, Decision and Control* published by Springer. This is the result of a scientific collaboration in the fields of Mathematics, Statistics, and Social Science, among Prof. Šárka Hošková-Mayerová from the *University of Defence* of Brno (Czech Republic), Profs. Cristina Flaut and Daniel Flaut from the *Ovidius* University of Constanța (Romania). The different studies included in this volume, selected after a peer-review process, are collected into four parts.

Part I *General Remark* is composed of the chapter of Syamal K. Sen and Ravi P. Agarwal, which is titled "Perusing the Minds Behind Scientific Discoveries." The authors record the essential discoveries of a few scientists and their interaction/reaction with their respective diverse environments. The chapter records the winding path of their scholarship throughout history, and most importantly, the thought process of each individual that resulted in the mastery of their subject. Despite there are many other legendary scientists with their gigantic contributions, the few instances mentioned in this contribution are good enough to get a feel of the trend of the mindset of the scientific community, in general.

The contribution of Cavallo, D'Apuzzo, Di Nola, Squillante, and Vitale opens the Part II of the volume, which is titled *Theories in Social Systems*. This chapter, namely "A General Framework for Individual and Social Choices," considers some models for both individual and social choices. Specifically, the proposed structures allow unifying some previous models analyzed in the literature and also overcoming or interpreting some critical issues or paradoxes. Moreover, relevant properties are illustrated in detail.

Afterward, Maturo, Migliori, and Paolone propose the research paper "Nationality Board Diversity in Organizations: A Brief Review and Future Research Directions." Because of the increasing interest in the literature on the impact of diversity on firms' performance, the authors focus on the issue of nationality board variety in organizations. The chapter highlights that the concept of diversity is multidimensional and can concern different aspects such as gender, nationality, educational, age, ethnicity, and race. However, the authors concentrate on the previous studies limiting their attention to nationality board diversity.

The final purpose of this contribution is to illustrate and discuss the results and limits of these studies, and understand why most of them lead to conflicting results.

Chapter "Mathematical Modeling of Some Physical Phenomena Through Dynamical Systems," by Olivia Ana Florea, considers that differential equations and system of differential equations represent the kernel of the mathematical modeling, offering tools to predict the natural phenomena from science, technics, medicine, biology, etc. The author analyzes the phase portraits of different dynamical systems linear and nonlinear, the Lagrangian formalism of a problem encountered in aerodynamics, and averaging method for nonlinear differential equation.

Carp, Popa, and Serban, in their Chapter "Methods for Improving the Quality of Image Reconstruction in Computerized Tomography," present several classes of methods which can rise from classical projection-based algorithms, such as Kaczmarz- and Cimmino-type algorithms, for algebraic reconstruction of images in computerized tomography.

Pipina Nikolaidou contributes with the Chapter "Questionnaires, Bar and Hyperstructures." She presents an application of the hyperstructure theory in the field of social sciences; specifically, the bar is suggested as a tool to be used in questionnaires instead of Likert Scale. The utility of this instrument is giving the opportunity of obtaining more accurate results, and also automatically saving them on computers.

The following chapter is titled "Micropolar Thermoelasticity with Voids Using Fractional Order Strain" and is presented by Lavinia Codarcea-Munteanu and Marin Marin. This study deals with thermoelasticity of micropolar materials with voids that use the fractional order strain, to determine some equations of this linear thermoelasticity theory, as well as of a reciprocity relation for the mentioned bodies.

Adina Chirila and Marin Marin are the authors of Chapter "Diffusion in Microstretch Thermoelasticity with Microtemperatures and Microconcentrations." They focus on the linear theory of microstretch thermoelasticity for materials whose particles have microelements that are equipped with microtemperatures and microconcentrations. Specifically, they derive the field equations and constitutive equations for isotropic and homogeneous bodies, introduce some dimensionless quantities and establish the continuous dependence of solutions upon initial data and body loads by means of the Gronwall inequality, and finally provide a rigorous mathematical model with various possible applications in materials science, engineering, and even biology.

Chapter "Axial-Symmetric Potential Flows," by Plaksa, consider axial-symmetric stationary flows of the ideal incompressible fluid as an important case of potential solenoid vector fields. He establishes relations between axial-symmetric potential solenoid fields and principal extensions of complex analytic functions into a special topological vector space containing an infinite-dimensional commutative Banach algebra.

Dušan Knežo and Alena Vagaská, in Chapter "Monte Carlo Method Application and Generation of Random Numbers by Usage of Numerical Methods," deal with the Monte Carlo method, which is often used for simulating systems with many coupled degrees of freedom, for simulation of experiments. The chapter presents some methods of generating random numbers by usage of standard numerical methods for various probability distributions types.

Part III of the book is titled *Models in Social Systems*. The first contribution of this part is Chapter "Rolling Circles of Motions: Yesterday and Today," which is authored by Murat Tosun and Soley Ersoy. In this chapter, they give a short historical survey of basic events which happened during the development of models depending on rolling circles around circles. Specifically, the study focuses on elliptic and cycloidal (epicycloid or hypocycloid) motion by use of the complex forms of Bottema's instantaneous invariants characterizing the infinitesimal properties of motion.

Chapter "Some Remarks on Social Life in Romanian Towns and Cities in the 1930s, Based on Statistical Data," by Daniel Flaut and Enache Tuşa, presents some aspects regarding the social life in some Romanian towns and cities, based on statistical data, in the fourth decade of the twentieth century, a period marked by economic crises, social problems, and the imminent outbreak of World War II.

Dan Vătăman proposes the study in Chapter "Developments in Decision-Making Process Within the European Union System." This research deals with developments in decision-making process within the European Union, the way in which the European legislation is adopted, and also the problem of clarifying principles, theories, and technical issues regarding decision-making process within the European Union system.

Eva Kellnerová, Kristýna Binková, and Šárka Mayerová, in their Chapter "Assessment of the Efficiency of Respiratory Protection Devices Against Lead Oxide Nanoparticles," evaluate the current state of health- and safety-related problems of those people exposed to environmental burden due to occupational requirements. Specifically, the authors focus on personal protective equipment used against inhalation of pollutants from the air. Filter efficiency is determined according to standardized methods given by the standardized Czech technical norms. However, such rehearsals are not specifically focused on an ultrafine aerosol with the content of nanoparticles in the range of 7.6–299.6 nm. The study evaluates permeability of one of the most often used protective filter OF-90 against ultrafine aerosol lead oxide with predetermined characteristics.

Chapter "Community Detection in Social Networks," by Fataneh Dabaghi Zarandi, and Marjan Kuchaki Rafsanjani, deals with the use of social, biologic, communication, and the World Wide Web networks. Particularly, the authors introduce several methods for community detection and their comparison.

Lepellere, Cristea, and Gubiani, in Chapter "The E-Learning System for Teaching Bridging Mathematics Course to Applied Degree Studies," present a comparison between the University of Udine (Italy) and the University of Nova Gorica (Slovenia). Specifically, this study focuses on the e-learning system for teaching the math bridge-course to applied degrees studies.

Anata-Flavia Ionescu and Dorin-Mircea Popovici propose the research paper "Applications of Multi-Agent Systems in Social Sciences: Virtual Enterprises as an Example." The study is a review of the most important applications of the multi-agent systems in social sciences, with a particular focus on virtual enterprises.

Following, Fabrizio Maturo, Viviana Ventre and Angelarosa Longo discuss the topic in Chapter "On Consistency and Incoherence in Analytical Hierarchy Process and Intertemporal Choices Models." The authors focus on two different approaches in decision-making processes, i.e., the analytical hierarchy process and intertemporal choices models, highlighting the consistency conditions usually adopted. After a general discussion on consistence and incoherence in the framework of these two different approaches, they show that sometimes it is preferable to weaken or reinforce coherence conditions according to the specific context.

Šárka Hošková-Mayerová and Antonio Maturo, in their Chapter "On Some Applications of Fuzzy Sets for the Management of Teaching and Relationships in Schools," analyze the problem of uncertainty on the result of an aggregation operation and on the degree to which a relation holds. They suggest the use of hyperoperations, which permit considering together many possible results of the interaction of any ordered pair of elements and fuzzy sets that give the possibility to measure the degree of belonging of an element to a set described by a linguistic property or the degree of a relation between individuals. The authors show some possible applications to social science at the aim to give an efficient tool for modeling of social phenomena.

The research article "Resources and Capabilities for Academic Spin-Offs' Development. An Empirical Analysis of the Italian Context," by Migliori and De Luca, investigates which resources can affect more than others the creation and successful development of university spin-offs (USOs). Using a sample of 100 Italian USOs, their analysis shows that spin-offs appear to be quite innovative but they generally need time and probably more funding to protect their innovation through patents issuing. In this sense, established spin-offs suffer for the difficulties in raising funds, high costs of developing ideas, and lack of governmental support.

Part IV of this volume, which is titled *Mathematical Methods in Social Sciences*, starts with Chapter "A Fixed Point Result on the Interesting Abstract Space: Partial Metric Spaces" by Erdal Karapınar. The author investigates the existence of fixed point of certain mappings via simulation functions in the framework of an interesting abstract space, namely partial metric spaces. The main results of this manuscript not only extend, but also generalize, improve, and unify several existing results on the literature of metric fixed point theory.

Dorina Raducanu, in her study in Chapter "Geometric Properties of Mittag-Leffler Functions", concentrates on certain geometric properties for two-parametric Mittag-Leffler function.

Following, in the contribution in Chapter "Special Numbers, Special Quaternions and Special Symbol Elements," Diana Savin proposes a study of some properties of quaternion algebras and symbol algebras and obtains a specific algebraic structure.

Cristina Flaut presents some applications of quaternions and octonions. Specifically, her Chapter "An Algebraic Model for Real Matrix Representations. Remarks Regarding Quaternions and Octonions" illustrates the real matrix representation for complex octonions and some of its properties which can be used in computations, where these elements are involved. Moreover, she gives a set of invertible elements in a split quaternion algebra and in a split octonion algebra.

Chapter "A Theory of Quaternionic G-Monogenic Mappings in E_3," by Kuzmenko and Shpakivskyi, considers a class of so-called quaternionic G-monogenic mappings and proposes a description of all mappings from this class by using four analytic functions of complex variable. For G-monogenic mappings, they generalize some analogues of classical integral theorems of the holomorphic function theory of the complex variable (the surface and the curvilinear Cauchy integral theorems, the Morera theorem), and Taylor's and Laurent's expansions. Moreover, they introduce a new class of quaternionic H-monogenic (differentiable by Hausdorff) mappings and establish the relation between G-monogenic and H-monogenic mappings. Finally, they prove the theorem of equivalence of different definitions of a G-monogenic mapping.

Serpil Halici, in Chapter "On Bicomplex Fibonacci Numbers and Their Generalization," consider bicomplex numbers with coefficients from Fibonacci sequence and give some identities. He demonstrates the accuracy of such identities by taking advantage of idempotent representations of the bicomplex numbers, and then, by this representation, he gives some identities containing these numbers. Then, the author proposes a generalization that includes these new numbers and calls them Horadam bicomplex numbers. Finally, the Binet formula, the generating function of Horadam bicomplex numbers, and two important identities that relate the matrix theory to the second order recurrence relations are obtained.

Caruso, Gattone, Balzanella, and Di Battista, in their Chapter "Cluster Analysis: An Application to a Real Mixed-Type Data Set," stress the importance of clustering mixed data, and propose an application on a real-world mixed-type data set regarding flight delays.

In Chapter "Ordering in the Algebraic Hyperstructure Theory: Some Examples with a Potential for Applications in Social Sciences," several examples of concepts of the algebraic hyperstructure theory, which are all based on the concept of *ordering* are included. The author Michal Novák pointed the fact that in many aspects related to social sciences the population, i.e., the elements of the carrier set, on which the operation or a hyperoperation is constructed, are somehow put in relations. A typical example of this is *family relations*, in which the set of individuals are linked in two ways: by mating operation (or hyperoperation) and in descendant—ancestor relation. It is also shown how these concepts could be linked. The reason why this selection was made is the fact that in social sciences, objects are often linked in two different ways, which can be represented by an operation (or a hyperoperation) and a relation. The algebraic hyperstructure theory is useful in considerations of social sciences because in this theory the result of an interaction of two objects is, generally speaking, a set of objects instead of one particular object.

The closing chapter of this book is titled "Classical and Weakly Prime *L*-Submodules" by Razieh Mahjoob and Shaheen Qiami. Let *L* be a complete lattice, the authors introduce and characterize classical prime and weakly prime *L*-submodules of a unitary module over a commutative ring with identity. Also, they topologize Cl.L-Spec(M), the collection of all classical prime *L*-submodules of M and investigate the properties of this topological space.

In summary, the book *Models and Theories in Social Systems* collects a broad range of models and theories regarding social systems. Because of the wide spectrum of topics that social systems cover, different issues related to Mathematics, Statistics, Teaching, Social Science, and Economics are discussed. Due to the large number of interests of the papers collected in this volume, the latter is addressed, in equal measure, to Mathematicians, Statisticians, Sociologists, Philosophers, and more generally to scholars and specialists of different sciences.

The second editor thanks for the support of her work provided within project maintained by the Ministry of Defence the Czech Republic—Project code: DZRO K-217.

Constanța, Romania	Cristina Flaut
Brno, Czech Republic	Šárka Hošková-Mayerová
Constanța, Romania	Daniel Flaut

Contents

Part I General Remark

Part II Theories in Social Systems

Part III Models in Social Systems

Part IV Mathematical Methods in Social Sciences

About the Editors

Cristina Flaut is a professor in the Department of Mathematics and Computer Science at the Ovidius University of Constanţa, Romania. She is the author and the co-author of more than 60 monographs, chapters of books, and papers in important journals (as for example in Taylor & Francis, in Springer or in the journals: Ann. Mat. Pura Appl., Adv. Appl. Clifford Algebras, Bull. Korean Math. Soc., Adv. Differ. Equ.-NY, J. Intell. Fuzzy Syst., Results Math., Chaos, Solitons & Fractals, Soft Computing, Algebr. Represent. Theor., etc.).

She is Editor-in-Chief of the journal *Analele Ştiinţifice ale Universităţii Ovidius Constanţa-Seria Matematica*, an ISI journal.

In 2016, she was considered the best researcher of the Ovidius University of Constanţa.

Areas of interest: algebra (nonassociative algebras, logical algebras), coding theory and cryptography.

Šárka Hošková-Mayerová is an associate professor in the Department of Mathematics and Physics at the University of Defense in Brno, Czech Republic. She is the co-author of one monograph, co-editor of 4 books published in Springer Publishing house, author or co-author of several chapters of books, and papers in important journals (e.g., Soft Computing, Computers and Mathematics with Appl., An. Şt. Univ. Ovidius Constanţa-Seria Matematica, Quality & Quantity, Iran. J. Fuzzy sets, Advances in Fuzzy Systems, International Journal of Production Research, Ital. J. Pure and Appl. Math., Deturope, J. of Security Sustainability Issues).

She is Editor in Chief of the journal *Ratio Mathematica*, indexed in various Databases, also member of editorial board of various journals, eg. *Ital. J. Pure and Appl. Math.* and *Advances in Military Technology*.

Areas of interest: algebraic hyperstructures and fuzzy structures, mathematical modelling and decision-making process.

Daniel Flaut is a professor in the Department of History and Political Science at Ovidius University of Constanţa, Romania. He is the author and the co-author of more than 50 books, chapters of books, and papers in important journals.

He is the Editor-in-Chief of the journal *Revista Română de Studii Eurasiatice*, indexed in various Databases. He is also member of editorial board of various journals.

He is Director of the Eurasian Studies Center of Faculty of History and Political Science of Ovidius University of Constanţa, Romania.

As the co-author of the book *Arheologie medievală română*, he received in 2006 the “George Potra” prize, awarded by the Cultural Foundation “Magazin Istoric”, Romania.

Areas of interest: medieval history, auxiliary sciences of history, history of international relations.

Cluster Analysis: An Application to a Real Mixed-Type Data Set

G. Caruso, S. A. Gattone, A. Balzanella and T. Di Battista

Abstract When you dispose of multivariate data it is crucial to summarize them, so as to extract appropriate and useful information, and consequently, to make proper decisions accordingly. Cluster analysis fully meets this requirement; it groups data into meaningful groups such that both the similarity within a cluster and the dissimilarity between groups are maximized. Thanks to its great usefulness, clustering is used in a broad variety of contexts; this explains its huge appeal in many disciplines. Most of the existing clustering approaches are limited to numerical or categorical data only. However, since data sets composed of mixed types of attributes are very common in real life applications, it is absolutely worth to perform clustering on them. In this paper therefore we stress the importance of this approach, by implementing an application on a real world mixed-type data set.

Keywords Clusters analysis · Numeric data · Categorical data · Mixed data
Cluster algorithm

1 Introduction

In order to discover interesting groups, cluster analysis has been applied to a variety of scientific areas (Caruso et al. 2018). In marketing, it can help to identify different customers clusters and to use this knowledge to create targeted campaigns (Valentini et al. 2011). In the field of crime prevention, clustering analysis is used in the search

G. Caruso (✉) · S. A. Gattone · T. Di Battista
University G. d'Annunzio, Pescara, Italy
e-mail: giulia.caruso@unich.it

S. A. Gattone
e-mail: gattone@unich.it

T. Di Battista
e-mail: dibattis@unich.it

A. Balzanella
University of Campania Luigi Vanvitelli, Caserta, Italy
e-mail: antonio.balzanella@unicampania.it

C. Flaut et al. (eds.), *Models and Theories in Social Systems*, Studies in Systems,
Decision and Control 179, https://doi.org/10.1007/978-3-030-00084-4_27

of credit card frauds or in monitoring criminal activities in electronic commerce (Nie et al. 2010; Peng et al. 2005). In the educational sector, cluster analysis can be used to identify specific common patterns among test items (Di Battista and Fortuna 2016). In medicine and in biology the clustering of shapes is routinely applied by practitioners in order to discover different structures in the set of objects (Brignell et al. 2010; Gattone et al. 2017). In environmental studies, clustering is applied in order to classify ecological communities on the basis of their diversity (Di Battista 2002; Di Battista and Gattone 2003; Fortuna and Maturo 2018; Maturo 2018).

To perform such analysis it is necessary to group a data set into homogeneous clusters, and to efficiently interpret them. Traditionally, cluster analysis only focus on purely numerical data. K-means method, due to its huge efficiency in processing large data sets, is the most popular clustering algorithm, especially for data mining (Di Battista and Gattone 2004; MacQueen 1967). Nevertheless, this algorithm has a big disadvantage: it is often limited to numerical attributes, since it is based on the Euclidean distance measure between data points and clusters-means (Everitt 1974). Furthermore, data often contains both numeric and categorical values. A way to overcome this problem is to transform the categorical values into quantitative ones, such as the binary strings, and then apply the numerical-value based clustering methods.

However, this approach would cause a loss of knowledge, ignoring the similarity information enclosed in the categorical attributes (Ahmad and Dey 2007). Hence, it is desirable to overcome this problem by finding a unified similarity metric for both categorical and numerical data, in order to eliminate the metric gap between continuous and categorical data. Some papers have tried to find a unified similarity metric for categorical and quantitative attributes, but a computational efficient similarity measure has yet to be implemented.

Huang (1997) presents a clustering algorithm to solve data partition problems. Whilst it is based on the K-means paradigm, it removes the numeric data only limitation, preserving, at the same time, its efficiency. This algorithm clusters objects with quantitative and categorical attributes, similarly to K-means. It is called K-prototypes algorithm, because objects are clustered against K prototypes instead of K means.

Cheung and Jia (2013) method is based on the concept of object-cluster similarity. They propose a new metric for both quantitative and qualitative attributes, so that the object cluster similarity for both of them has a uniform criterion. In this way, they (Cheung and Jia 2013) eliminate the need of transformation and parameter adjustment between categorical and numerical values.

The rest of the paper is organized as follows. In Sect. 2, two methods for clustering mixed-data are reviewed. In Sect. 3 we present an application on a real mixed-type data set to show the interaction between quantitative and qualitative attributes in the clustering process. Finally, in Sect. 4 we draw some conclusions and discuss some suggestions for future research.

2 Clustering Mixed Data

Let $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ indicate a set of n objects and $X_i = [x_{i1}, x_{i2}, \ldots, x_{iM}]$ denote an object constituted by M variables. We consider the case in which the M variables are both continuous and categorical. Let $M = Q + C$ where Q is the number of numeric variables and C is the number of categorical variables. Let us define the two indicator variables subsets which identify the different types of variables as follows: $\mathcal{C} = \{m_1^{\mathcal{C}}, \ldots, m_C^{\mathcal{C}}\}$ denotes the categorical variables and $\mathcal{Q} = \{m_1^{\mathcal{Q}}, \ldots, m_Q^{\mathcal{Q}}\}$ denotes the numeric variables. The aim of clustering is to divide the n objects contained in $\mathbf{X}$ into K separate clusters. Since for a given n the number of possible partitions is significant, it is not advisable to examine each of them to find a better one, but instead try to maximize (or minimize) a suitably chosen objective function (Everitt 1974; Huang 1997). When clustering mixed data sets the main problem is to determine *how close* or *how far apart* objects are from each other. In what follows we consider two approaches that present two different ways to combine in a single cost function distance measures for numeric variables and dissimilarity measures for categorical variables.

2.1 Huang Method

Huang (1997) presented a so-called *K-prototypes* algorithm, which is based on the *K-means* method, but overcomes its quantitative data limitation, preserving, at the same time, its efficiency. The algorithm groups the objects in clusters against K prototypes. The updates occurs in a dynamical manner so to minimize the following objective function:

$$E = \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik} s(X_i, P_k), \tag{1}$$

where u_{ik} is an element of a *partition matrix* $U_{n\times k}$, and s is a dissimilarity measure between the objects X_i and P_k. $P_k = [p_{k1}, p_{k2}, \ldots, p_{kM}]$ is the *prototype* or *representative vector* for cluster k. U represents a *hard partition* matrix where $u_{ik} \in \{0, 1\}$ and $u_{ik} = 1$ if X_i is allocated to cluster k.

The dissimilarity measure is defined as:

$$s(X_i, P_k) = \sum_{m\in\mathcal{Q}} (x_{im} - p_{km})^2 + \gamma_k \sum_{m\in\mathcal{C}} \delta(x_{im}, p_{km}), \tag{2}$$

where the first term is the squared Euclidean distance, while the second term is defined as $\delta(r, t) = 0$ for $r = t$ and $\delta(r, t) = 1$ for $r \neq t$. γ_k is a weight for categorical attributes in cluster k.

Let define the internal term in Eq. (1) as $E_k = \sum_{i=1}^{n} u_{ik} s(X_i, P_k)$. This term measures the total dissimilarity of objects in cluster k from their prototype P_k, otherwise known as the total cost of allocating $\mathbf{X}$ to cluster k. This term may be rewritten as:

$$\begin{aligned} E_k &= \sum_{i=1}^{n} u_{ik} \sum_{m \in \mathcal{Q}} (x_{im} - p_{km})^2 + \gamma_k \sum_{i=1}^{n} u_{ik} \sum_{m \in \mathcal{C}} \delta(x_{im}, p_{km}) \\ &= E_k^{\mathcal{Q}} + \gamma_k E_k^{\mathcal{C}}, \end{aligned} \tag{3}$$

where $E_k^{\mathcal{Q}}$ and $E_k^{\mathcal{C}}$ represent the dissimilarity of the objects in cluster k, coming from the numerical and the categorical variables, respectively. In order to minimize these two components, let $P_k^{\mathcal{Q}}$ and $P_k^{\mathcal{C}}$ be the prototype for cluster k for the numerical and categorical variables, respectively.

$E_k^{\mathcal{Q}}$ is minimized with the usual update of the K-means algorithm for continuous variables, i.e. the generic component of $P_k^{\mathcal{Q}}$ is calculated by

$$p_{km} = \frac{1}{n_k} \sum_{i=1}^{n} u_{ik} x_{im} \qquad\qquad m \in \mathcal{C}, \tag{4}$$

where n_k is the number of objects in cluster k.

Let $V_m = \{v_{m_1}, v_{m_2}, \ldots\}$ be the set enclosing the distinct values of the m-th categorical variable and let $\mathrm{pr}(v_{m_j}|k)$ be the probability that value v_{m_j} is observed in cluster k.

It is possible to rewrite $E_k^{\mathcal{C}}$ in (3) as

$$E_k^{\mathcal{C}} = \sum_{m \in \mathcal{C}} n_k \left[1 - \mathrm{pr}(p_{km} \in V_m|k)\right]. \tag{5}$$

From (5), $E_k^{\mathcal{C}}$ is minimized by selecting the categorical values of the prototype $P_k^{\mathcal{C}}$, such that

$$\mathrm{pr}(p_{km} \in V_m|k) \geq \mathrm{pr}(v_{m_j} \in V_m|k)$$

for $p_{km} \neq v_{m_j}$ for all categorical attributes.

2.2 Cheung and Jia Method

Cheung and Jia (2013) provide a unified similarity metric which can be used with mixed attributes. For the categorical variables they define the similarity between a categorical attribute value $x_{im}^{\mathcal{C}}$ and cluster k as:

$$s(x_{im}^{\mathcal{C}}, P_l) = \mathrm{pr}(x_{im}^{\mathcal{C}} \in P_k|k). \tag{6}$$

In the Huang method, the contribution of each categorical attribute is fixed in each cluster k and has to be chosen in a subjective way. Cheung and Jia propose an automatic procedure to compute the importance of each categorical attribute. In particular, the importance of any categorical attribute $V_m (m \in \mathcal{C})$ can be calculated by the average information content of all its possible values:

$$H_{V_m} = -\frac{1}{m_L} \sum_{m_l=1}^{m_L} p(v_{m_l}) \log p(v_{m_l}), \tag{7}$$

where $p(v_{m_l}) = p(v_{m_l} \in V_m)$. The weight of each attribute is then computed as

$$\gamma_m = \frac{H_{V_m}}{\sum_{m \in \mathcal{C}} H_{V_m}}. \tag{8}$$

Finally, the similarity for the categorical attributes between the i-th object and cluster k is given by

$$s(X_i^{\mathcal{C}}, P_k) = \sum_{m \in \mathcal{C}} \gamma_m s(x_{im}^{\mathcal{C}}, P_k). \tag{9}$$

The similarity metric for numerical attributes is rescaled, so to fall into the interval [0, 1]. The normalization is given by

$$s(X_i^{\mathcal{Q}}, P_k) = \frac{exp[-0, 5D(X_i^{\mathcal{Q}}, P_k)]}{\sum_{k=1}^{K} exp[-0.5D(X_i^{\mathcal{Q}}, P_k)]}, \tag{10}$$

where $D(\cdot)$ is any distance function suitable for numerical variables. According to Eqs. (9) and (10), the object-cluster similarity metric for mixed data is defined as

$$s(X_i, P_k) = \frac{C}{C+1} s(X_i^{\mathcal{C}}, P_k) + \frac{1}{C+1} s(X_i^{\mathcal{Q}}, P_k), \tag{11}$$

where $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, K$.

3 An Application on a Real Mixed-Type Data Set

Delay is one of the most relevant indicators of any transportation system. Flight delays provoke negative impacts, mainly economic, for airlines, airports and passengers. We used, as data source, the Bureau of Transportation Statistics database.

This dataset shows departure and arrival delays for US domestic flights and it contains both qualitative and quantitative information. For our illustrative example, we analyzed data regarding the 1st of August 2017.

Table 1 Description of the variables

Variable	Type	Description
Airline	CATEGORICAL NOMINAL	The code assigned by IATA to identify a unique airline, also known as carrier
Origin	CATEGORICAL NOMINAL	Geographic area of the origin airport
Destination	CATEGORICAL NOMINAL	Geographic area of the destination airport
Departure Performance, composed by		
DepDelay	CONTINUOUS	The difference (in minutes) between the scheduled and the actual departure time Early departures show negative numbers
Taxi time out	CONTINUOUS	Time between off-block and take off
Arrival Performance, composed by		
ArrDelay	CONTINUOUS	The difference (in minutes) between the scheduled and the actual arrival time Early arrival show negative numbers
Taxi time in	CONTINUOUS	Time between landing and in-block
Flight Summaries, composed by		
Distance	CONTINUOUS	The distance between airports, expressed in miles
Cause of delay	CATEGORICAL NOMINAL	It specifies the reason for cancellation: A: carrier B: weather C: national air system D: security

We considered each commercial flights throughout the United States and Canada from the State of New York. The flights data consists of 1426 instances and 9 features, which are summarized in Table 1.

We run a cluster analysis with a number of clusters equal to $K = 3$ and compared the results of the following three methods:

1. Huang method (described in Sect. 2.1)
2. Cheung and Jia method (described in Sect. 2.2)
3. Standard K-means method (applied on numerical variables only).

Table 2 displays, for each cluster, the mean value of numerical attributes, while Table 3 the distributions of categorical attributes. Figure 1 represents a scatter-plot of Departure Delay vs Arrival Delay and displays the cluster labels recovered by the three methods. Table 2 shows that the patterns between the Huang and the K-means methods are very similar between them.

Interestingly, the Huang method highlights a very strong clustering structure among all numerical attributes whereas, using the K-means method, the cluster means of the variable "Taxi time out" appear to be very similar between them. Furthermore, Fig. 1 illustrates how clusters resulting from the Huang method produce a better separation in the space defined by the Departure and Arrival delays.

Table 2 Variable mean values for each cluster

Method	Cluster	Size	DepDelay	Taxi time out	ArrDelay	Taxi time in	Distance
Huang	1	59	1.45	21.07	1.32	6.93	689
	2	586	0.03	30.62	0.01	10.51	1461
	3	780	0.01	16.36	−0.10	6.95	782
Cheung and Jia	1	472	0.10	22.33	0.06	7.63	883
	2	248	0.02	30.80	−0.02	11.12	2376
	3	705	0.09	16.84	−0.03	7.96	710
K-means	1	54	1.51	21.61	1.38	6.83	640
	2	309	0.02	23.17	−0.01	12.31	2187
	3	1062	0.02	22.24	−0.06	7.36	750

Table 3 Geographic area of the origin and destination airport for each cluster

Method	Cluster	Size	Origin				Destination			
			NE	NW	SE	SW	NE	NW	SE	SW
Huang	1	59	0.70	0.02	0.27	0.02	0.83	0	0.17	0
	2	586	0.68	0.06	0.12	0.14	0.53	0.08	0.20	0.19
	3	780	0.68	0.03	0.24	0.05	0.79	0.01	0.19	0.01
Cheung and Jia	1	472	0.49	0.15	0	0.36	0.51	0.15	0	0.34
	2	248	0.99	0	0.01	0	0.30	0.04	0.58	0.08
	3	705	0.54	0.02	0.39	005	0.99	0	0	0.01
K-means	1	54	0.70	0	0.28	0.02	0.85	0	0.15	0
	2	309	0.50	0.15	0.06	0.29	0.51	0.17	0.02	0.30
	3	1062	0.74	0	0.23	0.03	0.72	0	0.25	0.03

Table 3 shows that, using the Huang method, the 1st cluster has a prevalence of flights with origin and destination NE (North East) and SE (South East). In the 2nd group the relative frequency of flights with origin and destination SW (South West) is higher than in the other groups. The 3rd group is very similar to 1st group, while in Fig. 1 clusters appear very separate. The qualitative variables have a greater impact with the Cheung and Jia method. It is evident, indeed, in Table 3 that the distributions of the qualitative variables are very different in the three clusters. Finally, the Origin and the Destination distributions observed in the clusters recovered by the K-means method look similar to the ones observed in the Huang method.

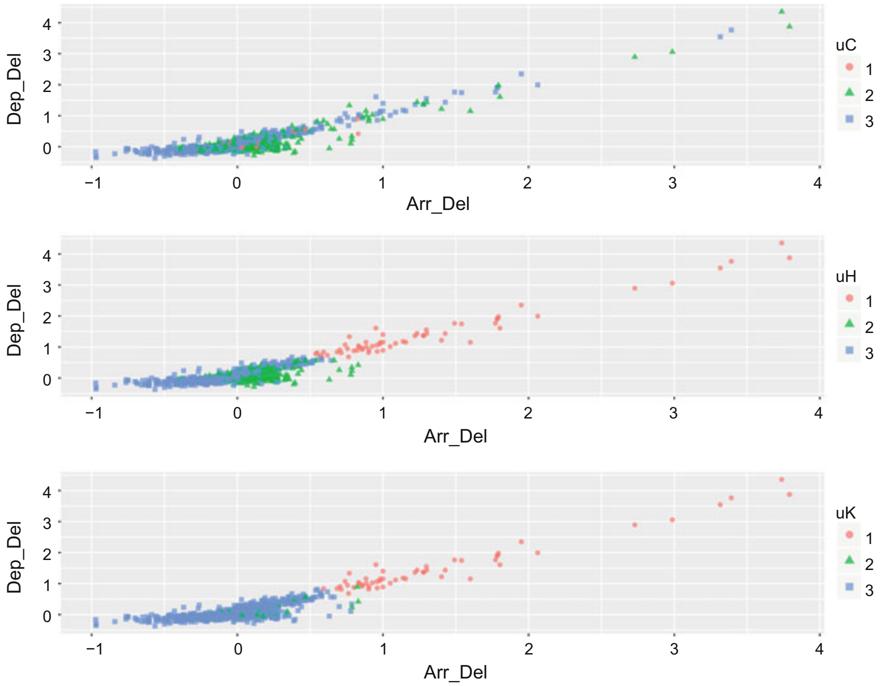


Fig. 1 Scatter-plot of Departure Delay vs Arrival Delay, together with cluster labels: Cheung and Jia (first panel), Huang (second panel) and K-means method (third panel)

4 Conclusions and Future Research

In this work we have shown a real application on clustering real mixed-data. The results have shown how both the Huang and Cheung and Jia algorithms retain the K-means efficiency, whilst removing, simultaneously, its quantitative data only limitation. In the application we have seen that the results obtained by the Huang method are better for the numerical attributes, while the Cheung and Jia results show a higher discrimination for the categorical attributes. Further work has to be implemented to provide a way to automatically choose a balance between numerical and categorical attributes.

References

Ahmad, A., Dey, L.: A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl. Eng. **63**, 503–527 (2007)

Brignell, C.J., Dryden, I.L., Gattone, S.A., Park, B., Browne, W.J.: Surface shape analysis with an application to brain surface asymmetry in schizophrenia. Biostatistics **11**(4), 1–22 (2010)

Caruso, G., Gattone, S.A., Fortuna, F., Di Battista, T.: Cluster analysis as a decision-making tool: a methodological review. In: Bucciarelli, E., Chen, S., Corchado, J.M., (eds.) Decision Economics: In the Tradition of Herbert A. Simon's Heritage. Advances in Intelligent Systems and Computing, vol. 618, pp. 48–55. Springer International Publishing (2018)

Cheung, Y., Jia, H.: Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recognit. **46**, 2228–2238 (2013)

Di Battista, T.: Diversity index estimation by adaptive sampling. Environmetrics **13**(2), 209–214 (2002)

Di Battista, T., Fortuna, F.: Clustering dichotomously scored items through functional data analysis. Electron. J. Appl. Stat. Anal. **9**(2), 433–450 (2016)

Di Battista, T., Gattone, S.A.: Multivariate bootstrap confidence regions for abundance vector using data depth. Environ. Ecol. Stat. **11**(4), 355–365 (2004)

Di Battista, T., Gattone, S.A.: Nonparametric tests and confidence regions for intrinsic diversity profiles of ecological populations. Environmetrics **14**(8), 733–741 (2003)

Everitt, B.: Cluster Analysis. Heinemann Educational Books Ltd. (1974)

Fortuna, F., Maturo, F.: K-means clustering of item characteristic curves and item information curves via functional principal component analysis. Qual. Quant. (2018). https://doi.org/10.1007/s11135-018-0724-7

Gattone, S.A., De Sanctis, A., Russo, T., Pulcini, D.: A shape distance based on the Fisher-Rao metric and its application for shapes clustering. Phisica A **487**, 93–102 (2017)

Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Proceedings in the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 21–34 (1997)

MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)

Maturo, F.: Unsupervised classification of ecological communities ranked according to their biodiversity patterns via a functional principal component decomposition of Hills numbers integral functions. Ecol. Indic. **90**, 305–315 (2018)

Nie, G., Chen, Y., Zhang, L., Guo, Y.: Credit card customer analysis based on panel data clustering. Procedia Comput. Sci. **1**(1), 2489–2497 (2010)

Peng, Y., Kou, G., Shi. Y., Chen, Z.: Improving clustering analysis for credit card accounts classification. In: Proceedings of the 5th International Conference on Computational Science—ICCS 2005, Part III, pp. 548–553. Springer Berlin Heidelberg (2005)

Valentini, P., Di Battista, T., Gattone, S.: Heterogeneity measures in customer satisfaction analysis. J. Classif. **28**, 38–52 (2011)