International Statistical Review

International Statistical Review (2024) doi: 10.1111/insr.12596

Handling Out-of-Sample Areas to Estimate the Unemployment Rate at Local Labour Market Areas in Italy

Roberto Benedetti[1], Federica Piersimoni[2], Monica Pratesi[3,4], Nicola Salvati[3] and Thomas Suesse[5]

[1]*Department of Economic Studies (DEc), 'G. d'Annunzio' University of Chieti-Pescara, Pescara, Italy*
[2]*Directorate for Methodology and Statistical Process Design, ISTAT, Rome, Italy*
[3]*Department of Economics and Management, University of Pisa, Pisa, Italy*
[4]*Department of Statistical Production, ISTAT, Rome, Italy*
[5]*National Institute of Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia*
Corresponding to: Nicola Universita degli Studi di Pisa Dipartimento di Economia e Management, Italy. Email: nicola.salvati@unipi.it

Summary

Unemployment rate estimates for small areas are used to efficiently support the distribution of services and the allocation of resources, grants and funding. A Fay–Herriot type model is the most used tool to obtain these estimates. Under this approach out-of-sample areas require some synthetic estimates. As the geographical context is extremely important for analysing local economies, in this paper, we allow for area random effects to be spatially correlated. The spatial model parameters are estimated by a marginal likelihood method and are used to predict in-sample as well as out-of-sample areas. Extensive simulation experiments are used to assess the impact of the auto-regression parameter and of the rate of out-of-sample areas on the performance of this approach. The paper concludes with an illustrative application on real data from the Italian Labour Force Survey in which the estimation of the unemployment rate in each Local Labour Market Area is addressed.

Key words: area level model; maximum marginalised likelihood; missing data; small area estimates; spatial dependency.

1 Introduction

The unemployment rate is one of the key indicators to plan a future for local administrative areas in which the distribution of resources is consistent with principles of social, economic and spatial justice. In recent years, employment inequalities among areas have widened as the post-2008 economic crisis and adoption of austerity policies, followed then by the pandemic crisis, had an uneven geographical impact. There is, accordingly, a pressing need to revalue the appropriateness and efficacy of existing policy instruments for reducing such spatial inequalities, and to consider and develop alternative mechanisms.

To enrich the empirical knowledge on such inequalities, a substantial increase of the number of data collection and analysis programmes conducted by National Statistical Offices (NSOs) is required. Alternatively, the use of estimators that can guarantee the production of more reliable statistics through the introduction of appropriate area level models need to be developed. This is carried out to extend the range of quantitative indicators available at multiple regional scales.

The Italian National Institute for Statistics (ISTAT) regularly provides estimates of unemployment indicators based on data obtained through the Italian Labour Force Survey (ILFS). The ILFS allows to obtain quarterly estimates of the main aggregates regarding the labour market that are important both at the local and the central government levels for the development of labour market policies. Direct estimates of the unemployment rate cannot be disseminated for local labour market areas (LLMAs), 611 unplanned domains obtained as clusters of municipalities. This is due to the presence of out-of-sample areas (158) and too many LLMAs having a small sample size, which leads to estimates with an unacceptably large coefficient of variation. For these reasons, ISTAT has implemented the use of indirect, model-based small area estimators to produce official yearly estimates of unemployment rate for Italian LLMAs (D'Alò *et al.*, 2017).

The increasing demand for estimates for new domains, like LLMAs, led to the development of a number of model-based small area estimation (SAE) methods (Jiang & Lahiri, 2006; Rao & Molina, 2015) including the empirical best linear unbiased predictor (EBLUP) based on a linear mixed model with data recorded at the area level. The EBLUP is a widely used solution recommended when the target is the average of a continuous response variable (Fay & Herriot, 1979). This traditional technique, employed by survey statisticians and empirical economists and geographers, often assumes independence of the observations: an hypothesis that is evidently violated in all geographical studies (Anselin, 1988). Empirical models that do not take into account spatial dependence may show serious misspecification problems. To take into account this propensity for nearby locations to influence each other, a general class of well-known models has been introduced in the statistical literature (Besag, 1974; Cressie, 1993) and in the last decade some research effort was dedicated to SAE models by using spatial dependence as a conceptual framework for spatial data modelling (Cressie, 1989; Petrucci & Salvati, 2006; Pratesi & Salvati, 2008; 2009; You & Zhou, 2011). For a detailed review, see Bertarelli *et al.* (2021) and Pratesi *et al.* (2023).

Surprisingly, in the SAE literature, we do not find a similar research effort dedicated to deal with out-of-sample areas, which is a relevant characteristic very frequent in real situations concerning existing surveys on families and individuals. This problem arises because there can be some (or often many) areas that do not have any statistical units selected in the sample and, consequently, no design-based estimation is possible for such areas. The ILFS is such an example because 158 LLMAs have 0 sample size. The model-based estimates can be computed by making the apparently incorrect assumption of no random effects for these areas. If random effects are uncorrelated between areas, then estimates of out-of-sample area random effects are zero, as there is no further information about the out-of-sample areas available. However, as correlation between small area effects should be the norm rather than the exception, models should allow for spatial correlation of area random effects. An immediate benefit of using such models is that prediction of random area effects for out-of-sample areas are possible and could become standard practice.

To the best of our knowledge, the problem of estimating and predicting target parameters spatially correlated in-sample and out-of-sample areas has never been addressed in the frequentist literature. Recently, Chung & Datta (2022) propose spatial Fay–Herriot models that effectively account for heteroscedasticity and spatial dependence of the small area effects. But the authors take a fully Bayesian approach by specifying a class of non-informative *priors* on

the model parameters and model spatial dependence of small area random effects by four widely used autocorrelation structures. Burgard *et al.* (2022) propose multivariate empirical best predictor based on the Fay–Herriot model for partially missing direct estimates, but the multivariate model does not take into account for the spatial correlation and needs that at least a component of the multivariate dependent variable is available in each small area. The presence of missing data and the estimates for small areas are two important aspects to be treated together (Longford, 2004; 2005), but at the moment, there are only some attempts made in the case of missing data in the auxiliary variables proposing an approach based on multiple imputation (Benedetti & Filipponi, 2010; Panzera *et al.*, 2016). An attempt to consider together the two aspects was carried out by Saei & Chambers (2005) who proposed a method based on the prediction of random area effects for out-of-sample areas.

The important concept that links the result of the selection process and the realisation of a set of values of a population generated according to a spatial model is the *informative missingness*. Under this concept, the probability that a domain is either missing or observed depends on the value of the spatial process in its position within the spatial population of the domains. The adaptation of spatial analysis methods to account for the *informative missingness* has already been discussed in the geo-statistical literature (Reich & Bandyopadhyay, 2010), typically through joint modelling of observed responses and their positions. These solutions are mainly based on the *shared-variable* approach, which assumes an explicit model for the selected positions and the outcome. This mitigates the prediction bias and allows valid inference for the parameters of the spatial model.

The problem of out-of-sample areas in SAE has been faced with no reference to the literature on spatial missing values. Indeed, out-of-sample areas describe spatial patterns that could or not be related with the values of the area target parameters (spatial informativeness). This paper is the first attempt in the frequentist framework to consider the effects of spatial informativeness to predict area target parameters in out-of-sample areas. In particular, in this paper, we focus on the maximum likelihood (ML) estimation of a mixed-effect model with spatially auto-correlated random effects, assuming that the data are observed only on a selection, not necessarily random, of $m_s = m - m_r$ domains among the m possible domains representing a partition of the reference population. Standard estimation approaches for such models, the so-called spatial EBLUP (SEBLUP) (Petrucci & Salvati, 2006; Pratesi & Salvati, 2008; 2009), are based on having data on all domains m.

The theoretical framework we develop here is based on criteria that derive from the application of well-known principles in the likelihood theory. In particular, we show how the parameter estimates of the underlying population generation process, following Suesse (2018), can be obtained by ML to estimate the regression parameters from the observed data using the marginalised log-likelihood. The proposed ML approach is similar to the ideas suggested in Chambers et al. (2012, chapter 5); however, their approach is an EM (Expectation, E-step, Maximisation, M-step) type algorithm, aiming at maximising the marginal likelihood indirectly.

The remainder of the paper is organised as follows. In Section 2, we summarise the methodology of small area estimation of a continuous variable when some areas do not have any observed unit. In Section 3, we motivate the use of the suggested models for estimating the small area parameters with some exploratory analysis of the data set. In particular, because some diagnostics indicate a deviation from the assumption of data independence and considering the simultaneous presence of missing data, the introduction of specific solutions for small area estimation may be required. In Section 4.1, a simulation study based on a well-known spatial data sets is used to compare the proposed estimation method with some alternative SAE methods in terms of their bias and root mean squared error (RMSE). The simulations themselves assume different estimation scenarios with varying spatial auto-regressions of the generated data and

of the number of out-of-sample domains. In Section 4.2, we will use Italian data from the 2012 Labour Force Survey conducted by ISTAT to estimate the unemployment rate for each Local Labour Market Areas (LLMAs) comparing some alternative solutions to the treatment of missing data with the proposed approach of maximising the marginalised likelihood. Finally, in Section 5 the main results of the paper are discussed and some important issues that require further research are identified.

2 Small Area Estimation Based on Fay–Herriot Models

Fay & Herriot (1979) proposed an area-level SAE model (hereafter the FH model) that relates small area direct survey estimates to area-level covariates. The FH model is widely used because of its flexibility in combining different sources of information with different error structures. Cressie (1991), Singh *et al.* (2005), Petrucci & Salvati (2006) and Pratesi & Salvati (2008) extend the FH model to allow for spatially correlated random effects using conditional auto-regressive (CAR) and simultaneous auto-regressive (SAR) specifications for random effects (Anselin, 1988).

In this section, we describe the FH model and its extensions needed to introduce the spatial dependence of the estimates in each area and to be able to adapt the FH model to situations in which direct estimates are not available for all areas, in order to allow the estimation of area-specific random effects for out-of-sample areas.

2.1 Fay–Herriot Model

Let $\hat{\boldsymbol{\theta}}$ be the design unbiased direct estimator of the m-vector of parameters of interest $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$, typically small area totals or means

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{e}$ is a vector of independent sampling errors with mean vector $\boldsymbol{\theta}_m$ of size $m \times 1$, and known diagonal variance matrix $\mathbf{R} = diag(\phi_i),\ i = 1, \ldots, m$ with ϕ_i representing the sampling variance of the direct estimator of the parameter of interest in the $i - th$ area, which usually are assumed to be known. However, in most practical applications, this assumption is not met because National Statistical Offices (NSOs) release these estimates for much more aggregated domains than those used in the FH model. Therefore, it is necessary to produce an estimate for these variances. When unit-level data are available, one approach might be to directly estimate the variance of the estimate $\hat{\boldsymbol{\theta}}$ using analytical or resampling methods. However, these variance estimates can be unstable due to small sample sizes in the areas or domains of interest. Consequently, smoothing methods are often employed through generalised variance functions (for details, see Wolter, 1985) even if, as empirically noted by Hidiroglou *et al.* (2019), the resulting estimators are often sensitive to deviations from the normality assumption. It is worth underlining, however, that this common practice of considering the variance of the direct estimator to be known could lead to an underestimation of the uncertainty of the estimates, especially in domains with very few sample observations (You & Chapman, 2006). An underestimation of ϕ_i is usually more problematic when the variance of the random effect of the domain is relatively small compared with ϕ_i, which is the most likely case in SAE models as it is assumed that the variance of the sampling error is higher in smaller domains (Bell, 2008).

The basic area level model assumes that an $m \times p$ design matrix $\mathbf{X}$, which contains typically area-specific auxiliary variables, is linearly related to $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \tag{2}$$

where the p-vector $\boldsymbol{\beta}$ contains the regression coefficients and the m vector $\mathbf{u}$ contains the independent random area-specific effects with zero mean and covariance matrix $\boldsymbol{\Sigma}_u = \sigma_u^2 \mathbf{I}_m$ with $\mathbf{I}_m$ being the identity matrix of size $m \times m$. The matrix $\mathbf{Z}$ is the design matrix of the random effects.

The combined Fay–Herriot model (Fay & Herriot, 1979) has the following form:

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{3}$$

and it is a special case of linear mixed model with the variance of $\hat{\boldsymbol{\theta}}$ being equal to $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \mathbf{R} + \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^{\top}$. Under this model, the Best Linear Unbiased Predictor (BLUP) $\hat{\boldsymbol{\theta}}^{FH}$ is extensively used to obtain model-based indirect estimators of small area parameters $\boldsymbol{\theta}$ and associated measures of variability. The empirical best linear unbiased predictor (EBLUP) of $\boldsymbol{\theta}$, that is, the parameter of interest, is

$$\hat{\boldsymbol{\theta}}^{FH} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}, \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ is the generalised least squares estimator of $\boldsymbol{\beta}$ and

$$\hat{\mathbf{u}} = \hat{\boldsymbol{\Sigma}}_u\mathbf{Z}^{\top}\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{5}$$

is the EBLUP of $\mathbf{u}$; $\hat{\sigma}_u^2$ is the asymptotically consistent estimator of σ_u^2 obtained by ML or Restricted Maximum Likelihood (REML) methods based on the normality assumption of the random effects. For details, see Rao & Molina (2015).

In practice, areas are unplanned domains, and many of them have zero sample sizes. These areas are referred to as non-sampled areas or out-of-sample areas. Suppose now that for subset s these estimates are available or observed but not for the remaining areas r. Let m_s denote the number of small areas in the sample, with $m_r = m - m_s$ denoting the number of out-of-sample areas. We can partition the matrices $\mathbf{X}$ and $\mathbf{Z}$ as $\mathbf{X} = \left[\mathbf{X}_s^{\top} \ \mathbf{X}_r^{\top}\right]^{\top}$ and $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_s & \mathbf{0} \\ \mathbf{0}^T & \mathbf{Z}_r \end{bmatrix}$ corresponding to sampled and non-sampled areas where $\mathbf{0}$ is a $m_s \times m_r$ matrix of zeros. Similarly the random effects vector can be partitioned as $\mathbf{u} = \left[\mathbf{u}_s^{\top}, \mathbf{u}_r^{\top}\right]^{\top}$. The conventional approach for estimating area means in this case is the EBLUP estimation for sampled areas and the synthetic estimation, based on a suitable model fitted to the data from the sampled areas, for out-of-sample areas

$$\hat{\boldsymbol{\theta}}_s^{FH} = \mathbf{X}_s\hat{\boldsymbol{\beta}} + \mathbf{Z}_s\hat{\mathbf{u}}_s, \tag{6}$$

$$\hat{\boldsymbol{\theta}}_r^{FH} = \mathbf{X}_r\hat{\boldsymbol{\beta}}. \tag{7}$$

Analytic estimation of the mean squared error (MSE) of the $\hat{\theta}^{FH}$ is usually carried out using the estimator suggested by Prasad & Rao (1990); see also Datta *et al.* (2005).

2.2 Spatial Fay–Herriot Model

As we might expect when there is an unexplained spatial or temporal correlation in the data, neglecting these in our model leads to erroneous inferences (Cressie, 1993) because the estimates will be insufficient due to the exclusion of essential parameters from the statistical model. In small area estimation, when areas represent a geographic partition of the target population, the closest areas tend to have more similar socio-economic characteristics. This partition offers valuable information that can be used to improve estimators, that is, it is possible for the model

to gain efficiency from the spatial configuration of the data. In this sense, Singh *et al.* (2005), Petrucci & Salvati (2006) and Pratesi & Salvati (2008) proposed the introduction of the spatial auto-regression in SAE under the FH model by specifying a linear mixed model with spatially correlated random area effects for the vector of the parameters of interest. When data from nearby areas are correlated, considering this type of spatial correlation in the model leads to more efficient small area estimators (Marhuenda *et al.*, 2013; Molina & Rao, 2010). The independence between random effects of the FH model can be an incorrect assumption because most of small areas boundaries are essentially arbitrary and there appears to be no good reason why areas on one side of such boundary should not be correlated with areas on the other side. It is widely known that both environmental and socio-economic phenomena have a spatial distribution conditioned by nature and by the action of man.

The spatial Fay–Herriot (SFH) model is valid for sampled areas and assumes that the random effects vector $\mathbf{u}_s$ follows a SAR model, that is,

$$\mathbf{u}_s \sim N(\mathbf{0}_{m_s}, \mathbf{G}_s), \text{ where } \mathbf{G}_s = \sigma_u^2 \left[\mathbf{A}_s^\top \mathbf{A}_s\right]^{-1} \text{and } \mathbf{A}_s = \mathbf{I}_{m_s} - \rho \mathbf{W}_s, \tag{8}$$

and $\mathbf{W}_s$ refers to a $m_s \times m_s$ spatial weight matrix constructed for the sample, the weight w_{sij} referring to the spatial influence of unit j on unit i. The matrix $\mathbf{W}_s$ might be row-normalised and be based on an initial $m \times m$ spatial neighbourhood matrix $\mathbf{N}$, which indicates whether the areas are neighbour or not (one way to define $\mathbf{N}$ is to set $N_{ij} = 1$ if small area i and j are neighbour or 0 otherwise). The parameter ρ is the spatial auto-regressive coefficient that defines the strength of the spatial relationship among the random effects associated with the neighbouring areas. Because estimates $\hat{\boldsymbol{\theta}}_s$ of $\boldsymbol{\theta}_s$ are available, Singh *et al.* (2005), Petrucci & Salvati (2006) and Pratesi & Salvati (2008) imposed the SAR model on $\hat{\boldsymbol{\theta}}_s$ with the model implied covariance of $\hat{\boldsymbol{\theta}}_s$ that is $\mathbf{V}_s(\hat{\boldsymbol{\theta}}_s) = \mathbf{R}_s + \mathbf{Z}_s \mathbf{G}_s \mathbf{Z}_s^\top$, where $\mathbf{R}_s$ is a partition of matrix $\mathbf{R}$ on sampled areas. On the observed m_s units, the model is fitted, then prediction is

$$\hat{\boldsymbol{\theta}}_s^{SFH} = \mathbf{X}_s \hat{\boldsymbol{\beta}} + \mathbf{Z}_s \hat{\mathbf{u}}_s, \tag{9}$$

for m_s sampled areas and

$$\hat{\boldsymbol{\theta}}_r^{SFH} = \mathbf{X}_r \hat{\boldsymbol{\beta}}, \tag{10}$$

for the m_r out-of-sample areas where $\hat{\boldsymbol{\beta}}$ is the generalised least squares estimator of $\boldsymbol{\beta}$ assuming the spatial structure.

Chung & Datta (2022) explored various spatial models besides SAR, resulting in different parameterisations of the variance and covariance matrix $\mathbf{V}_s$ as a function of the parameter ρ and the contiguity matrix $\mathbf{W}_s$. A common alternative to SAR is the conditional autoregressive model (CAR) for formulating dependency. For a detailed comparison and the relationship between these two specifications of spatial dependence, see Ver Hoef *et al.* (2018).

Concerning the MSE, an analytical estimator has been proposed by Singh *et al.* (2005); their proposal is a second order approximation of the MSE of the SEBLUP. Another analytical formula of the MSE can be found in Petrucci & Salvati (2006) and Pratesi & Salvati (2008). Analytical approximations may require strong model assumptions and many small areas to approximate well the true values; therefore, Molina *et al.* (2009) proposed parametric and non-parametric bootstrap procedures for estimation of the MSE under the SFH model. The authors provided *naïve* and bias-corrected bootstrap estimators.

The literature on Bayesian methods in spatial dependence models is quite extensive. Notable contributions include Sun *et al.* (1999), who studied a hierarchical model incorporating conditional and intrinsic autoregressive parameters on the random effects, and Moura &

Migon (2002), who proposed a logistic hierarchical model for small area prediction of proportions, accounting for both possible spatial and unstructured heterogeneity effects. Similar models were also explored by Speckman & Sun (2003) for applications beyond SAE. Specifically, in the context of SAE, You & Zhou (2011) suggested using a conditional autoregressive model to parameterise the variance and covariance matrix. Extending the classic FH model with repeated surveys over time, Torabi (2012) and Marhuenda *et al.* (2013) proposed a space-time autoregressive random effects model. Furthermore, Porter *et al.* (2014) extended the FH model by introducing functional covariates along with autoregressive random effects, and Porter *et al.* (2015) incorporated conditional autoregressive random effects into the FH model. These models allow for spatial correlation in the area effects while keeping the fixed effects parameters spatially invariant.

2.3 Spatial Fay–Herriot Model Extended to the Presence of Missing Data

The maximum likelihood estimation of model parameters discussed in the previous sections is typically straightforward when data covers the entire population or when missing data can be treated as non-informative. Under this highly restrictive assumption, the model can be fitted to the available data from certain areas without necessarily considering that they represent only a portion of the target population. While this *naïve* approach yields consistent estimates for many regression models, it may not hold true for certain models like spatial autoregressive models (Benedetti *et al.*, 2020). Both in maximum likelihood and Bayesian frameworks, conducting spatial predictions for areas lacking observations is relatively straightforward. Extensive literature exists on this topic, including the classic *kriging* method (Cressie, 1993). For a comprehensive review, see also Banerjee *et al.* (2014), which extends these methods to SAE applications. However, the core challenge addressed in this article is that if missing data are informative, as they often do, the correct likelihood function could vary significantly, diverging from the incorrect one concerning the entire target population.

To deal with this problem, Saei & Chambers (2005) used the SFH model for proposing a new approach by imposing a SAR model on the all areas, irrespective of whether sampled or not, that is,

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_s \\ \mathbf{u}_r \end{bmatrix} \sim N(\mathbf{0}_m, \mathbf{G}), \text{ where } \mathbf{G} = \sigma_u^2[\mathbf{A}^\top \mathbf{A}]^{-1} \text{ and } \mathbf{A} = \mathbf{I}_m - \rho\mathbf{W}. \tag{11}$$

The spatial weight matrix is obtained from the $m \times m$ spatial neighbourhood matrix $\mathbf{N}$, for example, by row-normalising. Based on this extended model the random area effects for out-of-sample areas ($\hat{\boldsymbol{u}}_r$) are predictions based on ML estimates. Combining the ideas of Harville (1977) and Henderson (1950), the log-likelihood is formed by three components: (i) the log-likelihood for $\boldsymbol{\beta}$ generated by $\hat{\boldsymbol{\theta}}$ given the value of the random component vector $\mathbf{u}_s$, (ii) the logarithm of the probability density of $\mathbf{u}_s$ given the value of the random component vector $\mathbf{u}_r$, (iii) the logarithm of the probability density of the random component vector $\mathbf{u}_r$. Let $\mathbf{G} = \begin{bmatrix} \mathbf{G}_{ss} & \mathbf{G}_{sr} \\ \mathbf{G}_{rs} & \mathbf{G}_{rr} \end{bmatrix}$ partition of matrix $\mathbf{G}$ corresponding to the in-sample and out-of-sample components of $\mathbf{u}$. We note that in general using this partitioning notation, the aforementioned matrices $\mathbf{W}_s$, $\mathbf{A}_s$ and $\mathbf{G}_s$ are not identical to the sub-matrices of $\mathbf{W}$, $\mathbf{A}$ and $\mathbf{G}$, that is, $\mathbf{W}_s \neq \mathbf{W}_{ss}$ and hence $\mathbf{A}_s \neq \mathbf{A}_{ss}$ and $\mathbf{G}_s \neq \mathbf{G}_{ss}$. This means that the model proposed by Saei & Chambers (2005) and the SFH model are different, in the sense that they impose a different model on the sample, or in other words the marginal distributions of the sample differ. Then the EBLUP of the area-specific random effects can be written as

$$\hat{\mathbf{u}}_s = \hat{\mathbf{T}}_{ss}\mathbf{Z}_s^\top \mathbf{R}_s^{-1}(\hat{\boldsymbol{\theta}}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}), \tag{12}$$

$$\hat{\mathbf{u}}_r = \hat{\mathbf{T}}_{sr}\mathbf{Z}_s^\top \mathbf{R}_s^{-1}(\hat{\boldsymbol{\theta}}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}), \tag{13}$$

where

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{ss} & \mathbf{T}_{sr} \\ \mathbf{T}_{rs} & \mathbf{T}_{rr} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_s^\top \mathbf{R}^{-1}\mathbf{Z}_s + \boldsymbol{\Lambda}_{s|r} & -\boldsymbol{\Lambda}_{s|r}\mathbf{G}_{sr}\mathbf{G}_{rr}^{-1} \\ -\mathbf{G}_{rr}^{-1}\mathbf{G}_{sr}\boldsymbol{\Lambda}_{s|r} & \mathbf{G}_{rr}^{-1} + \mathbf{G}_{rr}^{-1}\mathbf{G}_{rs}\boldsymbol{\Lambda}_{s|r}\mathbf{G}_{sr}\mathbf{G}_{rr}^{-1} \end{bmatrix}^{-1}, \tag{14}$$

where $\boldsymbol{\Lambda}_{s|r} = (\mathbf{G}_{ss} - \mathbf{G}_{sr}\mathbf{G}_{rr}^{-1}\mathbf{G}_{rs})^{-1}$. The EBLUP for small area means for in-sample and out-of-sample areas are

$$\hat{\boldsymbol{\theta}}_s^{SC} = \mathbf{X}_s\hat{\boldsymbol{\beta}} + \mathbf{Z}_s\hat{\mathbf{u}}_s, \tag{15}$$

$$\hat{\boldsymbol{\theta}}_r^{SC} = \mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}}_r. \tag{16}$$

ML and REML for the estimation of the regression coefficients, variance components and area-specific random effects are reported in Saei & Chambers (2005). The authors proposed also the mean cross-product error matrix for the estimation of the MSE of the predictions.

3 The Marginal Maximum Likelihood Approach for Unobserved Areas

In this section, we develop an extension of the marginal maximum likelihood (MML) approach proposed by Suesse (2018) to fit spatial auto-regressive models with missing data. In SAE framework, the missing data are the out-of-sample data. Instead of imposing a SAR model only on the observed areas, we now impose a SAR model on all m areas similar to Saei & Chambers (2005), that is, $\mathbf{G} = \sigma_u^2\left[\mathbf{A}^\top \mathbf{A}\right]^{-1}$ with $\mathbf{A} = \mathbf{I}_m - \rho\mathbf{W}$. The covariance of $\hat{\boldsymbol{\theta}}$ is

$$\begin{aligned} \mathbf{V}(\hat{\boldsymbol{\theta}}) \equiv \mathrm{Var}(\hat{\boldsymbol{\theta}}) &= \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^\top \\ &= \begin{bmatrix} \mathbf{R}_s & \mathbf{0} \\ \mathbf{0}^T & \mathbf{0}_r \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_s\mathbf{G}_{ss}\mathbf{Z}_s^\top & \mathbf{Z}_s\mathbf{G}_{sr}\mathbf{Z}_r^\top \\ \mathbf{Z}_r\mathbf{G}_{rs}\mathbf{Z}_s^\top & \mathbf{Z}_r\mathbf{G}_{rr}\mathbf{Z}_r^\top \end{bmatrix}, \end{aligned} \tag{17}$$

where $\mathbf{0}_r$ is a $m_r \times m_r$ matrix of zeros. To obtain ML estimates for complete data, the log-likelihood of the multivariate normal is maximised

$$\log f(\hat{\boldsymbol{\theta}}) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2\omega}\mathbf{r}^\top \mathbf{V}^{-1}\mathbf{r},$$

where $\mathbf{r} = \hat{\boldsymbol{\theta}} - \mathbf{X}\boldsymbol{\beta}$.

As we have partially observed the data, the marginal log-likelihood is of the following form:

$$\log f(\hat{\boldsymbol{\theta}}_s) = -\frac{m_s}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_{ss}| - \frac{1}{2\omega}\mathbf{r}_s^\top (\mathbf{V}_{ss})^{-1}\mathbf{r}_s, \tag{18}$$

where $\mathbf{V}_{ss}$ and $\mathbf{r}_s$ are the sub-matrix and sub-vector referring to the sample, as it is widely known that subvectors of multivariate normally distributed random vectors are also multivariate normally distributed.

As anticipated in Section 2.2, it is worth noting that the likelihood (18) can be utilised both by maximising it with respect to the parameters and, in a more strictly Bayesian

context, as one of the two essential components of the *posterior* distribution. In this case, the use of the marginalised likelihood (18) rather than the likelihood defined on the entire set of areas, both observed and out-of-sample, would lead to a clear reduction in the dimensionality of the *posterior* distribution, being (18) defined on a data space of size $m_s < m$. Furthermore, deriving it analytically would protect against issues associated with the numerical marginalisation of the distribution, which would otherwise be necessary. Such issues could include, for example, from a non-excessive number of samples generated by the *posterior* distribution.

We also implemented a REML version where the following expression is maximised:

$$\log f(\hat{\boldsymbol{\theta}}_s) = -\frac{1}{2}\log|\mathbf{X}_s^\top (\mathbf{V}_{ss})^{-1}\mathbf{X}_s| - \frac{m_s}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_{ss}| - \frac{1}{2\omega}\mathbf{r}_s^\top (\mathbf{V}_{ss})^{-1}\mathbf{r}_s, \tag{19}$$

using the REML estimates that can be obtained by adding the first term in (19) to the log-likelihood, see Lindstrom & Bates (1988) for details. Based on the ML or REML estimates, the random effects can be predicted using the standard BLUP equation, see (5), and the resulting EBLUP is

$$\hat{\mathbf{u}} = \hat{\mathbf{G}}[\,\mathbf{Z}_s \quad \mathbf{0}\,]^\top \hat{\mathbf{V}}_{ss}^{-1}(\hat{\boldsymbol{\theta}}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}) = \begin{pmatrix}\hat{\mathbf{G}}_{ss}\\ \hat{\mathbf{G}}_{rs}\end{pmatrix}\mathbf{Z}_s^\top \hat{\mathbf{V}}_{ss}^{-1}(\hat{\boldsymbol{\theta}}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}). \tag{20}$$

Given the same estimates are used, Equation (20) gives identical prediction for $\mathbf{u}_s$ and $\mathbf{u}_r$ as (12) and (13). However, the EBLUP approach ($\hat{\boldsymbol{\theta}}_r^{MML}$) yields different results because the MML method provides different parameter estimates than those obtained by the approach of Saei & Chambers (2005).

In principle, the method of Saei & Chambers (2005) claims to provide ML and REML estimates and hence should provide identical EBLUP values as our approach. However, in general, both methods give different parameter estimates, the smaller m_s is and the larger ρ, greater will be in general the difference between the two approaches.

Notice that the procedure of Saei & Chambers (2005) is similar to an EM algorithm by predicting the unobserved random effects $\mathbf{u}$ using the BLUP $\hat{\mathbf{u}}$, see (12) and (13), effectively implementing the E-step and then updating and maximising expressions (M-step) before the next EM-iteration follows until convergence. If the E- and M-step are both exact, then the EM algorithm yields ML estimates. However, the algorithm presented of Saei and Chambers (2005, p. 7), see step 8, supposedly maximised the log-density of $\mathbf{u}$ (the SAR model) denoted by $\log f(\mathbf{u})$, but replaces the unobserved $\mathbf{u}$ by $\hat{\mathbf{u}}$, hence maximing $\log f(\hat{\mathbf{u}})$. Unfortunately this does not correspond exactly to maximising the M-step in a EM algorithm. Suesse & Zammit-Mangion (2017) investigated a similar problem, where $\mathbf{u}_s$ was observed and $\mathbf{u}_r$ was unobserved. An exact E-step yields in this case

$$-\frac{m}{2}\log(2\pi) - \frac{m}{2}\log\sigma_u^2 + \frac{1}{2}\log|\mathbf{M}| - \frac{\hat{\mathbf{u}}^\top \mathbf{M}\hat{\mathbf{u}} + (\sigma_u')^2\mathrm{tr}\left\{\mathbf{M}_{uu}(\rho')^{-1}\mathbf{M}_{uu}(\rho)\right\}}{2\sigma_u^2}, \tag{21}$$

where $\mathbf{M}(\rho) = \mathbf{A}(\rho)^\top \mathbf{A}(\rho)$ (as $\mathbf{G} = \sigma_u^2\mathbf{M}^{-1}$) and ρ' and σ_u' are estimates from the previous step. Expression (21) needs to be maximised with respect to σ_u and ρ.

Lesage & Pace (2004) instead proposed to maximise the log-likelihood

$$-\frac{m}{2}\log(2\pi) - \frac{m}{2}\log\sigma_u^2 + \frac{1}{2}\log|\mathbf{M}| - \frac{\hat{\mathbf{u}}^\top \mathbf{M}^{-1}\hat{\mathbf{u}}}{2\sigma_u^2}. \tag{22}$$

The Equation 22 can be seen as an approximation of (21). Suesse & Zammit-Mangion (2017) investigated the impact of the approximation on the final parameters when fitting SAR models and found that the estimated parameters of the approximated M-step are often biased, the larger the set of missing units is the larger the bias usually becomes. In our case, the complete vector of random effects $\mathbf{u}$ is unobserved (corresponds to a large portion, i.e. 100%, of unobserved areas), and we expect this to yield a large bias in the estimation of the spatial auto-regression parameter ρ, hence yielding more inaccurate results for spatial prediction of small area estimates.

Despite the extensive literature on Bayesian methods for addressing missing data (see, for instance, Little & Rubin, 2019; Schafer, 1997), none of these studies attempt to incorporate nearby area dependencies in either the model estimation or prediction phases to account for unobserved values. Bayesian inference offers a straightforward means to conduct spatial predictions for areas lacking observations, as it naturally handles missing data without requiring imputation. This is because missing data can be treated as unobserved parameters that can be estimated from nearby data, forming part of the *posterior* distribution alongside the parameters of interest. Consequently, computational methods like Markov Chain Monte Carlo (MCMC) can marginalise the *posterior* distribution, yielding point estimates and variances for both missing data and model parameters simultaneously. This inherent quantification of uncertainty within the Bayesian framework obviates the need for bootstrapping. Chung & Datta (2022) advocate for this approach, suggesting that it could be expedited and enhanced by correctly specifying the likelihood on observed data only and analytically marginalising it instead of numerically summing samples generated by the *posterior* distribution, as proposed in this paper and formally elaborated in subsequent sections. This article primarily focuses on models and predictors defined under the maximum likelihood approach; hence, Bayesian-based spatial small area estimators are not employed in the application on real data or in simulations. The simulation study in Section 4.1 compares the performance of the standard SFH model, the proposed method of Saei & Chambers (2005) and our proposed MML approach, which we expect to give better results than the other two methods as it uses the correct and complete likelihood in the ML and REML estimators (see Equations 18 and 19), simultaneously taking into account the spatial correlation of all the random effects, referring to sampled and out-of-sample areas.

3.1 Estimation of the Mean Squared Error

A second-order approximation of the MSE of the proposed small area predictor under regularity conditions, together with the assumption of a large m and ignoring the terms of the order $O(m^{-1})$, can be written following Singh *et al.* (2005) and Pratesi & Salvati (2009). It can be obtained by substituting the estimates of variance components estimated (by ML or REML) of the SFH with the variance component σ_u^2, the spatial autocorrelation ρ and the random effects estimated by the MML approach used to fit SAR models with out-of-sample areas proposed in previous section. However, this approximation might produce too optimistic or conservative confidence intervals depending on the strength of the spatial correlation and on the values of the sampling variances (Pratesi & Salvati, 2009). Moreover, analytical approximations usually rely on strong model assumptions and require a large number of small areas to approximate the true values well. Resampling techniques are a good alternative to these analytical approximations. They are attractive for practitioners because of their conceptual simplicity and their easy application to complex statistical models. Moreover they require less assumptions and their performance is less depended on the number of small areas. For these reasons, the bootstrap-based estimation of the MSE proposed by Molina *et al.* (2009) for the SFH could also be used for the spatial small area predictor proposed in this paper. The parametric bootstrap procedure works as follows:

1 Fit model SFH under (11) to the original data obtaining estimates $(\hat{\sigma}_u^2, \hat{\rho})$ and $\hat{\boldsymbol{\beta}}$.
2 Generate a vector $t_1^\star$ whose elements are m independent copies of a $N(0,1)$. Construct bootstrap vectors $\mathbf{u}^\star = \hat{\sigma}_u^2(\mathbf{I} - \hat{\rho}\mathbf{W})t_1^\star$ and calculate the bootstrap quantity of interest $\boldsymbol{\theta}^\star$ by regarding $\hat{\boldsymbol{\beta}}$ and $(\hat{\sigma}_u^2, \hat{\rho})$ as the true values of the parameters.
3 Generate a vector $t_2^\star$ with m independent copies of a $N(0,1)$, independently of the generation of $t_1^\star$, and construct the vector of random errors $\mathbf{e}^\star = \mathbf{R}^{1/2}t_2^\star$. For out-of-sample areas, the sampling variance values can be estimated by generalised variance function (Wolter, 1985).
4 Construct bootstrap data from the model: $\hat{\boldsymbol{\theta}}^\star = \boldsymbol{\theta}^\star + \mathbf{e}^\star$.
5 Fit the SFH using the marginal maximum likelihood approach proposed in Section 3 on the bootstrap data $\hat{\boldsymbol{\theta}}^\star$ obtaining the estimates $(\hat{\sigma}_u^{2\star}, \hat{\rho}^\star)$, $\hat{\boldsymbol{\beta}}^\star$ of the 'true' $(\hat{\sigma}_u^2, \hat{\rho})$ and $\hat{\boldsymbol{\beta}}$.
6 Compute the bootstrap Spatial EBLUP $\hat{\boldsymbol{\theta}}^{\star MML}$.
7 Repeat steps 2-6 B times. In the b-th bootstrap replication, let $\theta_i^{\star(b)}$, be the quantity of interest and $\hat{\theta}_i^{\star MML(b)}$ the bootstrap Spatial EBLUP for area i. A bootstrap estimator of the MSE is

$$mse^{PB}(\hat{\theta}_i^{MML}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_i^{\star MML(b)} - \theta_i^{\star(b)}\right)^2. \tag{23}$$

In the case of absence of normality in the random effects and errors, the non-parametric approach introduced by Molina *et al.* (2009) that resamples both the random effects and the errors from the empirical distribution of their respective estimators could be the best bootstrap-procedure. For details on the steps of the parametric and non-parametric bootstrap for SFH models, see sections 4 and 5 in Molina *et al.* (2009).

4 Empirical Evidence

Next we aim at comparing the proposed MML method and its estimates with the Fay and Harriot estimates, see (6) and (7), the spatial Fay and Harriot estimates, see (9) and (10), and the Saei and Chambers proposed solution, see (15) and (16), described in Section 2. To compare the methods, we generate synthetic data consisting of $\mathbf{y}$ with known auto-regression parameter ρ of the SAR model, known contiguity matrix $\mathbf{W}$ and known auxiliaries $\mathbf{X}$ for the m areal units of the population.

It is important to point out that for the purpose of this paper the interest is mainly focused on the model-based properties of the estimation method when dealing with this artificial population; thus, in each simulation, the observed data are generated according to a stochastic process. We aim to evaluate the error committed by the various methods when varying, for example, the dependence of the data and the number of out-of-sample areas.

When, on the other hand, interest focuses on the properties of the MML when dealing with real data, we used a spatial population in which the variable $\mathbf{y}$ is the unemployment rate, collected by the 2012 ILFS, for which we want to produce estimates for each LLMA. In order to enhance the potential for applying these methods in various real or simulated scenarios, it is advisable to delve into specific technical considerations. Both the FH and SFH methods have been implemented in the `R` packages `sae` (Molina & Marhuenda, 2015) and `emdi` (Kreutzmann *et al.*, 2019). For the extensions to models that facilitate estimation and prediction in the context of missing data, namely, the SC and MML methods, the `R` functions are available from the authors upon request.

4.1 Experiment on an Artificial Spatial Population

Regarding the artificial population, which by its nature is not related to a specific phenomenon, for each area $i = 1, \ldots, m$ with $m = 200$, we have assumed that the data follow a linear mixed model with spatially correlated random area effects. The $m \times 2$ design matrix $\mathbf{X}$ has two columns, the first column is a vector of ones, referring to an intercept, and the second column refers to a covariate generated by a $N(0, 1)$ random variable. The intercept and slope parameters β_0 and β_1 are set to 1 and 2, respectively. The spatial auto-regression parameter is set to values $\rho = 0.2, 0.5, 0.8$ to account for low, medium and high spatial correlation. To have an explicit form of the covariance matrix; however, knowledge of this parameter is not sufficient but must be complemented with a contiguity matrix between the areas. For this purpose, we have borrowed the spatial structure of a well-known case study: the Boston Housing data-set. Originally published by Harrison & Rubinfeld (1978), it contains data on housing in the Boston area and has been widely used in the literature on spatial data modelling. In particular, in order to reduce the computational burden in the simulation study, only the first 200 of the original 506 areas have been selected as the study population, representing the northern portion of Boston (see Figure 1). In order to account for heterogeneity and heteroskedasticity, which are fairly common in many data sets, the population was divided into five groups of 40 contiguous areas with constant ϕ_i parameter values in each group and set equal to $\{0.3, 0.4, 0.5, 0.6, 0.7\}$, respectively, see Figure 1.

Furthermore, to verify the efficiency of the various methods three values $m_r = 20, 60, 100$ are considered to account for a varying degree of missingness of areas (10%, 30% and 50% receptively). In each simulation, these missing areas were selected through a simple random sample without replacement.

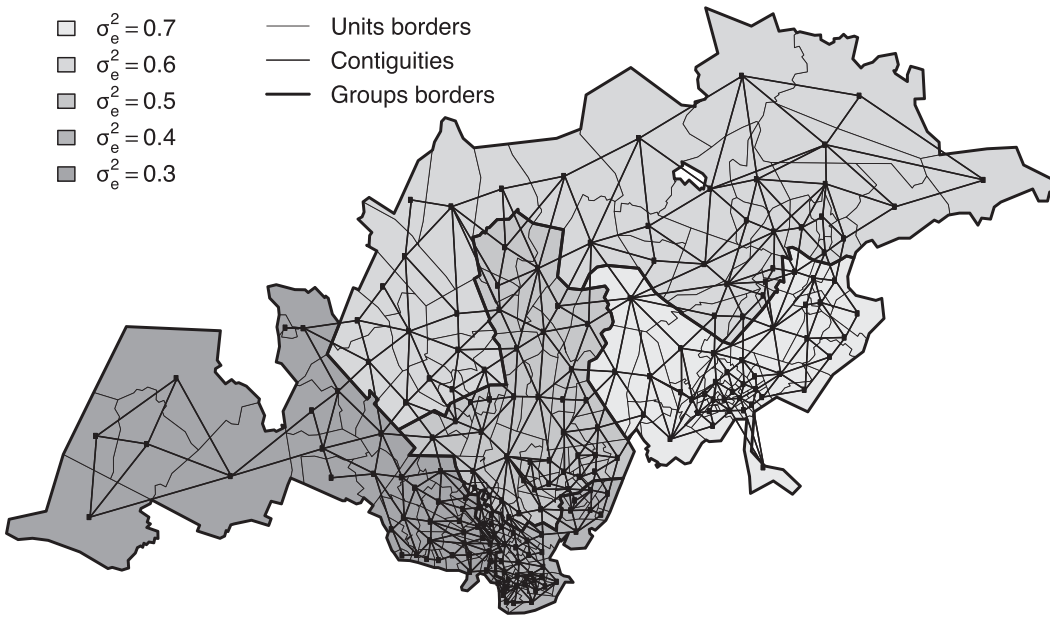

FIGURE 1. *Map of the northern portion of Boston with the characteristics of the artificial population used in the simulations with the spatial distribution of the areas, the contiguity graph and the homogeneous variance group to which each area belongs.*

For each simulated sample of areas indexed by t for $t = 1, \ldots, T$ with $T = 10{,}000$, the mean $\boldsymbol{\theta}_{it}^{M}$ of each area i was estimated using each of the four methods considered: the Fay–Herriot model ($M = FH$), the Spatial Fay–Herriot model ($M = SFH$), the Saei and Chambers proposed solution ($M = SC$) and the Marginal Maximum Likelihood method ($M = MML$).

The performances of the estimates $\hat{\boldsymbol{\theta}}_{it}^{M}$ for each method M are summarised by its average root mean square error (ARMSE, see Table 1) and average bias (AB, see Table 2), defined as follows:

$$ARMSE_M = \frac{\sum_{i=1}^{m} \sqrt{\frac{\sum_{t=1}^{T} \left(\hat{\boldsymbol{\theta}}_{it}^{M} - \boldsymbol{\theta}_{it}\right)^2}{T}}}{m}, \tag{24}$$

$$AB_M = \frac{\sum_{i=1}^{m} \sum_{t=1}^{T} \left(\hat{\boldsymbol{\theta}}_{it}^{M} - \boldsymbol{\theta}_{it}\right)}{Tm}, \tag{25}$$

where $\boldsymbol{\theta}_{it}^{M}$ are the (in practice usually unknown) generated values of the target variable for each of the $m = 200$ areas in each simulation.

The most relevant and interesting results can be summarised as follows. The bias of all the methods, evaluated both in the sampled and out-of-sample areas, is always very low and does not seem to show particular trends or configurations as the spatial dependence in the data varies. The only quite noticeable regularity is that, regardless of the method used, the bias tends to decrease as the number of out-of-sample areas increases (see Table 2). This effect could be due to the fact that none of the tested methods were subjected to benchmarking

Table 1. *Average root mean square error (ARMSE) and relative efficiency (RelEff) of the estimates in the in-sample areas only and in the out-of-sample areas only, for each method in 10,000 replicated simulations for different spatial autocorrelation parameter and number of out-of-sample areas.*

		$ARMSE$				$RelEff = \frac{ARMSE}{ARMSE_{FH}} \times 100$		
	Number of							
ρ	*out-of-sample areas*	$\hat{\boldsymbol{\theta}}_{it}^{FH}$	$\hat{\boldsymbol{\theta}}_{it}^{SFH}$	$\hat{\boldsymbol{\theta}}_{it}^{SC}$	$\hat{\boldsymbol{\theta}}_{it}^{MML}$	$\hat{\boldsymbol{\theta}}_{it}^{SFH}$	$\hat{\boldsymbol{\theta}}_{it}^{SC}$	$\hat{\boldsymbol{\theta}}_{it}^{MML}$
In-sample areas								
0.2	20	0.3368	0.3357	0.3358	0.3356	99.69	99.70	99.67
0.2	60	0.3381	0.3382	0.3388	0.3383	100.04	100.20	100.07
0.2	100	0.3355	0.3370	0.3452	0.3373	100.48	102.93	100.59
0.5	20	0.3554	0.3343	0.3342	0.3339	94.20	94.18	94.11
0.5	60	0.3593	0.3430	0.3445	0.3425	95.54	96.00	95.42
0.5	100	0.3582	0.3483	0.3746	0.3473	97.32	104.83	97.05
0.8	20	0.4183	0.3481	0.3406	0.3451	83.95	82.06	83.23
0.8	60	0.4223	0.3656	0.3644	0.3586	87.20	87.02	85.64
0.8	100	0.4192	0.3833	0.3902	0.3702	91.72	93.58	88.73
Out-of-sample areas								
0.2	20	1.0536	1.0525	1.0401	1.0361	99.90	98.74	98.36
0.2	60	1.0431	1.0434	1.0492	1.0449	100.03	100.60	100.19
0.2	100	1.0539	1.0543	1.1015	1.0661	100.04	104.56	101.18
0.5	20	1.3214	1.3147	1.0577	1.0514	99.49	80.42	79.93
0.5	60	1.3232	1.3184	1.1436	1.1223	99.65	86.86	85.28
0.5	100	1.3241	1.3185	1.3045	1.1697	99.59	98.67	88.62
0.8	20	2.9342	2.8678	1.1061	1.0986	98.10	40.82	40.39
0.8	60	2.8471	2.8200	1.3866	1.3435	99.18	51.29	50.46
0.8	100	2.9116	2.8606	1.7004	1.5167	98.36	60.37	54.72

Table 2. *Average Bias (AB) of the estimates in the in-sample areas only and in the out-of-sample areas only, for each method in 10,000 replicated simulations for different spatial autocorrelation parameter and number of out-of-sample areas.*

ρ	*Number of out-of-sample areas*	$\hat{\theta}_{it}^{FH}$	$\hat{\theta}_{it}^{SFH}$	$\hat{\theta}_{it}^{SC}$	$\hat{\theta}_{it}^{MML}$
In-sample areas					
0.2	20	0.0015	0.0013	0.0013	0.0013
0.2	60	0.0000	0.0000	0.0000	0.0000
0.2	100	−0.0018	−0.0016	−0.0016	−0.0016
0.5	20	0.0010	0.0009	0.0009	0.0009
0.5	60	−0.0001	−0.0003	−0.0003	−0.0002
0.5	100	−0.0016	−0.0018	−0.0021	−0.0018
0.8	20	0.0016	0.0011	0.0011	0.0011
0.8	60	0.0002	−0.0001	−0.0001	−0.0001
0.8	100	−0.0015	−0.0018	−0.0020	−0.0019
Out-of-sample areas					
0.2	20	0.0125	0.0122	0.0108	0.0100
0.2	60	−0.0012	−0.0002	0.0011	0.0011
0.2	100	−0.0099	−0.0100	−0.0086	−0.0102
0.5	20	0.0095	0.0101	0.0101	0.0111
0.5	60	0.0094	0.0085	0.0081	0.0075
0.5	100	0.0036	0.0024	0.0031	0.0020
0.8	20	0.0038	−0.0022	−0.0064	−0.0028
0.8	60	−0.0039	−0.0169	−0.0103	−0.0085
0.8	100	−0.0103	−0.0126	−0.0196	−0.0142

techniques that could, at least in theory, reduce this trend. Regarding the comparison between methods, none seems to be superior to the others in terms of bias, although we note that their behaviour is very similar in the sampled areas while some differences are found onlyin the out-of-sample areas.

However, the most relevant result emerges from the analysis of the efficiency of the estimates through the *ARMSE* coefficient and in particular from its transformation into *RelEff*, that is, its ratio with the same parameter calculated for the FH method considered as the conventional method and therefore as a benchmark, evaluated again both in the out-of-sample and in the sampled areas only (see Table 1). This efficiency parameter shows how crucial the role played by ρ is in reconstructing, through an appropriate use of contiguities, the estimates for those areas on which direct estimate is impossible as we do not have sample observations within it.

Interestingly, the simple introduction of the dependence parameter ρ, as in the SFH model, aids in estimating the sampled areas with a *RelEff* parameter consistently below 100. However, this is not the case for out-of-sample areas, where the *RelEff* parameter is close to 100. This phenomenon occurs because the bias introduced by missing data affects the estimation of the parameter ρ in the SFH model (Benedetti *et al.*, 2020), hindering the effective use of information from nearby areas, particularly in out-of-sample regions. This difference in the efficiencies of sampled and out-of-sample areas is not observed in the SC and MML methods. These methods correctly use the contiguity matrix between all the areas to estimate ρ without removing the non-sampled areas. The estimates proposed by FH, which make no use of spatial dependence, always produce higher relative efficiency values than the others with the only exception relating to the case of very low dependence and high number of out-of-sample areas. In this extreme situation, the other methods try to reconstruct a lot of data on the basis of their neighbourhood that, unfortunately, due to poor spatial dependence, do not produce estimates of sufficient quality. In contrast, for medium and high spatial dependence, the efficiency gain is evident not only

in the out-of-sample areas but also, with much smaller differences, in the sampled areas. In particular, a clear reduction of $ARMSE$ is observed when the method does not simply introduce the dependence in the statistical model, as in the SFH, but formally uses it for the estimation of the model considering the presence of missing data, as in the SC and MML methods. In fact, the mere introduction of the dependency in the model allows to increase efficiency by a few percentage points and almost exclusively in the sampled areas. While its more correct use in the estimation of each area, sampled or not, allows its efficiency to be reduced more effectively. This difference is even more evident in the out-of-sample areas, reaching values that can even be less than 50% for the case $m_r = 20$ (10% missingness) and $\rho = 0.8$. It seems counterintuitive that for a lower number of missing data the performance for out-of-sample is greater than for a high number of missing data. However, lower missingness also means that for a given out-of-sample area there are generally more observed or in-sample areas as neighbours, with higher spatial correlation with neighbours than with non-neighbours, from which information is available to produce more accurate estimates for the given out-of-sample area. This can be also seen from (20), where the submatrix $\mathbf{G}_{rs}$ represents the covariance between in-sample and out-of-sample areas, which is more strong for larger ρ but also for neighbouring areas.

Finally, between the two methods MML and SC, which correctly exploit the dependence of all the areas to reconstruct the missing data, the proposed MML method uses the proper spatial configuration also in the model estimation phase, thus avoiding a potential bias in the estimation of the spatial dependence parameter (Benedetti *et al.*, 2020). From this characteristic comes a further improvement in efficiency, which, although not very strong, is increasingly evident as both ρ and the number of out-of-sample areas increase. It is also worth noting that this additional efficiency gain of MML over SC comes from more precise estimate of the parameter ρ, which affects both the in-sample and out-of-sample area estimates equally. When applied to real survey data, these indicators demonstrate clear and tangible advantages of utilising the MML method over the simplistic approach of excluding unobserved regions, as commonly carried out in the FH and SHF methods. Particularly, in estimates concerning these regions, the benefits are evident. This recommendation, derived from simulations, becomes more pertinent with increased spatial dependence and the prevalence of unobserved areas. Therefore, to prevent significant inefficiencies in estimation, the potential presence of these two characteristics in geographical data, which frequently occurs in practical scenarios, should prompt their explicit incorporation into the estimation model, mirroring the approach of the MML method proposed herein.

4.2 Estimating Unemployment Rate for LLMAs in Italy

ISTAT regularly provides estimates of unemployment indicators using data from the ILFS. Its purpose is to provide information on the Italian labour market that can then be used to develop, manage, evaluate and report on labour market policies. The target population includes the members of all Italian households who regularly live within the national borders, have Italian or foreign citizenship and are regularly enrolled in the municipal lists. These estimates are reliable at a given, chosen *a priori*, geographical level. In particular, the official evaluations of the main aggregates of the job offer are produced and disseminated monthly at the national level and on a quarterly basis at the regional level; every year official evaluations for all provinces are also available.

The estimates of unemployment rate are not reliable at local labour market areas (LLMAs). This is due to the presence of out-of-sample areas (158) and too many LLMAs having a small sample size that leads to estimates with an unacceptable large coefficient of variation. LLMAs are 611 unplanned domains obtained as clusters of municipalities and defined at the Census on

the basis of daily working commuting flows. Auxiliary data at the LLMA level are required in order to apply an area-level random effect model. For this application, the logarithm of the number of unemployed persons obtained from 2011 Italian Census, the percentage of young people (less than 24 years old) and the percentage of people with diploma or higher are available at LMMA level. These explanatory variables are used to regress the variable of interest in this study, the unemployment rate that, by its definition, is a proportion and therefore has a range restricted to $[0, 1]$. Estimating a variable within such a limited range is a common challenge in all regression models for compositional data, whether binomial or multinomial. Modelling this type of data linearly, without considering its restricted range, not only clearly violates the normality assumption but also offers no assurance that the model will produce estimates within the specified range.

In the context of estimating small area proportions, Ha *et al.* (2014) proposed a normal-logistic model with a logistic distribution for the linking model. In addition, they extended the model by Liu *et al.* (2014) to general complex survey designs and denoted it as a normal-logistic random sampling variance model.

It is usually considered reasonable to assume that $m\hat{\theta}_i \sim Bin(m, \hat{\theta}_i)$ even if, under this hypothesis, the variance of $m\hat{\theta}_i$ would depend on the $\hat{\theta}_i$ themselves (Rao & Molina, 2015, Chapter 3.4). Efron & Morris (1975) proposed the well-known inverse sine transformation to stabilise the variance of this binomial distribution. This is an approach that has been followed by many authors in the field of SAE (Carter & Rolph, 1974; Jiang & Lahiri, 2001; Raghunathan *et al.*, 2007; Schmid *et al.*, 2017), and it is for this reason that in this article, we will adopt an inverse sine transformation for the unemployment rate. In particular, the direct estimate $\hat{\theta}_i$ is transformed via $\hat{\theta}_i^{trans} = \sin^{-1}\sqrt{\hat{\theta}_i}$ and the sampling variance of θ_i is approximated by $1/(4\tilde{n}_i)$, where $\tilde{n}_i$ is the effective sample size (Jiang *et al.*, 2001). The FH and the SFH models are applied to the transformed response variable and the estimates are in the interval $[0, \pi/2]$. Then the estimates are back-transformed to the original scale via $\sin^2(\hat{\theta}_i^a)$, where $\hat{\theta}_i^a$ is the estimate obtained fitting the FH and SFH models. For constructing the confidence intervals for the outcome, we use a parametric bootstrap procedure following Casas-Cordero *et al.* (2016) and Schmid *et al.* (2017). Note that the back-transformed confidence interval can be obtained because the inverse sine and sine-squared functions are monotonically increasing functions of the parameters in the ranges of interest. An alternative approach is to apply a jackknife method on the transformed scale proposed by Jiang *et al.* (2001), who considered an inverse sine transformation and showed that the bias of the jackknife MSE estimator is of order $o(m^{-1})$. The extension of the analytical MSE estimator presented in Section 3.1 to the transformed response variable using the inverse sine transformation remains an object of future research.

Table 3 reports the estimates of the model parameters for the different small area methods. In particular, the estimates of the regression coefficients obtained with different models show

Table 3. *Model parameter estimates for the different methods used.*

Coefficients	$\hat{\boldsymbol{\theta}}_i^{FH}$	$\hat{\boldsymbol{\theta}}_i^{SFH}$	$\hat{\boldsymbol{\theta}}_i^{SC}$	$\hat{\boldsymbol{\theta}}_i^{MML}$
Intercept	0.043	0.064	0.123	0.068
log of number of unemployed persons (Census)	0.010	0.007	0.009	0.006
Percentage of young people	1.193	1.066	1.042	1.000
Percentage of diploma or higher	−0.100	−0.058	−0.200	−0.030
σ_u^2	0.002	0.002	0.003	0.003
ρ	—	0.590	0.567	0.601

similar magnitude and same sign. The spatial auto-regression coefficient ρ is high for the three spatial models and shows the presence of a stationarity spatial effect and this high correlation can improve the performance of the MML and SC predictors, in particular the estimates of the area-specific random effects for out-of-sample areas.

Table 4 presents the value of the correlation coefficients between estimates obtained by MML and estimates computed by the other methods for sampled and out-of-sample areas. A high level of correlation between all methods for sampled areas can be seen. As expected, for out-of-sample areas the higher level of correlation is between the MML and the SC methods.

Figure 2 presents the maps of estimated values of unemployment rate for each LLMA in Italy, using the direct estimator, Fay–Herriot, spatial Fay–Herriot based estimators. Direct estimates are computed using Hájek-type estimators with adjusted weights that account for non-response and calibrate to population level information of demographic variables. First of all, we note that model-based estimates, especially the ones that incorporate spatial autocorrelation, appear to be consistent with direct estimates of the unemployment rate. These results indicate that the small area models that allow for more flexible incorporation of the spatial information produce overall consistent results.

As expected, the maps indicate that the LLMAs of the north-eastern part and north part of the Italy are characterised by the lowest estimates of unemployment rate. On the other hand, the estimates for the LLMAs in the south-western of Italy and in Sardegna Island show the highest level of unemployment rate. Thus, these areas can be considered the more critical in Italy. Moreover, we can note from these results that the lowest level of unemployment rate is usually in the LLMAs around the biggest cities.

The estimates of the unemployment rate at the LLMA level allow us to investigate the gap in terms of labour market between the different regions of Italy. In particular, we can note that the gap between regions in the north and in the centre of Italy is not very pronounced, in line with the relatively small differences in terms of unemployment rate between these two areas. Considering the estimates for the regions in the south of Italy and in the islands, we can notice instead a big gap with the central and northern regions: the lowest estimates of unemployment rate in the southern regions are comparable to the highest in the central and northern regions. These results confirm the existence of the so-called 'north-south divide' in Italy relating to the wealthy conditions of the population and they are in line with those obtained by Marino *et al.* (2019).

In general, we can note that the spatial distribution of estimates obtained with MML and SC appears to be more variable than that obtained with the traditional FH model. In particular, this occurs for out-of-sample areas. The range of estimates is larger and there is a wider diversification of the unemployment rate by LLMA. Given the same explanatory variables, the result is likely due to the additional spatial information included in our estimators. This moderates the spatial smoothing effect resulting from the application of the traditional FH and demonstrates the specific characteristics of the LLMAs. This happens without losing precision in the estimates.

Figure 3 shows the absolute and the relative (to the estimate) width of the confidence interval obtained for $\hat{\boldsymbol{\theta}}_i^{MML}$ using the bootstrap procedure proposed by Casas-Cordero *et al.* (2016) and

Table 4. *Correlation values between estimates obtained by the marginal maximum likelihood and estimates computed by the other methods.*

	$\hat{\boldsymbol{\theta}}_i^{FH}$	$\hat{\boldsymbol{\theta}}_i^{SFH}$	$\hat{\boldsymbol{\theta}}_i^{SC}$
In-sample areas	0.966	0.998	0.998
Out-of-sample areas	0.834	0.827	0.967

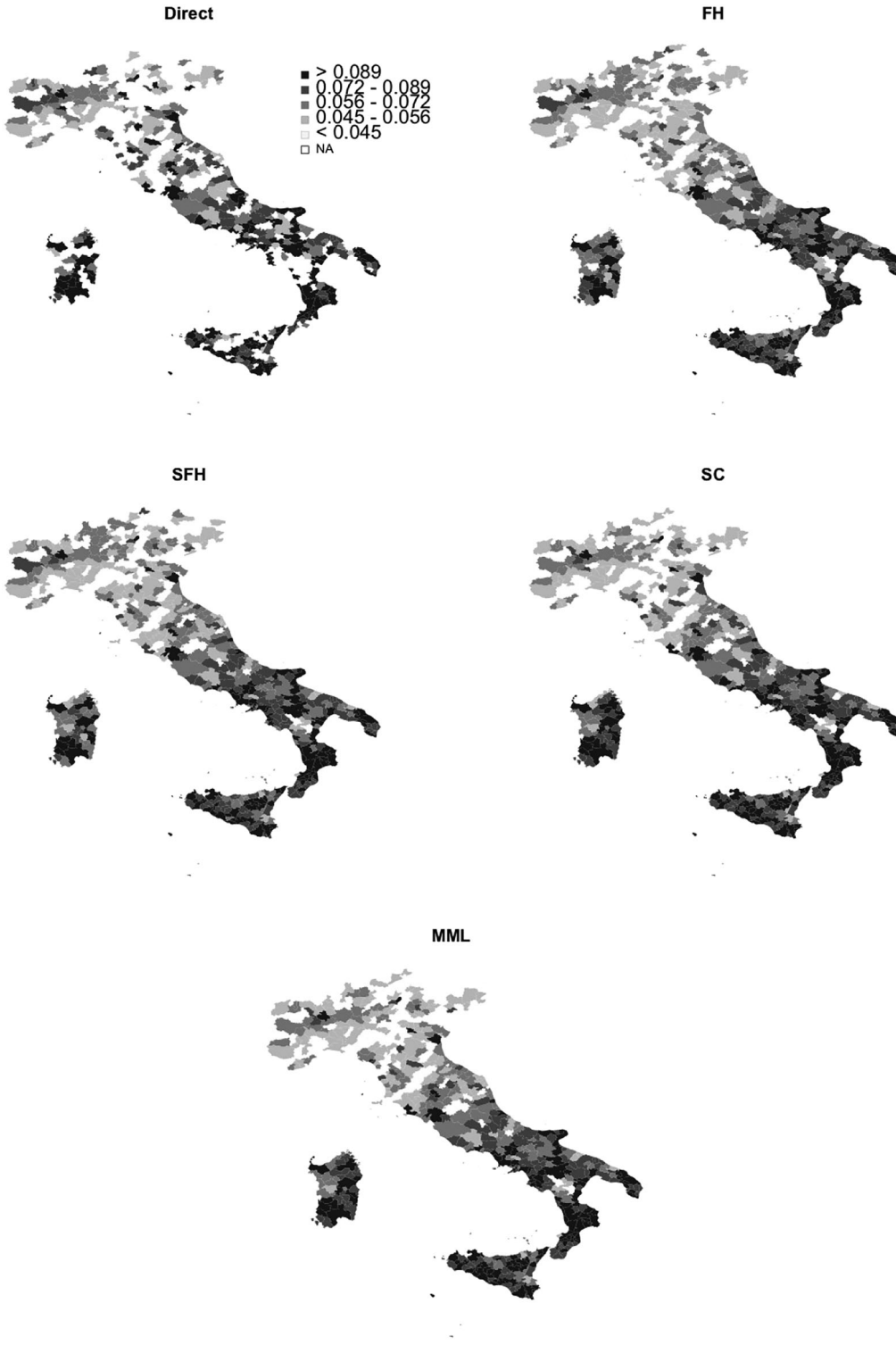

FIGURE 2. *Plots showing the spatial distribution of the estimates of the unemployment rate for the various methods. It should be noted that missing data (out-of-sample SLL—white polygons) exist only for direct estimates.*

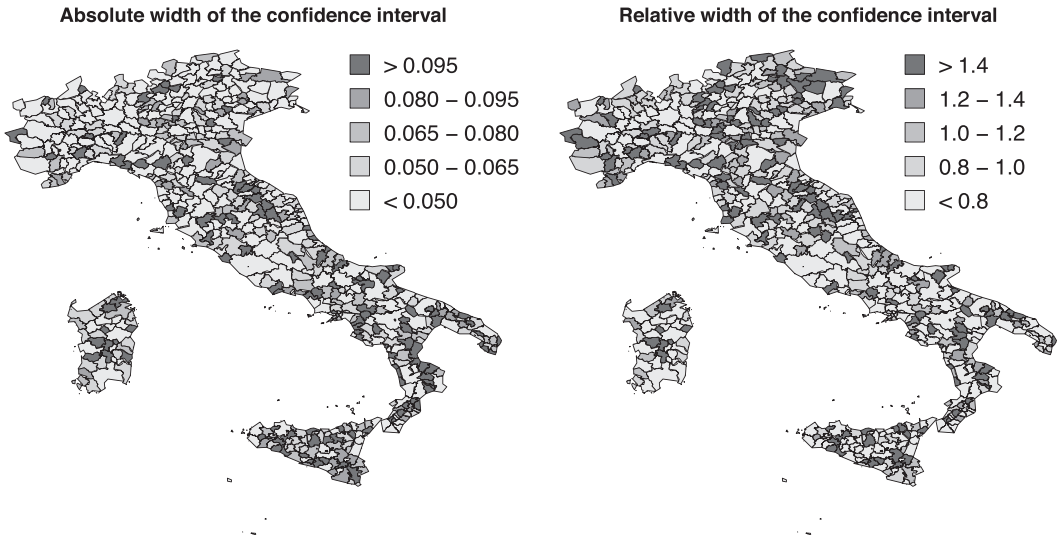

FIGURE 3. *Plots showing the spatial distribution of the width, absolute and relative to the estimates of the estimated confidence intervals for the MML.*

Schmid *et al.* (2017). The median value of the width of the confidence interval is 0.056, and it is smaller than the one of the traditional FH model (0.057). There is no evidence of spatial pattern in the distribution of the width, absolute and relative, of the estimated confidence intervals for the MML method. It is interesting to note that about 76% of the out-of-sample areas presents a smaller width of the confidence interval than the one obtained for the FH model. This indicates the gain in accuracy of the proposed estimator with respect to the FH based estimator.

Finally, it is important to note that our estimates provide no evidence that the width of the confidence intervals is affected by potential over-smoothing issues. Therefore, it might be more appropriate to consider the coverage probability rather than the interval widths. This is not surprising, given that we did not use kernel density estimation to approximate the distribution of the estimates. Instead, we focused on the asymmetry of the distributions introduced by the inverse non-linear transformation. As a result, we report the interval widths, calculated as the difference between the upper and lower limits, without assuming symmetry.

5 Final Remarks

We have proposed a method that allows to obtain estimates for small areas even on areas without sample data. Estimation domains of this kind, defined in the literature as out-of-sample areas, occurs frequently when the aim is to produce very detailed estimates of socio-economic aggregates. The small size of these estimation domains increases the probability that a random sample will not select any units in one or more of these areas. This can also happen in large surveys such as the labour force survey, not only when a sub-provincial level is needed, many areas will be sparsely populated and therefore not sampled but also because the most widely adopted design in these surveys is the two-stage sampling design that, notoriously, implies a strong concentration of families interviewed in only the municipalities (primary sampling units—PSU) selected in the first stage. This problem is typically addressed by NSOs such as ISTAT, which

publishes the time series of estimates of the unemployment rate for each quarter using Local Labour Markets in defining the estimation domain.

In practical applications of this type of small area estimation models, surprisingly, a lot of effort has been devoted to introducing the dependence of estimates, both spatial and temporal, but not to its correct use in the reconstruction of potential missing data in the target variable. To fill this gap, we have proposed a method of estimation for small areas that extends the classic Fay and Harriot area level model, not only recognising the existence and importance of spatial autocorrelation, as has already been carried out in previous papers, but by using it in conjunction with the generating process of the missing data so that the model can simultaneously exploit these aspects in both parameter estimation and prediction. The proposed method is based on maximising the marginal likelihood of the mixed-effects model.

It might be important to add that dependence is not the only peculiar characteristic of spatial socio-economic data but that the non-stationarity, or heterogeneity, of the model parameters is often present in the empirical analysis (Billé *et al.*, 2017). These parameters are usually locally approximately homogeneous, but they rapidly change as soon as natural, social or economic borders are crossed. As widely demonstrated by Anselin (1988), these two different typologies of spatial effects, autocorrelation and heterogeneity, both influence spatial economic data modelling. Neglecting one of them can lead to a misspecification of the statistical model and, thus, to biased parameter estimates. As the hypothesis of stationarity of the model parameters across space is usually violated by the empirical evidence, in the near future, we should foresee the introduction of a more cautionary conjecture regarding the heterogeneity of the model that should use some form of space varying parameters. For a detailed discussion and some solutions in the field of small area estimates, see Chandra *et al.* (2015).

Through the use of real and simulated data, we have shown how the proposed MML approach can provide much more efficient estimates than classical models that, although they may introduce dependency in the data, they do not effectively use it in considering the problem of the existence of missing data in the target variable. The experimental results suggest that, as was expected, this efficiency is particularly noticeable when we estimate the areas with missing data. However, greater efficiency is present, albeit to a lesser extent, even in those areas where it is possible to have a direct estimate.

It is also appropriate to underline the empirical evidence deriving from the simulated experiment: the spatial dependence plays a fundamental role in the application of the proposed method to any phenomenon observed through a real survey. This parameter allows to obtain information from the sampled neighbouring areas in order to be able to estimate more efficiently those that are out-of-sample. To increase the empirical evidence on the capabilities of the proposed method, it would be appropriate, if not necessary, to broaden the field of applications of the proposed method to phenomena, not necessarily socio-economic, with spatial distributions less concentrated in some sub-populations and less affected by the spatial trend such as, for example, those that arise from environmental surveys.

Finally, it is important to note that some issues relating to estimates of unemployment rate data have been detailed, such as estimates of confidence intervals, while others have been deliberately left out. We refer in particular to the well-known problem of benchmarking of the estimates that, as some argue (Pfeffermann & Tiller, 2006), can also protect estimates from potential model specification errors and be useful for reducing the over-shrinkage of model-based small area estimates. We did not address this issue in our application, because we did not find substantial differences between the estimates that were constrained to the regional estimates and those that were not. Their introduction would therefore have burdened the formal discussion without adding any practical contribution to the efficiency of the method.

ACKNOWLEDGEMENTS

Open access publishing facilitated by Universita degli Studi di Pisa, as part of the Wiley - CRUI-CARE agreement.

REFERENCES

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Kluwer Academic Publishers: Dordrecht.

Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2014). *Hierarchical modeling and analysis for spatial data*, 2. Chapman and Hall/CRC: New York.

Bell (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. In *Proceedings of the Survey Research Section American Statistical Association*, pp. 327–334.

Benedetti, R. & Filipponi, D. 2010. Estimation of land cover parameters when some covariates are missing. In *Agricultural Survey Methods*, John Wiley & Sons: Chichester, pp. 213-230.

Benedetti, R., Suesse, T. & Piersimoni, F. (2020). Spatial auto-correlation and auto-regressive models estimation from sample survey data. *Biometrical Journal*, **62**, 1494–1507.

Bertarelli, G., Schirripa, F., Salvati, N. & Pratesi, M. 2021. Small area estimation of agricultural data. In *In Spatial Econometric Methods in Agricultural Economics Using R*, Eds. Postiglione, P., Benedetti, R. & Piersimoni, F., CRC PRESS, Taylor & Francis Inc., pp. 202–233.

Besag, J. (1974). Spatial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, **36**, 192–236.

Billé, A.G., Benedetti, R. & Postiglione, P. (2017). A two-step approach to account for unobserved spatial heterogeneity. *Spatial Economic Analysis*, **12**(4), 452–471.

Burgard, J.P., Morales, D. & Wolwer, A. (2022). Small area estimation of socioeconomic indicators for sampled and unsampled domains. *AStA Advances in Statistical Analysis*, **106**, 287–314.

Carter, G. & Rolph, J. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, **69**, 880–885.

Casas-Cordero, C., Encina, J. & Lahiri, P 2016. Poverty mapping for the Chilean comunas. In *In Analysis of Poverty Data by Small Area Estimation*, Ed. Pratesi, M., John Wiley & Sons: Hoboken, pp. 379–403.

Chambers, R.L., Steel, D.G., Wang, S. & Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys*. Chapman and Hall/CRC: New York.

Chandra, H., Salvati, N. & Chambers, R. (2015). A spatially nonstationary Fay–Herriot model for small area estimation. *Journal of Survey Statistics and Methodology*, **3**, 109–135.

Chung, C.H. & Datta, S.G. (2022). Bayesian spatial models for estimating means of sampled and non-sampled small areas. *Survey Methodology*, **48**, 463–489.

Cressie, N. (1989). Empirical Bayes estimation of undercount in the decennial Census. *Journal of the American Statistical Association*, **84**(408), 1033–1044.

Cressie, N. (1991). Small-area prediction of undercount using the general linear model. In *Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada*, pp. 93–105.

Cressie, N. (1993). *Statistics for Spatial Data Revised Edition*. John Wiley & Sons: New York.

D'Alò, M., Falorsi, S. & Solari, F. (2017). Space-time unit-level EBLUP for large data sets. *Journal of Official Statistics*, **33**, 61–77.

Datta, G.S., Rao, J.N.K. & Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area-level model. *Biometrika*, **92**, 183–196.

Efron, B. & Morris, C. (1975). Data analysis using steins estimate and its generalizations. *Journal of the American Statistical Association*, **70**, 311–319.

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.

Ha, N.S., Lahiri, P. & Parsons, V. (2014). Methods and results for small area estimation using smoking data from the 2008 National Health Interview Survey. *Statistics in Medicine*, **33**, 3932–3945.

Harrison, D. & Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, **5**, 81–102.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–340.

Henderson, C.R. (1950). Estimation of genetic parameters (abstract). *The Annals of Mathematical Statistics*, **21**, 309–310.

Hidiroglou, M.A., Beaumont, J.-F. & Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, **45**(1), 101–126.
Jiang, J. & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, **53**, 217–243.
Jiang, J. & Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, **15**, 1–96.
Jiang, J., Lahiri, P., Wan, S.-M. & Wu, C.-H (2001). Jackknifing in the Fay–Herriot model with an example. In *In Proc. Sem. Funding Opportunity in Survey Research, Washington DC: Bureau of Labor Statistics*, pp. 75–97.
Kreutzmann, A.K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M. & Tzavidis, N. (2019). The R package EMDI for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, **91**(7), 1–33.
Lesage, J. & Pace, R. (2004). Models for spatially dependent missing data. *Journal of Real Estate Finance and Economics*, **29**(2), 223–254.
Lindstrom, M.J. & Bates, D.M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**(404), 1014–1022.
Little, R.J.A. & Rubin, D.B. (2019). *Statistical Analysis With Missing Data*, 3rd ed. John Wiley & Sons: Hoboken.
Liu, B., Lahiri, P. & Kalton, G. (2014). Hierarchical Bayes modeling of survey-weighted small area proportions. *Survey Methodology*, **40**, 1–13.
Longford, N.T (2004). Missing data and small area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society, Series A*, **167**, 341–373. New York: Springer.
Longford, N.T. (2005). *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*. Springer: New York.
Marhuenda, Y., Molina, I. & Morales, D. (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics and Data Analysis*, **58**, 308–325.
Marino, F., Ranalli, M.G., Salvati, N. & Alfò, M. (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, **13**, 1166–1197.
Molina, I. & Marhuenda, Y. (2015). SAE: An R package for small area estimation. *The R Journal*, **7**(1), 81–98.
Molina, I. & Rao, J.N.K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, **38** (3), 369–385.
Molina, I., Salvati, N. & Pratesi, M. (2009). Bootstrap for estimating the MSE of the spatial EBLUP. *Computational Statistics*, **24**, 441–458.
Moura, F. & Migon, H.S. (2002). Bayesian spatial models for small area estimation of proportions. *Statistical Modelling*, **2**(3), 183–201.
Panzera, D., Benedetti, R. & Postiglione, P. (2016). A Bayesian approach to parameter estimation in the presence of spatial missing data. *Spatial Economic Analysis*, **11**(2), 201–218.
Petrucci, M. & Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural Biological, and Environmental Statistics*, **11**(2), 169–182.
Pfeffermann, D. & Tiller, R. (2006). Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, **101**(476), 1387–1397.
Porter, A.T., Holan, S.H., Wikle, C.K. & Cressie, N. (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. *Spatial Statistics*, **10**, 27–27.
Porter, A.T., Wikle, C.K. & Holan, S.H. (2015). Small area estimation via multivariate Fay-Herriot models with latent spatial dependence. *Australian & New Zealand Journal of Statistics*, **57**, 15–29.
Prasad, N.G.N. & Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, **85**, 163–171.
Pratesi, M., Marchetti, S., Giusti, C. & Salvati, N (2023). The use of spatial information in area-level models: An evaluation based on auxiliary data availability. *Calcutta Statistical Association Bulletin*, **75**, 155–172. https://doi.org/10.1177/00080683231198589
Pratesi, M. & Salvati, N. (2008). Small area estimation: The EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, **17**(1), 113–141.
Pratesi, M. & Salvati, N. (2009). Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics*, **25**(1), 37–53.
Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. & Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, **102**, 474–486.
Rao, J.N.K. & Molina, I. (2015). *Small Area Estimation*, 2nd Edition. John Wiley & Sons: New York.
Reich, B.J. & Bandyopadhyay, D. (2010). A latent factor model for spatial data with informative missingness. *Annals of Applied Statistics*, **4**, 439–459.
Saei, A. & Chambers, R 2005. Out-of-sample estimation for small areas using area level data. Working Paper M05/11, Southampton Statistical Sciences Research Insitute, University of Southampton, UK. Downloadable at https://www.

ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/subnational-labour/model-based-estimates-of-ilo-unemployment-for-lad-uas-in-great-britain---guide-for-users.pdf

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Taylor & Francis Inc.

Schmid, T., Bruckschen, F., Salvati, N. & Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in Senegal. *Journal of the Royal Statistical Society Series A*, **180**, 1163–1190.

Singh, B.B., Shukla, G.K. & Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, **31**(2), 183–195.

Speckman, P.L. & Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, **90**, 289–302.

Suesse, T. (2018). Marginal maximum likelihood estimation of SAR models with missing data. *Computational Statistics and Data Analysis*, **120**, 98–110.

Suesse, T. & Zammit-Mangion, A. (2017). Computational aspects of the EM algorithm for spatial econometric models with missing data. *Journal of Statistical Computation and Simulation*, **87**(9), 1767–1786.

Sun, D., Tsutakawa, R.K. & Speckman, P.L. (1999). Posterior distribution of hierarchical models using CAR (1) distributions. *Biometrika*, **86**, 341–350.

Torabi, M. (2012). Hierarchical Bayes estimation of spatial statistics for rates. *Journal of Statistical Planning and Inference*, **142**, 358–365.

Ver Hoef, J.M., Hanks, E.M. & Hooten, M.B. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spatial Statistics*, **25**, 68–85.

Wolter, K. (1985). *Introduction to Variance Estimation*. Springer-Verlag: New York.

You, Y. & Chapman, B. (2006). Using area level models and estimated sampling variances. *Survey Methodology*, **32** (1), 97–103.

You, Y. & Zhou, Q. (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, **37**, 25–36.

[*Received October 2023; accepted August 2024*]