



Small area prediction of proportions and counts under a spatial Poisson mixed model

Miguel Boubeta¹ · María José Lombardía² · Domingo Morales³

Accepted: 30 September 2023
© The Author(s) 2023

Abstract

This paper introduces an area-level Poisson mixed model with SAR(1) spatially correlated random effects. Small area predictors of proportions and counts are derived from the new model and the corresponding mean squared errors are estimated by parametric bootstrap. The behaviour of the introduced predictors is empirically investigated by running model-based simulation experiments. An application to real data from the Spanish living conditions survey of Galicia (Spain) is given. The target is the estimation of domain proportions of women under the poverty line.

Keywords Small area estimation · Area-level models · Spatial correlation · Count data · Bootstrap · Living conditions survey · poverty proportion

AMS Subject Classification 62E30 · 62J12

1 Introduction

This paper introduces statistical methodology for estimating proportions and counts, with applications to poverty mapping in areas of Galicia. This is an autonomous community in the northwest of Spain with an economic activity strongly related to natural resources. The beginning of the 21th century has been characterized by the

Supported by the Instituto Galego de Estatística, by MICINN Grants PID2020-113578RB-I00 and PGC2018-096840-B-I00, by the Generalitat Valenciana Grant PROMETEO/2021/063 and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C 2020/14), and by GAIN (Galician Innovation Agency) and the Regional Ministry of Economy, Employment and Industry Grant COV20/00604 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01, all of them through the ERDF.

✉ María José Lombardía
maria.jose.lombardia@udc.es

¹ Universidade da Coruña, A Coruña, Spain

² CITIC, Universidade da Coruña, A Coruña, Spain

³ Centro de Investigación Operativa, Universidad Miguel Hernández de Elche, Elche, Spain

increase in the differences between the areas of the inner zone, which are poorer, and the coastal areas, which present greater development. One of the Spanish statistical sources to monitor poverty indicators is the living conditions survey. Unfortunately, the sample sizes of this survey do not allow obtaining precise direct estimators for the counties of Galicia. Small area estimation (SAE) deals with this problem by introducing indirect predictors based on statistical models. The monographs of Rao and Molina (2015); Pratesi (2016) and Morales et al. (2021) are introductory texts to SAE. Recent methodology for SAE is described in the review papers of Jiang and Lahiri (2006); Rao (2008); Pfeiffermann (2013) and Sugawara and Kubokawa (2020).

Developing statistical methodologies to model counting data and to predict domain counts and proportions in small areas is important to understand the phenomena studied (like poverty) and, consequently, to make decisions about public policies. This manuscript introduces a spatial extensions of the basic area-level Poisson mixed model, for better fitting the needs of real data and giving rise to increasingly complex and realistic models. The main contribution is the introduction of small area predictors that incorporates information from other domains, auxiliary variables and spatial correlation.

Without taking into account the spatial correlation at the domain level, the statistical literature on area-level mixed models for count data has interesting contributions to SAE. We can find univariate and multivariate models for count data. In the first case, regression models with Poisson distribution have been widely used. Some contributions in this field are Ghosh et al. (1998), Trevisani and Torelli (2017), Boubeta et al. (2016, 2017) or Reluga et al. (2021). In the second case, the multivariate models for the count data additionally borrow the strength of the correlation between the target variables to introduce new predictors. This was done by Ferrante and Trivisano (2010), López-Vizcaíno et al. (2013, 2015), Esteban et al. (2020) or Burgard et al. (2022), among others. These models link all the domains to enhance the estimation at a particular area, that is, they borrow strength from other areas. The models have random effects taking into account the between-domain variability that is not explained by the auxiliary variables, but they assume that the domain random effects are independent. However, in socioeconomic, environmental and epidemiological applications, estimates for areas that are spatially close may be more alike than estimates for areas that are further apart. In fact, Cressie (93) shows that not employing spatial models may lead to inefficient inferences when the auxiliary variables does not explain the spatial correlation of the study variable.

In other fields of statistics than SAE, such as econometrics or epidemiology, spatial models for count data have wide applicability. See for example the papers of Anselin (2001), Wakefield (2007), Mohebbi et al. (2011) or Glaser (2017) and references therein, or the books of LeSage and Pace (2009), Dubé and Legros (2014) and Banerjee et al. (2015). Among those models are the “spatial autocorrelation” or “simultaneous autoregressive” (SAR) Poisson regression models. However, SAR Poisson models do not play the most relevant role in those areas, unlike normal models, because there is no direct functional relationship between the dependent and the explanatory variables. On the other hand, the area-level approach to SAE looks for the relation between the intensity parameter (conditional expectation of

the dependent variable) and the regressors. The reason is that the dependent variable is an estimator obtained from unit-level survey data and not the area-level population parameter of interest. In the SAE setup, population parameters (such as poverty proportions) are directly related to the conditional expectation of the dependent variable.

In SAE, modelling the spatial correlation between data from different areas allows to borrow even more strength from the areas. This recommendation was applied to the basic Fay–Herriot model by Singh et al. (2005). Later, several authors have proposed new spatial area-level linear mixed models. Petrucci and Salvati (2006); Pratesi and Salvati (2008); Molina et al. (2009); Marhuenda et al. (2013) and Chandra et al. (2015) consider linear mixed models (LMM) that extend the Fay–Herriot model. Most of these papers assume that area effects follow a simultaneously autoregressive process of order 1 or SAR(1).

In the Bayesian framework, Moura and Migon (2002) and You and Zhou (2011) consider spatial stationary mixed models, Sugasawa et al. (2020) study an empirical Bayesian estimation method with spatially non-stationary hyperparameters for area-level discrete and continuous data having a natural exponential family distribution. Arima et al. (2012) propose a full Bayesian separable spatio-temporal hierarchical Bayesian model, which allows the integration of missing data imputation and pollutant concentration prediction. Choi et al. (2011) apply the spatio-temporal models to study chronic obstructive pulmonary disease at county level in Georgia.

Concerning nonparametric and robust methods, Opsomer et al. (2008) give a small area estimation procedure using penalized spline regression with applications to spatially correlated data. Ugarte et al. (2006) and Ugarte et al. (2010) study the geographical distribution of mortality risk using small area techniques and penalized splines. Chandra et al. (2012) introduce a geographical weighted empirical best linear unbiased predictor for small area averages. Baldermann et al. (2016) describe robust SAE methods under spatial non-stationarity linear mixed models. Chandra et al. (2017) introduce small area predictors of counts under a non-stationary spatial model. Chandra et al. (2018) develop a geographically weighted regression extension of the logistic-normal and the Poisson-normal generalized linear mixed models (GLMM) allowing for spatial nonstationarity.

The above cited papers introduce SAE procedures that borrow strength from spatial correlations. Assuming the frequentist parametric inference setup, they mainly apply spatial LMMs to small area estimation. However, few of them deal with empirical best predictors (EBP) under spatial GLMMs. This work partially fills that gap and studies an area-level Poisson mixed model containing SAR(1) spatially correlated domain effects. The final target is the estimation of domain counts and proportions.

This paper has a double starting point. On the one hand, it modifies the spatial LMM of Pratesi and Salvati (2008) by changing the conditional distribution of the target variable from normal to Poisson. On the other hand, it extends the model of Boubeta et al. (2016) by including random effects with SAR(1) distribution. The proposed model and the predictors derived from it are new. There is also no statistical software that allows estimating the parameters of the new model. For this reason, we adapt and implement the method of simulated moments (MSM) studied by Jiang (1998). Further, we give

approximate EBPs to avoid the problem of approximating multiple integrals and we empirically investigate their behavior.

The manuscript is organized as follows. Section 2 introduces the model and a fitting algorithm based on the MSM method. Section 3 gives empirical best and plug-in predictors of domain proportions and counts, provides approximations to the integrals that appear in the predictor formulas and proposes a parametric bootstrap method for estimating their mean squared errors (MSE). Section 4 empirically investigates the introduced predictors by means of simulation experiments. Section 5 gives a relevant application to real data in the socio-economic field. The target is the estimation of women poverty proportion in counties of Galicia. Section 6 collects the main conclusions. Finally, the paper contains a supplementary material with some appendixes. Appendix A provides the nonlinear system of MSM equations that must be solved to estimate the model parameters. Appendixes B and C describe algorithms to calculate the penalized maximum likelihood (PQL) and the maximum likelihood (ML) estimators of the model parameters. Appendix D shows two bootstrap algorithms to test hypotheses about the variance and correlation parameters. Appendix E gives some additional simulation results by considering deviations from normality in the data generating processes. Appendix F presents complementary results to the application to real data, including the analysis of the goodness-of-fit of alternative Poisson regression models and the use of EBLUPs based on a spatial Fay–Herriot model.

2 The model

This section introduces an area-level Poisson mixed model in the context of spatial correlation. Specifically, it assumes a SAR(1) process on the random effects. Let us consider a population partitioned into D domains and let us denote each particular domain by d , $d = 1, \dots, D$. Let $\mathbf{v} = (v_1, \dots, v_D)'$ be a vector of spatially correlated random effects following a SAR(1) process with unknown autoregression parameter ρ and known proximity matrix \mathbf{W} . This means that the vector of random effects \mathbf{v} fulfills the linear combination

$$\mathbf{v} = \rho \mathbf{W} \mathbf{v} + \mathbf{u}, \quad (2.1)$$

where $\mathbf{u} \sim N_D(\mathbf{0}, \mathbf{I}_D)$, $\mathbf{0}$ is the $D \times 1$ zero vector and \mathbf{I}_D denotes the $D \times D$ identity matrix. Assuming that $(\mathbf{I}_D - \rho \mathbf{W})$ is non-singular, the equation (2.1) can be expressed as

$$\mathbf{v} = (\mathbf{I}_D - \rho \mathbf{W})^{-1} \mathbf{u}. \quad (2.2)$$

For the proximity matrix \mathbf{W} , we assume that it is row stochastic, i.e. the elements of each row are positive and add up to 1. Then, the autoregression parameter ρ is a correlation and is called spatial autocorrelation parameter. Some of the most used proximity matrices are based on: (i) common borders, (ii) distances and (iii) k -nearest neighbours. In all cases, the proximity matrix \mathbf{W} is obtained from an original proximity matrix \mathbf{W}^0 with diagonal elements equal to zero and remaining entries depending on the employed option. In option (i), the non diagonal elements of \mathbf{W}^0

are equal to 1 when the two domains corresponding to the row and the column indices are regarded as neighbours and zero otherwise. In option (ii), the nondiagonal elements of the proximity matrix \mathbf{W}^0 are defined by applying a monotonically decreasing function to the domain distances; for example, by using the inverse function. Finally, the non diagonal elements of \mathbf{W}^0 in option (iii) are 1 if they correspond to the k -nearest neighbours of a given domain and zero otherwise. For each option, the row standardization is carried out by dividing each entry of \mathbf{W}^0 by the sum of the elements in its row. Consequently, \mathbf{W} is row stochastic. Spatial weights are usually standardized by row, especially in binary weights strategies (i) and (ii). Row standardization is used to create proportional weights in cases where domains have an unequal number of neighbours. Row standardization involves dividing the weight of each neighbour of a domain by the sum of the weights of all neighbours of that domain. For area-level models, the row standardization is recommended when the distribution of the target variable is potentially influenced due to the sampling design or to the aggregation scheme. More information on row standardization can be found in the monograph Viton 2010.

Equation (2.2) implies that $\mathbf{v} \sim N_D(\mathbf{0}, \mathbf{\Gamma}(\rho))$, where

$$\mathbf{\Gamma}(\rho) = (\gamma_{d_1, d_2}(\rho))_{d_1, d_2=1, \dots, D} = \mathbf{C}^{-1}(\rho) \tag{2.3}$$

and $\mathbf{C}(\rho) = (\mathbf{I}_D - \rho\mathbf{W})'(\mathbf{I}_D - \rho\mathbf{W})$. Therefore, the density function of the random effects is

$$f_{\mathbf{v}}(\mathbf{v}) = (2\pi)^{-D/2} |\mathbf{\Gamma}(\rho)|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{v}' \mathbf{\Gamma}^{-1}(\rho) \mathbf{v} \right\}.$$

The vector of response variables $\mathbf{y} = (y_1, \dots, y_D)'$ follows an area-level Poisson mixed model with a SAR(1) vector of domain random effects \mathbf{v} if its components y_1, \dots, y_D are independent conditionally on \mathbf{v} and the conditioned distribution of y_d , given \mathbf{v} , is

$$y_d | \mathbf{v} \sim y_d | v_d \sim \text{Pois}(\mu_d), \quad d = 1, \dots, D,$$

where μ_d denotes the mean of the Poisson distribution. We assume that μ_d can be expressed as $v_d p_d$, where v_d and p_d are size and probability parameters respectively. We find two advantages of using a Poisson mixed model instead of a binomial mixed model. The first is that we can avoid calculating combinatorial numbers with values outside the range of the computer. The second is that we can explicitly compute some integrals and this avoids the need to use approximation methods. This is particularly important when applying the MSM method. Besides, by the nature of our problem, v_d takes large values and p_d small values. Then, everything points to a good behavior of the Poisson model in addition to its computational advantages. As v_d is assumed to be known, the Poisson parameter, μ_d , is determined if and only if one knows the parameter p_d . In what follows, we will refer to p_d as target parameter. To finish the definition of the area-level Poisson mixed model with SAR(1) random effects (Model S1), we express the natural parameter $\log \mu_d$ in terms of a set of q covariates $\mathbf{x}_d = (x_{d1}, \dots, x_{dq})$ and the random effect v_d , i.e.

$$\log \mu_d = \log v_d + \log p_d = \log v_d + \mathbf{x}_d \boldsymbol{\beta} + \phi v_d, \quad d = 1, \dots, D,$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$ and ϕ are the regression and standard deviation parameters respectively. We denote the vector of all model parameters by $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi, \rho)'$.

Under Model S1, it holds that

$$\mathbb{P}(y_d | \mathbf{v}) = \mathbb{P}(y_d | v_d) = \frac{1}{y_d!} \exp\{-v_d p_d\} v_d^{y_d} p_d^{y_d},$$

where the target parameter is $p_d = \exp\{\mathbf{x}_d \boldsymbol{\beta} + \phi v_d\}$. The probability function (likelihood) of the response variable \mathbf{y} is

$$\mathbb{P}(\mathbf{y}) = \int_{\mathbb{R}^D} \mathbb{P}(\mathbf{y} | \mathbf{v}) f_{\mathbf{v}}(\mathbf{v}) d\mathbf{v} = \int_{\mathbb{R}^D} \prod_{d=1}^D \mathbb{P}(y_d | v_d) f_{\mathbf{v}}(\mathbf{v}) d\mathbf{v} = \int_{\mathbb{R}^D} \psi(\mathbf{y}, \mathbf{v}) d\mathbf{v}, \quad (2.4)$$

where

$$\begin{aligned} \psi(\mathbf{y}, \mathbf{v}) &= f_{\mathbf{v}}(\mathbf{v}) \prod_{d=1}^D \frac{\exp\{-v_d p_d\} v_d^{y_d} \exp\{y_d(\mathbf{x}_d \boldsymbol{\beta} + \phi v_d)\}}{y_d!} \\ &= f_{\mathbf{v}}(\mathbf{v}) \left(\prod_{d=1}^D y_d! \right)^{-1} \exp \left\{ \sum_{d=1}^D \left\{ -v_d \exp\{\mathbf{x}_d \boldsymbol{\beta} + \phi v_d\} + y_d \log v_d \right\} \right\} \\ &\quad \cdot \exp \left\{ \sum_{k=1}^p \left(\sum_{d=1}^D y_d x_{dk} \right) \beta_k + \phi \sum_{d=1}^D y_d v_d \right\}. \end{aligned}$$

To estimate the parameters of Model S1, the natural procedure would be to calculate the ML estimators. However, the likelihood (2.4) is a multiple integral over \mathbb{R}^D , which makes it necessary to combine an integral approximation procedure with another for maximization of multivariate functions. Since the size of the integral to be approximated coincides with the number of domains (sample size in the aggregated data set on which the area models are fitted), the optimization algorithms are inefficient in terms of stability (convergence) and computation time. For the sake of completeness of exposition, Appendix C of the supplementary material presents a likelihood optimization algorithm that combines the Laplace approximation of the multiple integral with the Newton–Raphson optimization algorithm. For the reasons stated, this ML-Laplace algorithm has not been selected to estimate the parameters of Model S1.

The scientific literature on GLMMs presents alternative procedures to the ML method. One of these methods consists of maximizing the joint likelihood of the fixed and random effects of the model, for each optimal selection of the variance and correlation parameters. This procedure, which we call PQL, alternately combines two optimization algorithms, the first on $\boldsymbol{\beta}$ and \mathbf{u} and the second on ϕ and ρ , such that the output of one algorithm feeds the input from the other and vice versa. It is a procedure that calculates ML estimators in the case of LMMs but does not guarantee consistent estimators for GLMMs. The method has a lower

computational cost than the ML-Laplace algorithm, but it presents instability problems (lack of convergence) in complex models, such as those of spatial correlation. Appendix B of the supplementary material describes the PQL method. Due to the mentioned drawbacks, we have not selected the PQL algorithm to estimate the parameters of the S1 model.

Jiang (1998) proposed the MSM method as a computationally efficient alternative to procedures based on maximizing likelihood. Under regularity conditions, Jiang (1998) proved that the MSM method gives consistent estimators of GLMM parameters. In addition, it is a computational stable and economic method that only requires solving a non linear system of $p + 2$ equations. This is why, we have decided to apply the MSM method to estimate the parameters of Model S1. In what follows, we describe how the MSM method estimates the vector of parameters, θ , of Model S1.

A natural set of equations for applying the MSM algorithm is

$$\begin{aligned}
 0 &= f_k(\theta) = M_k(\theta) - \hat{M}_k = \frac{1}{D} \sum_{d=1}^D \mathbb{E}_\theta[y_d]x_{dk} - \frac{1}{D} \sum_{d=1}^D y_d x_{dk}, \quad k = 1, \dots, p, \\
 0 &= f_{p+1}(\theta) = M_{p+1}(\theta) - \hat{M}_{p+1} = \frac{1}{D} \sum_{d=1}^D \mathbb{E}_\theta[y_d^2] - \frac{1}{D} \sum_{d=1}^D y_d^2, \\
 0 &= f_{p+2}(\theta) = M_{p+2}(\theta) - \hat{M}_{p+2} = \frac{1}{D(D-1)} \sum_{d_1 \neq d_2}^D \mathbb{E}_\theta[y_{d_1} y_{d_2}] - \frac{1}{D(D-1)} \sum_{d_1 \neq d_2}^D y_{d_1} y_{d_2},
 \end{aligned}
 \tag{2.5}$$

where $\sum_{d_1 \neq d_2}^D$ is a double sum with $d_1, d_2 = 1, \dots, D, d_1 \neq d_2$, in the $(p + 2)$ th equation. The MSM estimator of θ is the solution of the system of nonlinear equations (2.5). For solving this system, we may apply the Newton–Raphson algorithm with updating formula

$$\theta^{(m+1)} = \theta^{(m)} - \mathbf{H}^{-1}(\theta^{(m)})\mathbf{f}(\theta^{(m)}),
 \tag{2.6}$$

where

$$\theta = \underset{1 \leq k \leq p+2}{\text{col}} (\theta_k), \quad \mathbf{f}(\theta) = \underset{1 \leq k \leq p+2}{\text{col}} (f_k(\theta)), \quad \mathbf{H}(\theta) = \left(\frac{\partial f_k(\theta)}{\partial \theta_r} \right)_{k,r=1, \dots, p+2},
 \tag{2.7}$$

and the components of the vector θ are $\theta_1 = \beta_1, \dots, \theta_p = \beta_p, \theta_{p+1} = \phi, \theta_{p+2} = \rho$. Appendix A contains the calculations of the expectations appearing in $\mathbf{f}(\theta)$ and $\mathbf{H}(\theta)$. Appendixes B and C describe the PQL and the ML methods, and discuss the pros and cons of the three fitting algorithms.

As algorithm starting points for β and ϕ , we can take maximum likelihood estimates under Model S1 with $\rho = 0$ (denoted by Model 1). In the case of independent random effects ($\rho = 0$), we have functions to fit the model in different programming languages. For example, we can employ the glmer function of R. Concerning the parameter ρ , we take the Moran’s I measure of spatial autocorrelation based on the Pearson residuals obtained under Model 1, i.e.

$$I = \frac{D}{\sum_{d_1=1}^D \sum_{d_2=1}^D w_{d_1 d_2}} \frac{\sum_{d_1=1}^D \sum_{d_2=1}^D w_{d_1 d_2} (\tilde{e}_{d_1} - \bar{e})(\tilde{e}_{d_2} - \bar{e})}{\sum_{d=1}^D (\tilde{e}_d - \bar{e})^2}, \quad (2.8)$$

where \tilde{e}_d , $d = 1, \dots, D$, denote the Pearson residuals under Model 1, $\bar{e} = \frac{1}{D} \sum_{d=1}^D \tilde{e}_d$ and $w_{d_1 d_2}$, $d_1, d_2 = 1, \dots, D$, are the elements of the proximity matrix \mathbf{W} . Moran's I measures spatial autocorrelation based on the locations and values of the target variable simultaneously. Given a set of domains and a target variable, it evaluates whether that variable has a clustered, dispersed, or random pattern. Moran's I is also a statistic that allows testing the null hypothesis of no spatial correlation. The null hypothesis establishes that the variable being analysed is randomly distributed among the entities in the study area; that is, the spatial processes that promote the observed pattern of values is a random mechanism.

3 The predictors

This section gives the empirical best and plug-in predictors of p_d under Model S1. We focus on the calculation of the EBP of p_d , given its relationship with μ_d , and assuming that the size parameter v_d is known. The corresponding prediction of the expected count μ_d is straightforward. The EBP of p_d is obtained from the best predictor (BP) by replacing the vector of model parameters $\boldsymbol{\theta}$ by an estimator $\hat{\boldsymbol{\theta}}$. The BP of p_d is the unbiased predictor that minimizes the MSE and it is given by

$$\hat{p}_d(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[p_d|\mathbf{y}] = \frac{\int_{\mathbb{R}^D} \exp\{\mathbf{x}_d \boldsymbol{\beta} + \phi v_d\} \prod_{i=1}^D \mathbb{P}(y_i|v_i) f_v(\mathbf{v}) d\mathbf{v}}{\int_{\mathbb{R}^D} \prod_{i=1}^D \mathbb{P}(y_i|v_i) f_v(\mathbf{v}) d\mathbf{v}} = \frac{N_d(\mathbf{y}, \boldsymbol{\theta})}{D_d(\mathbf{y}, \boldsymbol{\theta})}, \quad (3.1)$$

where

$$N_d(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}^D} \exp \left\{ \sum_{i=1}^D \left[(y_i + \delta_{id})(\mathbf{x}_i \boldsymbol{\beta} + \phi v_i) - v_i \exp \{ \mathbf{x}_i \boldsymbol{\beta} + \phi v_i \} \right] \right\} f_v(\mathbf{v}) d\mathbf{v},$$

$$D_d(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}^D} \exp \left\{ \sum_{i=1}^D \left[y_i (\mathbf{x}_i \boldsymbol{\beta} + \phi v_i) - v_i \exp \{ \mathbf{x}_i \boldsymbol{\beta} + \phi v_i \} \right] \right\} f_v(\mathbf{v}) d\mathbf{v},$$

and δ_{ij} is the Kronecker's delta; i.e. $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. The EBP of p_d is $\hat{p}_d(\hat{\boldsymbol{\theta}})$ and it can be approximated by using an antithetic Monte Carlo algorithm. The steps are:

1. Generate $\mathbf{v}^{(\ell)}$ i.i.d. $N_D(\mathbf{0}, \boldsymbol{\Gamma}(\hat{\rho}))$ and calculate their antithetics $\mathbf{v}^{(L+\ell)} = -\mathbf{v}^{(\ell)}$, $\ell = 1, \dots, L$.
2. Calculate

$$\hat{p}_d(\hat{\boldsymbol{\theta}}) = \hat{N}_d / \hat{D}_d, \quad (3.2)$$

where

$$\hat{N}_d = \frac{1}{2L} \sum_{\ell=1}^{2L} \exp \left\{ \sum_{i=1}^D \left[(y_i + \delta_{id})(\mathbf{x}_i \hat{\boldsymbol{\beta}} + \hat{\phi} v_i^{(\ell)}) - v_i \exp \{ \mathbf{x}_i \hat{\boldsymbol{\beta}} + \hat{\phi} v_i^{(\ell)} \} \right] \right\},$$

$$\hat{D}_d = \frac{1}{2L} \sum_{\ell=1}^{2L} \exp \left\{ \sum_{i=1}^D \left[y_i (\mathbf{x}_i \hat{\boldsymbol{\beta}} + \hat{\phi} v_i^{(\ell)}) - v_i \exp \{ \mathbf{x}_i \hat{\boldsymbol{\beta}} + \hat{\phi} v_i^{(\ell)} \} \right] \right\}.$$

As the BP of p_d involves high-dimensional integrals on \mathbb{R}^D , we propose a computationally less demanding approximation. For that, let us divide the response variable \mathbf{y} and the vector of random effects \mathbf{v} into two parts (y_d, \mathbf{y}_{d-}) and (v_d, \mathbf{v}_{d-}) , where $\mathbf{y}_{d-} = \text{col}_{1 \leq i \leq D, i \neq d} (y_i)$ and $\mathbf{v}_{d-} = \text{col}_{1 \leq i \leq D, i \neq d} (v_i)$. The conditional distribution of \mathbf{y} , given \mathbf{v} , is

$$\mathbb{P}(\mathbf{y}|\mathbf{v}) = \prod_{i=1}^D \mathbb{P}(y_i|v_i) = \mathbb{P}(y_d|v_d) \prod_{i=1, i \neq d}^D \mathbb{P}(y_i|v_i) = \mathbb{P}(y_d|v_d) \mathbb{P}(\mathbf{y}_{d-}|\mathbf{v}_{d-}). \tag{3.3}$$

Using (3.3), the component $D_d(\mathbf{y}, \boldsymbol{\theta})$ in (3.1) can be rewritten as

$$D_d(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}^{D-1}} \mathbb{P}(\mathbf{y}_{d-}|\mathbf{v}_{d-}) f(\mathbf{v}_{d-}|v_d) d\mathbf{v}_{d-} \right] \mathbb{P}(y_d|v_d) f(v_d) dv_d,$$

and since $\mathbb{P}(\mathbf{y}_{d-}|\mathbf{v}_{d-}) f(\mathbf{v}_{d-}|v_d) = \mathbb{P}(\mathbf{y}_{d-}|\mathbf{v}_{d-}, v_d) f(\mathbf{v}_{d-}|v_d)$, the inner integral is

$$\int_{\mathbb{R}^{D-1}} \mathbb{P}(\mathbf{y}_{d-}|\mathbf{v}_{d-}, v_d) f(\mathbf{v}_{d-}|v_d) d\mathbf{v}_{d-} = \mathbb{P}(\mathbf{y}_{d-}|v_d).$$

Therefore

$$D_d(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}} \mathbb{P}(\mathbf{y}_{d-}|v_d) \mathbb{P}(y_d|v_d) f(v_d) dv_d.$$

Taking into account the relationship given in equation (3.3) and reasoning analogously with the component $N_d(\mathbf{y}, \boldsymbol{\theta})$, we have that

$$N_d(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}} \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \phi v_d \} \mathbb{P}(\mathbf{y}_{d-}|v_d) \mathbb{P}(y_d|v_d) f(v_d) dv_d.$$

Under the assumption that $\mathbb{P}(\mathbf{y}_{d-}|v_d) \approx \mathbb{P}(\mathbf{y}_{d-})$, $d = 1, \dots, D$, the BP of p_d , $\hat{p}_d(\boldsymbol{\theta})$, can be approximated by

$$\hat{p}_d^a(\boldsymbol{\theta}) = N_d^a(\mathbf{y}, \boldsymbol{\theta}) / D_d^a(\mathbf{y}, \boldsymbol{\theta}), \tag{3.4}$$

where

$$N_d^a(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}} \exp \{ (y_d + 1)(\mathbf{x}_d \boldsymbol{\beta} + \phi v_d) - v_d \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \phi v_d \} \} f(v_d) dv_d,$$

$$D_d^a(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}} \exp \{ y_d (\mathbf{x}_d \boldsymbol{\beta} + \phi v_d) - v_d \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \phi v_d \} \} f(v_d) dv_d.$$

Equating $\mathbb{P}(y_{d-}|v_d)$ to $\mathbb{P}(y_{d-})$ in the derivation of $\hat{p}_d^a(\boldsymbol{\theta})$ implies violating the dependency of the components of $\mathbf{v} = (v_1, \dots, v_D)'$ for each $d = 1, \dots, D$. This is to say, the approximate BP of p_d treats y_{d-} as independent of y_d . However, this predictor maintains the inner dependency structure of y_{d-} . On the other hand, the integrals involved in the calculation of the approximate BP, $\hat{p}_d^a(\boldsymbol{\theta})$, are on \mathbb{R} and not on \mathbb{R}^D . This is a great computational advantage. As before, the integrals on \mathbb{R} have a complex analytical solution. Therefore, they are approximated by using an antithetic Monte Carlo algorithm analogous to the previous case, but running the new Step 2'.

2'. Approximate the EBP of p_d as

$$\hat{p}_d^a(\hat{\boldsymbol{\theta}}) = \hat{N}_d^a / \hat{D}_d^a \quad (d = 1, \dots, D), \tag{3.5}$$

where

$$\hat{N}_d^a = \frac{1}{2L} \sum_{\ell=1}^{2L} \exp \left\{ (y_d + 1)(\mathbf{x}_d \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(\ell)}) - v_d \exp \left\{ \mathbf{x}_d \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(\ell)} \right\} \right\},$$

$$\hat{D}_d^a = \frac{1}{2L} \sum_{\ell=1}^{2L} \exp \left\{ y_d (\mathbf{x}_d \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(\ell)}) - v_d \exp \left\{ \mathbf{x}_d \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(\ell)} \right\} \right\}.$$

The approximate EBP of p_d depends on the predictor \hat{v}_d of the random effect v_d and of the estimator $\hat{\boldsymbol{\theta}}$ of the model parameters $\boldsymbol{\theta}$, under the assumed Model S1. This is to say, the spatial correlation plays an active role in the construction of $\hat{p}_d^a(\hat{\boldsymbol{\theta}})$ by the approximation the conditional distribution of \mathbf{v} given \mathbf{y} and by the incorporation of \hat{v}_d and $\hat{\boldsymbol{\theta}}$.

Another estimator of p_d , commonly used in this context, is the plug-in predictor. It is obtained by replacing, in the theoretical expression of p_d , the unknown parameters and random effects by their estimators and predictors, i.e.

$$\hat{p}_d^P(\hat{\boldsymbol{\theta}}) = \exp \left\{ \mathbf{x}_d \hat{\boldsymbol{\beta}} + \hat{\phi} \hat{v}_d \right\}. \tag{3.6}$$

Unlikely to the approximate EBP, the plug-in predictor takes advantage of the spatial correlation structure only through the predictors \hat{v}_d and the estimator $\hat{\boldsymbol{\theta}}$, but it does not incorporate information by means of the conditional distribution of \mathbf{v} given \mathbf{y} .

It is important to note that the MSM algorithm only provides estimates for the fixed effects $\boldsymbol{\beta}$, the standard deviation ϕ and the autocorrelation parameter ρ . However, for obtaining \hat{p}_d^P it is necessary to predict the vector of random effects $\mathbf{v} = (v_1, \dots, v_D)$. For this sake, we propose to use their EBPs that are obtained from the corresponding BPs. The BP of v_d is

$$\hat{v}_d(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[v_d|\mathbf{y}] = \frac{\int_{\mathbb{R}^D} v_d \prod_{i=1}^D \mathbb{P}(y_i|v_i) f_v(\mathbf{v}) \, d\mathbf{v}}{\int_{\mathbb{R}^D} \prod_{i=1}^D \mathbb{P}(y_i|v_i) f_v(\mathbf{v}) \, d\mathbf{v}} = \frac{N_{v,d}(\mathbf{y}, \boldsymbol{\theta})}{D_d(\mathbf{y}, \boldsymbol{\theta})},$$

where

$$N_{v,d}(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}^D} v_d \exp \left\{ \sum_{i=1}^D y_i(\mathbf{x}_i\boldsymbol{\beta} + \phi v_i) - v_i \exp\{\mathbf{x}_i\boldsymbol{\beta} + \phi v_i\} \right\} f_v(\mathbf{v}) dv.$$

If the assumption $\mathbb{P}(y_{d-}|v_d) \approx \mathbb{P}(y_{d-})$, holds for $d = 1, \dots, D$, similar mathematical developments as those presented above yield to an approximation to $\hat{v}_d(\boldsymbol{\theta})$ equivalent to that obtained for the target parameter p_d . The approximate BP of v_d is $\hat{v}_d^a(\boldsymbol{\theta}) = N_{v,d}^a(\mathbf{y}, \boldsymbol{\theta})/D_d^a(\mathbf{y}, \boldsymbol{\theta})$, where

$$N_{v,d}^a(\mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}} v_d \exp \{y_d(\mathbf{x}_d\boldsymbol{\beta} + \phi v_d) - v_d \exp\{\mathbf{x}_d\boldsymbol{\beta} + \phi v_d\}\} f(v_d) dv_d.$$

By plugging $\hat{v}_d^a(\boldsymbol{\theta})$ in the formula of p_d , we get a non calculable plug-in predictor. The approximate plug-in-BP predictor of p_d is

$$\hat{p}_d^{pa}(\boldsymbol{\theta}) = \exp \{ \mathbf{x}_d\boldsymbol{\beta} + \phi \hat{v}_d^a(\boldsymbol{\theta}) \}. \tag{3.7}$$

The EBP of v_d is $\hat{v}_d = \hat{v}_d(\hat{\boldsymbol{\theta}})$ and it can be approximated by $\hat{v}_d^a = \hat{v}_d^a(\hat{\boldsymbol{\theta}})$. We propose approximating the analytical integrals by using the same described antithetic Monte Carlo algorithm, but with the new Step 2”.

2”. Calculate $\hat{v}_d^a = \hat{v}_d^a(\hat{\boldsymbol{\theta}}) = \hat{N}_{v,d}^a / \hat{D}_d^a$ ($d = 1, \dots, D$), where

$$\hat{N}_{v,d}^a = \frac{1}{2L} \sum_{\ell=1}^{2L} v_d^{(\ell)} \exp \left\{ y_d(\mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(\ell)}) - v_d \exp \{ \mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(\ell)} \} \right\}.$$

The approximate plug-in predictor of p_d , which is the fully empirical version of (3.7) calculated from \hat{v}_d^a , is

$$\hat{p}_d^{pa}(\hat{\boldsymbol{\theta}}) = \exp \{ \mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{\phi}\hat{v}_d^a \}. \tag{3.8}$$

Concerning the prediction of the Poisson parameter $\mu_d = v_d p_d$ (expected count), the EBP is $\hat{\mu}_d(\hat{\boldsymbol{\theta}}) = v_d \hat{p}_d(\hat{\boldsymbol{\theta}})$, the approximate EBP is $\hat{\mu}_d^a(\hat{\boldsymbol{\theta}}) = v_d \hat{p}_d^a(\hat{\boldsymbol{\theta}})$, the plug-in predictor is $\hat{\mu}_d^P(\hat{\boldsymbol{\theta}}) = v_d \hat{p}_d^P(\hat{\boldsymbol{\theta}})$, and the approximate plug-in predictor is $\hat{\mu}_d^{pa}(\hat{\boldsymbol{\theta}}) = v_d \hat{p}_d^{pa}(\hat{\boldsymbol{\theta}})$.

The MSE is a measure of the accuracy of the predictors of p_d . Boubeta et al. (2016) showed that the analytical approach is computationally demanding in Model 1. This is why we recommend estimating the MSE of \hat{p}_d under Model S1 by using a parametric bootstrap procedure based on the ones given in González-Manteiga et al. (2008, 2010). The steps are

1. Fit Model S1 to the sample and calculate the estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\phi}, \hat{\rho})$.
2. For each domain $d, d = 1, \dots, D$, repeat B times ($b = 1, \dots, B$):
 - i) Generate the bootstrap random effects $\mathbf{v}^{*(b)} = (v_1^{*(b)}, \dots, v_D^{*(b)})' \sim N_D(\mathbf{0}, \boldsymbol{\Gamma}(\hat{\rho}))$, where $\boldsymbol{\Gamma}(\hat{\rho})$ is the plug-in version of the covariance matrix (2.3).
 - ii) Calculate the theoretical bootstrap parameter $p_d^{*(b)} = \exp\{\mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{*(b)}\}$.
 - iii) Generate the response variables $y_d^{*(b)} \sim \text{Pois}(v_d p_d^{*(b)})$.

- iv) Calculate the estimator $\hat{\theta}^{*(b)}$ and the EBP $\hat{p}_d^{*(b)} = \hat{p}_d^{*(b)}(\hat{\theta}^{*(b)})$.
3. Output: $mse^*(\hat{p}_d) = \frac{1}{B} \sum_{b=1}^B (\hat{p}_d^{*(b)} - p_d^{*(b)})^2$.

4 Simulation experiment

This section presents a simulation experiment for investigating the performance of the proposed predictors (EBP, approximate EBP and plug-in) based on Model S1, with SAR(1)-correlated random effects. It also studies the performance of the proposed predictors when model parameters are known (BP and approximate BP and plug-in-BP). In addition, we also consider the corresponding predictors based on Model 1, with independent random effects, to analyse the loss of efficiency when the spatial autocorrelation is not taken into account.

The simulations are based on the application to real data of poverty in Galicia during 2013 (see Sect. 5 for more details). We use the same explanatory variables as those used in the real case, i.e. proportions of unemployed (*lab2*) and of people with university level completed (*edu3*) by counties. We generate independent response variables $y_d|v_d \sim \text{Pois}(v_d p_d)$, where v_d and $p_d = \exp\{\beta_0 + lab_2\beta_1 + edu_3\beta_2 + \phi v_d\}$ are the sample size and target parameter, $d = 1, \dots, D$. The domain sample sizes and the model parameters $\beta_0, \beta_1, \beta_2, \phi$ and ρ are taken from the real data case. That is, we simulate the target variable from the Model S1 selected in the application to Galician data.

The domain random effects, $v_d, d = 1, \dots, D$, are generated according to a SAR(1) process with autocorrelation parameter ρ and the proximity matrix W given in Sect. 5. The number of domains (counties of Galicia) is $D = 49$. As the estimation of the autocorrelation parameter in the application to real data was $\hat{\rho} = 0.324$, we take $\rho = 0.1, 0.3, 0.5$. The number of Monte Carlo iterations (simulated data sets) is $K = 500$.

Tables 1 and 2 present the average across domains of the biases (BIAS) and the root-MSEs (RMSE), both multiplied by 10^2 , of the theoretical predictors BP (3.1), approximate BP (3.4) and approximate plug-in-BP (3.7) based on Model S1. They

Table 1 BIAS ($\times 10^2$) of predictors of p_d based on Models 1 and S1

ρ	Model 1				Model S1					
	BP	P_{BP}	EBP	P_{EBP}	BP	BP^a	P_{BP}^a	EBP	EBP^a	P_{EBP}^a
0.1	0.073	0.098	0.192	0.219	0.081	0.074	0.097	0.237	0.189	0.177
0.2	0.074	0.100	0.192	0.230	0.080	0.073	0.099	0.205	0.198	0.185
0.3	0.078	0.115	0.239	0.271	0.079	0.080	0.114	0.238	0.205	0.194
0.4	0.081	0.109	0.246	0.241	0.078	0.082	0.109	0.224	0.221	0.215
0.5	0.081	0.115	0.255	0.279	0.082	0.082	0.109	0.295	0.222	0.213
0.6	0.087	0.116	0.263	0.266	0.086	0.089	0.141	0.275	0.248	0.237
0.7	0.089	0.106	0.276	0.272	0.087	0.092	0.089	0.276	0.261	0.266

Table 2 RMSE ($\times 10^2$) of predictors of p_d based on Models 1 and S1

ρ	Model 1				Model S1					
	BP	P_{BP}	EBP	P_{EBP}	BP	BP^a	P_{BP}^a	EBP	EBP^a	P_{EBP}^a
0.1	1.806	1.794	2.238	2.232	2.140	1.805	1.793	2.370	2.170	2.159
0.2	1.845	1.822	2.177	2.161	2.165	1.843	1.820	2.516	2.206	2.190
0.3	1.876	1.845	2.297	2.271	2.153	1.873	1.842	2.415	2.236	2.226
0.4	1.919	1.901	2.261	2.247	2.237	1.910	1.892	2.642	2.299	2.284
0.5	2.020	2.001	2.468	2.455	2.336	1.997	1.977	2.650	2.380	2.354
0.6	2.152	2.223	2.477	2.462	2.431	2.108	2.088	2.889	2.525	2.511
0.7	2.361	2.341	2.616	2.607	2.578	2.247	2.231	3.105	2.672	2.659

are labelled as BP, BP^a and P_{BP}^a respectively. These tables also present the same performance measures for the empirical predictors EBP (3.2), approximate EBP (3.5) and approximate plug-in (3.8). They are labelled as EBP, EBP^a and P_{EBP}^a respectively. The plug-in predictors P_{BP}^a and P_{EBP}^a are obtained by predicting the vector of random effects \mathbf{v} by its approximate BP $\hat{\mathbf{v}}^a(\theta)$ and EBP $\hat{\mathbf{v}}^a(\hat{\theta})$ respectively. We implement the Monte Carlo algorithms for approximating integrals with $L = 5000$ iterations. For the sake of comparisons, Tables 1 and 2 also present the BIAS and the RMSE, both multiplied by 10^2 , of the theoretical predictors BP, and plug-in-BP based on Model 1. They are labelled as BP and P_{BP} respectively. These tables present the same performance measures for the EBP and plug-in predictors based on Model 1. They are labelled as EBP and P_{EBP} respectively. Again, the Monte Carlo algorithms for approximating integrals is carried out with $L = 5000$ iterations.

Table 1 shows an increase in bias when we consider empirical predictors. Regarding the comparison between Model 1 and Model S1, there are no substantial differences between the two models, although in general the approximate predictors based on Model S1 have slightly lower biases than the corresponding predictors of Model 1. Because of the approximation of multiple integrals, the BP and EBP based on Model S1 do not outperforms the BP and EBP based on Model 1. If the average across domains is ignored, the behaviour of the domain biases, B_d 's, shows that predictors of Model 1 are not centered in many domains (see Fig. 1 for more details). For low values of the correlation parameter ρ the predictors have lower biases than for high values. The plug-in predictors based on Model 1 have greater bias than the corresponding BP and EBP. When the variance components are known, the difference between the predictors BP, BP^a and P_{BP}^a based on Model S1 have the theoretical expected good behaviour with low biases. However, when we substitute the variance components by the their MSM estimators, the corresponding predictors EBP, EBP^a and P_{EBP}^a based on Model S1 have larger biases.

Table 2 presents the average across domains of the RMSEs ($\times 10^2$) of the BP, EBP and plug-in for both area-level Poisson mixed models: Model 1 and Model S1. It reveals an increase in the RMSE as the parameter ρ increases and also when one uses empirical versions instead of theoretical ones. Regarding the comparisons between predictors, the plug-in predictor P_{BP}^a has, in general, a slight lower

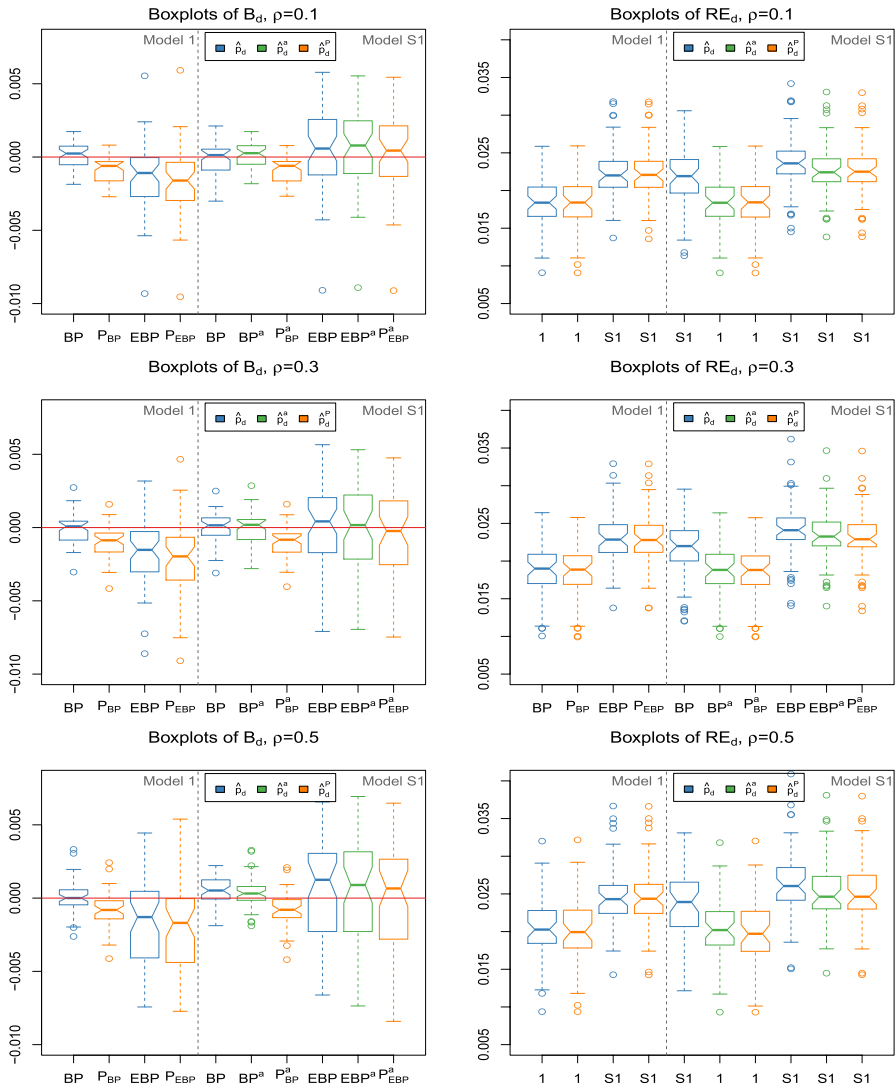


Fig. 1 Boxplots of B_d 's (first column) and RE_d 's (second column) for the predictors of p_d and values of ρ shown in Tables 1 and 2

RMSE than the best predictors BP and BP^a based on Model S1. On the other hand, approximate predictors BP^a and EBP^a reduce the RMSE with respect to BP and EBP respectively. Then, in terms of RMSE, it is preferable to use the approximate predictors based on Model S1. For any estimator, the variance is the most important term of the MSE since the bias is much smaller than the RMSE.

In summary, the biases of predictors EBP and P_{EBP} based on Model S1 are greater than the corresponding biases s (simulated data sets) of EBP^a and P_{EBP}^a based on Model S1. The same happens with root-MSEs if $\rho \leq 0.5$. For the cases $\rho \geq 0.6$, the fitting

algorithm becomes more unstable and produces an increase in the variance of predictors. Therefore, some improvement is achieved when using predictors based on model S1 that take advantage of the spatial correlation structure. In the application to real data it is worthwhile to apply predictors based on the Model S1, as $\hat{\rho} = 0.324$.

Figure 1 shows the boxplots of the domain biases, B_d 's, (first column) and the domain root-MSEs, RE_d 's, (second column) for the predictors and the values of ρ appearing in Tables 1 and 2. In each graph, the first four boxplots refer to the predictors based on Model 1 and the remaining six to the predictors based on Model S1. The BP's and EBP's ($\hat{p}_d(\theta)$ and $\hat{p}_d(\hat{\theta})$) are represented in blue, their approximations based on the Model S1 ($\hat{p}_d^a(\theta)$ and $\hat{p}_d^a(\hat{\theta})$) are plotted in green and the plug-in predictors ($\hat{p}_d^P(\theta)$ and $\hat{p}_d^P(\hat{\theta})$) are colored in orange.

The boxplots show an increase of the variability in both B_d 's and RE_d 's when one uses the empirical predictors. The bias of the predictors based on Model 1 has less variability, but these predictors are clearly biased (except the BP). This fact was not shown in Table 1. The predictors based on Model S1 are unbiased except the P_{BP}^a plug-in predictor. The behaviour of the RE_d 's for the predictors based on Model 1 is similar to the one based on Model S1, although for $\rho = 0.3$, the RE_d 's of the plug-in estimator are slightly lower. For predictors based on Model S1, the RE_d 's of the approximate BP $\hat{p}_d^a(\theta)$ and EBP $\hat{p}_d^a(\hat{\theta})$ are similar to those of $P_{BP}^P(\hat{p}_d^P(\theta))$ and $P_{EBP}^P(\hat{p}_d^P(\hat{\theta}))$ respectively, while the RE_d 's of the BP $\hat{p}_d(\theta)$ and EBP $\hat{p}_d(\hat{\theta})$ are generally bigger.

From Tables 1 and 2 and Fig. 1, we conclude that the approximate EBP $\hat{p}_d^a(\hat{\theta})$ (EBP^a) shows a competitive performance when there is an underlying spatial correlation structure in the data, since it is unbiased and its RE_d 's behave similarly to those of $\hat{p}_d^{Pa}(\hat{\theta})$ (P_{EBP}^a).

Appendix E of supplementary material extends the above simulation experiments to other data generating processes. Instead of generating u_1, \dots, u_D i.i.d. $N(0, 1)$, before applying the transformation $\mathbf{v} = (\mathbf{I}_D - \rho\mathbf{W})^{-1}\mathbf{u}$, with $\mathbf{u} = (u_1, \dots, u_D)'$, we generate u_1, \dots, u_D i.i.d from the distributions t -Student, Gumbel and skew normal. Appendix E shows some increase of bias and RMSE when deviating from the hypothesis of normality.

The system of MSM nonlinear equations (2.5) is solved by using the *nleqslv* package of R. We have also used the *mvtnorm* package to generate samples following a SAR(1) process and the package *spdep* to construct the proximity matrix \mathbf{W} and to test the null hypothesis of no spatial autocorrelation. For $\rho = 0.3$, the average runtime of the MSM fitting algorithm was 0.51 seconds. The computational average runtime of the approximate EBP, $\hat{p}_d^a(\hat{\theta})$, is 0.41 seconds. On the other hand, the EBP $\hat{p}_d(\hat{\theta})$ has a high computational burden compared to its competitors. Its average runtime was 48.73 seconds. The employed computer has a processor Intel© Core™ i7-8750 H CPU @ 2.20GHz \times 6, and 16 GBs of RAM memory.

5 Applications to real data

This section applies the developed methodology to the estimation of poverty proportions, p_d , in Galicia. The data are taken from the 2013 Spanish living conditions survey (SLCS). The Galician counties are the study territories. The domains of interest are constructed by crossing the variables county and sex = women. In Galicia there are 53 counties, but in four of them there are no available data. Therefore, the number of considered domains is $D = 49$. The performance of Model S1 depends on the choice of the proximity matrix \mathbf{W} . To the best of our knowledge, there is no a priori procedure to determine or estimate the proximity matrix in GLMM with random effects having SAR(1) spatial correlation. Therefore, the way to proceed is to propose several intuitively reasonable alternatives, based on knowledge of the socioeconomic situation of Galicia and on similar applications of SAR models to real data, and carry out the pertinent studies for the selection of explanatory variables, applying goodness-of-fit tools and performing a complete diagnosis of the model. Based on this statistical analysis, the most appropriate proximity matrix is chosen. Following this strategy, three different options are tested: common-borders, distances and k -nearest neighbours. In the first option (common-borders), two domains are neighbours if they have a common delimitation. The second option considers the Euclidean distance between the centroids of the counties and sets up a proximity measure by taking the inverse of the distance between domains. The last option applies k -nearest neighbours with $k = 3$. After analyzing the three possibilities in Appendix F of the supplementary material, we selected the first option because the common-borders proximity matrix makes Model S1 have a good fit to the data and present better diagnostics.

Figure 2 shows the proximity map that determines the proximity matrix \mathbf{W}^0 , i.e. it provides for each domain, which are its neighbours. See Sect. 2 for more details on the construction of the proximity matrix \mathbf{W}^0 and \mathbf{W} .

Fig. 2 Proximity map for each domain d ($d = 1, \dots, D$)

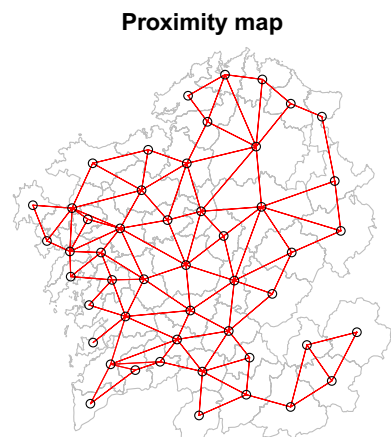


Table 3 Descriptive analysis

Variable	Mean	SD	Median	Min	Max	Cor
<i>log(y)</i>	3.13	1.15	3.18	0.00	5.51	–
<i>lab2</i>	0.10	0.04	0.10	0.02	0.21	0.32
<i>edu3</i>	0.15	0.07	0.14	0.03	0.33	0.42

Table 4 MSM estimates of regression parameters under Model S1

Variable	Estimate	SE	<i>z</i> -Value	<i>p</i> -Value
<i>Intercept</i>	–1.8803	0.1515	–12.4086	< 0.001
<i>lab2</i>	2.9848	1.2097	2.4689	0.0136
<i>edu3</i>	–1.3809	0.5033	–2.7445	0.0061

The target variable, y_d , counts the number of women under the poverty line in the domain d and v_d is the corresponding women sample size. The minimum, median and maximum values of v_d are 19, 152 and 1384, respectively. The minimum has been reached in the south east of the region, while the median and maximum belong to the south west. The auxiliary variables x_d are given in Table 4. We first fit Model 1 to the data $(y_d, v_d, x_d), d = 1, \dots, D$ and we apply Moran’s I test for spatial autocorrelation. As the obtained p -value is lower than 0.001, we assume that $y_d, d = 1, \dots, D$, follows Model S1 and we fit this model to the data.

Table 3 provides a descriptive analysis for the logarithm of the response variable, $\log(y)$, and the considered covariates (*lab2* and *edu3*). Specifically, it presents their mean, standard deviation (*sd*), median, minimum and maximum values (*min* and *max*) and the correlation (*cor*) between the covariates and $\log(y)$.

Table 4 presents the significant estimates (p -value < 0.05) of the fixed effect coefficients under Model S1 and their standard errors, z -values and p -values.

Taking into account the signs of the estimates, the auxiliary variable *lab2* (proportion of unemployed women), is directly related to the response variable while *edu3* (proportion of women with university level of education), helps to decrease the counts of women under the poverty line. Each domain $d, d = 1, \dots, D$, has a random intercept with distribution $N(0, \phi^2)$, where $\hat{\phi} = 0.130$. The 95% percentile bootstrap confidence interval for the standard deviation parameter is (0.001, 0.331). The estimated autocorrelation parameter is $\hat{\rho} = 0.324$. To test the null hypothesis $H_0 : \phi^2 = 0$, Algorithm 1 of Appendix B is applied. The obtained p -value is 0.018. Then, taking $\alpha = 0.05$, the null hypothesis is rejected. The Algorithm 2 of Appendix B is applied to test $H_0 : \rho = 0$. The obtained bootstrap p -value is 0.001. Taking $\alpha = 0.05$, the bootstrap test concludes that the autocorrelation parameter ρ is significantly different from 0. Therefore, this test recommends fitting Model S1 to the data, instead of Model 1.

Figure 3 plots the Pearson residuals of the EBP approximation under Model S1, i.e.

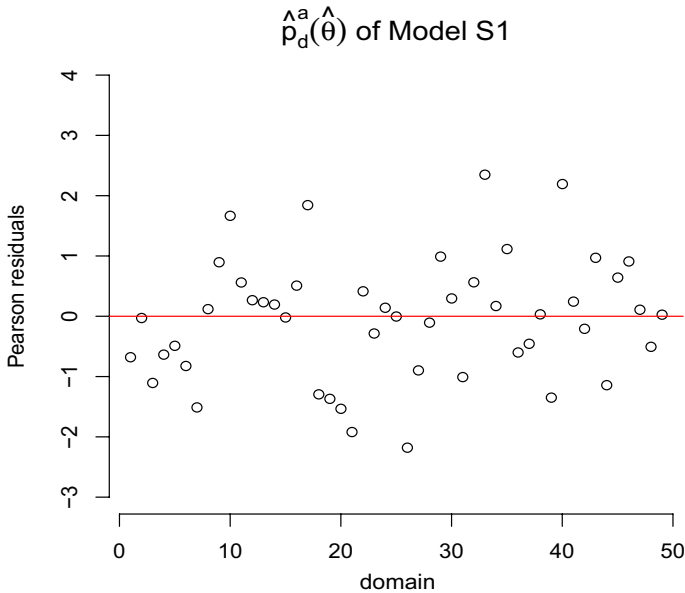


Fig. 3 Pearson residuals of the EBP approximation under Model S1

$$r_d^P = \frac{y_d - v_d \hat{p}_d^a(\hat{\theta})}{\sqrt{v_d \hat{p}_d^a(\hat{\theta})}}, \quad d = 1, \dots, D.$$

The distribution of the Pearson residuals is symmetrical around 0 and takes values in the interval $(-2, 2)$. In addition the tested hypotheses on ρ and ϕ , the conclusion is again that Model S1 is appropriate to fit the women poverty data in Galicia by counties in 2013. This is to say, the incorporation of the underlying spatial correlation structure to the inference process seems to be beneficial.

Concerning the prediction of the poverty proportions p_d , $d = 1, \dots, D$, the plug-in predictor based on Model S1 had a good performance in the simulations and lower computational cost than the approximate EBP. Therefore, it is a good candidate to be employed in this application to real data. On the other side, the plug-in predictor takes less amount of information from the spatial correlation structure than the approximate EBP. This is why we prefer to apply the approximate EBP.

Figure 4 (left) compares the behaviour of the approximate EBP (3.5) and direct estimations. The direct estimators are calculated by using the Hájek formula with the officially calibrated sampling weights. The domains are sorted by the sample sizes v_d 's. The direct estimators present oscillations of large amplitude, while the approximate EBPs have a smoother behaviour, which is something preferred by the statistical offices when publishing estimations. As the sample size increases, both sets of estimates tend to overlap.

Figure 4 (right) plots the relative root-MSEs (RRMSE) of the approximate EBPs based on Model S1 and the relative root-variances of the direct estimators. The

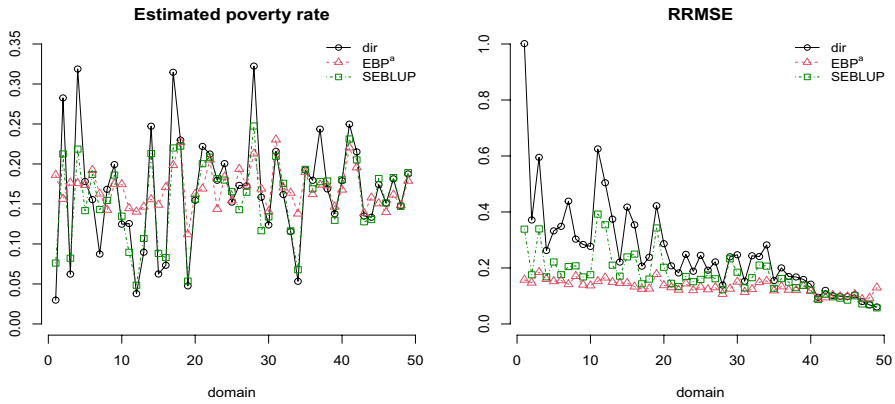


Fig. 4 Direct estimates and approximate EBPs of poverty proportions p_d (left) and relative root-MSEs (right) for women in 2013

direct estimates have high variability, specially for small sample sizes. As above, when the sample size v_d increases, both accuracy measures follow the same pattern. The RRMSEs of the approximate EBPs are estimated by using the bootstrap procedure of Sect. 3 with $B = 500$ replicates. The averages of the relative root-variances of the direct estimator and of the RRMSEs of the approximate EBP are 0.2595 and 0.1323, respectively. According to these results, we conclude that the approximate EBP performs better.

Figure 5 (left) maps the approximate EBP estimation of p_d for women in 2013. The regions where there is no data, are in white. Model S1 gives the following predictions of women poverty proportions: 1 county with poverty proportion $p_d \leq 0.12$, 12 counties with $0.12 < p_d \leq 0.15$, 24 counties with $0.15 < p_d \leq 0.18$ and 12 counties with $p_d > 0.18$. Highest levels of poverty are found in the south and west of the

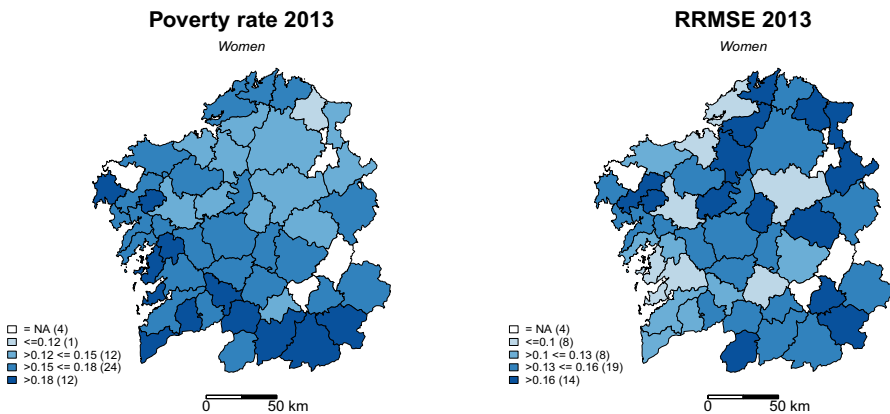


Fig. 5 Poverty proportion approximate EBPs for women based on Model S1 (left) and RRMSEs (right) in Galicia during 2013

community. On the other hand, the counties with the lowest estimated poverty proportion are located in the north-east of the region.

Figure 5 (right) maps the RRMSE estimates of the approximate EBP of p_d by counties in 2013. There are 8 counties with $\text{RRMSE} \leq 10\%$, 8 counties with $10\% < \text{RRMSE} \leq 13\%$, 19 counties with $13\% < \text{RRMSE} \leq 16\%$ and 14 counties with $\text{RRMSE} > 16\%$. The highest values are found in the north-east of the region. Their minimum and maximum are 8.82 and 18.49%, respectively. As the highest RRMSE is lower than 20%, these estimates could be accepted for publication by statistical offices.

6 Concluding remarks

This paper introduces an area-level Poisson mixed model with SAR(1) domain effects. The new model is a GLMM counterpart of the LMM considered by Pratesi and Salvati (2008) and generalises the area-level Poisson mixed model studied in Boubeta et al. (2016) to the context of spatial correlation. The MSM algorithm of Jiang (1998) is employed for estimating the model parameters. The empirical best predictor and a plug-in predictor of the target parameter p_d are given. As these predictors involve integrals in R^D , approximate predictors requiring the calculation of integrals in R are derived. As accuracy measure of the predictors, the MSE is considered and it is estimated by a parametric bootstrap approach.

For scenarios that mimic the application to real data, a simulation experiment studies the bias and the MSE of the new predictors. Specifically, this simulation investigates: (1) the behaviour of the EBPs and the plug-in predictors, (2) the performance of the suggested approximations to the multiple integrals, and (3) the loss of efficiency of BPs and plug-in predictors when model parameters are not known and estimated by the MSM method.

From the simulation experiment, we may give the following conclusions. First, the approximate EBPs and plug-in predictors have similar computational cost and behavior. They also have better performance than their theoretical counterparts. Second, the approximation of integrals in R^D introduces a source of error that can only be reduced with high computational cost; that is, by greatly increasing the number of Monte Carlo iterations. Third, the good properties of BPs do not necessarily carry over to EBPs, especially when the number of domains is not large.

We use the approximate EBP for estimating women poverty proportions in Galician counties. The data are taken from the 2013 SLCS. As the the Moran's I test indicates spatial correlation, we fit Model S1 to the data. In addition, the proposed predictions are compared against the direct estimates. The approximate EBPs of the women poverty proportion are smoother. As the RRMSE of the direct estimator is too high when the sample size v_d is small, it is preferable to use the approximate EBP. The predictions based on Model S1 suggests that the highest levels of women poverty are found in the south and west of the region. The average percentage of women poverty is 16.89% and its average error is 13.23%.

It is worth mentioning the programming problems of the predictors constructed under the introduced spatial model, which leads us to use numerical approaches that introduce

an additional source of error. However, we have observed in the simulations that the suggested approximate predictors (EBP and plug-in) have a lower computational cost and behave as well as their corresponding non approximate predictors.

The computations in this article have been performed entirely in the R programming language. The developed codes are available upon request.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10260-023-00729-7>.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anselin L (2001) Spatial econometrics. In: Badi H (ed) A companion to theoretical econometrics, Chapter 14. Blackwell Publishing Ltd., Baltagi
- Arima S, Cretarola L, Lasinio GJ, Pollice A (2012) Bayesian univariate space-time hierarchical model for mapping pollutant concentrations in the municipal area of Taranto. *Stat Methods Appl* 21:75–91
- Baldermann C, Salvati N, Schmid T (2016) Robust small area estimation under spatial non-stationarity. Discussion Paper, Economics, 22, N. 2016/5, School of Business and Economics
- Banerjee S, Carlin BP, Gelfand AE (2015) Hierarchical modeling and analysis for spatial data. CRC Press
- Boubeta M, Lombardía MJ, Morales D (2016) Empirical best prediction under area-level Poisson mixed models. *TEST* 25:548–569
- Boubeta M, Lombardía MJ, Morales D (2017) Poisson mixed models for studying the poverty in small areas. *Comput Stat Data Anal* 107:32–47
- Burgard JP, Krause J, Morales D (2022) Robust prediction of domain compositions from uncertain data using isometric logratio transformations in a penalized multivariate Fay-Herriot model. *Stat Neerl* 76:65–96
- Chandra H, Salvati N, Chambers R, Tzavidis N (2012) Small area estimation under spatial nonstationarity. *Comput Stat Data Anal* 56:2875–2888
- Chandra H, Salvati N, Chambers R (2015) A spatially nonstationary Fay-Herriot model for small area estimation. *J Surv Stat Methodol* 3:109–135
- Chandra H, Salvati N, Chambers R (2017) Small area prediction of counts under a non-stationary spatial model. *Spat Stat* 20:30–56
- Chandra H, Salvati N, Chambers R (2018) Small area estimation under a spatially non-linear model. *Comput Stat Data Anal* 126:19–38
- Choi J, Lawson AB, Cai B, Hossain MM (2011) Evaluation of Bayesian spatiotemporal latent models in small area health data. *Environmetrics* 22(8):1008–1022
- Cressie N (1993) Statistics for spatial data. Wiley, New York
- Dubé J, Legros D (2014) Spatial econometrics using microdata. Wiley, Hoboken
- Esteban MD, Lombardía MJ, López-Vizcaíno E, Morales D, Pérez A (2020) Small area estimation of proportions under area-level compositional mixed models. *TEST* 29(3):793–818

- Ferrante MR, Trivisano C (2010) Small area estimation of the number of firms recruits by using multivariate models for count data. *Surv Methodol* 36(2):171–180
- Ghosh M, Natarajan K, Stroud T, Carlin BP (1998) Generalized linear models for small-area estimation. *J Am Stat Assoc* 93(441):273–282
- Glaser S (2017) A review of spatial econometric models for count data. Hohenheim Discussion Papers in Business, Economics and Social Sciences, No. 19-2017, Universität Hohenheim, Fakultät Wirtschafts und Sozialwissenschaften, Stuttgart
- González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2008) Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. *Comput Stat Data Anal* 52:5242–5252
- González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2010) Small area estimation under Fay–Herriot models with nonparametric estimation of heteroscedasticity. *Stat Model* 10:215–239
- Jiang J (1998) Consistent estimators in generalized linear models. *J Am Stat Assoc* 93:720–729
- Jiang J, Lahiri P (2006) Mixed model prediction and small area estimation. *TEST* 15:1–96
- LeSage J, Pace RK (2009) Introduction to spatial econometrics. Chapman and Hall/CRC
- López-Vizcaíno E, Lombardía MJ, Morales D (2013) Multinomial-based small area estimation of labour force indicators. *Stat Model* 13:153–178
- López-Vizcaíno E, Lombardía MJ, Morales D (2015) Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *J R Stat Assoc, Ser A* 178:535–565
- Marhuenda Y, Molina I, Morales D (2013) Small area estimation with spatio-temporal Fay–Herriot models. *Comput Stat Data Anal* 58:308–325. The Third Special Issue on Statistical Signal Extraction and Filtering
- Mohebbi M, Wolfe R, Jolley D (2011) A poisson regression approach for modelling spatial autocorrelation between geographically referenced observations. *BMC Med Res Methodol* 11:133
- Molina I, Salvati N, Pratesi M (2009) Bootstrap for estimating the MSE of the spatial EBLUP. *Comput Stat* 24:441–458
- Morales D, Esteban MD, Pérez A, Hobza T (2021) A course on small area estimation and mixed models. Springer
- Moura FAS, Migon HS (2002) Bayesian spatial models for small area estimation of proportions. *Stat Model* 2(3):183–201
- Opsomer JD, Claeskens G, Ranalli MG, Kauermann G, Breidt FJ (2008) Nonparametric small area estimation using penalized spline regression. *J R Stat Soc Ser B* 70:265–286
- Petrucci A, Salvati N (2006) Small area estimation for spatial correlation in watershed erosion assessment. *J Agric Biol Environ Stat* 11:169–172
- Pfeffermann D (2013) New important developments in small area estimation. *Stat Sci* 28:40–68
- Pratesi M (2016) Analysis of poverty data by small area estimation. Wiley series in survey methodology. Wiley
- Pratesi M, Salvati N (2008) Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Stat Methods Appl* 17:113–171
- Rao JNK (2008) Some methods for small area estimation. *Riv Int Sci Soc* 4:387–406
- Rao JNK, Molina I (2015) Small area estimation, 2nd edn. Wiley, New York
- Reluga K, Lombardía MJ, Sperlich S (2021) Simultaneous inference for empirical best predictors with a poverty study in small areas. *J Am Stat Assoc*. <https://doi.org/10.1080/01621459.2021.1942014>
- Singh B, Shukla G, Kundu D (2005) Spatio-temporal models in small area estimation. *Surv Methodol* 31:183–195
- Sugasawa S, Kubokawa T (2020) Small area estimation with mixed models: a review. *Jpn J Stat Data Sci* 3(2):693–720
- Sugasawa S, Kawakubo Y, Ogasawara K (2020) Small area estimation with spatially varying natural exponential families. *J Stat Comput Simul* 90(6):1039–1056
- Trevisani M, Torelli N (2017) A comparison of hierarchical Bayesian models for small area estimation of counts. *Open J Stat* 7:521–550
- Ugarte MD, Ibáñez B, Militino AF (2006) Modelling risks in disease mapping. *Stat Methods Med Res* 15:21–35
- Ugarte MD, Goicoa T, Militino AF (2010) Spatio-temporal modeling of mortality risk using penalized splines. *Environmetrics* 21:270–289
- Viton PA (2010) Notes on spatial econometric models. City and regional planning 870.03. The Ohio State University

Wakefield J (2007) Disease mapping and spatial regression with count data. *Biostatistics* 8(2):158–183
You Y, Zhou QM (2011) Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Surv Methodol* 37:25–37

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.