Statistics and Its Interface Volume 14 (2021) 95–107

Spatial regression models for bounded response variables with evaluation of the degree of dependence

Sandra E. Flores*, Marcos O. Prates, Jorge L. Bazán, and Heleno B. Bolfarine

Bounded response variables such as percentages, proportions, or rates are common in applications involving social and educational datasets, including rates of poverty or rates of achievement by municipalities, counties or provinces. New regression models have been proposed in recent years by considering distributions such as the Beta, Simplex and Kumaraswamy models for this type of data. However, to this type of dataset, it is common to observe the spatial dependence of units. For instance, municipalities or counties are organized into states. For this case, the supposition of independence among observations in the same state removes relevant relations between neighboring provinces. In this paper, we present a model of spatially bounded distribution regression with a Bayesian estimation approach where spatial relations are modeled by a spatial random variable with a particular dependence structure, such as the intrinsic conditional autoregressive model or the Leroux definition. Additionally, the Bayesian inferential method and model comparison criteria are discussed. Simulation studies and an application in reading comprehension spatial data are used to illustrate the performance of the proposed model and the estimation method adopted.

Keywords and phrases: Bounded distribution, Bayesian inference, Proportions, Spatial models.

1. INTRODUCTION

Regression models for a bounded responses have been recently proposed in the literature, considering different distributions for the response variable, for instance, Beta [19, 11, 8], Simplex [33, 32], Kumaraswamy [6, 1], L-Logistic [9] and in a more general way, considering the CDF-quantile [31] and the Generalized Johnson System [20].

In regression analysis using the Beta and Simplex distributions, the mean and dispersion parameters are associated with a set of covariates considering proper link functions while for the Kumaraswamy, CDF-quantile, GJS, and L-Logistic distributions a quantile, preferentially the median, and dispersion parameters are considered to introduce the covariates. Additionally, mixed regression models have been proposed considering Beta, Simplex and Kumaraswamy distributions. Examples are Verkuilen and Smithson [37], Figueroa-Zuñiga, Arellano-Valle and Ferrari [12], Qiu, SONG and Tan [28], Bayes, Bazán and de Castro [1]. Those classes of models are more convenient when observations in data are measured, for instance, repeatedly through time, or if the existence of cluster between units is observed and it is necessary to incorporate heterogeneity among units considering random effects or hierarchical or multilevel structures.

On the other hand, spatial dependence among units is very common, considering, for instance, the proportion of students with a satisfactory level in reading comprehension in cities of a country or the proportion of people living in poverty in those same cities. In both cases, it is possible to consider that similar characteristics between neighboring cities, sharing those similar characteristics, can be associated with the variability observed in the units, and so an adequate spatial configuration based on neighbors must be incorporated in the regression analysis to achieve more realistic results.

At the present, there is only a spatial regression model for proportions available considering the Beta distribution. Specifically, Cepeda-Cuervo and Núñez-Antón [8] proposed a spatial double generalized beta regression model considering joint modeling approaches for the mean and dispersion parameters, which was applied to the analysis of the quality of education in Colombia. However, several authors have shown that not always assuming that the response variable follows the Beta distribution is the most appropriate model, so the, Simplex [22] and Kumaraswamy [1] can be good modeling alternatives. However, no spatial model has been proposed, to the best of our knowledge, considering Simplex and Kumaraswamy distributions using a Bayesian approach.

Thus, the main goal of this paper is to propose a spatial regression model with a Bayesian estimation approach for bounded response variables. In its formulation, the response variable can follow some bounded distribution such as the

*Corresponding author.

Simplex or Kumaraswamy distribution, and then by considering the intrinsic conditional autoregressive (ICAR) model [3, 2] or the Leroux definition [21], it is possible to measure dependency among neighbors to take into consideration the spatial relations in the data. Also, possible differences in comparisons with the spatial Beta regression model can be detected.

This work is organized as follows. A dataset and preliminary analysis are introduced in Section 2 to motivate the purpose of this work. In Section 3, three bounded distributions are presented for a response variable with support in the unit interval, which is considered in this approach. Section 4 proposes the general spatial regression model for bounded response models by considering the distributions discussed in the previous section and the spatial random effect to model the spatial dependence. In Section 5, a Bayesian approach is formulated and developed for the proposed regression model, including model comparison criteria to choose between alternative models. Section 6 presents results of simulation studies showing the recovery parameters of the proposed model and the model comparison criteria. To illustrate the proposed method, the real database introduced in Section 2 is re-analyzed considering alternative spatial models in Section 7. Additional comments and future developments are presented in Section 8.

2. DATA AND PRELIMINARY ANALYSIS

Since 2006 a Student Assessment Census (ECE in Spanish) have been carried out annually by the Ministry of Education in Peru, which aims to ascertain the level of student achievement in reading and mathematics in the second grade. In 2012, the ECE reached a school coverage rate of 97.7%, with five or more students, and 89.4% of the student population. Results are available at http://umc.minedu.gob.pe/evaluacion-censal-de-estudiantes-2012-ece-2012/.

In this work, we are interested in explaining variable y: the proportion of students with a satisfactory level in reading comprehension (RC) [10] in each one of the 195 Peruvian provinces (political subdivisions at the second level) during the year of 2012. Additional details about the RC Test are described in MINEDU [24]. As covariates, we considered some indexes that are part of the mentioned state density index [26], which measures the provision of essential services to support human development. Specifically, we consider the Health Index, which quantifies the number of medical doctors per 10,000 inhabitants and takes values between 0 and 60; higher values mean that there are more doctors and thus more access to health in the province. There is also the Sanitation Index, which is the percentage of dwellings with piped water and sewer service, taking values from 0 to 100. Finally, we consider the Electrification Index. This means the percentage of residences with electricity. The values for this index again range from 0 to 100. Details to compute these indicators can be found in PNUD [26] and the dataset is available at http://www.pe.undp.org/content/peru/es/home/library/poverty/Informesobredesarrollohumano2013/IDHPeru2013.html. In order to formulate a regression model, the covariates Health, Sanitation, and rate of Electrification are transformed regarding the proportion of each province and then named as x_1, x_2, and x_3 respectively. Some descriptive statistics for the response variable and covariates are presented in Table 1.

Table 1. Descriptive statistics of the variables for RC data

Descriptive statistics	y	x_1	x_2	x_3
Minimum	0.023	0.043	0.002	0.338
First quartile	0.099	0.108	0.410	0.613
Median	0.167	0.173	0.542	0.743
Mean	0.203	0.199	0.563	0.721
Third quartile	0.284	0.246	0.725	0.850
Maximum	0.616	0.746	0.995	0.995
Variance	0.017	0.015	0.042	0.025
Skewness	0.919	1.595	0.068	−0.29

Since the response variable is defined int the $(0, 1)$ interval, the following Beta regression model is a natural first option to fit the dataset:

$$(1)\qquad y_i|\boldsymbol{\beta}, \phi \overset{ind.}{\sim} \text{Beta}(\mu_i, \phi)$$
$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$
$$i = 1, \ldots, 195,$$

where μ_i and ϕ are respectively the mean and precision parameters of the Beta distribution, β_0 is the intercept and β_m, $m = 1, 2, 3$ are regression coefficients that are associated with the covariates. By considering a Bayesian approach, the parameters of the model are estimated using the integrated nested Laplace approximation (INLA) method, as implemented in the `R-INLA` package ([30], http://www.r-inla.org). The following default non-informative priors [see 4, cap. 5] are used to complete the model:

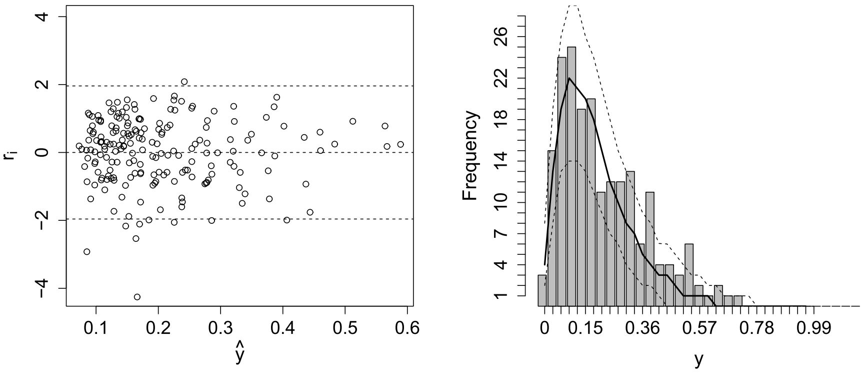
$$\log(\phi) \sim \text{logGamma}(1, 0.1)$$
$$\beta_m \sim \mathcal{N}(0, 10^6), \quad m = 0, 1, 2, 3.$$

Table 2 displays the posterior mean, standard deviation and the 95% credible interval (CI) for the parameters of the fitted Beta regression model. Looking at CI, from the regression coefficients where the zero value is not included, it is possible to conclude that the covariates are significant and have a positive effect on the response variable. In other words, if essential services in one province increase, then it is expected that the proportion of students with a satisfactory level in the RC will also increase.

To evaluate the fit of the model, we initially consider standardized residuals. These residuals are defined as: $r_i = (y_i^\star - y_i)/\sqrt{\hat{\text{var}}(y_i^\star)}$, where $y_i^\star$ is the posterior mean of the predictive distribution of the parameters [see 4, Sections 5.5 and 5.6]. The estimated variance is given

Table 2. Estimates of parameters for the RC data in the Beta regression model

Parameter	Mean	Standard deviation	95% Credible interval
β_0	-3.63	0.19	$(-4.01, -3.26)$
β_1	1.79	0.36	(1.08, 2.50)
β_2	1.27	0.28	(0.73, 1.82)
β_3	1.51	0.33	(0.86, 2.17)
ϕ	24.79	2.47	(20.21, 29.89)



(a)

(b)

Figure 1. Fit of the Beta regression model for RC data. a) Standardized residuals (r_i) vs. fitted values. b) Confidence band of the posterior predictive distribution with histogram of the observed dataset.

by $\widehat{\text{var}}(y_i^\star) = y_i^\star(1 - y_i^\star)/(1 + \hat{\phi})$, where $\hat{\phi}$ is the mean posterior distribution of the parameter ϕ. An example of R code for computing the standardized residuals can be found in Section 5 of the Supplementary Material, http://intlpress.com/site/pub/files/_supp/sii/2021/0014/0002/SII-2021-0014-0002-s001.pdf.

The left side of Figure 1 displays the standardized residuals against the posterior means for the predictive distribution $y_i^\star$. A band with $\Phi^{-1}(.025) = -1.96$ and $\Phi^{-1}(.975) = 1.96$ was added, where $\Phi(.)$ denotes the cumulative density function (cdf) of the standard normal distribution. It is possible to see that some residuals are outside the range indicating the model does not fit the data well, because of the existence of large residuals. The fitted model induces a distribution of future data which depend on the observations. This distribution is called the Posterior Predictive Distribution (PPD). The PPD is used to generate synthetic datasets. Considering values of $\hat{\mu}_i$ and $\hat{\phi}, i = 1, \ldots, n$, a total of $M = 1000$ datasets of size n were generated using the predictive distribution of y_i. To obtain a confidence band for the PPD, the data that were generated in each replication were grouped in frequency tables with each bin (bucket) of length 0.03. Values in quantiles 0.05 and 0.95 are shown in dashed lines and quantile 0.5 in solid line, on the right side of Figure 1, together with the histogram of the response variable observed. It is possible to see the length of the confidence band is big for response values around 0.15 and some observations are outside of the range, indicating the model does not fit the data well. Thus, it is possible to conclude that both results shown in Figure 1 suggest a lack of fit of the Beta regression model. No spatial configuration was considered in the previous Beta regression model.

Additionally, in order to identify a possible spatial configuration of the response variable, Figure 2 depicts residuals from the fitted model classified in seven intervals of size 0.91, where each interval has a specific color intensity. Provinces with a light shade mean negative residuals and a dark colors indicate a province with positive residuals. Thus, spatially, it is possible to observe the presence of groups of provinces with similar behavior not explained in the model, some of them are positive, others negative and also some with small residuals. As a consequence, a spatial configuration is identified for the provinces that need an investigation in order to formulate an adequate model for the RC data. This spatial configuration was not considered in the previous Beta regression model. This result was also observed considering alternative regression models without spatial effect, assuming the distribution Simplex or Kumaraswamy distributions, as will be shown in Section 7.

Motivated by the data, the main purpose of this paper is to formulate spatial regression models for proportions that can be used as alternatives to detect relations between neighboring cities.

This work will initially explore the use of the Beta-spatial regression model proposed by Cepeda-Cuervo and Núñez-Antón [8] to the RC data and at the same time, other spatial

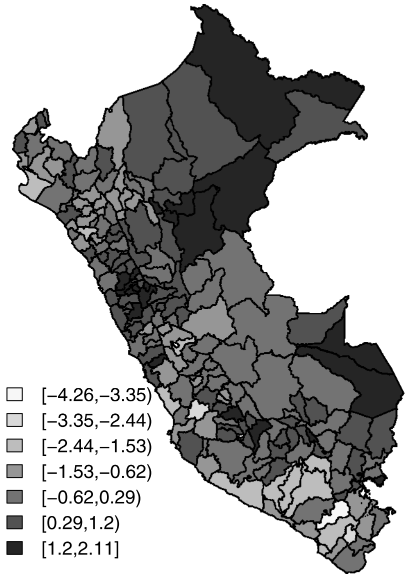


Figure 2. Map of the 195 *Peruvian provinces for RC data with the residuals of the Beta regression model, showing clusters with surrounding provinces.*

models with bounded response variables will be introduced, using other distributions and two types of spatial random effects.

3. BOUNDED RESPONSE VARIABLES

This section defines some distributions for bounded response variables with support in the unit interval $(0, 1)$.

- A random variable Y follows a *Beta distribution* with parameters μ and ϕ_1 if its probability density function (pdf) is given by:

  $$g(y|\mu,\phi_1)=\frac{\Gamma(\phi_1)}{\Gamma(\mu\phi_1)\Gamma[(1-\mu)\phi_1]}y^{\mu\phi_1-1}(1-y)^{(1-\mu)\phi_1-1}. \tag{2}$$

  Notation $Y \sim \text{Beta}(\mu, \phi_1)$ with $\mu \in (0, 1)$ as the location parameter and $\phi_1 > 0$ as the precision parameter is used. The mean and variance of this distribution are respectively:

  $$E[Y|\mu,\phi_1] = \mu \text{ and}$$
  $$Var[Y|\mu,\phi_1] = \frac{\mu(1-\mu)}{1+\phi_1}.$$

- A random variable Y follows a *Simplex distribution* with parameters μ and ϕ_2 if its pdf is given by:

  $$g(y \mid \mu,\phi_2) = \frac{\sqrt{\phi_2}}{\sqrt{2\pi\{y(1-y)\}^3}} \exp\left[-\frac{\phi_2}{2}d(y \mid \mu)\right] \tag{3}$$

  where $d(y \mid \mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}$ is a unitary deviance, which is nonnegative with value zero if only if $y = \mu$. Notation $Y \sim \text{S}(\mu, \phi_2)$ with $\mu \in (0, 1)$ as the location parameter and $\phi_2 > 0$ as the precision parameter is used. Following Jorgensen [18, pg 199], the mean and variance of this distribution are respectively

  $$E(Y|\mu,\phi_2) = \mu \quad \text{and}$$
  $$Var(Y|\mu,\phi_2) = \mu(1-\mu) - \sqrt{\frac{\phi_2}{2}} \exp\left\{\frac{\phi_2}{2\mu^2(1-\mu)^2}\right\} \Gamma\left(\frac{1}{2}, \frac{\phi_2}{2\mu^2(1-\mu)^2}\right),$$

  where $\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}dt$, which defines the incomplete gamma function.

- A random variable Y follows a *Kumaraswamy distribution* if its pdf is given by:

  $$g(y|\kappa,\phi_3) = -\frac{\log(1-q)\phi_3}{\log\left(1-e^{-\phi_3}\right)\log(\kappa)} y^{-\frac{\phi_3}{\log(\kappa)}-1} \times \left\{1 - y^{-\frac{\phi_3}{\log(\kappa)}}\right\}^{\frac{\log(1-q)}{\log(1-e^{-\phi_3})}-1}. \tag{4}$$

  Notation $Y \sim \text{K}(\kappa, \phi_3, q)$ is used, with the quantile $\kappa \in (0, 1)$ as the location parameter, $\phi_3 > 0$ as the precision parameter. $\kappa = G^{-1}(q)$ where the probability q is assumed to be fixed according to the quantile of interest and $G^{-1}(.)$ denotes the inverse cdf of Kumaraswamy distribution. Under this parameterization using κ and ϕ_3, the mean and variance of this distribution are given by:

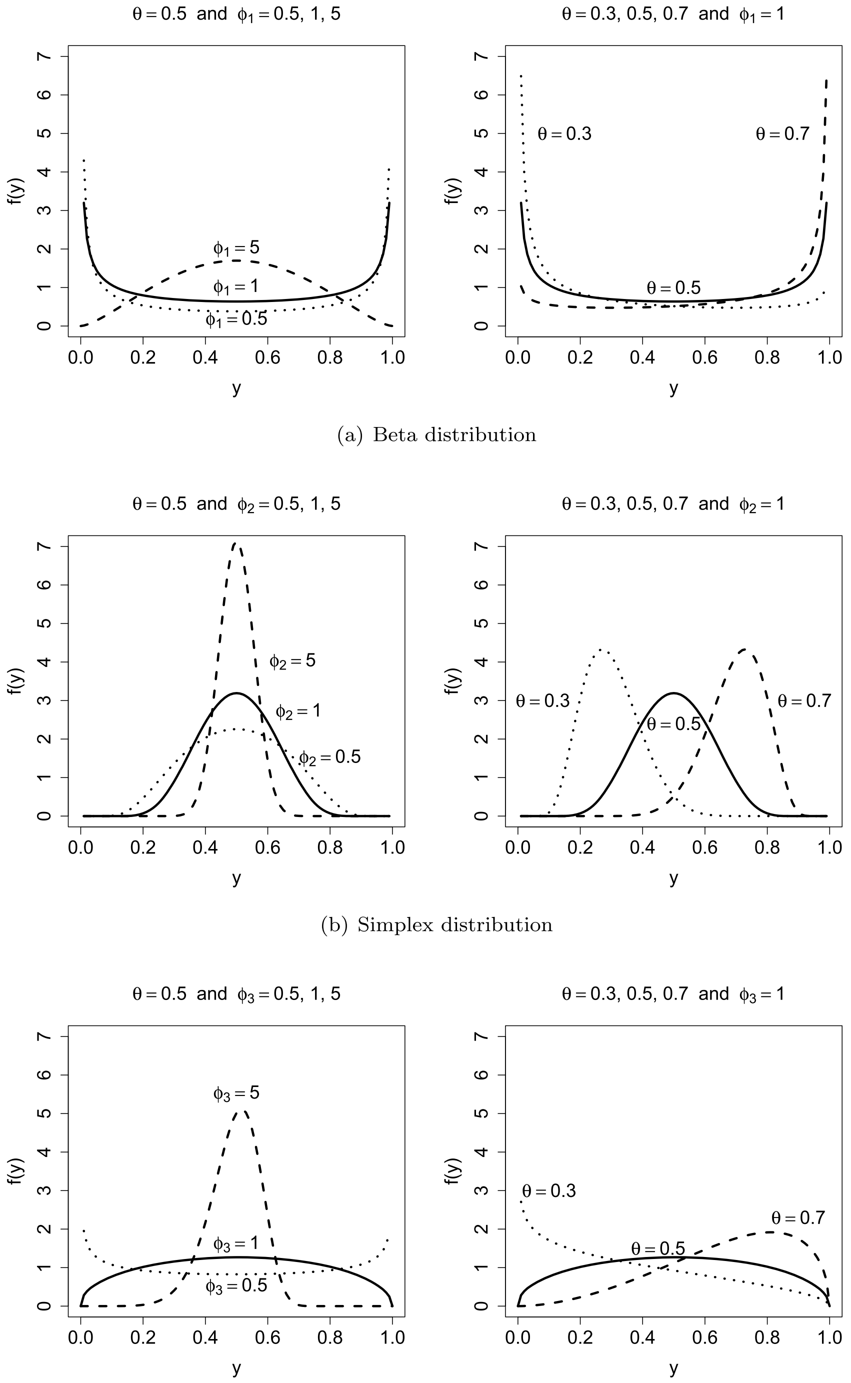
  $$E(Y|\kappa,\phi_3) = \frac{\log(1-q)}{\log(1-e^{-\phi_3})} B\left(1 - \frac{\log\kappa}{\phi_3}, \frac{\log(1-q)}{\log(1-e^{-\phi_3})}\right)$$

  and

  $$Var(Y|\kappa,\phi_3) = \frac{\log(1-q)}{\log(1-e^{-\phi_3})} \times B\left(1 - \frac{2\log\kappa}{\phi_3}, \frac{\log(1-q)}{\log(1-e^{-\phi_3})}\right) - (E(Y|\kappa,\phi_3))^2,$$

  where $B(\cdot, \cdot)$ denotes the beta function and κ denotes the quantile parameter which defines the median if $q = 0.5$.

In the remainder of the paper we consider a random variable y which follows a *bounded distribution*, denoted by $y \sim \pi(\theta, \phi)$, where θ is a location parameter (being μ in the Beta and Simplex distributions or κ corresponding to the median in the Kumaraswamy distribution), and ϕ is a precision parameter ϕ_1, ϕ_2, and ϕ_3 as previously described.



(a) Beta distribution

(b) Simplex distribution

(c) Kumaraswamy distribution

Figure 3. Pdf of the bounded distributions (Beta, Simplex and Kumaraswamy) with different values of the location parameter θ and the precision parameter ϕ.

Figure 3 shows the probability density function (pdf) of the *bounded distributions*. They are the Beta, Simplex and Kumaraswamy distributions, considering different values for the location parameter θ and the precision parameter ϕ. On the left side of Figure 3, different values of the precision parameter are shown with a fixed value of the location parameter $\theta = 0.5$. On the right side, the plot shows the pdf for different values of the location parameter with a fixed value of the precision parameter $\phi = \phi_1 = \phi_2 = \phi_3 = 1$. For a fixed value of θ, higher precision values of precision result in more concentrated values around the location parameter. Also, for a fixed value of ϕ, the distribution moves to the right or the left according to the value of the location parameter. Additional shapes of bounded distributions are included in Figures S1 to S3 of Section 2 of the Supplementary Material.

Additionally, Figure 4 plots, for a fixed value of the location parameter $\theta = 0.5$, the standard deviation (sd) of the *bounded distributions* as a function of ϕ. The value of *sd* decreases when ϕ increases. Also, for the same values of θ and ϕ, the Beta distribution has the highest value of *sd*, the Simplex distribution has the smallest value in one part ($\phi < 10$) and the Kumaraswamy in the other part ($\phi > 10$).

4. THE SPATIAL BOUNDED DISTRIBUTION REGRESSION MODEL

In this section, a spatial bounded distribution regression model (SBDR) is formulated to determine, beyond the effects of explanatory variables, how the spatial correlation between Peruvian provinces can explain the proportion of students with a satisfactory level in reading comprehension.

To formulate the SBDR model, we consider n observed variables $\boldsymbol{y} = \{y_i : i = 1, ..., n\}$ with y_i following a *bounded distribution* with support in the $(0, 1)$ interval, as defined in Section 3, with location and precision parameters θ_i and ϕ, respectively. Here, i denotes a region with a location in space (latitude and longitude coordinates). There is also a set of $p < n$ explanatory variables, thus the SBDR model is defined as follows:

$$\begin{aligned} y_i\,|z_i, \boldsymbol{\beta}, \phi &\overset{ind.}{\sim} \pi(\theta_i, \phi) \\ g(\theta_i) &= \boldsymbol{x_i}^\top \boldsymbol{\beta} + z_i \\ \boldsymbol{z}|\tau &\sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{R}) \\ i &= 1, \ldots, n, \end{aligned} \tag{5}$$

where $\pi(\theta_i, \phi)$ denotes the *bounded distribution* of the response variable, which can be the Beta, Simplex or Kumaraswamy distribution, where θ_i is the location parameter for locale i (the mean in the case of the Beta and Simplex distributions, and the median for the Kumaraswamy distribution) and ϕ is the precision parameter. The link function $g(.)$ is assumed to be the logit function, but other link functions can be considered. The unknown regression parameters or regression coefficients of fixed effects are denoted by a vector $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{X}$ is an $n \times (p+1)$ design matrix, which is assumed to have full rank, with information of covariates associated with the fixed effects including the intercept, $\boldsymbol{x_i}^\top = (x_{i0}, x_{i1}, \ldots, x_{ip})$ is the row i of $\boldsymbol{X}$ and $\boldsymbol{R}$ is an adequate precision matrix.

The random effects vector associated with locations $\boldsymbol{z} = (z_1, \cdots, z_n)^\top$ accommodates the spatial dependency. It is assumed that z_i are conditionally normally distributed such that $z_i|\boldsymbol{z}_{-i}$ denotes the conditional distribution of z_i given the all other values of $\boldsymbol{z}$ and n_i is the number of neighbors of region i. To attribute spatial association, let $z_i|\boldsymbol{z}_{-i}$ be represented by a Markov property such that $z_i|\boldsymbol{z}_{-i} = z_i|\boldsymbol{z}_{i\sim j}$ where $i \sim j$ denotes the neighbors of region i. This model is known as the intrinsic conditional autoregressive (ICAR) model [3]. Furthermore, τ is a spatial precision (inverse of the variance) parameter of $\boldsymbol{z}$. In other words, $\boldsymbol{z}$ encapsulates the relationship between region i with its neighbors $j's$ considering the joint distribution:

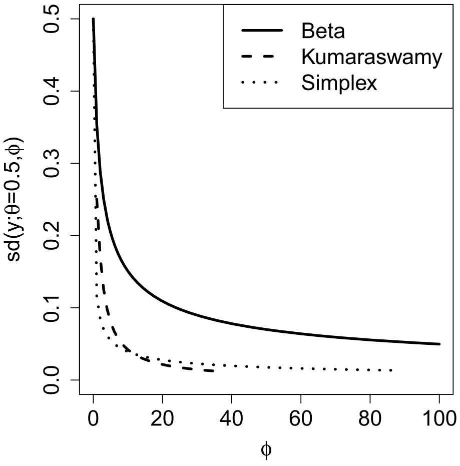


Figure 4. Standard deviation of one variable with bounded distribution, as a function of the precision parameter ϕ for a fixed value of the location parameter θ.

$$\pi(\boldsymbol{z}|\tau) \propto \tau^{r(\boldsymbol{R})/2} \exp\left(-\frac{\tau}{2}\boldsymbol{z}^\top \boldsymbol{R}\boldsymbol{z}\right), \tag{6}$$

where $r(.)$ is defined to be the rank of a matrix. The precision matrix $\boldsymbol{R}$, with dimension $n \times n$, has its entries defined as follows:

$$\boldsymbol{R}_{ij} = \tau \begin{cases} n_i, & \text{if } i = j, \\ -1, & \text{if } i \sim j, \\ 0, & \text{other case.} \end{cases}$$

This specification is called here as SBDR-Icar model.

From the ICAR parameterization, parameter τ can represent overdispersion and spatial dependence at the same time [21]. For this reason, we also employ the proposal of Leroux, Lei and Breslow [21] for random effect $\boldsymbol{z}$. Their proposal suggests a modification of the precision matrix of $\boldsymbol{z}$ as follows:

$$\boldsymbol{R}' = \tau'((1-\lambda)\boldsymbol{I} + \lambda\boldsymbol{R}) \tag{7}$$

where τ' is the precision parameter of matrix $\boldsymbol{R}'$, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{R}$ is according to the definition in (6) with precision one ($\tau = 1$) and λ is a spatial dependence parameter with values in the $[0, 1]$ interval. When $\lambda = 0$, $\boldsymbol{z}$ has independent components, resulting in a model named here SBDR-Ind (with no spatial relations) and retains spatial dependency when $\lambda = 1$, such that the SBDR-Icar model is obtained. In other cases, when $0 < \lambda < 1$, it is called the SBDR-Ler model. A bounded regression (BR) exists when a non-random effect is considered in the model, for instance the Beta regression presented in Section 2.

5. BAYESIAN INFERENCE

Bayesian inference was performed to fit the model defined in Equations (5), (6) and (7). In this case, the interest was to obtain the posterior distribution of $\mathrm{p}(\boldsymbol{z}, \boldsymbol{\beta}, \phi, \tau, \lambda|\boldsymbol{y})$.

Considering the definition in Equations (6) and (7), the augmented likelihood function takes the following form:

$$\begin{aligned} \mathrm{L}(\boldsymbol{\beta}, \phi, \tau, \lambda|\boldsymbol{y}, \boldsymbol{z}) &= \prod_{i=1}^{n} \mathrm{p}(y_i|\theta_i, \phi)\mathrm{p}(\boldsymbol{z}|\tau, \lambda) \\ &\propto \prod_{i=1}^{n} \mathrm{p}(y_i|\theta_i, \phi) \times \tau^{\frac{r(\boldsymbol{R}')}{2}} \exp\{-\frac{\tau}{2}\boldsymbol{z}^\top \boldsymbol{R}'\boldsymbol{z}\} \end{aligned} \tag{8}$$

where $\boldsymbol{\beta}$, ϕ, τ and λ are parameters to be estimated in the model, with $\theta_i = g^{-1}(\boldsymbol{x_i^T}\boldsymbol{\beta} + z_i)$.

Independent normal distributions, with mean zero and small precision, are considered as the prior for each one of components of $\boldsymbol{\beta}$ and the logit λ, $\log(\frac{\lambda}{1-\lambda})$, parameters. In other words, $\beta_j \sim \mathcal{N}(0, \tau_\beta)$, $j = 0, 1, \ldots, p$ and logit $\lambda \sim \mathcal{N}(0, \tau_\lambda)$. For the ϕ parameter, we adopt a prior Gamma(a, b) distribution with $a = 1$ and $b = 0.1$, giving a mean of 10 and variance of 100 [5]. Finally, for the τ parameter, following Blangiardo and Cameletti [4, pg. 182], a minimal informative prior is considered assuming a Gamma(a', b') distribution, where $a' = 1$ and $b' = 0.0005$, giving a mean of 2000 and a large variance.

Thus, the posterior distribution takes the following form:

$$\begin{aligned} \mathrm{p}(\boldsymbol{z}, \boldsymbol{\beta}, \phi, \tau, \lambda|\boldsymbol{y}) &\propto \mathrm{p}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\beta}, \phi)\mathrm{p}(\boldsymbol{z}|\tau, \lambda)\mathrm{p}(\boldsymbol{\beta})\mathrm{p}(\phi)\mathrm{p}(\tau)\mathrm{p}(\lambda) \\ &\propto \mathrm{L}(\boldsymbol{\beta}, \phi, \tau, \lambda|\boldsymbol{y}, \boldsymbol{z})\mathrm{p}(\boldsymbol{\beta})\mathrm{p}(\phi)\mathrm{p}(\tau)\mathrm{p}(\lambda), \end{aligned}$$

meaning that the posterior distribution is proportional to the likelihood function, $\mathrm{p}(\boldsymbol{y}|.)$, the random effects prior, $\mathrm{p}(\boldsymbol{z}|.)$, multiplied by the prior distribution of the parameters of interest. In this case, independent priors $\mathrm{p}(\boldsymbol{\beta})\mathrm{p}(\phi)\mathrm{p}(\tau)\mathrm{p}(\lambda)$ are assumed.

The posterior distribution can be approximated by MCMC algorithms; however, we opted to use the approximate Bayesian inference for latent Gaussian models using the `R-INLA` program. General information about the INLA procedure can be found in Section 1 of the Supplementary Material.

In this case, for the INLA approach the model proposed takes the latent field as $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{z})$ and hyperparameters as $\boldsymbol{\gamma} = (\phi, \tau, \lambda)$. Then, INLA returns the posterior marginal for the latent field and the posterior marginal for the hyperparameters.

5.1 Model comparison criteria

Since the proposal is a mixed effect model, several model comparison criteria can be adopted.

One measurement to select the best model is the deviance information criterion (DIC), which was introduced by Spiegelhalter et al. [34], defined as:

$$DIC = \hat{D} + p_D$$

where $\hat{D}$ is the posterior mean of the deviance and p_D is the effective number of parameters of the model. Following Rue, Martino and Chopin [30, Section 6.4], deviance is defined as:

$$D(\boldsymbol{\zeta}, \boldsymbol{\varrho}) = -2\sum_{i \in D} \log \mathrm{p}(y_i|\zeta_i, \boldsymbol{\varrho}) + \text{constant}.$$

where $\boldsymbol{\varrho} = (\phi, \boldsymbol{\beta}, \tau, \lambda)$ and the effective number of parameters is approximated by

$$p_D \approx n - \mathrm{Trace}\{\boldsymbol{Q} \times \boldsymbol{Q}^{*-1}\}.$$

where $\boldsymbol{Q}$ is the prior precision matrix of the latent field $\boldsymbol{\zeta}$ and $\boldsymbol{Q}^{*-1}$ is the posterior covariance matrix.

DIC is one of the most common criteria, but there is no consensus about its use. Therefore, other criteria are also used such, namely the widely applicable information criteria (WAIC) and the mean absolute error (MAE).

WAIC [38] is a Bayesian approximation to fit the expected out-of-sample log predictive density. First, it is necessary to calculate the posterior predictive density, then, to add a correction for the number of effective parameters. In the context of this work, WAIC is defined as two times the expression (9) [13]

$$\widehat{\mathrm{elppd}}_{\mathrm{WAIC}} = \mathrm{lppd} - \mathrm{p}_{\mathrm{WAIC}} \tag{9}$$

where lppd is the logarithm of the pointwise predictive density of y_i induced by the posterior distribution $p(\boldsymbol{\varrho}|\boldsymbol{y})$, given by

$$\sum_{i=1}^{n} \log \int \left(\int \mathrm{p}(y_i|\zeta_i, \boldsymbol{\varrho})\mathrm{p}(\zeta_i|\boldsymbol{\varrho}, \boldsymbol{y}_{-i})d\zeta_i \right) d\boldsymbol{\varrho}$$

and

$$\mathrm{p}_{\mathrm{WAIC}} = \sum_{i=1}^{n} \mathrm{Var}(\log \int \left(\int \mathrm{p}(y_i|\zeta_i, \boldsymbol{\varrho})\mathrm{p}(\zeta_i|\boldsymbol{\varrho}, \boldsymbol{y}_{-i})d\zeta_i \right) d\boldsymbol{\varrho}).$$

As suggested by Blangiardo and Cameletti [4], summary indexes can be computed using the posterior predictive distribution $p(y_i^\star|\boldsymbol{y}) = \int \mathrm{p}(y_i^\star|\zeta_i)\mathrm{p}(\zeta_i|\boldsymbol{y})d\zeta_i$, in other words, the likelihood of a replicated observation $y_i^\star$ given data $\boldsymbol{y}$. Then, with the purpose of model evaluation, MAE is defined as the absolute average difference between the observed value y_i and the corresponding estimated value $y_i^\star$ for $i = 1, \ldots, n$, with the following equation:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i^\star - y_i|}{n}.$$

Thus, MAE measures the deviation between the observed value y_i and its estimated value $y_i^\star$ providing model goodness of fit.

6. SIMULATIONS

This section presents a simulation study to analyze the performance of parameter recovery for SBDR models and model misspecification. Additionally, other studies are developed to evaluate the model comparison criteria defined in Section 5.1.

6.1 Recovery study and model misspecification

In this first study, nine scenarios with $Re = 100$ replicates for each one are simulated from the SBDR-Ler model, since this model is the more general one. The models in each scenario depend on three values of parameter τ: $(0.5, 1$ and $2)$ and on the three *bounded distributions*: Beta, Simplex and Kumaraswamy.

The reason for comparing these models using the three *bounded distributions* is to find characteristics of performance for each one of the distributions. At the same time, different values are used for parameter τ to identify characteristics of this parameter, representing the precision of the spatial relations between regions.

The amount of spatial dependence for the simulated scenarios is fixed with $\lambda = 0.9$. In order to preserve a comparison with the application of RC data, the neighborhood configuration is identical to the application dataset. This means that the work was perfomed with $n = 195$ regions. Each one has a location in the space with latitude and longitude values, which was used to elaborate the matrix $\boldsymbol{R}$ containing the relations of dependency between regions. Random vector $\boldsymbol{z}$ was simulated from a multivariate Normal distribution with zero mean and precision matrix $\boldsymbol{R}'$ defined in (7). Population parameters $\boldsymbol{\beta}$ were fixed with values similar to the RC data, $\boldsymbol{\beta} = (-3.0, 2.0, 0.7, 1.5)$. The values of $\boldsymbol{x}_i$, $i = 1, \ldots, n$, were drawn from three independent distributions, they are respectively Beta$(\mu_x = 0.2, \phi_x = 15)$, $\mathcal{N}(0, 1)$ and Bernoulli(0.7). The intention, of this selection, is to have variety of covariates representing proportion, continuous and categorical inputs. The ϕ parameter of each distribution was selected similar to its point estimate in real data application, thus $\phi_1 = 80$, $\phi_2 = 19$ and $\phi_3 = 12$.

For each scenario, the SBDR-Ler model was generated and then the SBDR-Ler, SBDR-Icar and BR models were fitted following the method described in Section 5. In order to evaluate parameter recovery of the "True" model and the models that are misspecified, some measures are considered for to evaluate the bias in the estimation of the parameters of the model. Specifically, we consider the absolute bias (AB), which is defined as: $\text{AB}_m = \frac{\sum_{r=1}^{Re} |\hat{m}_r - m|}{Re}$, where m identifies the parameter under evaluation; the standard deviation (Sd) and the number of times when the parameter is in the 95% credible interval (CI).

Table 3 shows results only for scenarios which were simulated from the SBDR-Ler model with $\tau = 1$ and the three *bounded distributions*. As can be seen, the point estimates of the regression parameters are very accurate when considering the true model in all distributions, but we find some bias in the estimation of the ϕ and τ parameters associated with spatial effects. Additionally, for the true model, empirical IC coverage for the different parameters seems adequate, especially for the Simplex distribution. These results are coherent with results reported in other studies where the parameters for the spatial effects are known to have identifiability issues, as to mentioned in [17], so usually it is not easy to estimate all the parameters simultaneously. The lower coverage of the ϕ parameter in the case of the SBDR-Ler model with Beta distribution can be explained because this distribution has a scale bigger than the case of Kumaraswamy and Simplex distributions. Also, Kumaraswamy distribution does not come quickly reach small precision compared with the corresponding precision parameter in the Simplex distribution, in particular for $\theta < 0.5$, see Figure 4 and Figures S1 to S3 in Section 2 of the Supplementary Material. By analyzing results of the models that are misspecified, as expected, we found that IC coverage are lower than in the true model. We found bias in the estimation of the parameters when considering the corresponding ICAR model, and even more for the non-random effects model.

Results of the scenarios with $\tau = 0.5$ and $\tau = 2$ are included in Section 3 of the Supplementary Material, showing similar results.

6.2 Model comparison criteria study

Two studies are demonstrated to shed light the capacity of the model comparison criteria to choose the true model. In both cases, 100 datasets are simulated from the SBDR-Ler model with one of the three *bounded distributions* and $\tau = 1$.

In the first study, inside the same response distribution, we compare the capacity of the model comparison criteria to select the true generated model analyzing SBDR-Ler, SBD-Icar and the model without z (BR). Table 4 presents the number of times (in 100 datasets) when the criterion selects the correspondingt fitted model as the best one. It is possible to see that all criteria arrive at the correct selection, making the SBDR-Ler model as the preferable one in comparison with SBDR-Icar and BR models.

In the second study, we compare, across different bounded response (that is, between different likelihoods), the capacity of the model comparison criteria to determine the model with true response distribution. As can be seen in Table 5, the model comparison criteria can be inconclusive to select the model considering the different bounded response distributions studied. In order to show this observation, we fix $\tau = 1$ and $\lambda = 0.9$ and perform two experiments: Simulation

Table 3. Parameter recovery study of the SBDR-Ler model considering three bounded distributions with spatial precision $\tau = 1$ and model misspecification considering alternative models

	Parameter	β_0	β_1	β_2	β_3	ϕ	τ	λ
Beta distribution								
	True	−3	2	0.7	1.5	80	1	0.9
SBDR-Ler	Estimated	−2.973	2.001	0.692	1.479	60.995	1.398	0.864
	Sd	0.309	0.433	0.041	0.101	16.655	0.321	0.095
	AB	0.218	0.351	0.034	0.083	19.72	0.417	0.062
	95% CI Coverage	94	95	95	91	79	83	96
SBDR-Icar	Estimated	−2.96	1.994	0.689	1.471	55.578	1.466	
	Sd	0.125	0.424	0.04	0.101	14.348	0.377	
	AB	0.219	0.353	0.035	0.084	24.772	0.484	
	95% CI Coverage	66	95	95	91	62	80	
BR	Estimated	−2.647	1.808	0.62	1.275	13.537		
	Sd	0.152	0.471	0.045	0.121	1.348		
	AB	0.37	0.5	0.083	0.225	66.463		
	95% CI Coverage	47	87	54	54	0		
Kumaraswamy distribution								
	True	−3	2	0.7	1.5	12	1	0.9
SBDR-Ler	Estimated	−3.004	2.035	0.7	1.507	8.888	1.299	0.862
	Sd	0.315	0.405	0.037	0.09	1.461	0.231	0.093
	AB	0.215	0.317	0.029	0.069	3.134	0.306	0.064
	95% CI Coverage	89	97	95	95	52	78	97
SBDR-Icar	Estimated	−3.005	2.036	0.7	1.509	8.51	1.27	
	Sd	0.11	0.398	0.036	0.09	1.379	0.241	
	AB	0.215	0.321	0.029	0.068	3.51	0.28	
	95% CI Coverage	59	97	94	95	32	83	
BR	Estimated	−2.97	2.011	0.697	1.497	3.107		
	Sd	0.18	0.521	0.048	0.15	0.162		
	AB	0.248	0.54	0.044	0.106	8.893		
	95% CI Coverage	72	86	92	99	0		
Simplex distribution								
	True	−3	2	0.7	1.5	19	1	0.9
SBDR-Ler	Estimated	−3.016	2.032	0.696	1.503	12.379	1.125	0.845
	Sd	0.259	0.345	0.034	0.077	7.338	0.212	0.106
	AB	0.205	0.271	0.026	0.06	6.639	0.168	0.081
	95% CI Coverage	89	96	94	93	93	93	94
SBDR-Icar	Estimated	−3.016	2.033	0.696	1.504	10.606	1.054	
	Sd	0.087	0.357	0.032	0.07	8.081	0.137	
	AB	0.205	0.27	0.027	0.06	8.407	0.132	
	95% CI Coverage	51	95	93	92	83	92	
BR	Estimated	−2.836	1.895	0.651	1.439	0.248		
	Sd	0.131	0.536	0.043	0.097	0.025		
	AB	0.262	0.515	0.068	0.12	18.752		
	95% CI Coverage	53	92	66	77	0		

A sets $\phi_3 = 12$, $\phi_2 = 19$ and $\phi_1 = 80$ for the Kumaraswamy, Simplex and Beta distributions, respectively, while Simulation B sets $\phi_3 = 1$ for the Kumaraswamy distribution, $\phi_2 = 5$ for the Simplex distribution and $\phi_1 = 10$ for the Beta distribution.

Table 5 shows that for larger ϕ's the Simplex adapts very well to the data and dominates the model selection for all criteria even when it is not the true distribution. However, for small values of ϕ the criteria seem to better select the correct model. The MAE presents a poor performance when the Beta distribution is the true model generated. This is evidence that for a small scale of ϕ, the distributions behave differently and one cannot nicely fit the data generated. This result can be explained in part, since the bounded distributions are very flexible and for some combination of parameters can really adapt to each other.

7. REAL DATA ANALYSIS

In this section, the RC data are revisited and a comparative regression analysis is performed considering or not the spatial component $\boldsymbol{z}$ for this data. Thus, the SBDR models (SDBR-Ler, SDBR-Icar) and the model without the $\boldsymbol{z}$ spatial component (BR) were fitted using the three *bounded distributions* studied in the paper.

The purpose is to explain the proportion of students with a satisfactory level in the reading comprehension (RC) Test (y) in each one of the 195 Peruvian provinces (political subdivisions at the second level), with three explanatory variables as covariates: health, sanitation and electrification indexes, which are represented by x_1, x_2 and x_3, respectively.

Additional details about the RC Test are described in the framework document [24]. Health, sanitation and electrification indexes are components of the state density index, which is described in Section 2 and additional details can be found in the PNUD report [26].

As described in Section 2 and Figure 2, the proportion of students with satisfactory level in the RC Test has a spatial configuration that is considered in the SBDR models as a random variable $\boldsymbol{z}$, which summarizes the relations in adjacent provinces.

Table 4. Evaluation of model comparison criteria (DIC, WAIC and MAE) to select the true model for simulated data of the SBDR-Ler model considering different bounded response variables with spatial precision $\tau = 1$. Values correspond to the number of times (in 100 datasets)

Criterion	SBDR-Ler	SBDR-Icar	BR (without $\boldsymbol{z}$)
Beta distribution			
DIC	99	1	0
WAIC	100	0	0
MAE	100	0	0
Kumaraswamy distribution			
DIC	100	0	0
WAIC	100	0	0
MAE	100	0	0
Simplex distribution			
DIC	68	32	0
WAIC	97	3	0
MAE	100	0	0

Specifically, for this application the SBDR models are defined as follows:

$$
\begin{gathered}
y_i|z_i,\phi \overset{ind.}{\sim} \pi(\theta_i,\phi) \\
g(\theta_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + z_i \\
z_i|\boldsymbol{z}_{-i},\tau',\lambda \sim \mathcal{N}\left(\tfrac{\lambda}{1-\lambda+\lambda n_i}\sum_{i\sim j} z_j, \tfrac{1}{(1-\lambda+\lambda n_i)\tau'}\right) \\
i = 1,\ldots,195,
\end{gathered} \tag{10}
$$

where $\boldsymbol{z} = (z_1,\ldots,z_{195})'$ is the random vector, which retains the spatial effect and follows the definition of Leroux, Lei and Breslow [21] called here the SBDR-Ler model. When $\lambda = 1$ SBDR-Icar model is defined and when $\lambda = 0$, the SBDR-Ind model is specified. The distribution $\pi(\theta_i,\phi)$ can take the three different *bounded distributions*: Beta, Simplex and Kumaraswamy, where θ_i is the location parameter and ϕ the precision parameter. Also, $g(s) = \log(s/(1-s))$ is the logit function. The non-spatial regression can also be represented by Equation (10), but without the spatial effect $\boldsymbol{z}$ and its prior distribution named the BR model. The code used for this application is included in Section 5 of the Supplementary Material.

Table 6 shows only parameter estimation for the possible SBDR-Ler models and the corresponding model comparison criteria (DIC, WAIC and MAE) of those models. Also, Table S1 and S2 in Section 3 from Supplementary Material include the same information for the SDBR-Icar model and the analogous BR model without the $\boldsymbol{z}$ spatial component. Considering the three model comparison criteria, the SBDR-Ler Simplex model is selected as being the best model to fit the RC data. Additionally, it can be noted that this model presents the best result among the alternative models in the Simplex family (the fitted information criteria for the competing models can be seen in Tables S1 and S2 in Section 3

Table 5. Number of times that each fitted model was selected as the best model by DIC, WAIC and MAE values, for the simulation of the SBDR-Ler model and with spatial precision $\tau = 1$

	Simulation A			Simulation B		
Criterion	Beta	Simplex	Kumaraswamy	Beta	Simplex	Kumaraswamy
SBDR-Ler Beta distribution $\phi = 80$				SBDR-Ler Beta distribution $\phi = 10$		
DIC	4	92	4	69	1	30
WAIC	6	92	2	79	1	20
MAE	4	92	4	19	1	80
SBDR-Ler Simplex distribution $\phi = 19$				SBDR-Ler Simplex distribution $\phi = 5$		
DIC	0	93	7	0	89	11
WAIC	0	90	10	0	85	15
MAE	0	80	20	0	68	32
SBDR-Ler Kumaraswamy distribution $\phi = 12$				SBDR-Ler Kumaraswamy distribution $\phi = 1$		
DIC	0	100	0	27	0	73
WAIC	0	100	0	35	0	65
MAE	0	100	0	10	0	90

Table 6. Posterior mean (mean), standard deviation (s.d.) and 95% credible interval (CI) for parameter estimation for the RC data in the SBDR-Ler models with different bounded response

SBDR-Ler Beta model			
	mean	s.d.	95% CI
β_0	-3.54	0.24	$(-4.01, -3.07)$
β_1	1.62	0.35	(0.93, 2.31)
β_2	0.84	0.26	(0.34, 1.35)
β_3	1.67	0.30	(1.09, 2.27)
ϕ	82.73	17.52	(53.36, 121.89)
τ	2.86	0.59	(1.87, 4.19)
λ	0.90	0.06	(0.74, 0.98)
DIC $= -628$			
WAIC $= -642$			
MAE $= 0.0209$			
SBDR-Ler Simplex model			
	mean	s.d.	95% CI
β_0	-3.59	0.23	$(-4.04, -3.14)$
β_1	1.74	0.38	(0.99, 2.49)
β_2	0.82	0.27	(0.29, 1.35)
β_3	1.68	0.30	(1.08, 2.28)
ϕ	19.52	12.13	(5.16, 50.85)
τ	1.40	0.18	(1.08, 1.77)
λ	0.81	0.09	(0.60, 0.95)
DIC $= -1022$			
WAIC $= -1055$			
MAE $= 0.0028$			
SBDR-Ler Kumaraswamy model			
	mean	s.d.	95% CI
β_0	-3.54	0.24	$(-4.01, -3.06)$
β_1	1.66	0.35	(0.98, 2.34)
β_2	0.77	0.25	(0.28, 1.27)
β_3	1.72	0.30	(1.14, 2.30)
ϕ	11.25	1.71	(8.24, 14.95)
τ	2.09	0.35	(1.49, 2.87)
λ	0.88	0.07	(0.69, 0.98)
DIC $= -723$			
WAIC $= -733$			
MAE $= 0.0133$			

of the Supplementary Material). Furthermore, the standardized residuals and PPD of the SBDR-Ler Simplex model are presented in Figure 5 and of the alternative models in Figures S4 to S6 of Section 3 of the Supplementary Material. These figures suggest that the residuals are smaller, and no spatial configuration is presented in Figure 6 in comparison with the preliminary analysis in Figure 2 from Section 2.

Recently, the issue of spatial confounding between fixed and spatial random effects was introduced and studied in the literature [29, 16]. Later, different solutions to alleviate it and discussions about the need for this adjustment have appeared [e.g., 17, 15, 36, 27]. The SPOCK method [27] is an alternative that can be fitted using INLA, so, a version of the proposed Leroux model combined with the SPOCK method is fitted to alleviate a possible spatial confounding between the RC responses and the covariates. This model is called SBDR-SPOCK. As it can be seen in Table S3 in the Supplementary Material, the SBDR-Ler Simplex still presents smaller DIC and WAIC values but equal MAE values, suggesting it is preferable. Thus, we continue the analysis with the SBDR-Ler Simplex.

Considering this model, the 95% CI coverage suggests that β_0 is negative. Moreover, the 95% CI for β_1, β_2 and β_3 do not have the zero value, meaning the existence of a positive effect of the health, sanitation and electrification indexes regarding the proportion of students with a satisfactory level in the RC Test in provinces. This means that higher values of these indexes indicate more students having success in the RC Test in the cities.

The results reveal the existence of a spatial effect $\lambda = 0.81$, where the precision associated with this effect is $\tau = 1.4$ meaning variation among values of z_i. Estimations of z_i are depicted in Figure 7 where the size of the triangles depends on the $\boldsymbol{z}$ values. It is possible to identify positive ($\triangle$) and negative ($\bigtriangledown$) effects from the direction of those triangles. The positive spatial effects are concentrated in the southern Peruvian provinces. The northeastern (the Amazon Forest) provinces have negative spatial effects.

8. FINAL COMMENTS

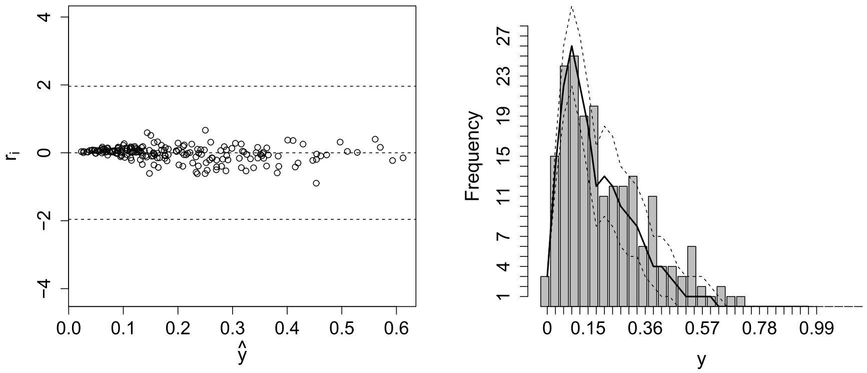
In this paper, a spatial *bounded distribution* regression (SBDR) model is proposed where the spatial effect follows Leroux's definition, returning a degree of spatial dependence $\lambda \in [0, 1]$, which has as particular cases the definition of Icar and the independent case.

Since the posterior distribution is not amenable to analytical treatment, the INLA method was chosen. Results from the simulation study show that the Bayesian proposal yields estimators with good performance. Furthermore, a real dataset is analyzed using the proposed methods. Since the value of λ is in $[0, 1]$ for the Leroux representation of the spatial effects, an advantage of this definition is that it provides the magnitude of the spatial dependence.

As it is traditional in real data, the proposal assumes equal dependence between neighboring regions, although, other dependence parameters can be used, e.g., using the distance between the centroids of the regions, (not considered here). Bayesian estimation using BUGS [23, 35, 14], JAGS [25], STAN [7] or other statistical Bayesian tools can be developed. Finally, zero/one inflated spatial regression considering bounded distributions can be a direction for future research.

ACKNOWLEDGMENTS

This study was financed in part by the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil" (CAPES) - Finance Code 001. The second author acknowledges financial support from CNPq grants 436948/2018-4



(a) SBDR Leroux

(b) SBDR Leroux

Figure 5. Fit of the SBDR-Ler Simplex model for RC data. a) Standardized residuals (r_i) vs fitted values. b) Confidence Band of the Posterior predictive distribution with histogram of the observed response variable.

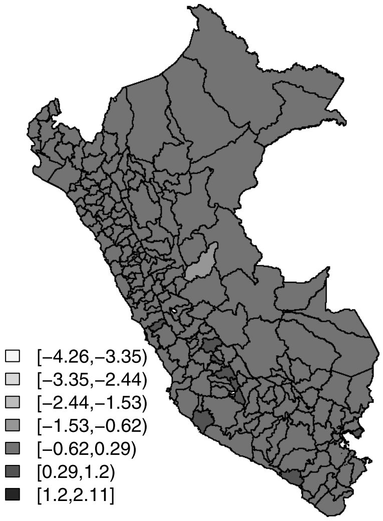


Figure 6. Map of the 195 Peruvian provinces for RC data with residuals of SBDR-Ler Simplex model.

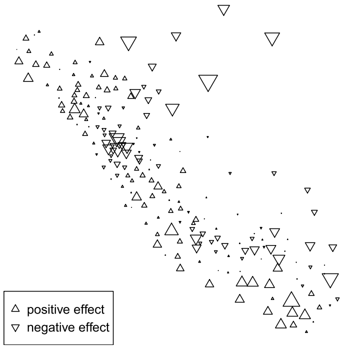


Figure 7. Posterior mean of the spatial effects for the RC data.

and 307547/2018-4 and FAPEMIG grant PPM-00532-16. The third author was partially supported by FAPESP-Brazil 2017/15452-5.

Received 2 May 2019

REFERENCES

[1] Bayes, C. L., Bazán, J. L. and de Castro, M. (2017). A quantile parametric mixed regression model for bounded response variables. *Statistics and Its Interface* **10** 483–493.

[2] Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811

[3] Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43** 1–20. MR1105822

[4] Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R INLA*. John Wiley & Sons. MR3364017

[5] Bonat, W. H., Ribeiro Jr, P. J. and Shimakura, S. E. (2015). Bayesian analysis for a class of beta mixed models. *Chilean Journal of Statistics* **6** 3–13. MR3350148

[6] Bonat, W. H., Riveiro Jr, P. J. and Zeviani, W. M. (2013). Regression models with responses on the unit interval: specification, estimation and comparison. *Rev. Bras. Biom* **20(1)** 1–10.

[7] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* **76** 1–32.

[8] Cepeda-Cuervo, E. and Núñez-Antón, V. (2013). Spatial double generalized beta regression models. Extensions and application to study quality of education in Colombia. *Journal of Educational and Behavioral Statistics* **38** 604–628.

[9] da Paz, R. F., Balakrishnan, N. and Bazán, J. L. (2018). L-Logistic regression models: Prior sensitivity analysis, robustness to outliers and applications. *Brazilian Journal of Probability and Statistics* **33** 455–479. MR3960271

[10] Unidad de Medición de la Calidad Educativa (2012). Resultados de la Evaluación Censal de Estudiantes 2012 Ministerio de Educación, Lima, Perú.

[11] Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31** 799–815. MR2095753

[12] Figueroa-Zuñiga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. P. (2013). Mixed beta regression: A Bayesian per-

spective. *Computational Statistics and Data Analysis* **61** 137–147. MR3063006

[13] Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24** 997–1016. MR3253850

[14] Goudie, R. J., Turner, R. M., De Angelis, D. and Thomas, A. (in press). MultiBUGS: Massively parallel MCMC for Bayesian hierarchical models. *Journal of Statistical Software*. arXiv:1704.03216.

[15] Hanks, E. M., Schliep, E. M., Hooten, M. B. and Hoeting, J. A. (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics* **26** 243–254. MR3340961

[16] Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* **64** 325–334. MR2758564

[17] Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 139–159. MR3008275

[18] Jorgensen, B. (1997). *The Theory of Dispersion Models*. Chapman and Hall, London. MR1462891

[19] Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling* **3** 193–213. MR2005473

[20] Lemonte, A. J. and Bazán, J. L. (2016). New class of Johnson distributions and its associated regression model for rates and proportions. *Biometrical Journal* **58** 727–746. MR3527412

[21] Leroux, B. G., Lei, X. and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* 179–191. Springer. MR1731684

[22] López, F. O. (2013). A Bayesian approach to parameter estimation in simplex regression model: a comparison With beta regression. *Revista Colombiana de Estadística* **36** 1–21. MR3075189

[23] Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10** 325–337.

[24] MINEDU (2016). *Marco de fundamentación de las pruebas de la Evaluación Censal de Estudiantes*. Ministerio de Educación del Perú, Lima.

[25] Plummer, M. (2012). JAGS Version 3.3. 0 user manual. *International Agency for Research on Cancer, Lyon, France.*

[26] PNUD (2013). *Informe sobre desarrollo humano Perú 2013: Cambio climático y territorio: Desafíos y respuestas para un futuro sostenible*. Programa de las Naciones Unidas para el Desarrollo. Lima.

[27] Prates, M. O., Assunção, R. M. and Rodrigues, E. C. (2019). Alleviating spatial confounding for areal data problems by displacing the geographical centroids. *Bayesian Analysis* **14** 623–647. MR3959875

[28] Qiu, Z., SONG, P. X.-K. and Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics* **35** 577–596. MR2468863

[29] Reich, B. J., Hodges, J. S. and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62** 1197–1206. MR2307445

[30] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 319–392. MR2649602

[31] Smithson, M. and Shou, Y. (2017). CDF-quantile distributions for modelling random variables on the unit interval. *British Journal of Mathematical and Statistical Psychology* **70** 412–438.

[32] Song, P. X.-K., Qiu, Z. and Tan, M. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal* **46** 540–553. MR2101142

[33] Song, P. X.-K. and Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics* **56** 496–502.

[34] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 583–639. MR1979380

[35] Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2014). OpenBUGS Version 3.2.3 User Manual. http://www.openbugs.net/Manuals/Manual.html.

[36] Thaden, H. and Kneib, T. (2018). Structural equation models for dealing with spatial confounding. *The American Statistician* **72** 239–252. MR3836447

[37] Verkuilen, J. and Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics* **37** 82–113.

[38] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11** 3571–3594. MR2756194

Sandra E. Flores
Instituto de Matemática e Estatística
Universidade de São Paulo
SP
Brazil
E-mail address: sefari@gmail.com

Marcos O. Prates
Departamento de Estatística
Universidade Federal de Minas Gerais
Belo Horizonte, MG
Brazil
E-mail address: marcosop@est.ufmg.br

Jorge L. Bazán
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos
SP
Brazil
E-mail address: jlbazan@icmc.usp.br

Heleno B. Bolfarine
Instituto de Matemática e Estatística
Universidade de São Paulo
SP
Brazil
E-mail address: hbolfar@ime.usp.br