Journal of Educational and Behavioral Statistics
2013, Vol. 38, No. 6, pp. 604–628
DOI: 10.3102/1076998613499779

Spatial Double Generalized Beta Regression Models: Extensions and Application to Study Quality of Education in Colombia

Edilberto Cepeda-Cuervo
Universidad Nacional de Colombia

Vicente Núñez-Antón
Universidad del País Vasco UPV/EHU

In this article, a proposed Bayesian extension of the generalized beta spatial regression models is applied to the analysis of the quality of education in Colombia. We briefly revise the beta distribution and describe the joint modeling approach for the mean and dispersion parameters in the spatial regression models' setting. Finally, we motivate the need for the innovative use of the generalized beta spatial regression model to the study of the performance of school children in Mathematics and Language, as well as in the analysis of illiteracy data, and present its results in the context of evaluating the quality of education in Colombia.

Keywords: *beta regression models; spatial econometric models; quality of education*

1. Introduction

This article studies and motivates the methodology required to analyze a specific real data set situation when the observations are associated with the beta distribution. Observations associated with the beta distribution are commonly found in the quality of education studies, that is, in the evaluation of the school's performance in Mathematics, Language, Arts, Natural Sciences, or in any other areas of study where a number between 0 and 5 (or another positive integer value) is provided as a measure of this performance. In these cases, the measure assigned to each student can be expressed as a number between 0 and 1. Therefore, it can be assumed that the level of a student's performance is a random variable following a beta distribution. The beta distribution is typically used to model uncertainty or random variation of a probability, fraction, or prevalence. Moreover, this distribution has many applications in areas such as Finance, Social Sciences, or Education, where random variables like school children's performance can be assumed to follow a continuous distribution and, in addition, their values are bounded by real numbers.

A review of the literature in this area of research indicates that many studies whose focus is to determine the quality of educational systems have been developed in quite different contexts. The Programme for International Student Assessment (PISA) was designed and launched by the Organization for Economic Cooperation and Development (OECD) at the end of the 1990s as an international, comparative, and public way of evaluating the performance of school children so that indicators of the various aspects describing the way educational systems function could be generated. These indicators would enable countries to implement the required policies that can improve the quality of their educational system, with a special focus on the learning results students obtain within the system. The 2009 PISA study involved 57 countries, including 30 OECD countries, as well as 27 additional countries, known as *partner countries*. In this case, the variable under study (i.e., dependent or response variable) could be, for example, the percentage of students in each of the countries in the study that are bad, regular, or good readers (see, e.g., Martínez, 2006). In addition, the variables whose effect on the response variable we wish to study (i.e., the explanatory or independent variables) provide information about the learning methodology, as well as about the students' social and family environments, students' ethos, characteristics of the teaching staff, and about teenagers' expectations and inclinations to learn (Cepeda, 2005; Coleman, 1969; Donoso, 2002; Sancho, 2006). Similar studies in countries like Colombia, Chile, Spain, or the United States, where the students' or the educational system's performance could be associated with a beta distribution, have included explanatory variables such as familiar, socioeconomic, or scholar variables. In order to determine the contribution or effect these variables have on the educational system's performance, a beta regression model can be used for all of them. That is, if Y is the response variable under study, $g(\mu_i) = \mathbf{x}'_i\beta$, where $\mu_i = E(Y_i)$, g is an appropriately selected real-valued function, β is the vector of regression parameters, and $\mathbf{x}_i$ is the ith vector of explanatory variables. These models were originally proposed by Ferrari and Cribari-Neto (2004), although a more general model had been previously proposed by Cepeda-Cuervo (2001). In this work, Cepeda proposed a joint modeling approach of the mean and variance (or precision) parameters in the two-parameter exponential family. In the specific case of the beta distribution, the general regression models $\text{logit}(\mu_i) = \mathbf{x}'_i\beta$ and $\log(\phi_i) = \mathbf{z}'_i\alpha$, were proposed. This model has also appeared in Cepeda and Gamerman (2005), and it was later studied by Smithson and Verkuilen (2006) and also by Simas, Barreto-Souza, and Rocha (2010).

In educational studies like the PISA study, the Latin American Laboratory for Assessment of the Quality of Education, or the National Systems Educational Evaluation in countries such as Colombia, Chile, or El Salvador, regional results are also of great interest for researchers in the area. Therefore, the mean performance or the percentage of students (by country or state) having a given reading level are obtained. These results could depend, for example, on variables such as regional policies, income taxes, or the level of regional development, and they

cannot be assumed to be independent of variables related to the regional structures. This is one of the reasons that motivates the use of a spatial econometric method that can incorporate spatial interactions and spatial structures into the proposed regression analysis. That is, countries, regions, or departments (departments in Colombia are different regions into which the country is divided, similar to what provinces represent in European countries or states do in the United States) having neighbors with well-developed educational systems would have higher probabilities of educational development than countries having neighbors with not such a developed educational system. Therefore, the application of spatial econometric models can be convenient and justifiable (see, e.g., Anselin, 1999; Anselin & Florax, 1995; Kelley Pace, Barry, & Sirmans, 1998; LeSage, Fischer, & Scherngell, 2007). Moreover, and given that the classical spatial models assume that the variable under study follows a normal distribution, which is clearly not a valid assumption for these studies, we propose to use a Bayesian version of the beta econometric spatial model proposed by Cepeda-Cuervo, Urdinola, and Rodríguez (2011). This model explicitly incorporates spatial structures and it has been mainly applied to study spatial land concentration in the context of generalized econometrics models. More specifically, and in the context of the data under study, in educational data analysis, when data of state or departments are considered, it is necessary to propose statistical models that take into account the spatial structure of the data. If, in addition, the data include bounded outcome variables, such as, for example, proportions or educational indexes of success or failure, a beta spatial regression model, such as the one we motivate and propose in this article, would be appropriate to analyze the real nature of the data under study.

Many applications of the beta regression models have been developed in the last few years, a clear sign of the motivation for its use within the aforementioned context. Some of these recent applications include, for example, Ferrari and Cribari-Neto (2004) and Rocha and Simas (2011) for the study of proportions; Espinheira, Ferrari, and Cribari-Neto (2008) for the analysis of a test of reading accuracy; and Verkuilen and Smithson (2012) for the analysis of the results of cognitive experiments. Therefore, there are many well-motivated reasons to use beta regression models instead of other alternative approaches (Smithson & Verkuilen, 2006). Moreover, these researchers stated in their findings that "one important advantage that beta regression shares with other GLMs over the ladder-of-powers transformations stems from the fact that the transformations transform raw data, whereas the link function in any GLM transforms expected values."

Besides being used to study the quality of education, many spatial applications of the beta distribution are also possible. For example, to study levels of illiteracy, levels of literacy, unemployment or infant mortality rates by regions, the analysis of the evolution and behavior of poverty and development indexes, study of concentration of wealth or land, and also in the study of risk or corruption indicators by departments, states, or countries. In all of the aforementioned examples, the beta distribution can be assumed as a model. However, and given

that the observations of these variables cannot be assumed to be independent, the use of a spatial beta regression model is well justified to model the aforementioned response variables as a function of the socioeconomical political variables that could have an influence on them. Therefore, spatial structures should be considered, so that the proposed model takes into account the existing neighborhood association. Spatial beta regression models can easily measure this association, and their results are very simple to understand from a practical point of view. This issue is addressed by proposing the use of a so-called *weights matrix* in the model, where its parameter estimates are associated with the corresponding lag variable modeling this neighborhood association. From the above, it is clear that the election of the weights matrix is very important because it should be the result of interpreting and understanding the spatial structure of the variable under study and, although many first-order time structures are considered, other spatial structures can also be assumed. Therefore, this model allows us to assess, in an intuitive and very natural manner, what real influence neighbor departments or states have on the educational system for specific sites in the data we are analyzing. As will be described in detail in Subection 3.2, we will use regression parameters that take into account this site-closeness effect. A brief description of the spatial econometric beta model definition is also given in Subsection 3.2, where it will be shown that this model is also easy to fit and its results simple to interpret in the context of the application.

The rest of the article is organized as follows. In Section 2, we describe the beta distribution. In Section 3, we introduce the general joint mean and dispersion beta regression model and also include a brief description of the econometric beta regression model. Sections 4–6 present the application of the spatial beta econometric model to the data sets under study, which are centered on educational research. Finally, Section 7 includes some conclusions and final recommendations. In addition, the Appendix includes the relevant WinBugs code used to fit some of the models proposed here, as well as a comparison of the fitting for the proposed joint beta regression models and transformation models, where the response variable is transformed. That is, in the Appendix, we have reanalyzed the data on the student's performance in Mathematics using three different models, the proposed beta regression model with a joint modeling of the mean and dispersion parameters, and two transformation models, a heteroscedastic normal regression model applied to the logit transformation of the response variable, and a heteroscedastic model applied after the logarithm transformation of the response variable.

2. The Beta Distribution

In real data set situations, many random variables can be assumed to have a beta distribution. For example, the income inequality or the land concentration is measured using the Gini index (Atkinson, 1970) or the students' performance

in areas such as Mathematics, Natural Sciences, or Literature. In the latter case, if performance Z takes on values in the real interval (a, b), the random variable $Y = (Z - a)/(b - a)$ can be assumed to have a beta, $B(p, q)$, distribution. As we have already mentioned in the previous section, there will be variables related to, for example, household socioeconomic characteristics, which will have a significant effect on the students' cognitive achievement. More specifically, the mean level of students' achievement is closely related to their parents' educational level and to the number of hours devoted to study a given subject.

Some reparameterizations of the beta distribution can be appropriate for specific studies. An initial reparameterization makes $\phi = p + q$, so that we have that $p = \mu\phi$, $q = \phi(1 - \mu)$, and $\sigma^2 = \frac{\mu(1-\mu)}{\phi+1}$, where $\mu = E(Y)$ and $\sigma^2 = \mathrm{Var}(Y)$. In this case, ϕ can be interpreted as a precision parameter in the sense that, for fixed values of μ, larger values of ϕ would correspond to smaller values of the variance of Y, σ^2. However, given that ϕ is not exactly a precision parameter, we refer to it as a dispersion parameter instead. This reparameterization, presented in Ferrari and Cribari-Neto (2004), appeared well before in the literature (see, e.g., Cepeda-Cuervo, 2001; Jorgensen, 1997), and, given that changes in the response variable's (i.e., in the students' school performance) variability can be explained by some explanatory variables, for example, the mother's educational level, the mean and dispersion parameters can now be modeled as functions of the explanatory variables (see, e.g., Cepeda-Cuervo, 2001). Therefore, the beta regression model can be appropriate to explain the behavior of the school development as a function of several explanatory variables, which are usually denoted as its associate factors.

The beta distribution can also be reparameterized as a function of the mean and variance parameters, as proposed in Cepeda-Cuervo (2012), where $p = \frac{(1-\mu)\mu^2 - \mu\sigma^2}{\sigma^2}$ and $q = \frac{(1-\mu)[\mu - \mu^2 - \sigma^2]}{\sigma^2}$. Although writing the beta density as a function of mean and variance can generate a more complex expression for the probability density function, the joint modeling of the mean and variance can be easily obtained by applying a generalization of the Bayesian methodology originally proposed in Cepeda-Cuervo (2001). Moreover, given that the regression parameters are better and more easily interpreted within this approach, the joint modeling of the mean and variance can be sometimes more appropriate than the joint modeling of the mean and the so-called *dispersion parameter*.

3. Beta Regression Models

3.1. Joint Modeling in Beta Regression

The aforementioned reparameterization of the beta distribution as a function of μ and ϕ is interesting and allows us to define double generalized linear models, as originally proposed in Cepeda-Cuervo (2001). In this sense, the joint beta

regression modeling approach of the mean and dispersion parameters is defined, and a very flexible Bayesian methodology to fit the parameters of the proposed model is described. In a more general framework, we can assume that we have a random sample $Y_i \sim B(\mu_i, \phi_i)$, $i = 1, \ldots, n$, where both the mean and the dispersion parameters are not assumed to be constant for all observations, and, in addition, they can be modeled as functions of the regression models. That is,

$$\text{logit}(\mu_i) = \mathbf{x}'_i\boldsymbol{\beta}, \ \log(\phi_i) = \mathbf{z}'_i\boldsymbol{\alpha}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)'$ are the vectors of the mean and dispersion model parameters, and $\mathbf{x}_i$ and $\mathbf{z}_i$ are the corresponding vectors of the mean and dispersion explanatory variables for the ith observation. Later on, Ferrari and Cribari-Neto (2004) proposed the same reparameterization of the beta distribution; that is, $\mu_i = p_i/(p_i + q_i)$ and $\phi_i = a_i + b_i$. In their article, they assumed that $g(\mu_i) = \mathbf{x}'_i\boldsymbol{\beta}$, where g is a strictly monotone and twice differentiable link real-valued function defined on the interval $(0, 1)$. However, they also assumed that the dispersion parameter remains constant for any value the explanatory variables take on. Even though they considered many possible link functions g in the application they used to illustrate their proposals, they assumed g to be the logit link function, so that the parameters of the mean model could be interpreted as a function of the odds ratio. Moreover, the original joint beta regression models proposed by Cepeda-Cuervo (2001) were later used by Smithson and Verkuilen (2006) and also studied by Simas et al. (2010). At around the same time, the nonlinear beta regression model was proposed by Cepeda-Cuervo and Achcar (2010) in the settings of double generalized nonlinear models. Nonlinear regression has also been considered by Simas et al.

An innovative and very useful generalization of the aforementioned joint mean and dispersion beta regression model includes the possibility of having a random error term in the dispersion model. The main purpose of including this random error term would be to be able to incorporate in the model any possible effect that has not been adequately or possibly explained by the independent variables already included in the dispersion model. Therefore, we have that, for this new joint mean and dispersion beta random effects models, the mean can be defined by Equation 1, whereas the dispersion model is now given by

$$\log(\phi_i) = \mathbf{z}'_i\boldsymbol{\alpha} + \varepsilon_i, \tag{2}$$

where $\varepsilon_i \sim N\left(0, \tau_\varepsilon^{-1}\right)$, with unknown constant precision τ_ε. The usefulness of this model will be illustrated in Section 5 when it is applied to the analysis of the performance of school children in Language in the context of evaluating the quality of education in Colombia. It has also been applied to the analysis of the performance of school children in Mathematics, but, for the sake of brevity, these results have not been included in Section 4.

3.2. Double Generalized Spatial Beta Regression

It is of great interest for states or regions within a given country, or for countries themselves, to investigate and find out if the implemented educational policies or particular sociocultural conditions have an influence over the students' performance in specific areas, such as, for example, Mathematics, Language, or Science. In order to further investigate this research question, a particular test (of Mathematics, Language, or Science) is given to a sample (or to the whole population) of students in a particular level within the formal educational process that is being evaluated. This specific test will be used to assess the students' performance in each subject, assigning a real number (i.e., a score) in the interval (a, b) to each student participating in the study. Moreover, for each of the states or regions within the countries, or countries themselves, the average performance of the students will also take on values in the same interval (a, b). Therefore, if we denote these average performances by $\bar{Z}$, the random variable $Y = (\bar{Z} - a)/(b - a)$ will take on values in the real interval $(0, 1)$ and, thus, it can be assumed that it follows a beta distribution.

Our initial interest will be focused on trying to explain how, within the settings of beta regression models, the students' average performance is influenced by a function of the available explanatory variables associated with each of the observational units (i.e., states or municipalities) in the study. However, and given that there is a cultural interaction and that sociocultural conditions do not change from one observational unit to another, it may not be reasonable or justified to assume independence between the different observational units' performances. That is, it does make perfect sense to assume that there will be a "contagious" or "border" effect in the quality of education among geographical neighbors within the same country. This effect is clearly expected in the Colombian case for several reasons. The main and most important reason can be motivated by the obvious relation neighbors sharing borders have, given that they share similar socioeconomic and cultural characteristics. For example, a poverty or economic boom can clearly be shared across neighbors' borders, as well as their similarity in the reasons that may have caused such poverty or temporary economic shock.

Modeling proposals that may be used to include such effects are the so-called *spatial econometric models* (Anselin, 1988), where a spatial lag specification introduced as explanatory variable in the regression model can be used to be able to capture the spatial neighbors' interaction effect. This new lag-specification variable is defined as the product of the $n \times n$ weights matrix, $\mathbf{W}$, and the n dimensional vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ of response or dependent variables (i.e., the mean average performance of the observational units). The spatial weights matrix $\mathbf{W}$ defines the neighbors' set for each of the observational units, with $w_{ii} = 0$, $i = 1, \ldots, n$. That is, their elements w_{ij} are nonzero when the observational unit j belongs to the neighborhood of a geographical unit i, or if there is geographical contiguity between regions i and j. Therefore, $w_{ij} = 1$ if

geographical unit i and j share their borders (i.e., are next to each other), and $w_{ij} = 0$ otherwise. In general, $\mathbf{W}$ is a weight matrix that is typically row normalized in such a way that $\sum_j w_{ij} = 1, i = 1, \ldots, n$. Alternative definitions of the neighbors' relations are also possible. The new explanatory variable $\mathbf{WY}$ in the regression models is referred to as a *spatially lagged dependent variable* or *spatial-lag variable*. For a row-standardized weights matrix $\mathbf{W}$, it consists of a weighted average of the values of $\mathbf{Y}$ in neighboring locations, with weights given by w_{ij} (see, Anselin & Bera, 1998).

Therefore, for this specific type of application, we propose the analysis using a spatial beta regression model assuming that the spatial variable under study Y_i, $i = 1, 2, \ldots, n$, given its values in all neighborhoods of the ith region, but not including the ith region itself (i.e., $Y_{\sim i}$), is assumed to have a beta conditional distribution

$$f(y_i|y_{\sim i}) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma(\phi_i(1-\mu_i))} y^{\mu_i\phi_i - 1}(1-y)^{\phi_i(1-\mu_i)-1} I_{(0,1)}(y), \tag{3}$$

where $\mathrm{E}(Y_i|Y_{\sim i}) = \mu_i$ and $\mathrm{Var}(Y_i|Y_{\sim i}) = \frac{\mu_i(1-\mu_i)}{1+\phi_i}$ represent the conditional mean and the conditional variance, respectively. This model includes, in addition to the regression structures already assumed in Equation 1, a spatial lag of the standardized values $\tilde{\mathbf{z}} = (\mathbf{y} - \bar{y}\,\mathbf{1}_n)/s$, which will be included both in the mean and in the dispersion models, where $\mathbf{y} = (y_1, \ldots, y_n)'$ is a realization of the aforementioned assumed beta random response variable under study $\mathbf{Y}$, y_i is a specific sample value taken by this random variable, $\bar{y}$ and s are the response sample mean and standard deviation, and $\mathbf{1}_n$ is an n vector of ones. Therefore, and given that $\mathbf{Wz} = c_1\mathbf{Wy} + c_2\mathbf{1}_n$, where c_1 and c_2 are constants, the mean and dispersion model can also be written in an equivalent way as:

$$h(\mu_i) = \lambda(\mathbf{W}_1\mathbf{y})_i + \mathbf{x}'_i\boldsymbol{\beta}, \tag{4}$$

$$g(\phi_i) = \rho(\mathbf{W}_2\mathbf{y})_i + \mathbf{z}'_i\boldsymbol{\alpha} + \varepsilon_i, \tag{5}$$

where $\varepsilon_i \sim N\left(0, \tau_\varepsilon^{-1}\right)$. In addition, in the mean model given in Equation 4, $\mathbf{x}'_\mathbf{i} = (\mathbf{x_{0i}}, \mathbf{x_{1i}}, \ldots, \mathbf{x_{Ki}})$ is the ith row of the $N \times (K+1)$ design matrix, $\mathbf{W_1}$ is an $N \times N$ spatial weights matrix, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)'$ is the K-dimensional mean parameter vector, and λ is the spatial autoregressive coefficient. In the dispersion model in Equation 5, $\mathbf{z}'_i = (z_{0i}, z_{1i}, \ldots, z_{Pi})$ is the ith row of the $N \times (P+1)$ design matrix, $\mathbf{W_2}$ is an $N \times N$ spatial weights matrix, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_P)'$ is the P-dimensional dispersion parameter vector, and ρ is the spatial autoregressive coefficient. In addition, h and g are appropriately chosen real functions, which are usually the logit function for the mean model and the logarithm function for the dispersion model. The latter function selection guarantees the positivity of the variance. In general, we assume that the first elements of $\mathbf{x}_i$ (i.e., x_{0i}) and

$\mathbf{z}_i$ (i.e., z_{0i}) are equal to 1, so that both the mean and the dispersion models include an intercept. In these models, $\mathbf{W_1}$ and $\mathbf{W_2}$ are square matrices whose entries, $w_{1,ij}$ and $w_{2,ij}$, reflect the intensity of the spatial interdependence between regions i and j. The elements $w_{k,ij}$, $(k = 1, 2)$ should be positive real numbers, but there is not a single definition for them. However, the most commonly used form is the one that assumes a first-order physical contiguity (see, e.g., Geary, 1954; Moran, 1948). Alternative definitions of $w_{k,ij}$ based on distances between regions are provided in Cliff and Ord (1973). We should mention that, in order to avoid any model unresolved simultaneities, the mean and dispersion parameters (i.e., μ_i and ϕ_i) in Equations 4 and 5 are really conditional mean and dispersion parameters, which are indeed conditioned, as previously motivated in this section, on all neighborhoods of the ith region, not including the ith region.

In the previously proposed model, the lag spatial variable $\mathbf{Wy}$ is included. However, other possibilities could also be considered. For example, the lag variable defined by $\mathbf{W}\tilde{\mathbf{y}}$ can be assumed, where $\tilde{\mathbf{y}}$ is the vector with individual elements given by $\tilde{y}_i = h(y_i)$ in the mean model in Equation 4, and $\tilde{y}_i = g(y_i)$ in the dispersion model in Equation 5, respectively. Therefore, the spatial lag variables $\mathbf{Wy}$ and $\mathbf{W}\tilde{\mathbf{y}}$ could be considered in the spatial analysis when using beta regression models. We will illustrate these two possibilities in the applications included in Sections 5 and 6, where the spatial beta regression model is applied to the study of the performance in the subject "Language" of Colombian school children and to the study of the variables affecting the proportion of illiteracy for a population with ages starting at 15 in Colombia.

Finally, in order to apply Bayesian methodology, very flexible and general independent normal prior distributions $N(0, 10^k)$, with $k = 5$, are assumed for all the parameters in the model; more specifically, for λ, ρ, and for each one of the components of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Given that the posterior distribution is unknown, we can obtain the posterior sampling distribution of the parameters of interest using the well-known and commonly used WinBugs statistical software package, where we previously determine the lag of the terms in the models using the *R* statistical software package.

4. Application: Quality of Education in Colombia

In this section, we use the aforementioned double generalized beta regression model to analyze a real data set obtained from the evaluation of the quality of education in Colombia, disaggregated by department. Although the average performance score in Mathematics takes on values in the closed real interval $[0, 100]$ (i.e., larger values are associated with high efficiency and smaller values to low efficiency), in this specific study the smallest value for the average performance score in Mathematics was 51.35 and the largest one was 69.82. Therefore, most of the scores are located in a shorter interval of performance scores and, without loss of generality, the proposed methodology is clearly able to handle this specific situation.

In this application, the response variable under study is the average development in Mathematics (for each one of the 31 departments) for students in their fourth year of secondary school in Colombia. As mentioned above, departments in Colombia are different regions into which the country is divided, similar to what provinces represent in European countries or states in the United States. The data have been kindly provided by the Colombian National Institute of Evaluation (ICFES) and the Colombian National Administrative Department of Statistics (DANE). The average performance in Mathematics for each department is denoted by P and takes on values in the closed real interval $[0, 100]$. Therefore, and in order to be able to assume a beta distribution for it, we define a new random variable for this average performance by $Y = (P - a)/(b - a)$, where a is the largest number smaller than $P_{(1)}$, the minimum value of the average performance in the sample, and b is the smallest value larger than $P_{(n)}$, the maximum average performance value in the sample. The available explanatory variables in the sample were obtained from the DANE and are the unmet basic needs (UBN) and the percentage of teachers having a postgraduate level of education (PORC). Goodness of fit for fitted models will be assessed by means of the Deviance Information Criterion (DIC) value. The DIC criterion is considered a natural way to compare models because it is based on a trade-off between the fit of the data to the model and the corresponding complexity of the model (Spiegelhalter, Best, Carlin, & Van der Linde, 2002). Therefore, models with smaller DIC value are better supported by the data, in terms of both the model's goodness of fit and the model's complexity (i.e., effective number of parameters in the model). We would like to mention that the DIC value can be negative. More specifically, and given that the deviance can be negative because a specific probability density function can take on values larger than 1, this especially occurs in the case of the beta distribution, a continuous distribution that takes on values larger than 1 on a bounded interval (i.e., the interval $(0, 1)$). Moreover, in the definition of the DIC, the deviance is given by $D(\theta) = -2\ log\ p(y|\theta)$ and, thus, if a given probability density function $p(y|\theta)$ is larger than 1, the DIC = "goodness of fit" + "complexity" should be negative. Moreover, the interpretation of the DIC value is the same as before. That is, smaller values are associated with better fitting models. For more details on DIC values, interested readers can consult the seminal article by Spiegelhalter, Best, Carlin, and Van der Linde (2002). In addition, there exist alternative model complexity measures that, in general, produce similar results to those obtained when using the DIC (see, e.g., Lu, Hodges, and Carlin, 2007).

4.1. Analysis of the Performance in Mathematics of Colombian School Children: The Colombian Border Plan

The Colombian border plan was a program of the Colombian government whose main objective was to strengthen and visualize the National Government Relations, as well as those relations between neighboring communities, by

TABLE 1
Parameter Estimates, Together With Their Standard Deviations and 95% Credible Intervals (CI), for Parameters in the Initial Beta Regression Model in Equation 6

Parameters	β_0	β_1	γ_0	γ_1	α_0
Estimates	0.717	−1.956	2.855	−2.288	−1.982
95% CI lower bounds	−0.549	−3.789	−0.214	−4.160	−2.443
95% CI upper bounds	2.049	−0.195	5.873	−0.352	−1.480
Standard deviations	0.649	0.918	1.532	0.947	0.249

Note: The value of minus 2 times the log-likelihood function is $-2 \log L = -31.618$ and the value of the corresponding Deviance Information Criterion (DIC) = −22.087.

strengthening governance in 13 departments of the Colombian border. Health, education, training, culture, sports, and other related programs were accordingly implemented in these regions. Therefore, it is of interest to compare how the quality of education depends upon the explanatory variables considered in this study, and, in addition, to compare the model for departments in the Colombian border with those not in the border. In order to implement this comparison, a new explanatory variable $\mathbf{V} = (v_1, \ldots, v_n)'$ is considered. This variable takes value 1 if the department has borders with other countries, and 0 otherwise. As mentioned in Section 4, we assume a beta distribution model, so that $Y_i \sim B(\mu_i, \phi_i)$, with mean and dispersion models now given by

$$\begin{aligned} \text{logit}(\mu_i) &= \beta_0 + \beta_1 \text{UBN}_i + \beta_{2i} \text{PORC}_i \\ \log(\phi_i) &= \alpha_0, \end{aligned} \tag{6}$$

where $\beta_{2i} = \gamma_0 + \gamma_1 v_i$. In order to be able to apply the innovative Bayesian methodology to fit this model for the quality of education data, independent normal $N(0, 10^k)$, with $k = 5$, prior distributions are assumed for all of the parameters in the proposed model. Table 1 includes the parameter estimates, together with their corresponding standard deviations and 95% credible intervals, as well as the value of minus 2 times the log-likelihood function and the corresponding DIC value.

We should mention that, given that the mean model in Equation 6 does include the indicator variable **V** for the covariate PORC, but not for the covariate UBN, this model assumes that the relationship between the mean performance and the covariate UBN is the same for all departments in the country. This is a well-accepted assumption by members of the Colombian educational institutions, which basically suggests that the indicator variable **V** will only have a real effect on the mean model through the covariate PORC. In any case, we have verified the validity of this assumption using it in the covariate UBN and arrive at the conclusion that, for this specific data set, this assumption holds. In addition, we have verified the need to include the indicator variable **V** or not in the dispersion model in Equation 6, concluding that it is not necessary to do so.

As a result of fitting model (Equation 6) and from the estimated parameters reported in Table 1, we can conclude that the mean model for departments in the Colombian border (i.e., $v_i = 1$) is given by $\text{logit}(\mu_i) = 0.717 - 1.956\,\text{UBN}_i + 0.567\,\text{PORC}_i$, and that the corresponding one for departments that are not in the Colombian border (i.e., $v_i = 0$), is given by $\text{logit}(\mu_i) = 0.717 - 1.956\,\text{UBN}_i + 2.855\,\text{PORC}_i$. Therefore, we can see that, when the explanatory variable PORC increases, the mean increases more rapidly for departments that are not in the Colombian border than for those that are in the Colombian border.

Some of the alternative models that were considered in this study included NBI, PORC, as well as the lag **Wy** variables in the mean model. However, and based on the obtained DIC values for each of these models, none of them provided a better fit than that obtained using Equation 6. Therefore, Equation 6 was the best fitted model for this data set.

5. Spatial Analysis of the Performance in Language of Colombian School Children

In this second application, the response variable under study is the average development in Language (by department) for students in their fourth year of secondary school studies. As in the previous example, data were provided by the ICFES and the DANE. The average performance in Language for each department is denoted by P and it takes on values in the closed real interval between 0 and 100 (i.e., $[0, 100]$), with the smallest value in the data equal to 53, and the largest one equal to 67. In order to be able to assume a beta distribution, and as in Section 4 for the score in Mathematics, a similar transformation was performed to the raw scores so that students' scores performance was in the $(0, 1)$ interval. Using the same explanatory variables that we have previously used to study the academic performance in Mathematics (i.e., the UBN; and the percentage of teachers having a PORC), we initially suggest the use of the following spatial models for the mean and dispersion parameters, where we assume that $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W}$.

$$\text{logit}(\mu_i) = \lambda(\mathbf{Wy})_i + \beta_0 + \beta_1\,\text{UBN}_i + \beta_2\,\text{PORC}_i, \tag{7}$$

$$\log(\phi_i) = \rho(\mathbf{Wy})_i + \alpha_0 + \alpha_1\text{UBN}_i + \alpha_2\text{PORC}_i + \varepsilon_i, \tag{8}$$

where $\varepsilon_i \sim N\left(0, \tau_\varepsilon^{-1}\right)$.

As in previous sections, independent normal $N(0, 10^k)$, with $k = 5$, prior distributions are also assumed for the mean and dispersion regression parameters, and a prior gamma $G(0.001, 0.001)$ distribution for the precision τ_ε parameter. Table 2 includes the parameter estimates, together with their corresponding standard deviations and 95% credible intervals for the parameters in the spatial models for the mean and dispersion parameters given by Equations 7 and 8, and the value of minus 2 times the log-likelihood function as well as their corresponding DIC value.

TABLE 2
Parameter Estimates, Together With Their Standard Deviations and 95% Credible Intervals (CI), for the Mean Parameters in the Beta Regression Model in Equation 7 and Dispersion Parameters in Equation 8, With $\varepsilon_i \sim N(0, \tau_\varepsilon^{-1})$

Mean Parameters	λ	β_0	β_1	β_2	
Estimates	−0.6136	0.4504	−2.658	4.378	
95% CI lower bounds	−2.4160	−0.1777	−4.114	2.161	
95% CI upper bounds	0.6942	1.5630	−1.545	6.958	
Standard deviations	0.6429	0.4143	0.777	1.190	
Dispersion parameters	ρ	α_0	α_1	α_2	τ_ε
Estimates	−9.519	7.202	−1.722	4.762	176.80
95% CI lower bounds	−16.550	2.332	−5.443	−1.100	0.7545
95% CI upper bounds	−2.335	11.650	2.403	10.400	1302.0
Standard deviations	3.620	2.407	1.995	2.929	389.40

Note: The value of minus 2 times the log-likelihood function is $-2 \log L = -50.140$ and the corresponding Deviance Information Criterion = −33.271.

TABLE 3
Parameter Estimates, Together With Their Standard Deviations and 95% Credible Intervals (CI), for the Mean Parameters in the Beta Regression Model in Equation 9 and Dispersion Parameters in Equation 8, With $\varepsilon_i \sim N(0, \tau_\varepsilon^{-1})$

Mean Parameters	β_0	β_1	β_2		
Estimates	0.2959	−2.540	3.769		
95% CI lower bounds	−0.2254	−3.676	2.080		
95% CI upper bounds	0.8381	−1.491	5.389		
Standard deviations	0.2583	0.5424	0.830		
Dispersion parameters	ρ	α_0	α_1	α_2	τ_ε
Estimates	−10.54	7.901	−2.276	5.225	171.5
95% CI lower bounds	−16.87	3.726	−5.830	−0.324	0.834
95% CI upper bounds	−4.175	11.90	0.909	11.19	1343.0
Standard deviations	3.240	2.020	1.751	2.972	356.5

Note: The value of minus 2 times the log-likelihood function is $-2 \log L = -50.619$ and the corresponding Deviance Information Criterion = −35.251.

In addition and given the results reported in Table 2, we have considered the model that does not include the spatial effect in the mean model. That is, we also consider the beta regression model with mean given by

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{UBN}_i + \beta_2 \text{PORC}_i, \tag{9}$$

and dispersion model given by Equation 8. Table 3 includes the parameter estimates, together with their corresponding standard deviations and 95% credible intervals for the parameters in the spatial models for the mean and dispersion parameters given by Equations 9 and 8, and the value of minus 2 times the log-likelihood function as well as their corresponding DIC value, which is a smaller value than that associated with the more general model in Equations 7 and 8. Based on the results obtained from the fitting of the aforementioned models and on their DIC values, the latter model will be the best one to study students' Language performance. It is important to mention that these fitted models clearly motivate and justify the usefulness of including the spatial effect structure in the model. In our view, and given that language development is expected to be associated with socioeconomic factors, this is a natural fact. In the results reported in Tables 2 and 3, the resulting estimate for the coefficient associated with the variable UBN is negative and statistically significant, implying that larger values of the variable UNB are associated with poorer students' Language performance scores. Along the same lines and given that the estimated coefficient associated with the variable PORC (i.e., the percentage of teachers having a PORC) is positive and statistically significant, this implies that larger values of the variable PORC are associated with better students' Language performance scores. Finally, in order to better justify the need and usefulness of the proposed spatial beta regression model, we have also fitted the model in Equations 7 and 8 (detailed results not reported for brevity), but without including its spatial terms (i.e., with $\rho = 0$ and $\lambda = 0$), providing a larger DIC value (i.e., DIC $= -22.169$). As can be seen, this reported DIC value is larger than either that of the best-fitting spatial beta regression model (i.e., the one with results reported in Table 3 and DIC $= -35.251$), which clearly provides a better fit for the data set under study, or that of the spatial beta regression model in Equations 7 and 8 (i.e., the one with results reported in Table 2 and DIC $= -33.271$).

6. Spatial Analysis of Illiteracy in Colombia

In this application, we study the illiteracy proportion for a population aged 15 and over in Colombia. Data were collected by DANE. Illiteracy proportions are lower than 16%, except for the Department of Choco, which was not considered in this study and reported, in 2005, a illiteracy proportion of 22.1%. In this study, the department with smaller illiteracy proportion is Valle, reporting an illiteracy proportion of 4.8%. Given the social and economic specific characteristics of Colombia, it is clear that the "illiteracy" proportion response variable includes a spatial association and, therefore, it should be incorporated into the model. We also consider, as in the previous applications, some relevant explanatory variables, such as the UBN, mainly because it is evident that the possibility of having no access to a given education is determined by the lack of a set of individual basic living conditions. This specific application intends to illustrate some

TABLE 4
Parameter Estimates, Together With Their Standard Deviations and 95% Credible Intervals (CI), for the Mean Parameters in the Beta Regression Model in Equation 10 and Dispersion Parameters in Equation 11

Mean Parameters	λ	β_0	β_1
Estimates	8.072	−3.303	0.008766
95% CI lower bounds	0.1167	−3.927	−0.002238
95% CI upper bounds	15.54	−2.646	0.02079
Standard deviations	3.818	0.3249	0.005849
Dispersion parameters	α_0	α_1	
Estimates	5.532	−0.02844	
95% CI lower bounds	4.105	−0.06699	
95% CI upper bounds	6.887	0.006621	
Standard deviations	0.7109	0.01875	

Note: The value of minus 2 times the log-likelihood function is $-2 \log L = -131.190$ and the corresponding Deviance Information Criterion = −121.409.

relevant topics related to the different possible specifications of the spatial beta regression model. The first model we have fitted to the illiteracy data set includes mean and dispersion models given by Equations 10 and 11. That is,

$$\text{logit}(\mu_i) = \lambda(\mathbf{Wy})_i + \beta_0 + \beta_1 \text{UBN}_i, \tag{10}$$

$$\log(\phi_i) = \alpha_0 + \alpha_1 \text{UBN}_i. \tag{11}$$

Table 4 includes the parameter estimates, together with their corresponding standard deviations and 95% credible intervals, as well as the value of minus 2 times the log-likelihood function and the corresponding DIC value.

The second spatial beta regression model we have considered is similar to the previous one in the specification of the dispersion model, but with a mean model now given by

$$\text{logit}(\mu_i) = \lambda \text{ logit}[(\mathbf{Wy})_i] + \beta_0 + \beta_1 \text{UBN}_i. \tag{12}$$

We should mention that the latter model includes the lag variable having the same structure than that in the link function. Table 5 includes the parameter estimates, together with their corresponding standard deviations and 95% credible intervals, as well as the value of minus 2 times the log-likelihood function and the corresponding DIC value. Parameter estimates reported in Table 5 suggest several very interesting issues to be carefully analyzed: There is a clear agreement between the regression parameter estimates obtained for the explanatory variable and also in their DIC values. Moreover, the credible intervals obtained

TABLE 5
Parameter Estimates, Together With Their Standard Deviations and 95% Credible Intervals (CI), for the Mean Parameters in the Beta Regression Model in Equation 12 and Dispersion Parameters in Equation 11

Mean Parameters	λ	β_0	β_1
Estimates	0.6634	−1.064	0.008626
95% CI lower bounds	0.05396	−2.655	−0.001691
95% CI upper bounds	1.165	0.2369	0.019910
Standard deviations	0.2792	0.7254	0.005492
Dispersion parameters	α_0	α_1	
Estimates	5.533	−0.02875	
95% CI lower bounds	4.095	−0.06664	
95% CI upper bounds	6.866	0.00693	
Standard deviations	0.7074	0.01868	

Note: The value of minus 2 times the log-likelihood function is $-2 \log L = -130.616$ and the corresponding Deviance Information Criterion $= -121.409$.

for the regression model parameter λ in both models do not include the value zero. Therefore, we can conclude that in both models the lag variable has a significant contribution in explaining the illiteracy proportion variable under study.

In addition, to further illustrate the contribution made by the aforementioned lag variable in the spatial beta regression model, a third model was fitted. This model does not include any spatial component in the mean model, and it only includes the UBN variable as explanatory variable. The dispersion model is given by Equation 11. Thus, the mean model is now given by

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{UBN}_i. \tag{13}$$

Table 6 includes the parameter estimates, together with their corresponding standard deviations and 95% credible intervals, as well as the value of minus 2 times the log-likelihood function and the corresponding DIC value. The DIC value for this model (i.e., DIC $= -118.932$) is larger than those from the previous two models that included a spatial structure in the beta regression model.

From the results of the aforementioned models, we can conclude that the variable UBN may not be significant in the mean and dispersion models, and that the lag variable is clearly significant in the mean model. There still remains the question about the need to include or not the lag variable in the dispersion model. Therefore, as a final model for the illiteracy data, we fit mean and dispersion models that do not include the UBN explanatory variable, but that include the corresponding lag variables. That is,

TABLE 6
Parameter Estimates, Together With Their Standard Deviations and 95% Credible Intervals (CI), for the Mean Parameters in the Beta Regression Model in Equation 13 and Dispersion Parameters in Equation 11

Mean Parameters	β_0	β_1
Estimates	−2.785	0.01531
95% CI lower bounds	−3.153	0.00415
95% CI upper bounds	−2.396	0.02702
Standard deviations	0.1905	0.005749
Dispersion parameters	α_0	α_1
Estimates	5.653	−0.03525
95% CI lower bounds	4.290	−0.07409
95% CI upper bounds	7.001	−0.00126
Standard deviations	0.696	0.01842

Note: The value of minus 2 times the log-likelihood function is $-2 \log L = -126.984$ and the corresponding Deviance Information Criterion = −118.932.

TABLE 7
Parameter Estimates, Together With Their Standard Deviations and 95% Credible Intervals (CI), for the Mean Parameters in the Beta Regression Model in Equation 14 and Dispersion Parameters in Equation 15

Mean Parameters	λ	β_0
Estimates	10.51	−3.279
95% CI lower bounds	3.87	−3.909
95% CI upper bounds	17.16	−2.654
Standard deviations	3.426	0.3264
Dispersion parameters	ρ	α_0
Estimates	−16.13	6.069
95% CI lower bounds	−40.81	3.574
95% CI upper bounds	7.874	8.479
Standard deviations	12.57	1.251

Note: The value of minus 2 times the log-likelihood function is $-2 \log L = -129.734$ and the corresponding Deviance Information Criterion = −122.0.

$$\text{logit}(\mu_i) = \lambda(\mathbf{Wy})_i + \beta_0, \tag{14}$$

$$\log(\phi_i) = \rho(\mathbf{Wy})_i + \alpha_0. \tag{15}$$

Table 7 includes the parameter estimates, together with their corresponding standard deviations and 95% credible intervals, as well as the value of minus 2

times the log-likelihood function and the corresponding DIC value. The DIC value for this model (i.e., DIC $= -122.0$) is smaller than those from the two previous best fitting models that included a spatial structure only in the mean model for the beta regression model. Moreover, alternative models to those in Equations 14 and 15, but that included the term $\text{logit}(\mathbf{Wy})$ instead of $\mathbf{Wy}$ in both the mean and the dispersion models, were also considered but did not provide a better fit than that of the best fitting model. Therefore, this is clearly suggesting that it is more convenient, in terms of goodness of fit, to consider the lag variable instead of its logit transformation in order to model the spatial structure in both the mean and the dispersion models. Finally, from the results obtained in the four models fitted to the illiteracy data, we conclude that, based on the DIC values, the best fitting model is the one that included the spatial structure in both the mean and the dispersion models and that, in addition, the competing models were those only including the different spatial structures in the mean model.

7. Conclusion

We have motivated an innovative application of the generalized beta spatial regression models to the analysis of the quality of education in Colombia. In order to do so, the beta distribution was briefly described and the joint modeling approach for the mean and dispersion parameters in the spatial regression models' setting was introduced. Our proposed methodology suggests the use of a Bayesian version of the beta econometric spatial model recently proposed by Cepeda-Cuervo et al. (2011), especially of those models that consider the inclusion of a random effect term in the dispersion model. Results obtained from the analysis of the generalized beta spatial regression model to the study of school children performance in Mathematics have suggested that factors such as the UBN and the teachers' training are significant factors to be taken into account by the government in order to improve the quality of education in Colombia.

Finally, when the model takes into account the information about departments belonging or not to the border of Colombia, we could see that, when the explanatory variable measuring the teacher's training, PORC, increases, the mean increases more rapidly for departments that are not in the Colombian border than for those that are in the Colombian border. Proposed models have shown their usefulness and applicability within this area of research.

Even though the models associated with the performance in Mathematics did not include the spatial factor $\mathbf{Wy}$ as explanatory variable, the models that explained the performance in Language did so, and these parameters were statistically significant. This is a natural fact because researchers have commonly associated language development with socioeconomic factors.

In the applications used in this article to illustrate the proposed methodology, inferences from the posterior distribution samples were obtained using the well-known WinBugs software. However, the generalized beta distribution models

can be fitted by applying different methodologies. A classic approximation can be applied by maximizing the logarithm of the likelihood function with the use of the Newton Raphson algorithm. A Bayesian approach can be applied as well by appropriately adapting and using the Bayesian methodology proposed in Cepeda-Cuervo (2001).

Appendix

In this section, we include the WinBugs code that can be used by interested readers to fit the joint mean and dispersion beta regression models. In this specific WinBugs code, it is assumed that Y_i, $i = 1, \ldots, n$, follows a beta distribution with parameters p_i and q_i. In addition, it is assumed that the mean and dispersion parameters follow the models given by $\text{logit}(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$, with $\beta_0 = b0$, $\beta_1 = b1$, and $\beta_2 = b2$, and $ln(\phi_i) = \alpha_0 + \alpha_1 X_{2i} \Rightarrow \phi_i = exp(\alpha_0 + \alpha_1 X_{2i})$, with $\alpha_0 = c0$, $\alpha_1 = c1$, and $\phi_i = \tau_i$, respectively. Models are fitted by assuming normal prior distributions for all of the regression parameters.

Joint mean and dispersion beta regression model

```
model
    {
        for( i in 1 : N ) {
            Y[i] ~ dbeta(p[i],q[i])
            p[i]<-mu[i]*tau[i]
            q[i]<-tau[i]-mu[i]*tau[i]
            logit(mu[i])<-b0+  b1*x1[i]+b2*x2[i]
            tau[i] <-exp(c0+ c1*x2[i])

                        }
        b0 ~ dnorm(0.0,1.0E-2)
        b1 ~ dnorm(0.0,1.0E-2)
        b2 ~ dnorm(0.0,1.0E-2)
        c0 ~ dnorm(0.0,1.0E-2)
        c1 ~ dnorm(0.0,1.0E-2)
            }
```

```
Data
 list(Y=c(0.53,0.82,0.69,0.62,0.47,0.94,0.87,0.78,0.82,0.68,0.68,
 0.48,0.53,0.83,0.56,0.34,0.74,0.53,0.53,0.81,0.06,0.74,0.58,0.53,
 0.80,0.95,0.64,0.71,0.58,0.68,0.24),
x1=c(0.441,0.23,0.356,0.247,0.466,0.308,0.177,0.415,0.355,0.466,
0.447,0.792,0.591,0.213,0.602,0.399,0.326,0.651,0.477,0.25,0.436,
0.303,0.345,0.162,0.173,0.219,0.549,0.298,0.156,0.548,0.668),
x2=c(0.188,0.185,0.428,0.159,0.09,0.46,0.413,0.342,0.312,0.327,
0.185,0.278,0.283,0.271,0.362,0.178,0.3918,0.225,0.131,0.392,
0.411,0.424,0.336,0.332,0.272,0.383,0.204,0.3978,0.169,0.133,0.188),
 N=31)
Inits


list(b0=-1,b1=-1,b2=0,c0=-3,c1=0)
```

Graphical Model

Figure A1 presents a graphical representation describing the sequencing of the *Markov chain Monte Carlo* sampling procedure used by the WinBugs code provided above. In this diagram beta0, beta1, and beta2 represent the parameters in the mean model for mu[i], and alpha0 and alpha1 the corresponding parameters in the dispersion models for phi[i]. In the Bayesian settings, these parameters are assumed to be random and, thus, in the program, a normal prior distribution is assumed for each of them. In Figure A1, Y[i] is the response variable, X1[i] and X2[i] represent the values of the explanatory variables included in the mean model, and Z[i] the corresponding vales for the explanatory variables included in the dispersion model. Stochastic dependence and functional dependence are denoted by single-edged arrows and double-edged arrows, respectively. For example, between the model parameter beta0 and the mean model mu[i] there is functional dependence and between p[i] and Y[i] there is stochastic dependence.

Model Comparisons: Joint Beta Regression Models Versus Transformation Models

Transformations of the response variable or of the variable under study are commonly used with bounded outcome scores; that is, when the response or dependent variable takes on values in closed intervals. These approaches are

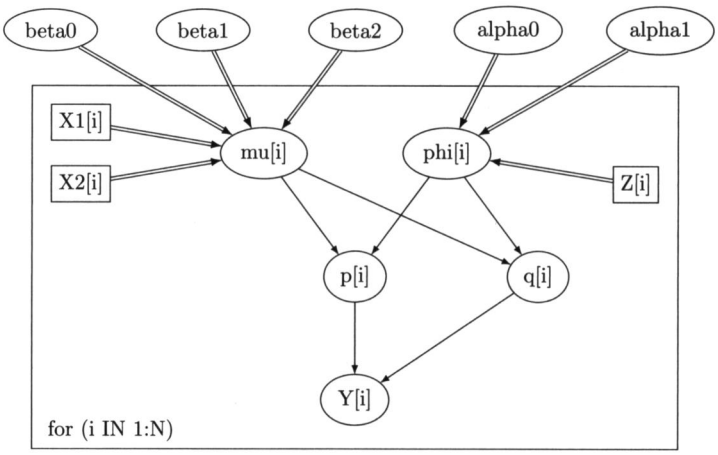

FIGURE A1. *Directed graph for the joint beta regression models.*

commonly denoted as transformation models. Therefore, transformation models may be an alternative to be considered before even proposing the fit of the joint beta regression models, whose use we have motivated and illustrated in this article. This is one reason that has led us to consider response variables given by the logit and logarithm transformations of the students' mean performance in Mathematics, with the main objective of comparing the performance of the joint beta and normal regression model approaches. In order to do so, we consider three possible models. The first one, the joint beta regression model with mean and dispersion models given by

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{UBN}_i + \beta_2 \text{PORC}_i \tag{A1}$$

$$\log(\phi_i) = \alpha_0 + \alpha_1 \text{PORC}_i. \tag{A2}$$

The second one, a heteroscedastic normal regression model $\tilde{Y}_i \sim N(\mu_i, \sigma_i^2)$, with μ_i and σ_i^2 given by

$$\mu_i = \beta_0 + \beta_1 \text{UBN}_i + \beta_2 \text{PORC}_i \tag{A3}$$

$$\log(\sigma_i^2) = \alpha_0 + \alpha_1 \text{PORC}_i, \tag{A4}$$

where the observations of the response variable, $\tilde{Y}_i$, are given by the expression $\tilde{y}_i = \text{logit}[y_i/(1-y_i)]$, $i = 1, \ldots, n$, and y_i represents the mean observed performance in Mathematics. The third model is a heteroscedastic normal regression model, where the observations of the response variable, $\tilde{Y}_i$, are given by $\tilde{y}_i = log(y_i)$, $i = 1, \ldots, n$. Even though DIC values are reported for all of the three aforementioned models, and given that the response variable is not the same for each of these models, comparison should be done in terms of the sum

TABLE A1

Parameter Estimates, Together With Their Standard Deviations, for Parameters in the Joint Beta Regression Model Given by Equations A1 and A2, and the Heteroscedastic Normal Regression Models Given by Equations A3 and A4 With Logit and Logarithm Transformations

Model	Parameter	β_0	β_1	β_2	α_0	α_1
Beta regression	*M*	0.907	−2.056	1.448	3.555	−5.527
	SD	0.484	0.765	1.230	0.716	2.282
DIC = −22.280 SS = 0.810						
Normal logit	*M*	−0.889	−2.321	2.357	−2.160	6.460
	SD	0.573	0.890	1.518	0.754	2.417
DIC = 83.186 SS = 5.266						
Normal logarithm	*M*	0.278	−0.923	0.399	−3.847	7.472
	SD	0.272	0.428	0.757	0.753	2.385
DIC = 39.584 SS = 6.115						

Note: DIC = Deviance Information Criterion; SS = sum of squares.

of squares (SS) given by $SS = \sum(y_i - \hat{y}_i)^2$, where $\text{logit}(\hat{y}_i) = \mathbf{x}'_i\hat{\beta}$ for the first two models being considered here. For the heteroscedastic normal regression model with logarithm transformation of the response variable, the SS is computed using $log(\hat{y}_i) = \mathbf{x}'_i\hat{\beta}$ instead. In addition, we should mention that SS corresponds to the SS for errors in the joint beta regression model, but not in the heteroscedastic normal regression models. Table A1 includes the parameter estimates for the models being compared here, as well as their standard deviations and the corresponding DIC and SS values for all of the three models above. As can be seen, the SS value for the joint beta regression model (i.e., SS = 0.810) is much smaller than those for the heteroscedastic normal regression model with logit transformation (i.e., SS = 5.266) and for the heteroscedastic normal regression model with logarithm transformation (i.e., SS = 6.115). Therefore, we can see that the proposed joint beta regression model seems to be the one providing a better fit to the data under study and, in addition, it has a better performance than that of the two transformation models considered.

Finally and also in order to better motivate the need for the use and usefulness of the spatial beta regression model with mean and dispersion models given by Equations 9 and 8, respectively, we compare the results obtained from this best

TABLE A2
Goodness-of-Fit Criteria Results Comparing the Fit for the Joint Spatial Beta Regression Model Given by Equations 9 and 8, and Those for the Heteroscedastic Normal Regression Models Given by Equations A5 and A6 With Logit and Logarithm Transformations

Model	DIC	SS
Beta regression	−35.251	0.692
Normal logit	73.933	0.756
Normal logarithm	19.488	0.801

Note: DIC = Deviance Information Criterion; SS = sum of squares.

fitting model to those from a spatial normal regression model (with logit and logarithm transformations) with mean and variance models, respectively, given by:

$$\mu_i = \beta_0 + \beta_1 \mathrm{UBN}_i + \beta_2 \mathrm{PORC}_i \tag{A5}$$

$$\log(\sigma_i^2) = \rho(\mathbf{Wy})_i + \alpha_0 + \alpha_1 \mathrm{UBN}_i + \alpha_2 \mathrm{PORC}_i + \varepsilon_i, \tag{A6}$$

where $\varepsilon_i \sim N\left(0, \tau_\varepsilon^{-1}\right)$, with τ_ε being an unknown parameter. Table A2 includes the goodness-of-fit criteria (i.e., DIC and SS) used to compare the fit of the joint spatial beta regression model given by Equations 9 and 8, and those for the heteroscedastic spatial normal regression models given by Equations A5 and A6 with logit and logarithm transformations. Therefore, we can see that the proposed spatial joint beta regression model seems to be the one providing a better fit to the data under study and, in addition, it has a better performance than that of the two normal transformation models considered.

Acknowledgments

The authors wish to thank the editor and two anonymous referees for providing thoughtful comments and suggestions which have led to substantial improvement of the presentation of the material in this article.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work of the first author was supported by the Department of Statistics of Universidad Nacional de Colombia and by Ministerio de Ciencia e Innovación and FEDER under research grant MTM2010-14913. The work of the second autor was supported by Ministerio de Ciencia e Innovación, FEDER, the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group), and Universidad del País Vasco UPV/EHU under research grants MTM2010-14913, IT-334-07, UFI11/03, US12/09, and IT-642-13.

References

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

Anselin, L (1999). The future of spatial analysis in the social sciences. *Geographic Information Sciences*, *5*, 67–76.

Anselin, L., & Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah & D. E. A. Giles (Eds.), *Handbook of applied economic statistics* (pp. 237–289). New York, NY: Marcel Dekker.

Anselin, L., & Florax, R. (Eds.). (1995). *New directions in spatial econometrics*. Berlin, Germany: Springer-Verlag.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, *2*, 244–263.

Cepeda, C. E. (2005). Factores asociados al logro cognitivo en matemáticas. *Revista de Educación*, *336*, 503–514.

Cepeda, C. E., & Gamerman, D. (2005). Bayesian methodology for modeling parameters in the two parameters exponential family. *Revista Estadística*, *57*, 93–105.

Cepeda-Cuervo, E. (2001). *Modelagem da Variabilidade em Modelos Lineares Generalizados* (Unpublished PhD thesis). Universidade Federal de Rio de Janeiro, Brazil.

Cepeda-Cuervo, E. (2012). *Beta regression models: Joint mean and variance modeling* (Reporte Técnico). Universidad Nacional de Colombia: Bogotá, Colombia. Retrieved from: http://www.bdigital.unal.edu.co/6207/

Cepeda-Cuervo, E., & Achcar, J. A. (2010). Heteroscedastic nonlinear regression models. *Communications in Statistics—Simulation and Computation*, *39*, 405–419.

Cepeda-Cuervo, E., Urdinola, B. P., & Rodríguez, D. (2011). Double generalized spatial econometric models. *Communication in Statistics—Simulation and Computation*, *45*, 671–685.

Cliff, A., & Ord, J. (1973). *Spatial autocorrelation*. London, England: Pion.

Coleman, J. S. (1969). *Equal educational opportunity*. Cambridge, MA: Harvard University Press.

Donoso, D. S. (2002). School efficiency and socioeconomic differences: On the results of assessment exams of the Quality of Education in Chile. *Educaçao e Pesquisa*, *28*, 11–23.

Espinheira, P. L., Ferrari, S. L. P., & Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, *35*, 407–419.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, *31*, 799–815.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, *5*, 115–145.

Jorgensen, B. (1997). Proper dispersion models (with discussion). *Brazilian Journal of Probability and Statistics*, *11*, 89–140.

Kelley Pace, R., Barry, R., & Sirmans, C. F. (1998). Spatial statistics and real estate. *Journal of Real Estate Finance and Economics*, *17*, 5–13.

LeSage, J., Fischer, M. M., & Scherngell, T. (2007). Knowledge spillovers across Europe: Evidence from a Poisson spatial interaction model with spatial effects. *Papers in Regional Science*, *86*, 393–421.

Lu, H., Hodges, J. S., & Carlin, B. P. (2007). Measuring the complexity of generalized linear hierarchical models. *The Canadian Journal of Statistics*, *35*, 69–87.

Martínez, R. F. (2006). PISA en América Latina: Lecciones a partir de la experiencia de México de 2000 a 2006. *Revista de Educación*, extraordinario, 153–167.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Biometrika*, *35*, 255–260.

Rocha, A. V., & Simas, A. B. (2011). Influence diagnostics in a general class of beta regression models. *Test*, *20*(1), 95–119.

Sancho, G. J. (2006). Aprender a los 15 años: Factores que influyen en este proceso. *Revista de Educación*, extraordinario, 171–193.

Simas, A. B., Barreto-Souza, W., & Rocha, A. V. (2010). Improved estimators for a general class of beta regression model. *Computational Statistics and Data Analysis*, *54*, 348–366.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*, 54–71.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society—Series B*, *64*, 583–616.

Verkuilen, J., & 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.

Authors

EDILBERTO CEPEDA-CUERVO is an associate professor of statistics at Departamento de Estadística of Universidad Nacional de Colombia, Carrera 45 No 26-85 - Edificio Uriel Gutiérrez, Bogotá D.C., Colombia; e-mail: ecepedac@unal.edu.co. His research interests include Bayesian statistics, longitudinal data analysis, generalized linear models, and overdispersion models.

VICENTE NÚÑEZ-ANTÓN is a professor of statistics at Departamento de Econometría y Estadística (Economía Aplicada III) of Universidad del País Vasco UPV/EHU, Avenida Lehendakari Aguirre 83, E-40015 Bilbao, Spain; email: vicente.nunezanton@ehu.es. His research interests include longitudinal data analysis, survival data analysis, nonparametric and semiparametric estimation, goodness of t-testing, language acquisition in multilingual settings, and health related quality of life studies.

Manuscript received October 31, 2011
Revision received March 10, 2013
Accepted June 10, 2013