Generalized Spatial Dispersion Models

Dani Gamerman

Departamento de Métodos Estatísticos, Instituto de Matemática

Universidade Federal do Rio de Janeiro

Caixa Postal 68530, 21941-909 Rio de Janeiro, Brazil

Edilberto Cepeda-Cuervo

Departamento de Estadística, Facultad de Ciencias

Universidad Nacional de Colombia

Carrera 45 No 26-85, Bogotá, Colombia

Abstract

Standard generalized spatial models assume that spatial components affect the mean structure of observations. This is a useful starting point but, in some applications, further spatial structure may still remain. Models should include these structures in order to adequately describe spatial heterogeneity. In this paper regression models that account for spatial heterogeneity in both the mean and the dispersion are introduced. The models consider both spatially structured and unstructured random effect components in a bi-parametric family of models. These models are defined in the framework of lattice (regional summary) data. A number of possibilities are considered for the spatial effects, including conditional autoregressive models, and extension to geostatistical contexts are also discussed. This is a generalization that takes into account extra variability and also strengthens the model in terms of spatial dependence. The Bayesian paradigm is adopted to perform inference. Samples of the posterior distribution are drawn using standard MCMC procedures. The relevance of the methodology when compared against currently used methods is highlighted. A number of biomedical situations that require the use of these spatial dispersion models are presented.

Key words: Bayesian, Biparametric exponential family, CAR, regression.

1 Introduction

This paper focuses on the analyses of life expectancy analysis in Colombia and of the postpartum screening of Colombian mothers. The variables of interest in this study are the lifespan of Colombian people and the proportion of mothers that underwent postpartum screening, broken down by the country's political departments (states). In both studies, the response is affected by the economic, cultural and political spatial structure of Colombia. Observation of the variable of interest by departments imposes consideration of the spatial neighborhood structure (see Moran, 1948, or Geary, 1954). Standard procedures in these cases are to study the data with regression models such as GLM with addition of spatially structured and unstructured random effects in the mean possibly after a suitable link transformation (Besag, 1974). The questions addressed in this paper are: is this procedure adequate? does it adequately describe the heterogeneity present in the data?

Life expectancy is the number of years that a set of newborns will live on average in each department if mortality conditions do not change throughout life. The life expectancy of people in each of the 31 continental departments of Colombia from 2005 to 2010 is related to socioeconomic and political conditions. A population that has unsatisfied basic needs (UBN) is expected to have shorter life expectancy and higher infant mortality, given that UBN is an indicator of inadequate access to housing and services such as water, electricity and sanitation, as well as high levels of economic dependence and school-age children not attending school (Soto et al., 2012). Similarly, intra-familiar violence is likely to be related with life expectance. Thus, the aim of this study is to determine the impact of UBN and violence on life expectancy, taking into account the social and geographical structure of the country and the political and economic distribution of the land.

This paper also studies the proportion of mothers who underwent postpartum screening and the effect of factors that explain the assistance, including unsatisfied basic needs, proportion of women over 18 who suffered any type of physical abuse, percentage of the population without sufficient basic services and percentage of mothers who had to pay the total cost of postnatal

screening. Discovering the proportion of Colombian mothers who underwent postpartum screening, by department, is a second interest. Good health of mothers is very important for the development of healthy children. However, the first postpartum year is a critical period characterized by the development of affective disorders like anxiety and depression (see Giakoumaki et al., 2009). Postpartum disorders can have grave consequences for a mother and her infant, including hallucinations, confusion, inability to sleep or eat and some times suicide and infanticide. However, these are not the only reasons that mothers should make periodic visits to a physician after a child is born. It is important to verify that the mother's blood pressure is appropriate, that her body and cervix are recovering appropriately and that her breasts do not show any physical or medical abnormality. The last factor is very important to detect mastitis, an illness that commonly causes pain in a mother's breasts and, on some occasions, is accompanied by blushing of the affected area and fever. This is the main reason a mother should make an early visit to the doctor so that the possibility of mastitis will be minimized (Mathew, 2004 or Johnson et al., 2005).

2 Basic concepts

2.1 Spatial structure

Let $\{A_i,\ i = 1, \ldots, n\}$ be a partition of an area $S \in R^2$ into geographical units. These units can be state, countries, counties but will be called regions hereafter. This partition produces a neighborhood structure $\{N_i : i = 1, ..., n\}$, where N_i denotes the set of all regions that are neighbors of region i. The most common neighbor definitions are given by the physical first-order contiguity or by the distance between regions. However, being neighbors does not necessarily mean geographic proximity (Case et al., 1993). If two subregions A_j and A_k, are in the neighborhood of region A_i, it does not mean that the dependence between A_j and A_i, and between A_k and A_i are the same. Dependence between A_i and A_j, $j \neq i$, is characterized here by nonnegative real numbers w_{ij}, $j \in \{1, 2, \ldots, n\} - \{i\}$ such that $\sum_{j \neq i} w_{ij} = 1$ and $w_{ii} = 0$. Specifically, each

component of the spatial variable $\mathbf{Y}$ is associated with a vector $\mathbf{w}_i$ that indicates the importance of their neighborhoods. Thus, the intensity of spatial dependence between a region i and its neighborhoods is reflected by the point product $\mathbf{w}_i\mathbf{y}$, where $\mathbf{y}$ is the vector of observed values of $\mathbf{Y}$, that is by

$$[\mathbf{Wy}]_i = \sum_{j=1}^{n} w_{ij} y_j, \tag{1}$$

where $\mathbf{W}$ is a spatial lag operator, specifying for each geographical unit i (state, country,...), in the rows, the neighbors as the columns corresponding to non-zero elements w_{ij} in a fixed and positive $n \times n$ spatial weights matrix. A usual weight matrix is used to define the neighborhood structure given by the $n \times n$ symmetric matrix $\mathbf{A} = (a_{ij})$, where $a_{ij} = 1$ if i and j are neighbors, and $a_{ij} = 0$ otherwise. With the neighborhood thus defined, the weight matrix is given by the product $W = GA$, where $G = diag(n_i^{-1})$ and n_i^{-1} is the number of neighbors of region i. However, there is no unique definition for the spatial lag (see, for example, Moran, 1948, or Geary, 1954). A definition of w_{ij}, based on the distance between regions, was given by Cliff and Ord (1973). Other proposals were made by Dacey (1968), Bodson and Peeters (1975) and Gamerman and Moreira (2004).

The choice of the spatial weights matrix is a crucial decision for researchers using georeferenced data. Some recommendations for researchers applying spatial models regarding model selection and weights matrix specification are given in Getis and Aldstadt (2004), Aldstadt and Getis (2006), and Stakhovych and Bijmolt (2008). It is also possible to consider nonlinear specifications of the mean target model to obtain a nonlinear mean spatial model.

2.2 Spatial autoregressive models

One of the first approaches to analyze a geostatistics normal dataset was the mixed regressive-autoregressive model proposed by Ord (1975). In this proposal, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{WY} + \mathbf{e}$, where $\mathbf{X}$ is an $n \times p$ matrix with ones in the first column, $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \rho\}$ is the set of parameters of the model,

$\mathbf{Y}$ is the $n \times 1$ variable of interest and $\mathbf{e}$ the random error vector. This model can be rewritten as $\mathbf{Y} = (I - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (I - \rho\mathbf{W})^{-1}\mathbf{e}$. Assuming that $\mathbf{e} \sim N(0, \sigma^2\mathbf{I}_n)$ gives

$$\mathbf{Y} \sim N((I - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}, \sigma^2(I - \rho\mathbf{W})^{-2}). \tag{2}$$

A second approximation assumes an autoregressive dependence for the error. In this case, the model is given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \mathbf{e}$ and can be rewritten as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{e} \tag{3}$$

and thus, $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2(I - \rho\mathbf{W})^{-2})$. Model (3), called simultaneously autoregressive (SAR), is extensively used in the literature. More theoretical and applied information on SAR models can be found, for example, in Cressie and Wikle (2011) or in Wall (2004).

Another common spatial autoregressive class of models is the conditional autoregressive structure CAR models

$$[Y_i|\mathbf{Y}_{\sim i}] \quad = \quad [\mathbf{x}_i\boldsymbol{\beta} + \rho(\mathbf{w}_i\mathbf{y} - \mathbf{x}_i\boldsymbol{\beta}) + \nu_i] \tag{4}$$

where ν_i, $i = 1, 2, \ldots, n$, are independent random variables, such that $\nu_i \sim N(0, \sigma_i^2)$, $Y_{\sim i}$ denote all the components of $\mathbf{Y}$ except Y_i, $\mathbf{y}$ is the observed value of $\mathbf{Y}$ (Song and De Oliveira, 2012 and De Oliveira, 2012) and $[Z]$ denotes the distribution of Z. In this case, given that $\mathbf{D} = \text{diag}(\sigma_i^2)$, if $\mathbf{D}^{-1}\mathbf{W}$ is symmetric and $\mathbf{D}^{-1}(\mathbf{I}_n - \mathbf{W})$ is definite positive, the joint distribution of $\mathbf{Y}$ is given by

$$\mathbf{Y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, (\mathbf{I_n} - \mathbf{W})^{-\mathbf{1}}\mathbf{D}). \tag{5}$$

Some usual CAR models are defined with a particular weight matrix structure. For example, W can be defined by the product $\mathbf{W} = \mathbf{GA}$, as in section 2.1, or by the product $\mathbf{W} = \delta\mathbf{W}^*$, where δ is an unknown parameter and $\mathbf{W}^*$ is a nonnegative ($w_{i,j} \geq 0$), symmetric neighbor with known weight matrix (Cressie and Wikle, 2011). If $\mathbf{D} = \sigma^2\mathbf{I}_n$, the variance-covariance matrix is equal to $\sigma^2(\mathbf{I} - \mathbf{W})^{-1}$.

In the cases where Y is not normally distributed, data transformations may be used to develop an approximate statistical analysis. For example, if $Y > 0$ a log-normal distribution may be

assumed. In this case, after the log-transformation of the data, usual CAR or SAR models can be used in the data analysis. If Y is scored from 0 to M, a logit or probit transformation of the data could be appropriate and the analysis developed assuming normal distribution. A general method to perform this type of analysis is to combine the spatial correlation structure of Y with Box-Cox transformation (Box and Cox, 1964). In this case it is assumed that $(Y_i^{\lambda} - 1)/\lambda$, $i = 1, 2, \ldots, n$ follows the usual SAR or CAR models.

This and other families of transformations are used to normalize random variables. Some examples of these families and their application in the context of spatial data analysis are given in De Oliveira et al. (1997) and Lai (2010). These direct approaches include in the mean process a spatial autoregressive component taking a spatial area location into account, through a weight matrix W. In a latent approach, the spatial variation is incorporated into the model via an unobserved component. A general spatial structure is assumed for this latent component, as for example a CAR structure (Gamerman and Moreira, 2004).

2.3 CAR structure

In this section we present one of the spatial structures used in the definition of a generalized spatial dispersion model (GSDM, in short). The spatial structure is taken into account assuming latent random variables ν_i, $i = 1, 2, \ldots, n$, associated with regions $A_1, ..., A_n$, and let $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)'$. Assuming that $\boldsymbol{\nu}$ is a multivariate normal random variable, the full conditional distributions are

$$\nu_i | \nu_{\sim i}, \tau_i \sim N(\alpha \sum_{\sim i} b_{ij} \nu_j, \tau^{-1}), \quad i, j = 1, 2, \ldots, n. \tag{6}$$

where $j \in N_i$ is used to mean that the addition is carried out over all regions A_j that are neighbors of A_i. Thus,

$$\nu \sim N(0, [\mathbf{D}_\tau(\mathbf{I} - \alpha \mathbf{B})]^{-1}) \tag{7}$$

where $\mathbf{B}$ is an $n \times n$ matrix with $b_{ii} = 0$ and $\mathbf{D}_\tau = diag(\tau_i)$ (Jin et al., 2005). Usually, it is assumed that $\mathbf{D}_\tau = \tau \mathbf{D}$, where $\mathbf{D}$ is an $n \times n$ diagonal matrix. The parameter α is called a smoothing

parameter, taking values between 0 and 1. Even tough $\alpha = 0$ corresponds to an independent model, α cannot be interpreted as a spatial correlation (Jin et al., 2007).

The quantities α, $\mathbf{D}$ and $\mathbf{B}$ can be chosen to obtain CAR model structures. In the intrinsic autoregressive model $\alpha = 1$, $\mathbf{D} = \text{diag}(n_i)$, where n_i is the number of neighborhoods of region i and $\mathbf{B} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{W}$ is the adjacent matrix, with entries $w_{ij} = 1$, if j is a neighbor of i; and $w_{ij} = 0$, if not. Thus, model (7) can be rewritten as

$$\nu \sim N(0, [\mathbf{D}_\tau(\mathbf{D} - \mathbf{W})]^{-1}) \tag{8}$$

In these models, called intrinsic autoregressive (IAR) models, $\mathbf{D}_\tau(\mathbf{D} - \mathbf{W})$ is singular and thus (8) is improper and contains no parameter to control the strength of spatial dependence (Jin et al., 2005). This CAR structure corresponds to

$$\nu_i|\nu_{\sim i} \sim N\Big(\frac{\alpha}{n_i}\sum_{\sim i}\nu_j, \frac{\tau}{n_i}\Big), \quad i, j = 1, 2, \ldots, n, \tag{9}$$

(see Jin et al., 2007). The ICAR models reduce to an independent model if $\alpha = 0$, assuming independence between spatial observations.

It is quite common to find discrete count data, generally assumed to follow either a binomial or a Poisson distribution, depending on the particular data characteristics. If the researcher is interested in modeling such data and studying their dependence on some given covariates, then generalized linear models (GLMs) with a binomial or a Poisson response are the most commonly used data analysis methods. In practice, the data variance is larger than that assumed by the model given the spatial correlation of the data. The existence of latent heterogeneity is one of the main causes of overdispersion. So, models that incorporate spatial structure should be considered.

2.4 Spatial geostatistical structure

Two possibilities to extend the spatial structure defined in Section 2.3 to geostatistical analysis are considered in this section. The first assumes that a place s_j is in the neighborhood of s_i if the

distance is smaller than a real number r and that s_j does not belong to the neighborhood of s_i, otherwise. Thus, the geostatistical models have similar structure of the spatial areal data.

The second one assumes that the random effects $\boldsymbol{\nu}$ can be specified as a Gaussian process with a multivariate normal distribution with an n-dimensional mean vector with all components equal to zero and an $n \times n$ variance-covariance matrix $\boldsymbol{\Sigma}_1 = Cov(\boldsymbol{\nu})$ with entries defined as a function of the distance between geographic coordinates of the observation (Wang and Wall, 2003). One possible choice of the geostatistical model for the variance covariance matrix is to assume that the entries $\sigma_{ij}^2 = c\exp(-a|s_i - s_j|)$, where $|s_i - s_j|$ is the distance between site s_i and site s_j, c is the sill, representing the variance in the absence of spatial correlations, and a is the range parameter, representing the speed of decrease in correlation between two locations as the distance increases.

2.5 Double generalized regression models

The class of double generalized regression models was defined by Cepeda (2001) and Cepeda and Gamerman (2005). This class of models assumes that the distribution of the variable of interest belongs to the biparametric exponential family of distributions defined by Gelfand and Dalal (1990). That is, it is assumed the family of distributions for the response y is given by

$$p(y|\theta, \tau) = b(y) \exp\{\theta y + \tau T(y) - \rho(\theta, \tau)\} \tag{10}$$

where if y is continuous (discrete), p is assumed to be a density with respect to the Lebesgue measure (to the counting measure, respectively). Typically, $E(y) = \mu$ depends on θ and τ and the variance $Var(y) = \sigma^2$ depends only on τ. Gelfand and Dalal (1990) showed that if (10) is integrable over $y \in Y$, and if $T(y)$ is convex, then for a common mean, σ^2 increases in τ. The normal, gamma and reparameterized beta distributions are examples of distributions belonging to this family. A special case of this family occurs when $\tau = 0$, corresponding to the well known one-parameter exponential family.

In this family of distributions, the double generalized regression models are defined assuming

regression structures for the pair of parameters (θ, τ). Some models belonging to this family are:

1. The heteroscedastic normal regression models defined by Aitkin (1987), where mean and variance or mean and precision parameters are modeled as linear or nonlinear functions of the explanatory variables (see Cepeda and Gamerman, 2001, and Cepeda and Achcar, 2010).

2. The gamma regression models, where mean and dispersion, mean and variance, mean and shape, or shape and scale parameters are modeled as functions of regression structures. A particular case of this class of model is the joint mean and variance gamma regression models, with mean and variance regression models given by

$$h(\mu_i) = \mathbf{x}_i^{'}\boldsymbol{\beta} \quad \text{and} \quad g(\sigma_i^2) = \mathbf{z}_i^{'}\boldsymbol{\gamma}, \tag{11}$$

   where h and g are appropriate real functions, $\mathbf{x}$ and $\mathbf{z}$ are the mean and precision explanatory variables, respectively, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the respective parameter vectors (Cepeda, 2001, and Cepeda and Gamerman, 2005).

3. The beta regression models, where mean and another parameter representing dispersion, such as precision or variance, is modeled as a function of the explanatory variables, including regression structures, as in (11). See Cepeda (2001) and Cepeda and Gamerman (2005).

Other classes of models belonging to the double generalized linear models are generalized linear models defined by McCullagh and Nelder (1989), and nonlinear regression models and the overdispersed models defined by Dey, Gelfand and Peng (1997). Joint modeling of the parameters as regression models can be easily defined and fitted using the Bayesian methods.

2.6 Aim of the paper

The aim of this paper is to take into account spatial structure and extra variability in the data, in the framework of double generalized models. This can be achieved by the introduction of spatially structured and unstructured random components. Bayesian methods can be implemented to

incorporate prior information into the model, using Monte Carlo Markov chains (MCMC) to obtain samples of the posterior distribution (Gamerman and Lopes, 2006). Two applications are presented to highlight the importance of these proposals.

After these introductory sections, this paper has the following structure. In Section 3, generalized spatial dispersion models are defined. Section 4 presents the use of these models to model life expectancy and postnatal period screening. Finally, Section 5, contains our concluding remarks.

3 Model definition

The class of a generalized spatial dispersion models (GSDM) is now presented. Let Y_i, $i = 1, 2, \ldots, n$, be spatially dependent random variables with distribution in the biparametric exponential family. The GSDM is characterized by the following components.

1. *The random component*: components Y_i has a distribution belonging to the biparametric exponential family, with means μ_i and dispersions τ_i, $i = 1, 2, \ldots, n$, conditional on the random effects.

2. *The systematic component*: the linear predictor $\boldsymbol{\eta}_i = (\eta_{1i}, \eta_{2i})$, given by $\eta_{1i} = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + \nu_i + \xi_i$, and $\eta_{2i} = \boldsymbol{z}'_i\boldsymbol{\gamma} + \psi_i + \varepsilon_i$, where $\boldsymbol{x}_i$ is the i^{th} vector of the mean explanatory variables, $\boldsymbol{z}_i$ is the i^{th} vector of the variance explanatory variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^{'}$ is the vector of the mean regression coefficients, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_r)^{'}$ is the vector of the dispersion regression coefficients, ν_i and ψ_i are structured random effects, following one of the spatial structures given in the previous Section, and ξ_i and ε_i are unstructured independent random effects, with zero mean and constant variance.

3. *The link functions between the random and systematic components*: $\mu_i = h^{-1}(\eta_{1i})$ and $\tau_i = g^{-1}(\eta_{2i})$, where h and g are monotonic twice differentiable functions.

Dependence between these components may be introduced for the sake of parsimony. The dependence can lead, in extreme cases, to degeneracy, eg $\nu_i = \phi\psi_i$, for all i. In these cases a single structured effect for each observational unit suffices. Our model are used in the sequel with independent effects, to capture all possible spatial features present in the data.

3.1 Special cases

1. **Heteroscedastic geostatistical models.** Standard geostatistical models assume a linear regression on the mean with addition of a spatially structured Gaussian noise. Palacios and Steel (2006) introduced spatial heteroscedastic models by assuming further that dispersion is spatially structured possibly after a logarithmic transformation and inclusion of explanatory variables. They assumed that $Y \sim N(\mu, \lambda)$ and $\mu = \mathbf{x}'\boldsymbol{\beta} + \nu$ and $\log\lambda = \mathbf{z}'\boldsymbol{\gamma} + \varphi$. These models belong to the biparametric exponential family and are therefore included in the class of GSDM.

2. **Spatial gamma dispersion models**. In these models it is assumed that responses come from a gamma distribution. That is, it is assumed that $Y \sim G(\alpha, \lambda)$ with mean $\mu = \alpha\lambda$ and variance $\sigma^2 = \alpha\lambda^2$, and that a pair of parameters is modeled as a spatial regression. For example, the mean and shape parameters can follow the models given by

$$h(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + \nu_i + \xi_i \quad \text{and} \quad g(\alpha_i) = \mathbf{z}_i'\boldsymbol{\gamma} + \psi_i + \varepsilon_i. \tag{12}$$

   Often the link functions $h(\mu_i) = \log(\mu_i)$ and $g(\alpha_i) = \log(\alpha_i)$ are chosen and assumed known. Also, in many applications, the identity and reciprocal link functions are considered for the mean. If $\nu_i = \psi_i = \varepsilon_i = \xi_i = 0$, the double generalized gamma regression model is obtained, where both parameters of the gamma distribution are modeled only as a function of the explanatory variables.

3. **Spatial beta dispersion models**. Here the case where the random variable of interest comes from a beta distribution is considered. Let $Y \sim B(\alpha, \lambda)$, with mean $\mu = \alpha/(\alpha + \lambda)$ where $\alpha, \lambda > 0$. Taking $\phi = \alpha + \lambda$, this density can be rewritten as

$$f(y|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} I_{(0,1)}(y), \tag{13}$$

where $I_{(0,1)}(y)$ is the indicator function, equal to 1 if y belongs to the real open interval $(0, 1)$ and 0 otherwise. Joint mean and precision beta regression models was proposed in Cepeda (2001) by assuming that the mean and precision models are given by $\text{logit}(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$ and $\log(\phi_i) = \mathbf{z}_i^{'}\boldsymbol{\gamma}$, respectively. This model was considered by Smithson and Verkuilen (2006) and by Simas et al. (2010), under a classic perpective. A nonlinear beta regression model was proposed by Cepeda and Achcar (2010). The generalized spatial beta regression models are defined by

$$h(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + \nu_i + \xi_i \quad \text{and} \quad g(\phi_i) = \mathbf{z}_i'\boldsymbol{\gamma} + \psi + \varepsilon_i \tag{14}$$

where h and g are appropriate real functions. Often the link functions $h(\mu_i) = \text{logit}(\mu_i)$ and $g(\phi_i) = \log(\phi_i)$ are considered. Although a spatial joint mean and precision model is considered in (14), many other alternatives are possible, such a spatial joint mean and variance beta regression, where the mean is modeled as in (14) and the variance follows the model $g(\phi_i) = \mathbf{z}_i'\boldsymbol{\gamma} + \psi_i + \varepsilon_i$.

One parameter spatial models are obviously members of this family of models, obtained when the dispersion is fixed. These include the spatial Poisson models of Besag et al. (1991), the Poisson normal model, spatial binomial models and spatial exponential models (Quintero-Sarmiento et al., 2012; Gamerman, 1997). There are many options to characterize spatial dependence in the spatial components in all models above, including the options provided in the previous section and combinations of them.

3.2 Inference

Bayesian inference is performed to the models. This implies specification of prior distribution for all unknowns. The unknowns in the models are the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ the random effects $\{\nu_i\}$, $\{\xi_i\}$, $\{\psi_i\}$ and $\{\varepsilon_i\}$ and the collection of hyperparameters θ, associated with the specification of the random components.

Thus, one must obtain the posterior distribution via Bayes theorem as

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \{\nu_i\}, \{\xi_i\}, \{\psi_i\}, \{\varepsilon_i\}, \theta \mid \mathbf{y}, \mathbf{X}, \mathbf{Z}) \propto \prod_{i=1}^{n} p(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \{\nu_i\}, \{\xi_i\}, \{\psi_i\}, \{\varepsilon_i\}, \mathbf{x}_i, \mathbf{z}_i)$$
$$p(\boldsymbol{\beta})p(\boldsymbol{\gamma})p(\{\nu_i\} \mid \theta) \prod_{i=1}^{n} p(\xi_i \mid \theta) p(\{\psi_i\} \mid \theta) \prod_{i=1}^{n} p(\varepsilon_i \mid \theta) p(\theta). \qquad (15)$$

The first terms in the right hand side is the likelihood, obtained from (10). The remaining terms constitute the prior distribution for all unobserved model components.

The regression coefficients $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are usually given a joint normal distribution butr correlation between them can be introduced. Many options are available here: the variance matrix can be block diagonal if one wants to impose prior independence between the two sets of coefficients; the variance can be made large, eg $10^3\mathbf{I}$, if one wants to represent vague prior information.

The spatially structured random effects $\{\nu_i\}$ and $\{\psi_i\}$ may be given one of the spatial distributions of the previous sections. They are assumed to be mutually independent but correlation between them can be introduced by using one the multivariate version of spatial effects (Gamerman and Moreira, 2004). The unstructured random effects $\{\xi_i\}$ and $\{\varepsilon_i\}$ are usually given independent zero mean normal distributions. Model is completed with a prior for hyperparameters θ that will depend on the choices made for the random effects.

The posterior distribution in (15) is too complicated for analytical treatment and approximations must be used to summarize it. Among the methods currently available, MCMC was chosen due to its simplicity and ease of use. So, results of the data analyses of the next section are based on approximate samples from the posterior distribution.

For each proposed model, we simulated 10,000 initial Gibbs samples. After this burn-in period,

another 40,000 Gibbs samples were drawn and recorded every 150th sample, to have approximately uncorrelated samples. WinBugs (Spiegelhalter et al. 2003) was used in these simulation procedures. Convergence of the Gibbs sampler algorithm was monitored using standard existing methods as the trace plots of the simulated samples and parallel chains starting from different initial values, to provide indication of stationarity. In all applications, the posterior samples showed the same behavior for all chains, providing strong indication of convergence.

4 Results

In this section the results of the analysis of life expectancy and postnatal screening data are reported. Each data analysis will require a different observational specification and will thus provide an illustration of the capabilities of the models proposed here.

In order to apply the Bayesian method in all models, independent normal prior distributions $N(0, 10^k)$, with $k = 3$, for the regression parameters were assumed, to represent lack of prior information. Larger values of k were also considered but made no difference in the results. Vague gamma distributions $G(0.0001, 0.0001)$ were assumed for the precision parameters of normal distributions included in the models, given that we do not have information about these parameter values.

4.1 Life expectancy

In this section, the variable of interest is life expectancy of people in each of the 31 mainland departments of Colombia from 2005 to 2010. Life expectancy is the number of years that a cohort of newborns will live on average in each department assuming that mortality conditions do not change through his lifespan. One possible reason to believe that life expectancy between departments is not independent is their geographic location and the common characteristics that they may have. That is, because of the different geographic regions in Colombia (i.e., mountainous,

coastal and prairies), it is expected that the different departments in the country may be spatially correlated. Each region has several departments and some of them have specific characteristics, such as having a large proportion of black or indigenous population and similar climatic conditions. Figure 1 shows a map of Colombia, in which the life expectancy by departments is represented. It can be seen that shorter life expectancy is associated with more isolated departments of the country.

PLACE FIGURE 1 ABOUT HERE

Life expectancy is related to socioeconomic conditions. Thus, a population that has unsatisfied basic needs (UBN) is expected to have shorter life expectancy. Likewise the presence of violence is related with life expectance: more intra-familiar violence should be related with life expectance, given that women abuse should increase infant mortality. Consequently, in the proposed model we consider UBN, in percentage, and violence ($VIOL$), the percentage of women over 18 years who had suffered any type of physical abuse from their current partners. Data regarding life expectancy, UBN and $VIOL$ were provided by the National Statistics Office (DANE) of Colombia. Figure 1 shows the map of these auxiliary variables.

In order to apply the proposed spatial regression models, a gamma model with mean and precision given by $log(\mu_i) = \beta_0 + \beta_1 UBN_i + \beta_2 VIOL_i + \nu_i + \xi_i$ and $\log(\phi_i) = \gamma_0 + \gamma_1 UBN_i + \gamma_2 VIOL_i + \psi_i + \varepsilon_i$, respectively, was considered. For this model, the likelihood function and DIC values are $2logL = -12.976$ and $DIC = 57.392$. However, their regression coefficients are not significatively different from zero at a 95% level, except to β_1 in the mean model. Thus, the gamma regression models considered in the sequel assume

$$\log(\mu_i) = \beta_0 + \beta_1 UBN_i + \nu_i + \xi_i \quad \text{and} \tag{16}$$

$$\log(\phi_i) = \gamma_0 + \psi_i + \varepsilon_i. \tag{17}$$

The model is denoted by M1 and its parameter estimates and their standard deviations are given in Table 1. Model M2 differs from M1 by the removal of spatial effects ξ_i. Further removal of the random effect ε_i from the dispersion regression leads to model M3. Finally, removal the structured

random effect ψ from the dispersion regression in model M3 leads to model M4. The parameter estimates and their standard deviations for these models are also given in Table 1.

PLACE TABLE 1 ABOUT HERE

For all fitted models, the estimates of the regression coefficient showed a strong, inverse relation between the response life expectancy and unsatisfied basic needs (UBN), as expected. The higher the unsatisfied basic needs, the shorter are life expectations. This relation shows a stable pattern, as it seems to remain basically constant across models. According to the criteria used, Model M3 was the best among all models considered. Note that it includes spatially structural errors in the mean and in the precision. The presence of the spatially structural errors in the models is explained by the spatial distribution of poverty and sociocultural conditions and the resulting violence.

In this analysis, geographic characteristics and economic conditions of the country seem to be more adequately captured by the presence of spatially structured random effects in the mean and in the dispersion. Figure 2 shows the estimates of these spatial effects. They seem to provide similar but not same information with important differences noted for example in the southwestern regions. These regions exhibit small mean spatial effects but sizeable positive dispersion effects, that increasing the dispersion of these regions. This possibly reflects the scarcity of information in these more remote regions in the Amazonian rain forest. In any case, the dispersion spatial effect are clearly relevant. Standard spatial models, that do not consider these dispersion effect, seem to be missing an important component and capture the remaining heterogeneity only partially.

PLACE FIGURE 2 ABOUT HERE

4.2 Modeling postnatal period screening

Good health of mothers is obviously very important for the health of their children. Thus, it is relevant that mothers have proper health care after their children are born. Physicians should verify that blood pressure of the mothers is appropriate, that their body and cervix are recovering

appropriately and that their breasts do not show any physical or medical abnormality. The last factor is very important to detect mastitis, an illness that commonly causes pain in a mother's breasts and, on some occasions, is accompanied by blushing of the affected area and fever. This is one of the main reasons a mother should make an early visit to the physician after her child is born, to increase the chance of early detection and treatment. Additional scientific evidence on this can be found in Mathew (2004) and Johnson et al. (2005). Figure 3 shows a map of Colombia, in which the response variable is depicted. It seems to indicate some spatial structure, specially for the lower values of the response.

PLACE FIGURE 3 ABOUT HERE

A beta model is proposed to describe the variation in the proportion of mothers. The behavior of this random variable will be tentatively explained by the proportion of women over 18 who suffered any type of physical abuse from their current partners ($VIOL$) in the i-th department, the percentage of the population that had basic services not being satisfactorily attended (UBN) and the percentage of mothers who had to pay for the total cost of the postnatal screening (PAY).

Relevance of the covariates was tested at a credibility level of 95% and they are all removed but for UBN in the mean model. Thus, the beta regression model

$$\begin{aligned} \text{logit}(\mu_i) &= \beta_0 + \beta_1 UBN_i + \nu_i + \xi_i \quad (18) \\ \log(\phi_i) &= \gamma_0 + \psi_i + \varepsilon_i \quad (19) \end{aligned}$$

is considered and denoted as Model 1. Results of the five additional beta regression models considered in this analysis are reported in Table 2, all of them including spatial structures: Model 2 - including structural errors in both mean and precision and unstructured error in the precision, Model 3 - including structural error in the mean model and unstructured error in the precision, Model 4 - removed structural and unstructured errors from the mean in model 1, Model 5 - with structural errors in the mean and precision and Model 6 - only with structural error in the mean.

PLACE FIGURE 4 ABOUT HERE

PLACE TABLE 2 ABOUT HERE

For all fitted models, the estimates of the regresssion coefficient showed a strong, inverse relation between the response and unsatisfied basic needs (UBN), as also observed in the previous application. This relation shows a stable pattern, as it seems to remain basically constant across models.

The model with both structural and unstructured errors in the mean and in the precision seems to fare best among the models considered. This shows again the relevance of including spatial effects also in the precision to adequately capture the unexplained data heterogeneity. Figure 4 shows estimates of the mean and dispersion spatial effects. They seem to show a similar pattern, but relevant difference may still be observed.

5 Extensions

In this paper, generalized spatial dispersion models are proposed after an initial introduction of different possibilities for incorporation of spatial components into regression models. The models were applied the to the studies of life expectancy and relevance of postpartum screening. The ideas were introduced at the more basic levels of linear regression and spatial components in additive form. The data applications showed the relevance of the incorporation of spatial components at both mean and dispersion levels.

Among the many possible extensions, one can single out the use of GSDM in the context of non-linear regression and also in non-additive forms, where they could be for example interacting with the covariates. This could be achieved by allowing the regression coefficients also to vary in space as in Gamerman, Moreira and Rue (2003) for areal data or Gelfand et al. (2003) for geostatistical data. This is a working in development.

Acknowledgments

Gamerman's work was supported by grants from CNPq-Brazil and CAPES-Brazil. Cepeda's work was supported by a grant from the Research Division of the National University of Colombia.

References

[1] Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Applied statistics*, 332-339.

[2] Aldstadt, J. and Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, **38**(4), 327-343.

[3] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**(2), 192-236.

[4] 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), 1-20.

[5] Bodson, P. and Peeters, D. (1975). Estimation of the coefficients of a linear regression in the presence of spatial autocorrelation. An application to a Belgian labour-demand function *Environment and Planning A*, **7**(4), 455-472.

[6] Box G. E. P. and Cox D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211-246

[7] Case, A. C., Rosen, H. S. and Hines, J. R. (1993). Budget spillovers and fiscal policy interdependence. *Journal of Public Economics*, **52**(3), 285-307.

[8] Cepeda, E.C. (2001). Variability modeling in generalized linear models, *Unpublished Dr.Sc. Thesis. Mathematics Institute, Universidade Federal do Rio de Janeiro.*

[9] Cepeda, C. E. and Gamerman, D. (2001). Bayesian modeling of variance heterogeneity in normal regression models. *Brazilian Journal of Probability and Statistics*, **14**, 207-221.

[10] Cepeda, C. E. and Gamerman, D. (2005). Bayesian methodology for modeling parameters in the two parameter exponential family. *Estadística*, **57**(168-169), 93-105.

[11] Cepeda-Cuervo E. and Achcar, J. (2010). Heteroscedastic nonlinear regression models. *Communications in Statistics-Simulation and Computation*, **39**(2), 405-419.

[12] Cliff, A., Ord, J. (1973). Spatial autocorrelation, monographs in spatial enviromental systems analysis. London: Pion.

[13] Cressie, N. and Wikle, C. K. (2011). Statistics for spatio-temporal data. *Wiley Series in Probability and Statistics.*

[14] Dacey, M. (1968). A review of measures of contiguity for two and k-color maps. In: Berry, B., Marble, D. F., eds. *Spatial Analysis: A Reader in Statistical Geography. New York: Prentics-Hall*, pp. 479-495.

[15] De Oliveira, V., Kedem, B. and Short, D. A. (1977). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association*, **92**(440), 1422-1433.

[16] De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*, **64**(1), 107-133.

[17] Dey, D. K., Gelfand, A. E. and Peng, F. (1997). Overdispersed generalized linear models, *Journal of Statistical Planning and Inference*, **64**, 93-107.

[18] Gamerman D. (1997). Sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing*, **7**(1), 57-68.

[19] Gamerman D., Moreira A. R. and Rue H. (2003). Space-varying regression models: specifications and simulation. *Computational Statistics & Data Analysis* **42**, 513-533.

[20] Gamerman D., Moreira ARB. (2004). Multivariate spatial regression models. *Journal of Multivariate Analisis*, **91**(2), 262-281.

[21] Gamerman D. and Lopes H.F. (2006). *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference. Chapman and Hall/CRC.*

[22] Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5**(3), 115-145.

[23] Gelfand A. E. and Dalal S. R.(1990). A note on overdispersed exponential families, *Biometrika*, **77** (1), 55-64.

[24] Gelfand A.E., Kim H-.J., Sirmans C. F. and Banerjee S. (2003) Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association* **98**(462), 387-396.

[25] Getis, A. and Aldstadt, J. (2010). Constructing the spatial weights matrix using a local statistic. *Perspectives on Spatial Data Analysis (pp. 147-163).* Springer Berlin Heidelberg.

[26] Giakoumaki O., Vasilaki K., Lili L., Skouroliakou M. and Liosis G. (2009). The role of maternal anxiety in the early postpartum period: screening for anxiety and depressive symptomatology in Greece, *Journal of Psychosomatic Obstetrics and Gynecology*, **30**(1), 21-28.

[27] Jin X., Carlin B. P. and Banerjee S. (2005). Generalized hierarchical multivariate CAR models for areal data, *Biometrics*, **61**(4), 950-961.

[28] Jin, X., Banerjee, S., and Carlin, B.P. (2007). Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(5), 817- 838.

[29] Johnson, C. C., Ownby, D. R., Alford, S. H., Havstad, S. L., Williams, L. K., Zoratti, E. M., Peterson, E. L. and Joseph, C. L. M. (2005). Antibiotic exposure in early infancy and risk for childhood atopy, *J. Allergy Clin. Immunol.*, **115**(6), 1218-1224.

[30] Lai, D. (2010). Box-Cox transformation for spatial linear models: a study on lattice data. *Statistical Papers*, **51**(4), 853-864.

[31] Mathew, J. L. (2004). Effect of maternal antibiotics on breast feeding infants, *Postgraduate Medical Journal*, **80**, 196-200.

[32] McCullagh, P. and J.A. Nelder. (1989). Generalized Linear Models (Vol. 37). Chapman & Hall/CRC.

[33] Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, **10**(2), 243-251.

[34] Nelder, J. A. and Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments, *Appl. Stochast. Mod. Data Anal*, **7**(1), 107-120.

[35] Ord K.(1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, **70**(349), 120-126.

[36] Palacios, M. B. and Steel, M. F. J. (2006). Non-gaussian bayesian geostatistical modeling. *Journal of the American Statistical Association*, **101**(474), 604-618.

[37] Quintero-Sarmiento, A., Cepeda-Cuervo, E., and Núñez-Antón, V. (2012). Estimating infant mortality in Colombia: some overdispersion modelling approaches. *Journal of Applied Statistics*, **39**(5), 1011-1036.

[38] Simas, A. B., Barreto-Souza, W. and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, **54**(2), 348-366.

[39] Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, **11**(1), 54-71.

[40] Spiegelhalter D. J., Best N. G., Carlin B. P. and van der Linde, A. (2002). Bayesian measures of complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583-639.

[41] Song, J.J. and De Oliveira, V.(2012). Bayesian Model selection in spatial lattice models. *Statistical Methodology*, **9**(1), 228-238.

[42] Soto V. E., Farfan, M. I. and Lorant, V.(2012). Fiscal decentralisation and infant mortality rate: the Colombian case. *Social Science and Medicine* **74**, (9),1426-34.

[43] Stakhovych, S., Bijmolt, T. H. A. (2009). Specification of spatial models: A simulation study on weights matrices. *Papers in Regional Science*, **88**, No. 2, 389-408.

[44] Wall, M. M.(2004). A close look at the spatial structure implied by the CAR and SAR models, *Journal of Statistical Planning and Inference*, **121**, 311-324.

[45] Wang, F. and Wall M. M. (2003). Generalized common spatial factor model. *Biostatistics*, **4**(4), 569-582.

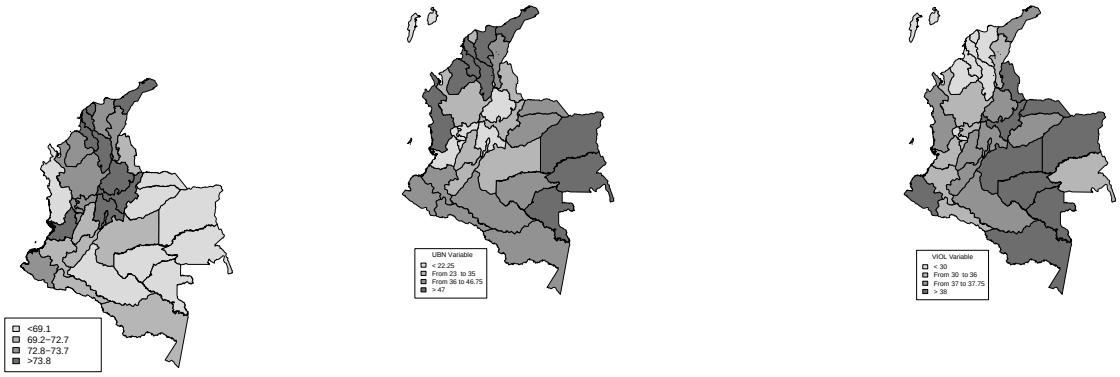

Figure 1: Response variable: left - Mean life expectancy by department. Explanatory variables: center - UBN by department; right - $VIOL$ by department.

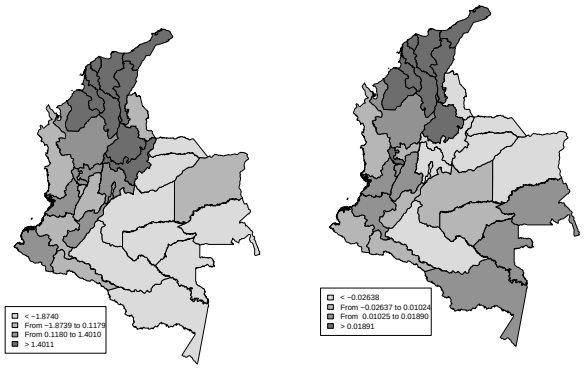

Figure 2: Posterior mean of spatial effects in M3: left - mean; right - dispersion.

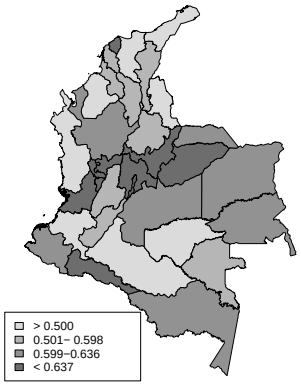

Figure 3: Proportion of mothers that underwent postnatal screening among those that had their last child between 1999 and 2005, by department.

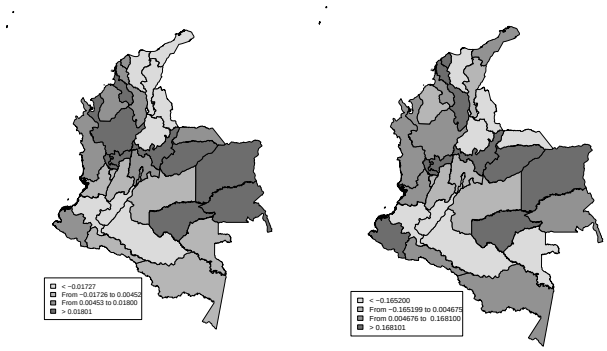

Figure 4: Posterior mean of spatial effects, model 1

Table 1: Parameter estimates for gamma spatial models.

Param	β_0	β_1	γ_0	τ_ν	τ_ξ	τ_ψ	τ_ε	2 log L	DIC
M1	4.311	-9.834E-4	2.248	395.2	752.3	20.48	21.84	13.415	46.940
	(0.026)	(6.355E-4)	(2.038)	(152.3)	(240.5)	(42.52)	(41.96)	—	—
M2	4.311	-9.83E-4	2.588	535.9	—	18.64	22.8	19.211	34.372
	(0.012)	(3.083E-4)	(1.694)	(149.5)	—	(35.23)	(46.42)	—	—
M3	4.31	-9.764E-4	3.725	532.4	—	20.27	—	55.764	-2.805
	(0.011)	(2.804E-4)	(2.298)	(144.0)	—	(39.02)	—	—	—
M4	4.311	-9.915E-4	2.338	1001.0	—	—	—	4.299	20.051
	(0.009)	(2.378E-4)	(1.855)	(398.1)	—	—	—	—	—

Table 2: Parameter estimates for beta spatial regression models.

Param	β_0	β_1	γ_0	τ_ν	τ_ξ	τ_ψ	τ_ε	2 log L	DIC
M1	1.066	-0.019	5.297	29.33	21.96	18.64	18.83	133.018	-108.020
	(0.240)	(0.006)	(1.691)	(45.89)	(33.78)	(39.81)	(36.41)	—	—
M2	1.035	-0.018	3.946	18.39	—	20.77	25.58	87.988	-70.584
	(0.227)	(0.006)	(1.178)	(34.19)	—	(39.87)	(46.99)	—	—
M3	1.03	-0.018	3.855	21.36	—	—	25.72	85.456	-71.310
	(0.231)	(0.006)	(1.182)	(37.93)	–	—	(44.7)	—	—
M4	1.028	-0.018	3.095	—	—	26.32	28.75	59.511	-47.728
	(0.214)	(0.005)	(0.293)	—	—	(45.34)	(46.69)	—	—
M5	1.03	-0.01798	3.548	25.14	—	23.30	—	76.906	-60.300
	(0.222)	(0.005)	(0.847)	(41.22)	—	(45.66)	—	—	—
M6	1.019	-0.017	3.173	172.8	—	—	—	63.526	-51.430
	(0.216)	(0.005)	(0.485)	(364.2)	—	—	—	—	—