On computational aspects of Bayesian spatial models

† Vinicius D. Mayrink and Dani Gamerman
DME, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Summary
This study applies computationally intensive methods for Bayesian analysis of spatially distributed data. It is assumed that the space of study is divided in contiguous and disjoint regions. Different neighboring structures may be available and can be compared, from the ones which contain few neighbors per region to structures where all sites are neighbors of each other. The main aim of this work is to evaluate the influence of neighborhood on results of Markov Chain Monte Carlo (MCMC) methods. Proper and improper prior specification for state parameters are compared. Three schemes proposed in the literature for sampling from the joint posterior distribution are also compared. The comparison criteria is based in the autocorrelation structure of the chains. Two classes of models are studied: one is characterized by a simple modelling without any explanatory variables and the other one is an extension with multiple regression components. Initially, interest is concerned with the performance of the spatial models in terms of their inference results based on different prior distributions. Finally, extensive analyses confront the outcomes obtained by different neighboring arrangements of the units.

Key Words: Bayesian inference, Neighboring influence, Sampling Schemes, Multiple Regression.

1 Introduction

Depending on the problem under study, different types of models, involving spatially distributed data, can be considered. In this work the spatial modelling considers the division of the space in n regions. In each region i, the variable of interest y_i is observed and $y = (y_1, y_2, ..., y_n)$ is the complete set of observations. The neighboring structure, establishing the dependence among observations, is an important aspect to be considered. Assume, for example, a spatial configuration where the regions are ordered in a line and the neighbors are the two regions immediately adjacent. In this example, the neighboring arrangement is the simplest non-degenerate one, where degeneracy implies independence. Cases where regions have three or more neighbors can also be evaluated.

The spatial structure can be based on a map over a region that does not need to be restrict to $\mathbb{R}$ or $\mathbb{R}^2$. Any spatial structure in $\mathbb{R}^d$, $\forall\ d \in \mathbb{Z}^+$ is applicable in the models that will be defined. The main question addressed here is whether the quantity of neighbors would influence the results of MCMC algorithms. Also, if this influence actually exists, which behavior can be observed in the comparison of results? Which neighboring structures provide the best performance? The search for these questions answers became the target motivating the development of this work.

† *Address for correspondence:* Universidade Federal do Rio de Janeiro, Instituto de Matemática, Departamento de Métodos Estatísticos, CP: 68530, Rio de Janeiro - RJ, CEP: 21945-970, Brazil. E-mail: viniciusm@dme.ufrj.br, dani@im.ufrj.br.

A Bayesian analysis will be performed and three schemes, proposed in the literature, will be used to sample from the posterior distribution. The blocking structure of parameters, determining which ones will be jointly or individually sampled, is the main aspect that differentiates these schemes. The performance of the MCMC algorithms, used by the sampling schemes, will be studied in the presence of different neighboring structures. Reis, Salazar and Gamerman (2006) developed a study, applied to linear dynamic models, comparing four sampling schemes. Gamerman, Moreira and Rue (2003) studied the multiple regression spatial model. They described three sampling schemes, that differ in the way blocks are formed, but did not provide any comparison. Knorr-Held and Rue (2002) developed a study where several algorithms were proposed to sample the parameters. Knorr-Held (1999), Rue (2001), Knorr-Held, Raber and Becker (2002), Fernandez and Green (2002) and Rue, Steinsland and Erland (2004) is a partial list of papers dealing with discussion on algorithms to jointly update parameters.

The outline of this study is: Section 2 describes the simplest spatial model used in this work. Proper and improper approaches will be specified for the prior distribution carrying the spatial information. The section is finished with the description of the data generating process. Section 3 describes the multiple regression spatial model. A model considering regressors is defined and only an improper approach is used for the prior distribution. Section 4 presents three sampling schemes used to define algorithms to sample from the joint posterior distribution. Section 5 shows inference results to check the performance of the spatial models. An optimization study is developed to choose tuning parameters defined in Metropolis-Hastings steps. In Section 6, distinct neighboring configurations are applied to the spatial models and the chains, obtained by MCMC methods, analyzed. Each configuration assumes a different number of neighbors per region. Differences in the algorithms are shown by studying the behavior of their autocorrelation structure. Section 7 shows the main conclusions and proposals for future works.

2 Simple spatial model

Many important aspects, common to spatial models in general, can be studied through the simple model presented here. In this section we introduce notation, elements specifying the model and derive properties.

A suitable model to represent spatial situations is based in the Markov Random Field (MRF). In simple terms, a collection of random values $X = (X_1, X_2, \ldots, X_n)$ is said to form a MRF if the joint distribution of X satisfies the property $(X_i \mid X_{-i}) \sim (X_i \mid X_{\delta_i})$, where $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and $\delta_i = \{j : i \sim j\}$, where $i \sim j$ means region i and j are neighbors, for $i = 1, \ldots, n$. Particularly, for the neighboring structure that orders regions in a line, $\delta_i = \{i-1, i+1\}$ and the MRF property simplifies to $(X_i \mid X_{-i}) \sim (X_i \mid X_{i-1}, X_{i+1})$, for all $i \neq 1, n$.

It is assumed that each observation y_i is associated to a mean value θ_i and is perturbed by independent errors v_i, for $i = 1, ..., n$. In this study, v_i is assumed normally distributed with mean 0 and variance σ^2, fixed along the space. In summary,

$$y_i = \theta_i + v_i, \quad v_i \sim N(0, \sigma^2), \text{ for } i = 1, ..., n. \tag{2.1}$$

The spatial structure information is inserted through a prior distribution for $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$. Besag, York and Mollié (1991) suggested the following prior specification considering pairwise differ-

ences

$$p(\theta) \propto \exp\left\{-\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} Z_{ij}h(\theta_i - \theta_j)\right\}. \tag{2.2}$$

These authors proposed in their work several possibilities and interpretations for the weights Z_{ij} and for function h. A typical choice for the weights is given by

$$Z_{ij} = \begin{cases} 1, & \text{if } i \sim j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

Consider that $h(x) = x^2/(2W)$. This specification for function h allow us to rewrite expression (2.2) as

$$p(\theta) \propto \exp\left\{-\frac{1}{2W}\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} Z_{ij}(\theta_i - \theta_j)^2\right\}. \tag{2.4}$$

The prior distribution (2.4) establishes the dependence between neighbors in the spatial structure. The weights Z_{ij}, as defined in (2.3), determine that the sum observed in (2.4) contemplates only differences between θ_i and its neighbors. This prior distribution for θ follows a MRF.

The components σ^2 and W are called hyperparameters and these elements are assumed fixed along the space. Assume prior independence between hyperparameters. The full prior conditional density $p(\theta \mid \sigma^2, W)$, carrying the spatial information, can be easily determined as

$$p(\theta \mid \sigma^2, W) \;\propto\; \exp\left\{-\frac{1}{2W}\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} Z_{ij}(\theta_i - \theta_j)^2\right\} \exp\left\{-\frac{1}{2R}(\theta_1 - a)^2\right\}. \tag{2.5}$$

The first exponential term is proportional to the pairwise difference prior (2.4). The second term is associated to an initial information $\theta_1 \sim N(a, R)$, where $R \in \mathbb{R}^+$. The choice of a and R will influence the model. Whenever we inform the prior distribution for θ is proper, the specification (2.5) is assumed. When an improper prior distribution is mentioned, R is assumed to be infinite and (2.4) is retrieved.

Proceeding with the model prior specification, it is important to express the full prior conditional density (2.5) in the matrix form. This strategy will facilitate the identification of the associated probability distribution. Consider the $(n \times n)$ matrix K with elements

$$K_{ij} = \begin{cases} Z_{i+}, & \text{if } i = j, \\ -Z_{ij}, & \text{if } i \sim j, \\ 0, & \text{otherwise,} \end{cases} \tag{2.6}$$

where $Z_{i+} = \sum_{j=1}^{n} Z_{ij}$. If (2.3) is valid, Z_{i+} represents the number of neighbors of region $i = 1, 2, ..., n$.

Matrix K carries the spatial information to be considered by the model. Note that the sum of values in any row or column is equal 0. This kind of matrix is singular and thus (2.4) is improper. Nevertheless, the full prior conditional density (2.5) can be rewritten as

$$p(\theta \mid \sigma^2, W) \;\propto\; \exp\left\{-\frac{1}{2W}\,\theta' K \theta\right\} \exp\left\{-\frac{1}{2R}(\theta_1 - a)^2\right\}. \tag{2.7}$$

The expression (2.7) simplifies to

$$p(\theta \mid \sigma^2, W) \;\propto\; \exp\left\{-\frac{1}{2W}\,\theta' K \theta\right\} \tag{2.8}$$

for an improper specification. This is the kernel of $N_n(\vec{0}, WK^{-1})$, where $\vec{0}$ represents the null vector.

The prior density (2.7) can be rewritten as

$$p(\theta \mid \sigma^2, W) \propto \exp\left\{-\frac{1}{2}\left[\theta'\left(\frac{1}{W}K\right)\theta + R^{-1}\theta_1^2\right]\right\}, \tag{2.9}$$

if $a = 0$. This is the kernel of the multivariate normal distribution with mean given by the null vector and full rank covariance matrix. The specification $R \in \mathbb{R}^+$ establishes that (2.9) is a proper prior distribution, consequently the respective posterior distribution will also be.

There are other approaches to approximate (2.4) by a proper distribution. These include the work of Ferreira, Higdon, Lee and West (2005), Fernandez and Green (2002) and Gamerman, Moreira and Rue (2003). They are all based on the inclusion of an extra term that ties down components of θ to a specific location.

2.1 Model properties

2.1.1 Full prior conditional joint distribution for θ_i

It can be shown that

$$(\theta_1 \mid \theta_{-1}, \sigma^2, W) \sim (\theta_1 \mid \theta_{\delta_1}, \sigma^2, W) \sim N\left[\bar{\theta}_{\delta_1}\left(\frac{RZ_{1+}}{RZ_{1+}+W}\right), \frac{WR}{RZ_{1+}+W}\right].$$

If an improper prior distribution ($R \to +\infty$) is assumed for θ, the following limits are obtained for mean and variance:

$$\lim_{R\to+\infty} \bar{\theta}_{\delta_1}\left(\frac{RZ_{1+}}{RZ_{1+}+W}\right) = \bar{\theta}_{\delta_1} \quad \text{and} \quad \lim_{R\to+\infty} \frac{WR}{RZ_{1+}+W} = \frac{W}{Z_{1+}},$$

where $\bar{\theta}_{\delta_i} = \left(\sum_{j=1}^{n} Z_{ij}\theta_j\right)/Z_{i+}$. Therefore, $(\theta_1 \mid \theta_{\delta_1}, \sigma^2, W) \sim N\left(\bar{\theta}_{\delta_1}, \frac{W}{Z_{1+}}\right)$ for R infinite. When $i = 2, ..., n$, $(\theta_i \mid \theta_{\delta_i}, \sigma^2, W) \sim N(\bar{\theta}_{\delta_i}, W/Z_{i+})$, irrespective of whether the proper prior (2.7) or improper prior (2.8) was used.

The neighboring structure determines a dependence of each θ_i on its neighbors. It determines which components of θ will enter the average $\bar{\theta}_{\delta_i}$. The more neighbors region i possesses, the more elements will compose this average and more information is available about this component since the number of neighbors of region i is part of the variance expression. A greater number of neighbors will determine a smaller uncertainty about θ_i.

2.1.2 Band diagonal neighboring matrix

The neighboring matrix (2.6) can be a band diagonal matrix or it can be band diagonalized. When this matrix has many null value entries, it is called sparse matrix. There are numerical techniques to organize elements of a sparse matrix in such a way it becomes band diagonal. In a band diagonal matrix, the null values concentrate in a band along the main diagonal. The band width depends of the quantity of neighbors.

The procedure to organize a sparse matrix is known as band diagonalization. Permutations are performed between rows and between columns without changing the matrix information. Rue (2001) developed a study where the applied algorithms are based in the use of band diagonalized sparse matrices. The paper showed the benefit from the matrix organization in the band diagonal format. Matrix operations such as Choleski decomposition become orders of magnitude faster. Knorr-Held and

Rue (2002), Rue, Steinsland and Erland (2004) and Rue (2005) use the numeric procedures proposed by Rue (2001). Knorr-Held, Raber and Becker (2002) also use techniques proposed by Rue (2001), however, the matrices studied were not sparse and the diagonalization process was unnecessary.

2.2 Joint posterior distribution

Definition: If $\Omega \sim IG(\alpha, \Sigma)$, where $\alpha > 0$ and $\Sigma > 0$, then:

$$p(\Omega) = \frac{\Sigma^{\alpha}}{\Gamma(\alpha)}\ \Omega^{-(\alpha+1)}\ \exp\left\{-\frac{\Sigma}{\Omega}\right\}, \quad \text{for} \quad \Omega > 0.$$

The full prior conditional distribution $p(\theta \mid \sigma^2, W)$ is given in (2.9). The following prior specifications will be used here: $\sigma^2 \sim IG(n_v, S_v)$ and $W \sim IG(n_w, S_w)$, where n_v, S_v, n_w and S_w are values greater than 0. To perform a conjugate analysis, the Inverted Gamma distribution is chosen to describe the prior uncertainty about the hyperparameters. As will be seen below, the full conditional posterior distributions for the hyperparameters will also follow the Inverted Gamma distribution. A prior independence is assumed to the hyperparameters. Using the Bayes theorem, the joint posterior density can be written by

$$\begin{aligned} p(\theta, \sigma^2, W \mid y) \quad \propto \quad & \left(\sigma^2\right)^{-\left(\frac{n}{2}+n_v+1\right)} \exp\left\{\ -\frac{1}{2\sigma^2}\left[(y-\theta)'(y-\theta)+2S_v\right]\right\} \\ & \times (W)^{-\left(\frac{n-1}{2}+n_w+1\right)} \exp\left\{-\frac{1}{2W}\left[\theta' K \theta + 2S_w\right]\right\} \\ & \times \exp\left\{-\frac{1}{2R}\theta_1^2\right\}. \end{aligned} \tag{2.10}$$

It is not possible to identify in (2.10) the kernel of a known probability distribution. Indirect methods will be required to generate values from this distribution.

Improper prior specification for θ will lead to the posterior density (2.10) without the exponential term in its third line. Despite its prior impropriety, the joint posterior distribution for (θ, σ^2, W) is proper (Besag, Green, Higdon and Mengersen 1995). Once again, it is not possible to identify the kernel of a known probability distribution and indirect methods will be needed to sample from this distribution.

2.3 Simulated data

The simulation of a dataset allows the analysis of different scenarios for variation of the number of neighbors. It also allows the verification of the performance of the models in terms of estimates. The procedure described in this subsection will be applied to generate the data used in the analysis of the simple spatial model. The first requirement is to obtain vector θ. In the second step we generate y_i as indicated in (2.1).

We assume a model with an improper prior distribution for θ but one has to consider a proper prior distribution to generate samples for θ. Therefore, the data is generated from an approximating proper prior distribution. From expression (2.9), the matrix K^* can be defined through an alteration in the precision matrix $\frac{1}{W}K$ as

$$K^* = \frac{K}{W} + R^{-1}J, \tag{2.11}$$

where $J = \text{diag}(1, 0, \ldots, 0)$ is a $(n \times n)$ matrix.

The matrix K^* is not singular and, therefore, θ can be generated from $N_n\left[\vec{0}, (K^*)^{-1}\right]$. The greater the value specified for R, the smaller will be the modification imposed in the precision matrix $\frac{1}{W}K$.

3 Multiple regression spatial model

This spatial model is an extension of the simple spatial model. The covariates and coefficients are inserted in the place of parameter θ_i.

Suppose the set $(y_i, x_{1,i}, x_{2,i}, ..., x_{d-1,i})$, $d = 2, 3, 4, \ldots$, is observed along the space, for $i = 1, ..., n$. Assume further that one wishes to model the association between the response variable y_i and the $d-1$ regressors $(x_{1,i}, x_{2,i}, ..., x_{d-1,i})$. The multiple regression model could be applied to explain the response variable through the regressors by the following relation: $y_i = \beta_0 + \beta_1 x_{1,i} + ... + \beta_{d-1} x_{d-1,i} + v_i, \quad v_i \sim N(0, \sigma^2)$. A more flexible alternative is allowing the regression coefficients to vary in space. This model assumes the relation, between response variable and regressors, is only local and not fixed to all observed data. The model can be rewritten as

$$y_i = \beta_i' X_i + v_i, \quad v_i \sim N(0, \sigma^2), \tag{3.1}$$

where $\beta_i' = (\beta_{0,i}, \beta_{1,i}, \ldots, \beta_{d-1,i})$, $X_i' = (1, x_{1,i}, \ldots, x_{d-1,i})$ and σ^2 is a scalar.

Assume the following notation,$\beta' = (\beta_1', \beta_2', \ldots, \beta_n')$, grouping all regression coefficients. The spatial structure information is introduced by a prior distribution defined for $(\beta \mid \sigma^2, W)$. There are several prior proposals for β that follows a MRF. An interesting example is given by the multivariate form

$$p(\beta) \propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} Z_{ij}(\beta_i - \beta_j)' W^{-1}(\beta_i - \beta_j)\right\}. \tag{3.2}$$

This proposal was used by Assunção, Gamerman and Assunção (1999) and by Moreira and Migon (1999) as prior distribution for regression coefficients varying along the space. Similar expressions can be found in Bernardinelli *et al.* (1995) and Gelfand and Vounatsou (2001).

In line with comments from the previous section, the more general family of priors can be proposed with density

$$\begin{aligned} p(\beta \mid \sigma^2, W) \quad \propto \quad & \exp\left\{-\frac{1}{2}\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} Z_{ij}\left(\beta_i - \beta_j\right)' W^{-1}\left(\beta_i - \beta_j\right)\right\} \\ & \times \exp\left\{-\frac{1}{2}\left(\beta_1 - a\right)' R^{-1}\left(\beta_1 - a\right)\right\}. \end{aligned} \tag{3.3}$$

The multivariate form for the pairwise difference prior, indicated in (3.2), can be identified in the first exponential function of (3.3). Taking into account the definition (2.3) for the weights Z_{ij}, and (2.6) for matrix K, expression (3.3) can be rewritten as

$$p(\beta \mid \sigma^2, W) \propto \exp\left\{-\frac{1}{2}\beta'\left(K \otimes W^{-1}\right)\beta\right\} \exp\left\{-\frac{1}{2}\left(\beta_1 - a\right)' R^{-1}\left(\beta_1 - a\right)\right\}, \tag{3.4}$$

where $K \otimes W^{-1}$ denotes the Kronecker product between K and W^{-1}.

The improper prior distribution in (3.2) is recovered by letting $R = \text{diag}(r, \ldots, r)$, $(d \times d)$, where $r \to +\infty$. In this section only this prior specification will be used. Thus, the remaining expression in

(3.4) is the kernel of $N_{dn}[\vec{0}, (K \otimes W^{-1})^{-1}]$ where $\vec{0}$ is the null vector of order dn. As already discussed, the neighboring matrix K is singular. Thus the covariance matrix $\left(K \otimes W^{-1}\right)^{-1}$, indexing the Normal distribution indicated here, is also singular. Therefore, the prior distribution for $(\beta \mid \sigma^2, W)$ is improper.

Simple extension of results of the previous section shows that $(\beta_i|\beta_{\delta_i}, W) \sim N_d\left[\bar{\beta}_{\delta_i}, W/Z_{i+}\right]$, where $\bar{\beta}_{\delta_i} = \left(\sum_{j=1}^{n} Z_{ij}\beta_j\right)/Z_{i+}$. The neighboring matrix K determines which of the vectors $\beta_1, \beta_2, ...,$ β_n will enter the sum in the expected value of the normal distribution. The larger number of neighbors of region i, the larger number of vectors entering this sum and less uncertainty about β_i.

3.1 Joint posterior distribution

Definition: Suppose Ω is a symmetric $(d \times d)$ positive definite matrix. If $\Omega \sim IW(\alpha, \Sigma)$ then

$$p(\Omega) = \frac{|\Sigma|^{\alpha}}{\pi^{\frac{d(d-1)}{4}} \prod_{i=1}^{d} \Gamma\left(\alpha + \frac{1-i}{2}\right)} \; |\Omega|^{-\alpha - \frac{d+1}{2}} \; \exp\left\{-\text{tr}\left(\Sigma\Omega^{-1}\right)\right\},$$

where $\alpha > (d-1)/2$ and Σ is a symmetric $(d \times d)$ positive definite matrix.

Consider the following prior specifications for the hyperparameters: $\sigma^2 \sim IG(n_v, S_v)$ and $W \sim IW(n_w, S_w)$, where $n_v > 0$, $S_v > 0$, $n_w > (d-1)/2$ and S_w is symmetric $(d \times d)$ positive definite matrix. In addition, note that $(y_i \mid \beta_i, \sigma^2) \sim N(\beta_i' X_i, \sigma^2)$ and $(\beta \mid \sigma^2, W) \sim N_{dn}[\vec{0}, (K \otimes W^{-1})^{-1}]$. A block diagonal matrix $X = \text{diag}(X_1', X_2', \ldots, X_n')$ can be formed with diagonal blocks $X_1', \ldots, X_n'$. This matrix notation is inserted to deal with regressors in the model. Based on these results, the use of Bayes theorem determines

$$\begin{aligned} p(\beta, \sigma^2, W \mid y) \quad \propto \quad & \exp\left\{-\frac{1}{2\sigma^2}\left[y'y - 2y'X\beta + \beta'\left[X'X + \sigma^2(K \otimes W^{-1})\right]\beta + 2S_v\right]\right\} \\ & \times \exp\left\{-\text{tr}\left(S_w W^{-1}\right)\right\} \; \left(\sigma^2\right)^{-\frac{n+2n_v+2}{2}} \; |W|^{-\frac{n+2n_w+d}{2}}. \end{aligned} \tag{3.5}$$

Despite the impropriety of the prior specification, results from Besag, Green, Higdon and Mengersen (1995) ensure the posterior distribution is proper.

It is not possible to recognize the kernel of a known probability distribution in expression (3.5), therefore, indirect methods will be necessary to sample from this posterior distribution.

3.2 Simulated data

The problem faced with simple spatial model is experienced again. The precision matrix, $K \otimes W^{-1}$, is not invertible. Therefore, the prior distribution for $(\beta \mid \sigma^2, W)$ is improper and, thus, it is not possible to generate β. The prior (3.4) is easily shown to be a $N(\vec{0}, (K^*)^{-1})$ distribution where

$$K^* = K \otimes W^{-1} \; + \; J \otimes R^{-1},$$

$J = \text{diag}(1, 0, \ldots, 0)$ is a $(n \times n)$ matrix. Componente a is equal $\vec{0}$ and the $(d \times d)$ covariance matrix R is $\text{diag}(r, \ldots, r)$ where $r \in \mathbb{R}^+$. The smaller value attributed for r, the smaller the change verified in the precision matrix $\left(K \otimes W^{-1}\right)$.

Matrix K^* is not singular, hence we can generate β from (3.4). After obtaining β, the data generating process is accomplished by generation of y, using equation (3.1). It is worth noting again that the procedure used here to generate regression coefficients is just one of the options that can be implemented to overcome the difficulties faced by the use of an improper prior distribution. Gamerman, Moreira and Rue (2003) discussed alternatives.

4 Sampling schemes

Given the analytic intractability of the posterior distributions, approximating methods must be used. In this paper, MCMC methods are considered, therefore, transition kernels must be proposed and compared. These will be detailed below. Three sampling schemes are proposed to sample from the joint posterior distribution. These schemes differ in the way they group parameters, establishing blocks. The literature indicates blocking parameters is beneficial for the sampling schemes. Liu, Wong and Kong (1994) studied the Gibbs Sampling method considering correlation structure and convergence rates. They showed estimates obtained by blocking components are more accurate than those obtained considering each component separately.

4.1 Full posterior conditional distributions for the simple spatial model

The full posterior conditional distribution for θ_i is obtained by the following application of the Bayes theorem

$$p(\theta_i \mid \theta_{-i}, \sigma^2, W, y) \;\propto\; p(y_i \mid \theta_i, \sigma^2)\, p(\theta_i \mid \theta_{\delta_i}, \sigma^2, W).$$

Straightforward calculations show that the full posterior conditional of θ_1 is $N[(y_1 V^* + \sigma^2 M^*)/(V^* + \sigma^2), (\sigma^2 V^*)/(V^* + \sigma^2)]$ where $V^* = WR/(RZ_{1+} + W)$ and $M^* = Z_{1+}\bar{\theta}_{\delta_1}[R/(RZ_{1+} + W)]$. This parameter is the only component of θ whose full posterior conditional distribution depends on R. For θ_i, $i = 2, ..., n$, the full posterior conditional is $N\left[(y_i W + \sigma^2 Z_{i+}\bar{\theta}_{\delta_i})/(W + \sigma^2 Z_{i+}), (\sigma^2 W)/(W + \sigma^2 Z_{i+})\right]$, irrespective of a proper or improper specification.

Components of θ can be individually sampled, using the full conditional distributions previously presented, but can also be jointly sampled. It is not difficult to show that the full posterior conditional of θ is $N(\mu, \sigma^2\Sigma)$ where the expression for μ and Σ vary with prior specification. When a proper prior is assumed, $\mu = \left(\sigma^2 K^* + I_n\right)^{-1} y$, $\Sigma = \left(\sigma^2 K^* + I_n\right)^{-1}$ and K^* is defined in (2.11). When an improper prior is assumed, $\mu = \left[(\sigma^2/W)K + I_n\right]^{-1} y$ and $\Sigma = \left[(\sigma^2/W)K + I_n\right]^{-1}$.

The full posterior conditional distributions for σ^2 and W are not influenced by the prior specification for θ. Consider the factorization $p(\theta, \sigma^2, W \mid y) = p(\sigma^2 \mid \theta, W, y)\, p(\theta, W \mid y)$. From this result, the density of interest $p(\sigma^2 \mid \theta, W, y)$ is easily obtained as

$$p(\sigma^2 \mid \theta, W, y) \propto \left(\sigma^2\right)^{-\left(\frac{n}{2} + n_v + 1\right)} \exp\left\{-\frac{1}{2\sigma^2}\left[(y-\theta)'(y-\theta) + 2S_v\right]\right\}.$$

This is the kernel of the distribution $IG[(n/2) + n_v, (y-\theta)'(y-\theta)(1/2) + S_v]$.

The full posterior conditional distribution of hyperparameter W is easily obtained by noting that $p(\theta, \sigma^2, W \mid y) = p(W \mid \theta, \sigma^2, y)\, p(\theta, \sigma^2 \mid y)$. The density of interest $p(W \mid \theta, \sigma^2, y)$ is

$$p(W \mid \theta, \sigma^2, y) \propto (W)^{-\left(\frac{n-1}{2} + n_w + 1\right)} \exp\left\{-\frac{1}{2W}\left[\theta' K \theta + 2S_w\right]\right\}.$$

This is the kernel of $IG[(n-1)(1/2) + n_w, (\theta' K \theta)(1/2) + S_w]$.

4.2 Full posterior conditional distributions for the multiple regression spatial model

Considering only the improper prior distribution for $(\beta \mid \sigma^2, W)$, the full posterior conditional distribution for β_i can be obtained by the following Bayes theorem application: $p(\beta_i \mid \beta_{-i}, \sigma^2, W, y) \propto p(y_i \mid \beta_{-i}, \sigma^2)\; p(\beta_i \mid \beta_{\delta_i}, W)$. It is immediate to obtain that $(\beta_i \mid \beta_{-i}, \sigma^2, W, y) \sim N_d(M, V)$ where $V^{-1} = (1/\sigma^2) X_i X_i' + Z_{i+} W^{-1}$ and $M = V^{-1}(y_i/\sigma^2) X_i + Z_{i+} W^{-1} \bar{\beta}_{\delta_i}$. Similarly, the full

posterior conditional distribution for β can be obtained as $(\beta \mid \sigma^2, W, y) \sim N_{dn}(\Sigma\mu, \Sigma)$ where $\Sigma = \left[(1/\sigma^2)X'X + \left(K \otimes W^{-1}\right)\right]^{-1}$ and $\mu = (1/\sigma^2)X'y$.

The full posterior conditional of σ^2 is an $IG[(n/2) + n_v, (y'y - 2y'X\beta + \beta'X'X\beta + 2S_v)/2]$ distribution. To consider the full posterior conditional for W, define the $(n \times d)$ matrix B containing all regression coefficients: $B = (\beta_1', \beta_2', \ldots, \beta_n')'$. Note that each row contains d regression coefficients associated to an observation y_i in the dataset. In the jth column, all coefficients β_{j-1}, for $j = 1, \ldots, d$, are found. β is the column vectorization of matrix B. The full posterior conditional of W is an Inverted Wishart distribution with parameter $[(B'KB) + 2S_w]/2$ and $[(n-1)/2] + n_w$ degrees of freedom.

4.3 Sampling schemes description

There are several options to sample from the joint posterior distribution. Three of them will be studied here. To unify the presentation, a new notation is introduced: $\Omega_i = \theta_i$ and $\Omega = \theta$ when the simple spatial model is considered and $\Omega_i = \beta_i$ and $\Omega = \beta$ when the multiple regression spatial model is considered.

Scheme 1 samples the parameters $\Omega_1, \ldots, \Omega_n$, σ^2 and W separately. Scheme 2 jointly samples $\Omega_1, \ldots, \Omega_n$, therefore, three blocks are formed: Ω, σ^2 and W. Gibbs sampling method is applied in these sampling schemes since all full posterior conditional distributions have already been presented and are available for easy sampling.

Scheme 3 jointly samples all the parameters in a single block (Ω, σ^2, W). Consider the posterior distribution factorization $p(\Omega, \sigma^2, W \mid y) = p(\Omega \mid \sigma^2, W, y)\ p(\sigma^2, W \mid y)$. The terms, in the right side of this expression, summarize the two steps to be taken when we apply Scheme 3: first, jointly generate the hyperparameters from $p(\sigma^2, W \mid y)$; then, generate Ω from $p(\Omega \mid \sigma^2, W, y)$. We have already applied the second step in Scheme 2. A new problem appears when considering the first step. Assume the simple spatial model where a proper prior distribution is specified for $p(\theta \mid \sigma^2, W)$. It can be shown that

$$\begin{aligned} p(\sigma^2, W \mid y) \quad \propto \quad & \left(\sigma^2\right)^{-\left(\frac{n}{2}+n_v+1\right)} \exp\left\{-\frac{1}{2\sigma^2}\left[y'y - \mu'\Sigma^{-1}\mu + 2S_v\right]\right\} \\ & \times \left|K^* + \frac{1}{\sigma^2}I_n\right|^{-\frac{1}{2}} (W)^{-\left(\frac{n-1}{2}+n_w+1\right)} \exp\left\{-\frac{S_w}{W}\right\}, \end{aligned}$$

where $\mu = \Sigma y$, $\Sigma = \left(\sigma^2 K^* + I_n\right)^{-1}$ and K^* is the matrix defined in (2.11). Small changes are required when considering the improper prior case.

Analogously, for the multiple regression spatial model,

$$\begin{aligned} p(\sigma^2, W \mid y) \quad \propto \quad & \left(\sigma^2\right)^{-\left(\frac{n+2n_v+2}{2}\right)} (W)^{-\left(-\frac{n+2n_w+d}{2}\right)} \exp\left\{-\mathrm{tr}\left(S_w W^{-1}\right)\right\} \\ & \times |\Sigma|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[2S_v + y'y - \sigma^2\mu'\Sigma\mu\right]\right\}, \end{aligned}$$

where $\mu = (1/\sigma^2)X'y$ and $\Sigma = \left[(1/\sigma^2)X'X + (K \otimes W^{-1})\right]^{-1}$.

Note that in all the cases showed, it is not possible to recognize the kernel of a known probability distribution for the joint posterior distribution of the hyperparameters. Metropolis-Hastings algorithm is necessary to sample from these distributions.

The first step is define the probability distribution used to generate proposals for the hyperparameters. In the simple spatial model, propose $\sigma^2_{(\mathrm{prop})}$ from $IG(K_1, K_1\sigma^2_{(\mathrm{prev})})$ and $W^2_{(\mathrm{prop})}$ from

$IG(K_2, K_2 W_{(\text{prev})})$. In the multiple regression spatial model generate $\sigma^2_{(\text{prop})}$ from $IG(K_1, K_1\sigma^2_{(\text{prev})})$ and $W^2_{(\text{prop})}$ from $IW(K_2, K_2 W_{(\text{prev})})$. The notation (prev) is used to indicate the previous value of the parameter. Tuning parameters index the proposal generating distributions. These constants are positive integer values denoted by K_1 and K_2. Their specifications will influence the acceptance rate of the algorithm. Assume $q\left[(a_1, a_2) \rightarrow (b_1, b_2)\right]$ is the proposal to generate values (b_1, b_2) from the product of densities indexed by (a_1, a_2).

When considering the Simple spatial model, $q[(\sigma^2_{(\text{prop})}, W_{(\text{prop})}) \rightarrow (\sigma^2_{(\text{prev})}, W_{(\text{prev})})]$ is the product of the densities $IG(K_1, K_1\sigma^2_{(\text{prop})})$ and $IG(K_2, K_2 W_{(\text{prop})})$. In the Multiple regression spatial model, this is the product of $IG(K_1, K_1\sigma^2_{(\text{prop})})$ and $IW(K_2, K_2 W_{(\text{prop})})$. The probability to accept or reject a proposal in a Metropolis-Hastings step is

$$\min\left\{1, \frac{p(\sigma^2_{(\text{prop})}, W_{(\text{prop})} \mid y)}{p(\sigma^2_{(\text{prev})}, W_{(\text{prev})} \mid y)} \frac{q\left[\left(\sigma^2_{(\text{prop})}, W_{(\text{prop})}\right) \rightarrow \left(\sigma^2_{(\text{prev})}, W_{(\text{prev})}\right)\right]}{q\left[\left(\sigma^2_{(\text{prev})}, W_{(\text{prev})}\right) \rightarrow \left(\sigma^2_{(\text{prop})}, W_{(\text{prop})}\right)\right]}\right\}.$$

Next step of Scheme 3 is to generate Ω from $p(\Omega \mid \sigma^2, W, y)$. There is no difficulty in this task as this distribution is known and was obtained in Scheme 2.

5 Applications and results

Programs were implemented to develop each sampling scheme procedure, using MCMC methods. The *Ox* environment was used here, see Doornik and Ooms (2006) for further information.

5.1 Efficiency comparison

The sampling schemes, proposed in the last section, provide chains with different characteristics. One of the main problems, in the analysis of serial data, is decide whether observations are originated from a process considering independent random variables or not. In practical terms, chains obtained via MCMC have an autocorrelation structure, and samples extracted from these chains will not be independent. Autocorrelation has influence over estimators, harming the quality of its estimates. Consider a sequence of quantities $z_1, \ldots, z_n$. The (assumedly constant) autocorrelation of lag j can be estimated by

$$\hat{\rho}_j = \frac{\sum_{i=j+1}^{n}(z_i - \bar{z})\,(z_{i-j} - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}, \quad \text{where } \bar{z} = \sum_{i=1}^{n}\frac{z_i}{n}.$$

The process of sampling from the posterior distribution, using MCMC methods, is an approximation and, therefore, has an associated error. Let f denote any real function applied to vector α containing all parameters in the model. Function f can be defined from the parametric space to any space $\mathbb{R}^d$, $d = 1, 2, \ldots$. Assume $\alpha^{(i)}$ a vector containing the ith elements in the Markov chain generated for α. The ergodic theorem ensures $\bar{f}_n \overset{\text{a.s.}}{\rightarrow} E[f(\alpha)]$, as $n \rightarrow \infty$, where $\bar{f}_n = (1/n)\sum_{i=1}^{n} f(\alpha^{(i)})$ is a Monte Carlo estimator of $E[f(\alpha)]$ and n represents the size of the chain. The variance of estimator $\bar{f}_n$ is

$$Var[\bar{f}_n] = \frac{\sigma_f^2}{n}\left(1 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\rho_k\right), \tag{5.1}$$

where σ_f^2 is the posterior variance of $f(\alpha)$ and ρ_k is the autocorrelation of lag k for the chain of $f(\alpha)$. For further details about this result, see Gamerman and Lopes (2006).

Autocorrelation plays an important role in the determination of efficiency of Monte Carlo estimators obtained from samples of the chain. For a random sample, expression (5.1) will be σ_f^2/n, because $\rho_k = 0$ for all $k = 1, \ldots, n-1$. Summarizing the autocorrelation information in the generated chains for each parameter, consider the efficiency factor

$$\mathrm{dn}_{eff} = 1 + 2\sum_{i=1}^{L} \hat{\rho}_i,$$

where $\hat{\rho}_i$ is the autocorrelation estimate of lag $i = 1, \ldots, L$. In this study consider $L = 50$.

The further away from 1 the value of dn_{eff}, the more autocorrelated the chain is, less efficient the sampling scheme is, and the further away from a random sample the studied sample is.

Another statistic that will be useful to interpret results, related to the chain autocorrelation structure, is called effective sample size. This statistic, denoted by n_{eff}, is obtained dividing the chain size n by the efficiency factor

$$\mathrm{n}_{eff} = \frac{n}{1 + 2\sum_{i=1}^{L} \hat{\rho}_i}.$$

In a random sample extracted from the chain, the autocorrelations would be all equal 0, hence the statistic n_{eff} would assume the sample size value. In practical terms, the efficiency factor assumes values larger than 1, thus the effective sample size is smaller than n. As an example, suppose this denominator equal 2 which determines $\mathrm{n}_{eff} = n/2$. The following interpretation can be made: $2n$ observations are needed from the chain to obtain a sample equivalent, in terms of information, to a random sample of size n.

5.2 Search for the tuning parameters

This subsection describes a searching strategy for optimal tuning parameters in Scheme 3. Only the simple spatial model, where a improper prior distribution is specified for $(\theta \mid \sigma^2, W)$, will be studied.

When considering a large value for K_1 the expected value and mode, of σ^2 proposal generating function, are close the previous value. In terms of variance, the larger K_1, the smaller variance. This conclusion can be extended for W since the expected value, mode and variance of its proposal generating function is equally structured, changing only the involved parameters. Hence, large tuning parameters provide a greater dependence between values in the chain.

Reis, Salazar and Gamerman (2006), in the first order dynamic linear models context, used Scheme 3 with the same proposal generating distribution. They chose the tuning parameters $K_1 = K_2 = 35$ which ensure accepting rates between 50% and 70% in Metropolis-Hastings, but provide high chain autocorrelations. It configures a disadvantage for Scheme 3 when a comparative analysis is performed among the proposed sampling schemes. Choice of appropriate values for K_1 and K_2 are based in the optimization algorithm described as follows.

Assume a space divided in 100 regions, ordered along a line, where the neighboring structure establishes 2 neighbors for all but the limiting (first and last) regions. In the data generating process, $R = 100$ was set and the true values of the hyperparameters are $\sigma^2 = W = 2$. The $IG(2.1, 6.2)$ was chosen as prior distribution for both hyperparameters. Its mode is 2 (real value), expected value is 5.63 and its variance is 317.68. In the MCMC algorithm consider the seeds: $\sigma^{2(0)} = 2$ and $W^{(0)} = 2$. Chains of size 1500 were generated. Values considered for tuning parameters were 0.5, 1, 1.5, 2,

..., 35, thus determining 4900 pairs (K_1, K_2). Scheme 3 was applied for each pair, and the effective sample size is calculated.

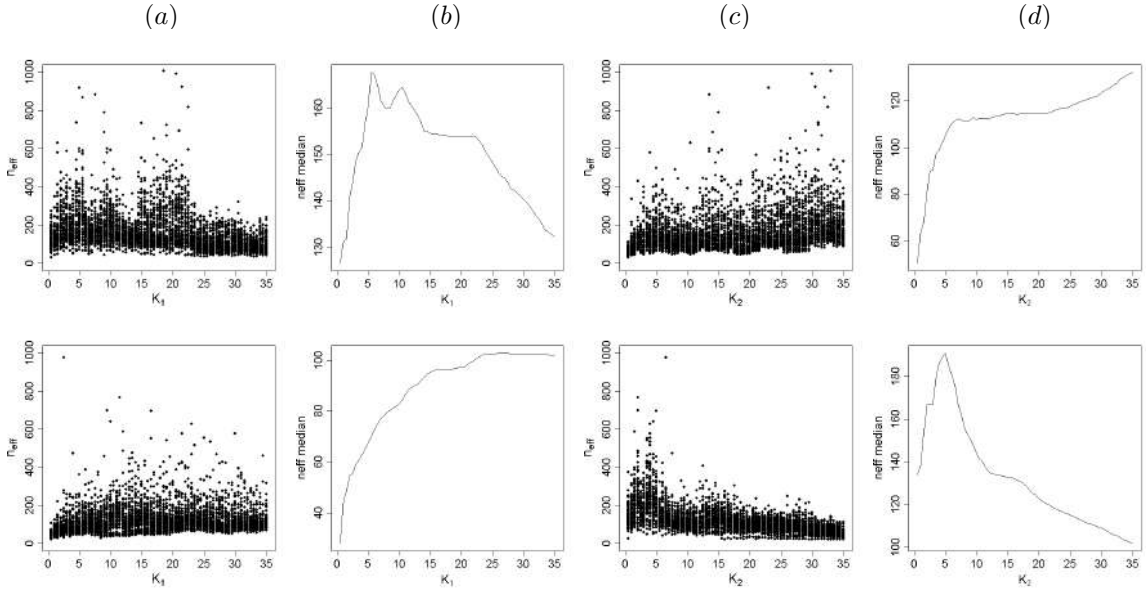

Figure 1: *n_{eff} vs. tuning parameters, σ^2 in the first row and W in the second row. (a) - for each K_1, points represent different values of K_2; (b) - for each K_1, median of points of panel (a); (c) - for each K_2, points represent different values of K_1; (d) - for each K_2, median of points of panel (c).*

Figure 1, in the first row, shows the behavior of the effective sample size for chains of σ^2. For a fixed K_1, 70 pairs (K_1, K_2) are formed by varying K_2. Note that, for chains of σ^2, the largest effective sample size is obtained when $K_1 = 5$. The graphic for K_2 indicates this parameter should be large to provide a less autocorrelated chain. The second row of Figure 1 shows the same analysis for W. Comparing to σ^2 panels the observed behavior is inverted. The larger the value of K_1, the larger effective sample size is obtained for chains of W. Graphic (d) indicates the less autocorrelated chain is obtained when $K_2 = 5$. Graphics for other parameters are inconclusive, therefore, the decision about choosing the tuning parameters is based in the chains of σ^2 and W. The chosen configuration for Scheme 3 hereafter is $K_1 = K_2 = 5$, which gives accepting rates around 10% in Metropolis-Hastings but is optimal in terms of the efficiency results above.

5.3 Inference for the simple spatial model

The focus here is to compare the behavior of the estimates provided by the simple spatial model, assuming proper and improper prior distribution for $(\theta \mid \sigma^2, W)$. Another purpose is to show that this model provides good estimates for the parameters. Consider a space with 500 regions where $(i-1)$ and $(i+1)$ are neighbors of region $i = 2, \ldots, 499$. The data generating process assumes $R = 100$ and the real values $\sigma^2 = 10$ and $W = 0.1$. The model considers the prior distributions: $IG(5, 60)$ for σ^2 and $IG(2.001, 0.3001)$ for W. The mode of these distributions is the real value specified for the hyperparameter. In addition, $E(\sigma^2) = 15$, $Var(\sigma^2) = 75$, $E(W) = 0.30$ and $Var(W) = 89.88$. When the model with a proper prior distribution is considered, it is assumed that $R = 100$ is known as specified in the data generating process. In the MCMC algorithms, the respective true value was attributed as initial value for the hyperparameter chain. Scheme 1 requires the specification of seeds

for each component of θ. In this case, 0 was considered as starting point of the chain.

An important parameter function to be monitored is the logarithm of the joint posterior density related to the model being considered. The logarithm function is used to avoid the analysis of values with high magnitude. Denote by $C_1(\theta, \sigma^2, W)$ the expression indicated in the right hand side of (2.10). Let $C = 1/\int_W \int_{\sigma^2} \int_\theta C_1(\theta, \sigma^2, W)\ d\theta\ d\sigma^2\ dW$ be the normalizing constant. Assume the following notation in the simple spatial model: $\log(\pi) = log[p(\theta, \sigma^2, W \mid y)] - log(C)$.

For each parameter of interest, 7000 iterations of the MCMC algorithm were performed to generate chains. The burn-in period involves the first 2000 iterations. A visual analysis, based on trace plots, was performed to study convergence. Chains with different starting points were considered. Similar variability, and traces located in the same position can be observed through these graphics. This behavior is indicated in all cases leading us to conclude about the convergence of the chains.

Table 1: Estimates for the simple spatial model, 500 regions and 2 neighbors.

Sampling	proper prior				improper prior			
	σ^2		W		σ^2		W	
schemes	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
Scheme 1	10.35	0.48	0.09	0.001	10.44	0.47	0.09	0.001
Scheme 2	10.34	0.48	0.10	0.002	10.35	0.47	0.10	0.002
Scheme 3	10.32	0.48	0.10	0.002	10.41	0.51	0.09	0.002

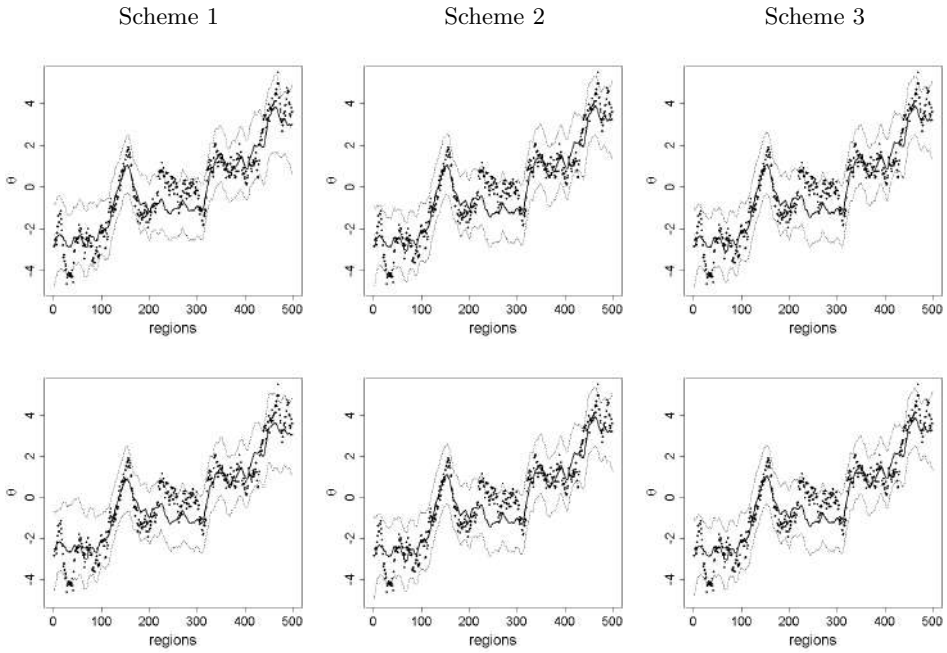

Figure 2: 95% *credibility interval (dashed line), mean (continuous line) and real values (dots) of the* θ *components. Proper prior distribution for* $(\theta \mid \sigma^2, W)$ *in the first row, improper prior distribution in the second row.*

Table 1 informs the estimates for each parameter based in the 5000 observations extracted from the respective chain. A comparison between the results of the model with proper and improper prior distribution for $(\theta \mid \sigma^2, W)$ can be made. All estimates are close to the real value and, thus, they are similar to each other. The simple spatial model with improper prior distribution has an alike

performance not showing any noticeable disadvantage when compared to the model with proper prior specification. The estimates obtained for each sampling scheme are quite similar, as expected after convergence.

Figures 2 show the 95% credibility interval for the real values of θ components. The first row is related to the simple spatial model with proper prior distribution, and the second row represents the model with improper prior distribution. When comparing both rows, we can not find any significative difference. Both approaches have a satisfactory performance in terms of θ estimates. The trace follows the behavior of the real values and most dots are inside the credibility interval. The model considering improper prior distribution for $(\theta \mid \sigma^2, W)$ does not have any disadvantage when compared to the approach where a proper prior distribution is specified.

5.4 Inference for the multiple regression spatial model

The larger number of parameters and the quantity of regions influence the execution time of the algorithms. Calculations involving matrices take longer, hence an analysis with 500 regions is not performed here.

Assume a space containing 150 regions and 2 regressors for the model, hence 3 regression coefficients are defined for each region. The sets of regressor values were generated from $N(0, 1)$. Assume that region i has 2 neighbors, $(i-1)$ and $(i+1)$ for $i = 2, \ldots, 149$. To generate the data we considered the proper prior distribution with $R = \text{diag}(100, 100, 100)$. In addition the following real values were specified for the hyperparameters: $\sigma^2 = 10$ and W is a symmetric (3×3) matrix where $W_{1,1} = W_{2,2} = W_{3,3} = 0.10$, $W_{1,2} = -0.05$, $W_{1,3} = 0$ and $W_{2,3} = 0.05$. Consider the prior specifications as follows: $\sigma^2 \sim IG(5, 60)$, as in Subsection 5.3, and $W \sim IW(2, S_w)$ where $S_w = \text{diag}(2/15, 2/20, 2/15)$.

In the MCMC algorithms, initial values for the hyperparameters were set at their real values. Particularly, Scheme 1 requires the definition of a seed for each β component. As before, value 0 was assigned as starting point of the chain. Consider $C_1(\beta, \sigma^2, W)$ the expression in the right hand side of (3.5). Denote $C = 1/\int_W \int_{\sigma^2} \int_\beta C_1(\beta, \sigma^2, W)\, d\beta\, d\sigma^2\, dW$. Let $\log(\pi) = \log[p(\beta, \sigma^2, W \mid y)] - \log(C)$.

As before, the chains of each parameter are formed by 7000 iterations of the MCMC algorithms. The first 2000 iterations were considered the burn-in period and then discarded. The visual analysis, based on trace plots, was developed in this case to study the behavior of the chains. Again, different starting points were considered and similar variability observed. The same trace position is registered after the burn-in period. Convergence was verified for all generated chains, and therefore, the sample of observations is valid for inference.

Table 2: Multiple regression spatial model estimates, 150 regions and 2 neighbors. The real values for those elements are 10 and 0.0005, respectively.

Sampling	σ^2			$\det(W)$		
schemes	Mean	Median	Var.	Mean	Median	Var.
Scheme 1	10.97	10.88	1.93	0.0005	0.0003	4×10^{-7}
Scheme 2	10.96	10.85	1.88	0.0005	0.0003	3×10^{-7}
Scheme 3	10.83	10.85	1.60	0.0005	0.0003	5×10^{-7}

Consider the estimates indicated in Table 2, for σ^2 and for $\det(W)$. Note that throughout the table the mean and median estimates are close the real value. This aspect shows the model is satisfactory in terms of inference for those components. Comparing sampling schemes, their estimates are similar,

including variability. Again, this result was expected because the three schemes are sampling from the same joint posterior distribution.

Figure 3 the real value is inside the credibility interval in all panels and in most cases the mean is very close to the real value. These results together with the estimates of Table 2 confirm the good performance of the multiple regression spatial model in terms of inference and the irrelevance of the proper prior approximation.

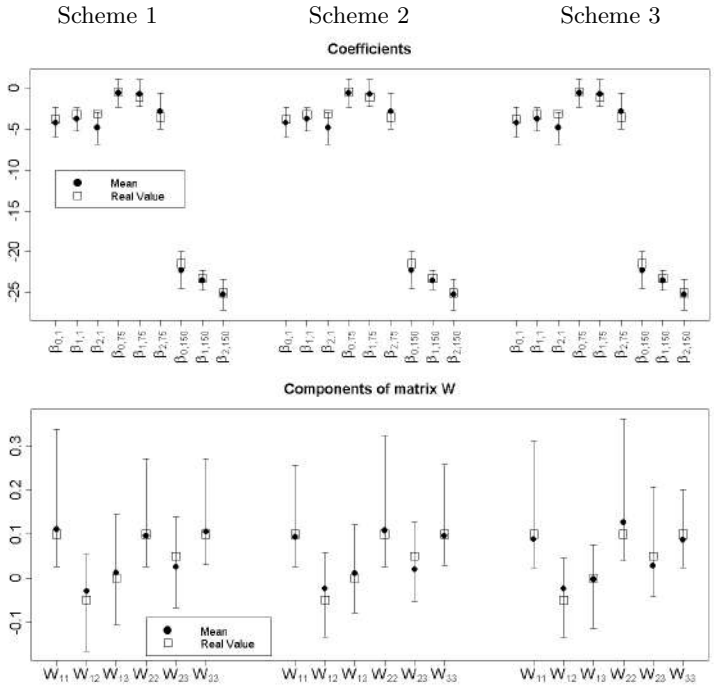

Figure 3: 95% *credibility interval, posterior mean estimate and real value for components of* β *and matrix* W*. The analysis is developed for regression coefficients of regions 1, 75 and 150 and for 6 components of the symmetric matrix* W*.*

6 Neighboring influence on the MCMC algorithms

A spatial model can be used to analyze any neighboring structure, for example those establishing 2 neighbors for each region or those where the amount of neighbors is greater. Our focus is to compare the efficiency in each scheme based on the chain autocorrelation obtained with different neighboring structures. For each combination model/scheme, 50 replications were considered.

6.1 Simple spatial model analysis

A space divided in 500 regions will be analyzed. Several neighboring configurations can be studied, particularly assume the specifications with 2, 10, 20, 50, 100, 200, 300, 400 and 499 neighbors. The last configuration represents a space where all regions are neighbors of each other.

Each neighboring structure requires the generation of a distinct dataset. The same real values are specified for the hyperparameters ($\sigma^2 = 10$ e $W = 0.1$) to ensure an homogeneous data generating process. The prior distribution with $R = 100$ was used in all neighboring configurations to generate θ.

The prior specifications adopted for hyperparameters are $\sigma^2 \sim IG(5, 60)$ and $W \sim IG(2.001, 0.3001)$, used in the analyses of the previous section.

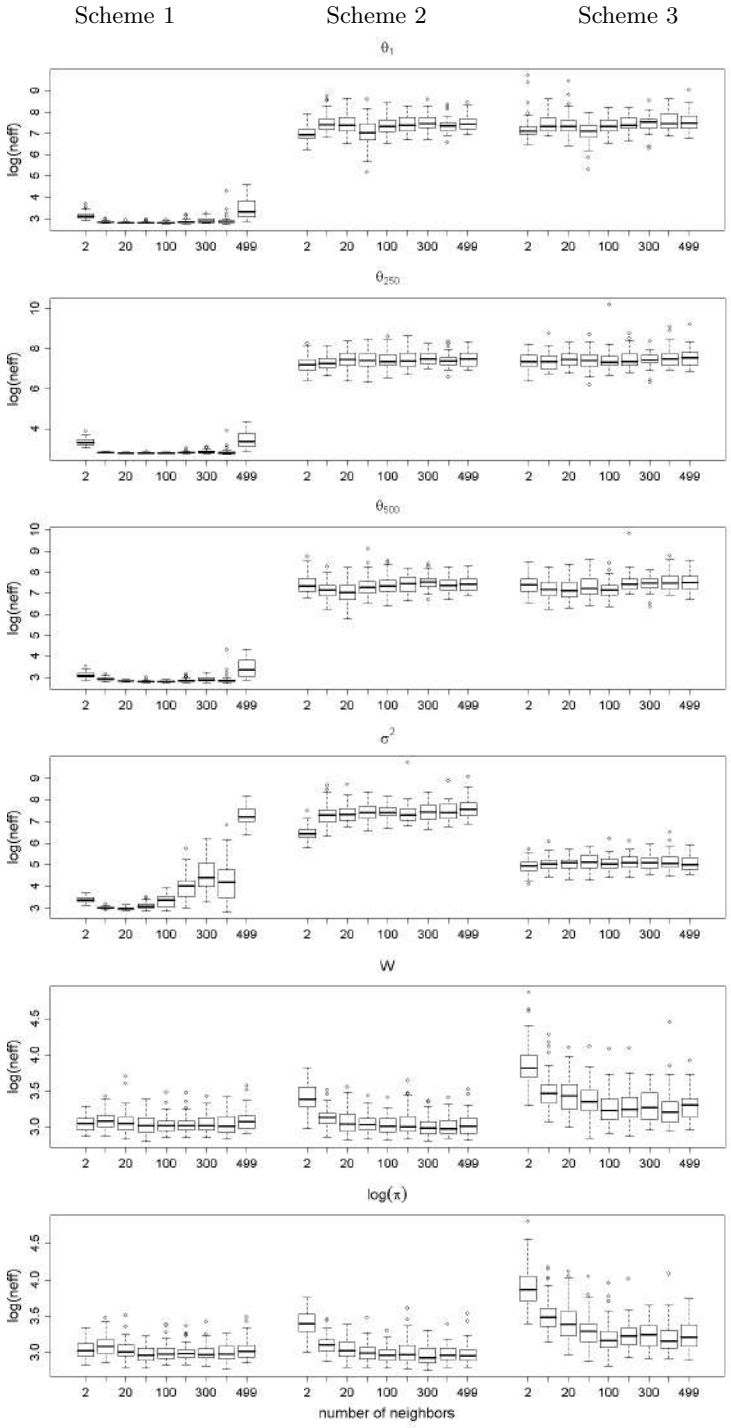

Figure 4: *Box-plots over the 50 replications of n_{eff} (in log scale) vs. number of neighbors: θ_1, θ_{250}, θ_{500}, σ^2, W and* $\log(\pi)$.

In the MCMC algorithms, the real values previously informed are the starting points for the

chains of hyperparameters. Scheme 1 assumes value 0 as seed of the generated chain, for each θ_i. The amount of 1500 iterations of the algorithms will be performed to build the chains. Analysis is based on 6 parameter functions: θ_1, θ_{250}, θ_{500}, σ^2, W and $\log(\pi)$.

The statistic n_{eff} is calculated for each chain replication. The autocorrelation structure of chains, obtained for each neighboring configuration, is evaluated through this criteria. Figure 4 shows a comparison of *box-plots* summarizing the information of samples of effective sample size.

For Scheme 1, the graphics related to the extremes configurations (2 and 499 neighbors) indicate less autocorrelation, larger *box-plot*'s, larger n_{eff}. Schemes 2 and 3 apply the same procedure to sample θ. This fact is observed through the similar levels of effective sample size. Comparing schemes, in most cases is evident the worst performance of Scheme 1 in terms of chain autocorrelation, as expected. Schemes 2 and 3 are undistinguishable in terms of θ's.

The hyperparameter σ^2 for Scheme 1 provides the most evident case of difference in efficiency along the distinct neighboring configurations. A greater number of neighbors per region determines a greater variability in the sample and a better performance of Scheme 1 in terms of less autocorrelated chains for σ^2. Chains for W and $\log(\pi)$ in Scheme 1 do not indicate difference in efficiency among the distinct neighboring configurations.

Results for σ^2 seems more erratic and indicate a small increase in efficiency as the number of neighbors increases for Scheme 2. Results of Scheme 3, for chains of this hyperparameter, do not exhibit a clear neighboring influence. Efficiency measures for Scheme 3 are lower than those of Scheme 2, indicating a better performance of the latter in terms of autocorrelation. The improvement in efficiency for Scheme 1 for large numbers of neighbors makes it achieve the performance of Scheme 2 and overcome Scheme 3.

Schemes 2 and 3 register a clear decrease in efficiency for parameters W and $\log(\pi)$ as the number of neighbors increases. Considering these chains, we can assert that smaller autocorrelation is obtained for structures with fewer neighbors for each region. These conclusions seem more evident for Scheme 3. Also, Scheme 3 seems to outperform the other schemes for all neighboring structures considered.

Given the more volatile behavior of the results for σ^2 and given the importance of the transformation $\log(\pi)$ as a prime summarizer of the posterior distribution, it seems safer to rely more heavily on the results obtained for the latter, confirmed by the results obtained with W.

6.2 Multiple regression spatial model analysis

Assume a space containing 150 regions. The following neighboring configurations are applied: 2, 10, 20, 50, 100 and 149 neighbors. The latter case represents a space where all regions are neighbors of each other.

A simulated dataset must be generated for each neighboring configuration. For an homogeneous data generating process the following real values are specified in the 6 procedures: $\sigma^2 = 10$ and W is a symmetric (3×3) matrix, where $W_{1,1} = W_{2,2} = W_{3,3} = 0.1$, $W_{1,2} = -0.05$, $W_{1,3} = 0$ and $W_{2,3} = 0.05$. The approximating prior distribution with $R = \text{diag}(100, 100, 100)$ allows the generation of β. The prior specification adopted for the hyperparameters: $\sigma^2 \sim IG(5, 60)$ and $W \sim IW(2, S_w)$ where $S_w = \text{diag}(2/15, 2/20, 2/15)$. These choices were previously used in the multiple regression spatial model.

The MCMC algorithms generate chains for the hyperparameters considering as starting points the real values previously indicated. In Scheme 1, specify 0 as initial value for components of β. The coefficients of regions 1, 75 and 150 will have their chains analyzed. Among the other parameters, this

study considers σ^2, $\log(\pi)$ and the parameter transformation $\det(W)$ summarizing the information about W.

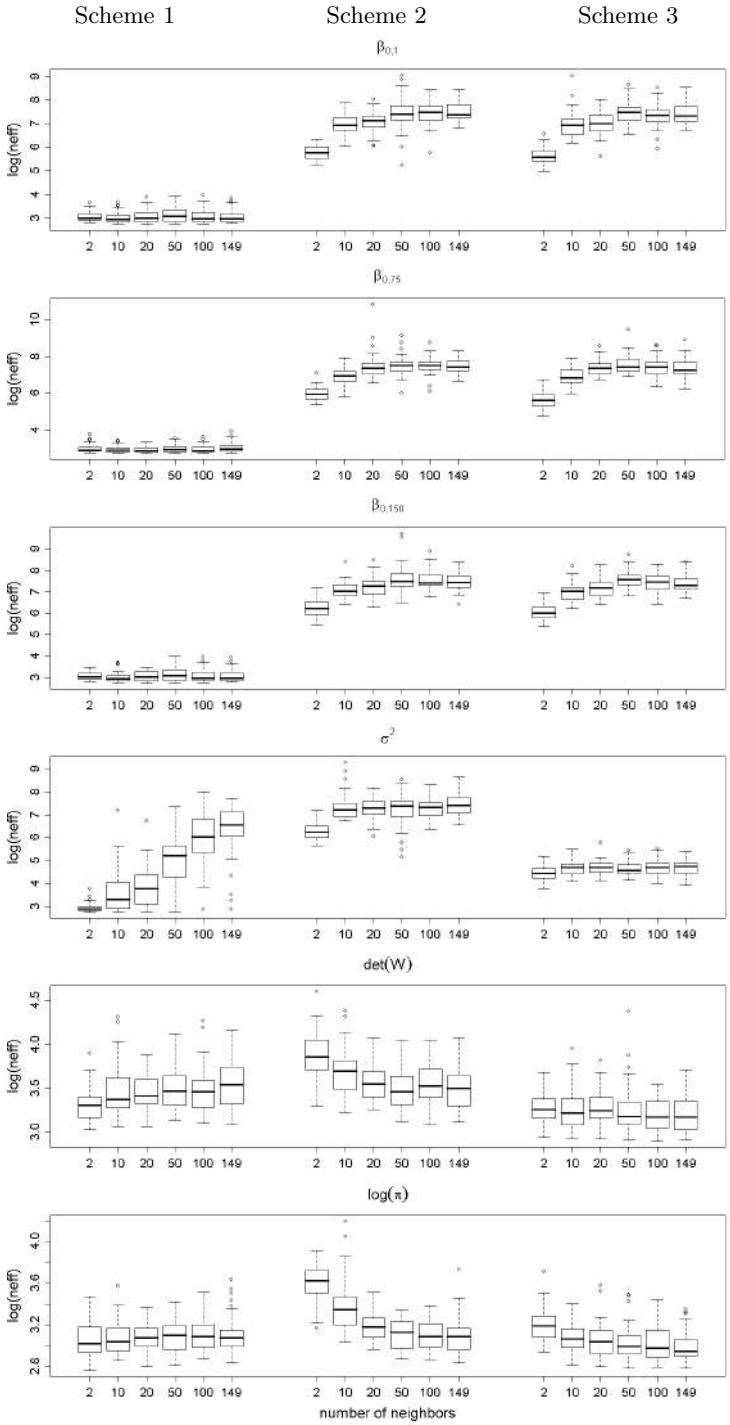

Figure 5: *Box-plots over the 50 replications of n_{eff} (in log scale) vs. number of neighbors: $\beta_{0,1}$, $\beta_{0,75}$, $\beta_{0,150}$, σ^2, $\det(W)$ and $\log(\pi)$.*

The comparative results for the regression coefficients are quite similar. Thus this section presents

only an analysis for the intercept $\beta_{0,i}$. Results for the other regression coefficients and the components of matrix W are shown in the Appendix.

Figure 5 shows a comparison of the different neighboring configurations. Neighboring influence, on the chain autocorrelation level, can not be observed for the regression coefficients in Scheme 1. The efficiency is smaller than for Schemes 2 and 3. Schemes 2 and 3 show a growth in the effective sample size as the number of neighbors is increased.

The same behavior registered in the simple spatial model analysis, can be observed for σ^2. Chains for $\det(W)$ and $\log(\pi)$ in Scheme 1 do not seem to indicate a neighboring influence. For $\det(W)$ and $\log(\pi)$, smaller number of neighbors provides chains less autocorrelated for Scheme 2. This characteristic is registered in the simple spatial model and can be also noted for $\log(\pi)$ in Scheme 3. Although influence can not be noted in this case for $\det(W)$, Figure A.2 in Appendix shows there is neighboring influence on the autocorrelation of chains for W components. The smallest autocorrelation levels are determined by structures with few neighbors per region.

7 Conclusions

This paper is concerned with the study of the simple spatial model and the multiple regression spatial model. It was shown that appropriate approximation distributions do not alter substantially the posterior distribution and can be used to replace proper by improper prior distributions or vice-versa.

An extensive computational exercise was also performed to provide guidance about the choice of appropriate proposals. The optimality was based on efficiency measures taking into account the chain autocorrelation instead of the commonly used but misleading figures of acceptance rates.

The most important contribution was to provide some light into the influence of the neighboring structures on the efficiency of the MCMC approximations. It was shown that neighboring structure influences the results, both for simple and regression spatial models.

The results obtained do not provide a unified point of view but seem to indicate a few directions. One of them is that, if suitable proposals can be found, it is generally beneficial to block parameters (in a single block, if possible). If a summarizer is required, results based on the (log) posterior density could be used. It contains information from all parameters, appropriately weighed. The result seem to indicate that the autocorrelation is smaller when the number of neighbors get smaller.

The spatial models, used in this study, belong to the class of normal observations. One of the main areas to extend results here is to consider non-normal observations. In the literature there are several applications involving spatial models considering other distributions, for example, studies using the Poisson distribution to analyze count data. Most of these studies are related to disease mapping, see for example Knorr-Held and Rue (2002) and Fernandez and Green (2002). An interesting extension of this study is analyze the neighboring influence on the MCMC methods considering non-normal spatial models. It can be anticipated that similar results will be observed if the conditional independence of observations is retained.

Appendix: additional graphics for neighboring influence analysis

Box-plots over the 50 replications of n_{eff} (in log scale) vs. number of neighbors. Graphics for Scheme 1, 2 and 3 are respectively in the left, middle and right hand side.

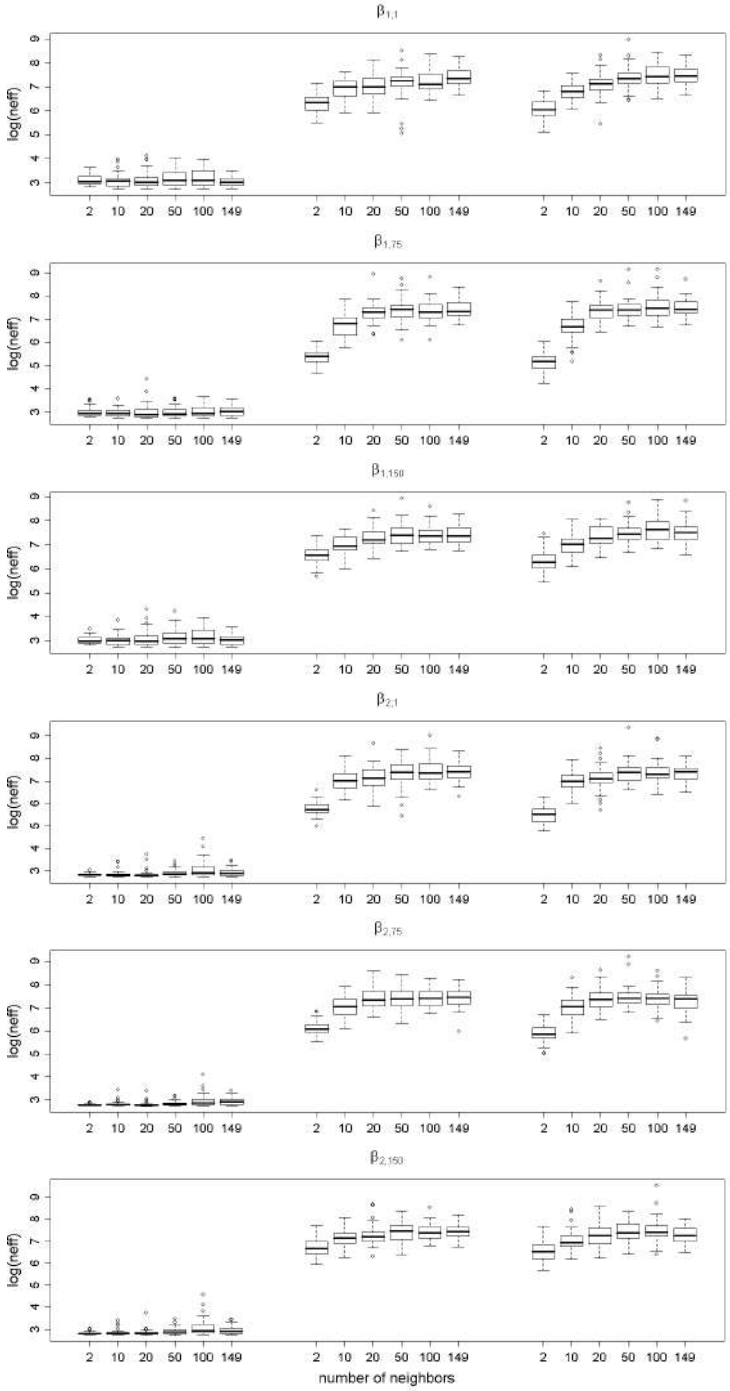

Figure A.1: $\beta_{1,1}$, $\beta_{1,75}$, $\beta_{1,150}$, $\beta_{2,1}$, $\beta_{2,75}$ *and* $\beta_{2,150}$.

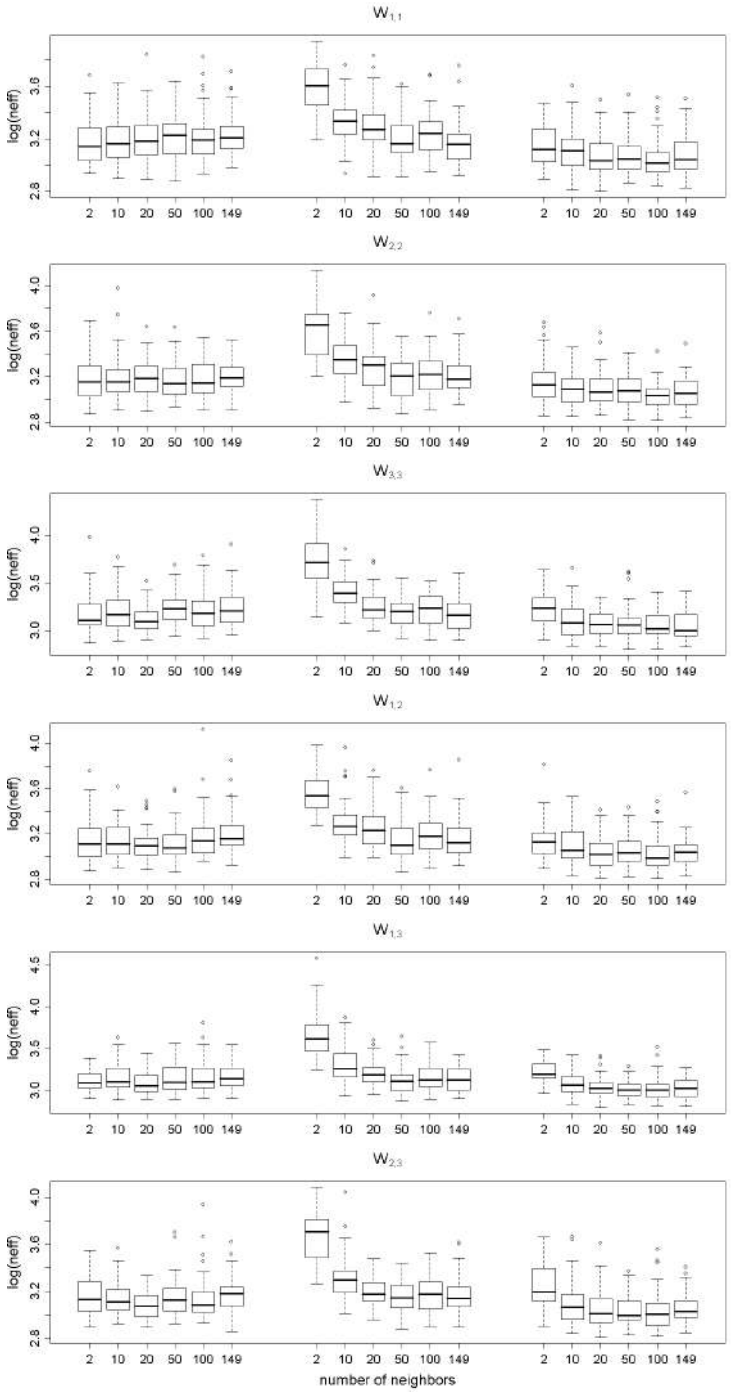

Figure A.2: $W_{1,1}$, $W_{2,2}$, $W_{3,3}$, $W_{1,2}$, $W_{1,3}$ *and* $W_{2,3}$.

Acknowledgements

The research of the first author was supported by a graduate studies grant from CAPES, Brazil. The research of the second author was supported by a grant from CNPq, Brazil.

References

Assunção, J. J., Gamerman, D., and Assunção, R. M. (1999), "Regional Differences in Factor Productivities of Brazilian Agriculture: a Space Varying Parameter Approach," in *Proceedings of the XVII Latin American Meeting of the Econometric Society*.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995), "Bayesian Analysis of Space-Time Variation in Disease Risk," *Statistics in Medicine*, 14, 2433-2443.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic System," *Statistical Science*, 10, 1, 3-66.

Besag, J., York, J., and Mollié, A. (1991), "Bayesian Image Restoration, With Two Applications in Spatial Statistics," (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Doornik, J. A., and Ooms, M. (2006), *Introduction to Ox: An Object-Oriented Matrix Language*, London: Timberlake Consultants Press.

Fernandez, C., and Green, P. (2002), "Modelling Spatially Correlated Data via Mixtures: a Bayesian Approach," *Journal of the Royal Statistical Society*, Ser. B, 64, 4, 805-826.

Ferreira, M. A. R., Higdon, D., Lee, H. K. H., and West, M. (2005), "Multi-Scale Random Field Models," Technical Report 186, Universidade Federal do Rio de Janeiro, Departamento de Métodos Estatísticos, Rio de Janeiro.

Gamerman, D., and Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (Vol. 68, 2nd ed.), eds. B. P. Carlin, C. Chatfield, M. Tanner and J. Zidek, London: Chapman & Hall/CRC.

Gamerman, D., Moreira, A. R. B., and Rue, H. (2003), "Space Varying Regression Models: Specifications and Simulation," *Computational Statistics & Data Analysis*, 42, 513-533.

Gelfand, A. E., and Vounatsou, P. (2001), "Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis," Mimeo, Research Report 301-315, University of Connecticut.

Knorr-Held, L. (1999), "Conditional Prior Proposals in Dynamic Models," *Scandinavian Journal of Statistics*, 26, 129-144.

Knorr-Held, L., Raber, G., and Becker, N. (2002), "Disease Mapping of Stage-Specific Cancer Incidence Data," *Biometrics*, 58, 492-501.

Knorr-Held, L., and Rue, H. (2002), "On Block Updating in Markov Random Field Models for Disease Mapping," *Scandinavian Journal of Statistics*, 29, 597-614.

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Correlation Structure and Convergence Rate of the Gibbs Sampler: Applications to the Comparison of Estimators and Augmentation Schemes," *Biometrika*, 81, 27-40.

Moreira, A. R. B., and Migon, H.S. (1999), "Spatial Heterogeneity of the Agricultural Productivity: Brazil - 1970 - 1996," in *Proceedings of the XXI Brazilian Meeting on Econometrics*, Brazilian Econometrics Society, Rio de Janeiro.

Reis, E. A., Salazar E., and Gamerman, D. (2006), "Comparison of Sampling Schemes for Dynamic Linear Models," *International Statistical Review*, 74, 2, 203-214.

Rue, H. (2001), "Fast Sampling of Gaussian Markov Random Fields," *Journal of the Royal Statistical Society*, Ser. B, 63, 2, 325-338.

Rue, H. (2005), "Marginal Variances for Gaussian Markov Random Fields," Statistics Report 1, Department of Mathematical Sciences, Norwegian University of Sciences and Technology, Trondheim.

Rue, H., Steinsland, I., and Sveinung, E. (2004), "Approximating hidden Gaussian Markov random fields," *Journal of the Royal Statistical Society*, Ser. B, 66, 4, 877-892.