ELSEVIER Computational Statistics & Data Analysis 42 (2003) 513–533

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Space-varying regression models: specifications and simulation

Dani Gamerman[a,*,1], Ajax R.B. Moreira[b], Håvard Rue[c]

[a] *Universidade Federal do Rio de Janeiro, Instituto de Matemática, Caixa Postal 68530, 21945-970, Rio de Janeiro, RJ, Brazil*

[b] *Instituto de Pesquisa Econômica Aplicada, Brazil*

[c] *Norwegian University of Science and Technology, Norway*

Received 1 December 2001; received in revised form 1 March 2002

Abstract

Space-varying regression models are generalizations of standard linear models where the regression coefficients are allowed to change in space. The spatial structure is specified by a multivariate extension of pairwise difference priors, thus enabling incorporation of neighboring structures and easy sampling schemes. Bayesian inference is performed by incorporation of a prior distribution for the hyperparameters. This approach leads to an untractable posterior distribution. Inference is approximated by drawing samples from the posterior distribution. Different sampling schemes are available and may be used in an MCMC algorithm. They basically differ in the way they handle blocks of regression coefficients. Approaches vary from sampling each location-specific vector of coefficients to complete elimination of all regression coefficients by analytical integration. These schemes are compared in terms of their computation, chain autocorrelation, and resulting inference. Results are illustrated with simulated data and applied to a real dataset. Related prior specifications that can accommodate the spatial structure in different forms are also discussed. The paper concludes with a few general remarks.

Keywords: Bayesian; Hyperparameters; Markov chain Monte Carlo; Markov random fields; Metropolis-Hastings algorithm; Sampling schemes

* Corresponding author. Tel.: 55-21-2562-7911; fax: 55-21-2290-1095.

E-mail address: dani@im.ufrj.br (D. Gamerman).

[1] Partially supported by grants from CNPq, FAPERJ, IPEA and Ministério de Ciência e Tecnologia (PRONEX).

0167-9473/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-9473(02)00211-6

1. Introduction

Consider the multiple linear regression model in which independent observations $y_1,\ldots,y_n$ follow $y_i \sim N(\mu_i,\sigma^2)$, for $i=1,\ldots,n$. In standard regression, the means μ_i are described by a linear relation $x_i'\beta$, where x_i is the r-dimensional vector of explanatory variables for the ith observational unit ($i=1,\ldots,n$) and β is the vector of regression coefficients. Inference on β provides information about the strength of the x's explanation of the variability of the observations y.

The extensions we consider below allow the regression coefficients to change with the observation unit. Hence, we will have $\mu_i = x_i'\beta_i$, for $i=1,\ldots,n$ and unknown quantities (β,ϕ), where $\beta = \mathrm{vec}(B)$, the column vectorization of the $n\times r$ matrix B given by

$$B = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1r} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2r} \\ & & \vdots & \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nr} \end{pmatrix} = \begin{pmatrix} \beta_1' \\ \beta_2' \\ \vdots \\ \beta_n' \end{pmatrix} \stackrel{\text{def}}{=} (\beta^1 \quad \beta^2 \quad \cdots \quad \beta^r)$$

and $\phi = 1/\sigma^2$. The likelihood is

$$l(\beta,\phi) = (2\pi)^{-n/2}\phi^{n/2}\exp\left\{-\frac{\phi}{2}\sum_{i=1}^{n}(y_i - x_i'\beta_i)^2\right\}. \tag{1}$$

Under a Bayesian formulation, the model must be completed by a prior for (β,ϕ), where assumptions about the relation between the components of β are made. Note also that this prior may also depend on other unknown quantities. They are typically variance parameters and are denoted by Ψ. Priors for these hyperparameters must also be specified. In general, these are taken as reasonably vague, reflecting the difficulty of incorporating substantial knowledge about them.

There are many possible routes to be taken, associated with qualitatively different assumptions about the relationship among the β_i's. They are in general based on weak quantitative assumptions, generally through vague priors on the hyperparameters. These models can only lead to precise posterior inference if the dimension r of the β_i's is substantially smaller than the number n of observational units.

One possibility is to allow the regression coefficients to vary across units without any particular structure. Hierarchical models introduced by Lindley and Smith (1972) suggest in their simplest forms that the β_i's should be allowed to vary as a random sample from a given distribution, indexed by Ψ. Another possibility, popularized by West and Harrison (1997), is to use a dynamic model.

For spatially distributed data, a number of possibilities are available. One can use the geostatistical approach (Cressie, 1991, Chapters 2–5), where the regression errors $y_i-\mu_i$ are spatially correlated, or the discrete approach of pairwise different priors (Besag et al., 1991), where unit-specific random effects have their distribution specified according to the neighboring arrangement of the units. We will pursue the second route through a multivariate generalization.

Even though the likelihood does not depend on Ψ, the posterior distribution for all model parameters must be evaluated. This is given by

$$\pi(\beta, \phi, \Psi | y) \propto l(\beta, \phi) f(\beta \mid \phi, \Psi) p(\phi) p(\Psi), \tag{2}$$

where it is assumed that the joint prior density f for β depends on Ψ and ϕ, and ϕ and Ψ are prior independent. It will be sometimes useful to define $\theta \equiv (\phi, \Psi)$.

The observational precision ϕ is generally given a Gamma prior with density $\phi^{(v_\phi/2)-1} \exp\{-v_\phi S_\phi \phi/2\}$, denoted $G(v_\phi/2, v_\phi S_\phi/2)$, and, when Ψ is a precision matrix, it is generally given a Wishart prior with density $|\Psi|^{(v_\psi/2)-(r+1)/2} \exp\{-tr(v_\psi S_\psi \Psi)/2\}$, denoted $W(v_\psi/2, v_\psi S_\psi/2)$.

The rest of the paper is organized as follows. The next section introduces the spatial model used in the regression context and derives the posterior distribution. Section 3 presents four different schemes to do sampling-based inference using Markov chain Monte Carlo (MCMC) technique. Section 4 provides some results from a simulation study and a real dataset. Section 5 presents other forms of prior specification, and Section 6 provides concluding comments.

2. Model definition

A model suitable for representing spatial situations is provided by Markov random fields (MRF). In simple terms, a collection $X = (X_1, \ldots, X_n)$ of random quantities is said to form a MRF if the joint distribution of X satisfies the property that $(X_i | X_{-i}) \sim (X_i | X_{\partial i})$, where $\partial i = \{j\colon\ j \text{ is a neighbor of } i\}$, for $i = 1, \ldots, n$. When the regions are ordered, $\partial i = \{i-1, i+1\}$ and the condition reduces to $(X_i | X_{-i}) \sim (X_i | X_{i-1}, X_{i+1})$, for all $i \neq 1, n$.

There are many possible prior models for β that follow a MRF. An interesting example is the pairwise difference prior

$$f_{pd}(\beta | \phi, \Psi) \propto \phi^{nr/2} |\Psi|^{n/2} \exp\left\{ -\frac{\phi}{2} \sum_{i,j=1}^{n} w_{ij} (\beta_i - \beta_j)' \Psi (\beta_i - \beta_j) \right\}, \tag{3}$$

where w_{ij} are weights associated with the neighboring structure. For example

$$w_{ij} = \begin{cases} 1 & \text{if } i \sim j, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $i \sim j$ means that regions i and j are neighbors. Of course, the matrix Ψ need not be scaled by ϕ, but this can be useful since it allows for easier calculations later on and provides a comparative scale to measure prior precision. Removal of ϕ from (3) implies trivial changes in calculation. The form (3) is proposed by Moreira and Migon (1999) and Assunção et al. (1999), without scaling on ϕ. It is basically a generalization of the univariate pairwise difference prior used by Besag et al. (1991). Similar forms have been proposed by Mardia (1988), Bernardinelli et al. (1995), Kim et al. (2001) and Gelfand and Vounatsou (2001). Pairwise differences are used as a model for regression coefficients by Assunção et al. (1998).

This prior attributes larger probability to regions of the β space that have similar values for neighboring β_i's. It is an improper prior because the variance of β is $\phi^{-1}W^{-1} \otimes \Psi^{-1}$, where $W = (k_{ij})$ and

$$k_{ij} = \begin{cases} w_{i+} & \text{if } i = j, \\ -w_{ij} & \text{if } i \sim j, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and } w_{i+} = \sum_{j \sim i} w_{ij}.$$

Consequently, the rows of the $n \times n$ matrix W add up to zero. However, this is a useful prior representation of spatial structure and leads to proper posteriors and sensible results, provided proper priors are specified for ϕ and Ψ.

Simple manipulation shows that (3) can be rewritten as

$$f_{pd}(\beta|\phi, \Psi) \propto \phi^{nr/2}|\Psi|^{n/2} \exp\left\{-\frac{\phi}{2} Q(\beta)\right\},$$

where $Q(\beta)$ has different expressions. The most useful ones are

$$Q(\beta) = \sum_{i,j=1}^{n} k_{ij}\beta_i'\Psi\beta_j = \beta'(W \otimes \Psi)\beta = \text{tr}[Q_s(\beta)\Psi],$$

where $Q_s(\beta) = B'WB = \sum_{i,j=1}^{n} k_{ij}\beta_i\beta_j' = \sum_{i,j=1}^{n} w_{ij}(\beta_i - \beta_j)(\beta_i - \beta_j)'$. It can also be shown that the full prior conditional distribution of β_i is given by

$$\beta_i|\beta_{-i}, \phi, \Psi \sim \beta_i|\beta_{\partial i}, \phi, \Psi \sim N\left(\bar{\beta}_{\partial i}, \frac{1}{\phi w_{i+}}\Psi^{-1}\right), \tag{5}$$

where $\bar{\beta}_{\partial i}=(1/w_{i+})\sum_{j\sim i} w_{ij}\beta_j$ is the average of the w_{i+} neighboring β's, for $i=1,\ldots,n$.

Combination of all model assumptions gives

$$\begin{aligned}\pi(\beta, \phi, \Psi|y) \propto{}& \phi^{n/2} \exp\left\{-\frac{\phi}{2}\sum_{i=1}^{n}(y_i - x_i'\beta_i)^2\right\} \phi^{nr/2}|\Psi|^{n/2} \\ &\times\exp\left\{-\frac{\phi}{2}\beta'(W \otimes \Psi)\beta\right\} \phi^{(v_\phi/2)-1} \exp\{-\tfrac{1}{2} v_\phi S_\phi \phi\}|\Psi|^{(v_\psi/2)-r} \\ &\times\exp\{-\tfrac{1}{2} tr(v_\psi S_\psi \Psi)\} \\ \propto{}& \exp\left\{-\frac{\phi}{2}[v_\phi S_\phi + \beta'(X'X + W \otimes \Psi)\beta - 2\beta'X'y + y'y]\right\} \\ &\times\phi^{[(v_\phi+n+nr)/2]-1}|\Psi|^{[(v_\psi+n)/2]-r} \exp\{-\tfrac{1}{2} tr(v_\psi S_\psi \Psi)\},\end{aligned} \tag{6}$$

where the design matrix $X = \text{diag}(x_1', \ldots, x_n')$ is in a slightly unusual block diagonal form. If Ψ is not scaled by ϕ in (3), then the posterior becomes

$$\begin{aligned}\pi(\beta, \phi, \Psi|y) \propto{}& exp\left\{-\frac{\phi}{2}[v_\phi S_\phi + \beta'X'X\beta - 2\beta'X'y + y'y] - \frac{1}{2}\beta'(W \otimes \Psi)\beta\right\} \\ &\times\phi^{[(v_\phi+n)/2]-1}|\Psi|^{[(v_\psi+n)/2]-r} \exp\left\{-\tfrac{1}{2} tr(v_\psi S_\psi \Psi)\right\}.\end{aligned} \tag{7}$$

These distributions are not easily summarized and special approximating schemes are needed. We shall concentrate here on the MCMC methodology, where samples are repeatedly taken from Markov chain kernels to reproduce a stationary trajectory towards chain equilibrium (Gamerman, 1997). A central issue is the determination of fast sampling schemes based on full conditionals. Several such schemes are detailed in the next section.

3. Sampling schemes

It should be noted that many possibilities are available for sampling from (6). We consider the following sampling schemes:

(A) Sampling from $\beta_1, \ldots, \beta_n$, ϕ and Ψ,
(B) Sampling from (β, ϕ) and Ψ,
(C) Sampling from (β, ϕ, Ψ) jointly.

It is generally believed in applied MCMC work that blocking is beneficial. This ranks the schemes in the sequence C > B > A, but no theoretical results support this belief. Similarly in dynamic (or state space) models, it has been shown by extensive empirical evidence that B > A, as expected (Shephard, 1994). Gamerman and Moreira (2002) uses scheme C for these models but provides no comparison with the other schemes.

Multimodality is likely to occur for complicated posterior forms. These spatial models are known to have convergence problems (Knorr-Held and Rue, 2002). Sampling schemes should converge to the relevant mode(s) of the posterior distribution.

Note that, for scheme (C), samples must be drawn from the marginal posterior of $\Psi|y$. Samples from β and ϕ are then obtained by noting

$$\pi(\beta, \phi, \Psi|y) = \pi(\beta, \phi|\Psi, y)\pi(\Psi|y).$$

Draws from (β, ϕ) are obtained by drawing a value Ψ^* from the marginal posterior $\pi(\Psi|y)$ and then sampling from the tractable distribution $\pi(\beta, \phi|\Psi^*, y)$.

For schemes A and B, the full posterior conditionals for ϕ and for precision matrices Ψ are trivially obtained from (6) as $G\{(v_\phi+n+nr)/2, [\beta'(X'X+W\otimes\Psi)\beta-2\beta'X'y+y'y]/2\}$ and $W\{(v_\psi+n)/2, [v_\psi S_\psi+\phi Q_s(\beta)]/2\}$ distributions, respectively. We, therefore, concentrate on sampling from the regression coefficients β. The case when Ψ is not a full precision matrix is dealt with in the next section, but note that it is always true that $\pi(\Psi|\beta, \phi, y) \propto p(\Psi)W[v_\psi/2, \phi Q_s(\beta)/2]$.

When the pairwise difference precision Ψ is not scaled by ϕ, the full conditionals for ϕ and Ψ are $G\{(v_\phi+n)/2, [\beta'X'X\beta-2\beta'X'y+y'y]/2\}$ and $W\{(v_\psi+n)/2, [v_\psi S_\psi + Q_s(\beta)]/2\}$ distributions, respectively.

3.1. Scheme A

The full posterior conditional of β_i is easily obtained as

$$\pi(\beta_i|\beta_{-i}, \phi, \Psi, y) \propto l(\beta_i)f_{pd}(\beta_i|\beta_{-i}, \phi, \Psi).$$

The likelihood term is simply $\exp\{-\phi(y_i - x_i'\beta_i)^2/2\}$ and the prior term is given by (5). Combining these results gives $(\beta_i|\beta_{-i}, \phi, \Psi, y) \sim (\beta_i|\beta_{\partial i}, \phi, \Psi, y) \sim N(a_i, \phi^{-1}R_i)$, where $a_i = R_i(x_i y_i + w_{i+}\Psi\bar{\beta}_{\partial i})$ and $R_i = (x_i x_i' + w_{i+}\Psi)^{-1}$, for $i = 1, \ldots, n$. When Ψ is not scaled by ϕ, the expressions for a_i and R_i change to $a_i = R_i(\phi x_i y_i + w_{i+}\Psi\bar{\beta}_{\partial i})$ and $R_i = (\phi x_i x_i' + w_{i+}\Psi)^{-1}$, $i = 1, \ldots, n$. These distributions are easily sampled and the most computationally demanding task is the inversion of $r \times r$ variance matrices. Therefore, sampling cost is not an issue here as r is usually rather small (in the order of 10^1) but convergence rate may be.

Neighboring β_i's are expected to be highly correlated due to their prior form and if this correlation is large this may cause considerable delay in reaching chain equilibrium. This problem may be severe since there may be many β_i's (typically in the order of 10^3–10^4).

3.2. Scheme B

Define $R = (X'X + W \otimes \Psi)^{-1}$ and $a = RX'y$. The full posterior conditional of (β, ϕ) is obtained from (6) as

$$\pi(\beta, \phi|\Psi, y) \propto \phi^{[(v_\phi + n + nr)/2]-1} \exp\left\{-\frac{\phi}{2}[v_\phi S_\phi + \beta' R^{-1}\beta - 2\beta' R^{-1} a + y'y]\right\}$$

$$\propto \phi^{[(\hat{v}_\phi + nr)/2]-1} \exp\left\{-\frac{\phi}{2}[(\beta - a)' R^{-1}(\beta - a) + \hat{S}_\phi]\right\}, \qquad (8)$$

where $\hat{v}_\phi = v_\phi + n$ and $\hat{S}_\phi = v_\phi S_\phi + (y - Xa)'(y - Xa)$. It is clear from (8) that $(\beta, \phi|\Psi, y) \sim NG(a, R, \hat{v}_\phi/2, \hat{S}_\phi/2)$, that is, $(\beta|\phi, \Psi, y) \sim N(a, \phi^{-1}R)$ and $(\phi|\Psi, y) \sim G(\hat{v}_\phi/2, \hat{S}_\phi/2)$.

When Ψ is not scaled by ϕ, the full posterior conditional of (β, ϕ) is no longer in closed *NG* form. Nevertheless, it is straightforward to obtain the full conditional of β for (7) as $(\beta|\phi, \Psi, y) \sim N(a, R)$, where the expression for the moments are changed to $R = (\phi X'X + W \otimes \Psi)^{-1}$ and $a = \phi RX'y$.

These distributions are also simple to sample from, but now the computational demand is substantially increased since we need to invert $nr \times nr$ dimensional matrices to obtain R. Fortunately, this is not required here. Rue (2001) shows that substantial computational savings are obtained by exploring MRF properties of R^{-1}. A summary of Rue's strategy in our setting is given in Gamerman et al. (2001).

The advantage of this approach over scheme A is its ability to enact block sampling over the possibly highly correlated components β_i. The disadvantage is its computational cost: the permutations and the Cholesky decomposition. This scheme is used by Besag and Higdon (1999).

A variation of schemes A and B is to sample from the β^j's. The similarity between line and column vectorization of B can be invoked to obtain the full prior conditional for β^j induced from (3). This similarity cannot be used in the posterior given the asymmetry in the way B enters the likelihood. Conditional distributions are easily obtained (Gamerman et al., 2001).

3.3. Scheme C

It is easy to see from (8) that the constant of proportionality required to complete the expression for the full posterior conditional density of (β, ϕ) is

$$(2\pi)^{nr/2}|R|^{-1/2}\,\frac{(\hat{S}_\phi/2)^{\hat{v}_\phi/2}}{\Gamma(\hat{v}_\phi/2)}.$$

Note, however, that $\pi(\Psi|y) = \pi(\beta, \phi, \Psi|y)/\pi(\beta, \phi|\Psi, y)$. Combining the above with (6) and (8), and discarding constants, gives

$$\begin{aligned}
\pi(\Psi|y) &\propto |R|^{1/2}(\hat{S}_\phi)^{-\hat{v}_\phi/2}|\Psi|^{n/2}\,p(\Psi)\\
&\propto \frac{|\Psi|^{n/2}}{|R^{-1}|^{1/2}}\,(\hat{S}_\phi)^{-\hat{v}_\phi/2}\,p(\Psi)\\
&\propto \frac{|\Psi|^{n/2}}{\prod_{i=1}^{nr} l_{ii}}\,(\hat{S}_\phi)^{-\hat{v}_\phi/2}\;p(\Psi),
\end{aligned} \tag{9}$$

where we make use of the facts that $|R^{-1}| = |L|^2 = (\prod_{i=1}^{nr} l_{ii})^2$ and $R^{-1}a = X'y$. Note that $\hat{S}_\phi$ depends on Ψ.

When Ψ is not scaled by ϕ in (3), only β can be integrated out in (7). Using the results obtained for scheme B, it is not difficult to obtain that

$$\begin{aligned}
\pi(\phi, \Psi|y) &\propto \phi^{[(v_\phi+n)/2]-1}\exp\left\{-\frac{\phi}{2}\,\hat{S}_\phi\right\}|\phi X'X + W\otimes\Psi|^{-1/2}\\
&\quad\times|\Psi|^{(v_\psi+n-r-1)/2}\exp\{-\tfrac{1}{2}\,tr(v_\psi S_\psi \Psi)\}.
\end{aligned} \tag{10}$$

In this case, ϕ cannot be integrated out analytically as before.

Although it is possible to obtain (9) or (10) analytically up to a proportionality constant, it is not easy to devise direct sampling schemes. Indirect sampling schemes such as SIR (Rubin, 1988) or adaptive SIR (Schmidt et al., 1999) may be applied here if the dimensions are not large. As a general purpose scheme, MCMC with Metropolis–Hastings proposals (Metropolis et al., 1953) are used here. Two possible proposal forms are random walks and sampling from an approximating density such as the Wishart for Ψ and Gamma for ϕ, where applicable.

4. Applications

Two applications are used to illustrate the models and sampling schemes: one with simulated and one with a real data. The spatial structure used in the applications is given by the microregions (simulated data) and counties (real data) of Brazil, as defined by the official Brazilian Statistics Institute, IBGE. The neighboring structure is defined by the existence of a common border of any length between regions; $w_{ij} = 1$ for such pairs and 0 otherwise and Ψ is not scaled by ϕ.

The MCMC algorithms are run with two parallel chains. Initial values are generated from the hyperparameters' prior distribution for all sampling schemes. Given that this prior is proper but has very large variances, these values provide reasonable reassurance that different regions of the parameter space are explored. It should be unlikely for both chains for the same scheme to converge to the same local mode. Convergence is diagnosed according to the test of Geweke (1992) within chains and the Gelman and Rubin (1992) shrinkage factor.

The analyses use results obtained from the last 1000 values from each chain, totalling 2000 values for all model parameters. Scheme C is run with a random walk form for the proposal described above and d.o.f. tuned to acceptance rates around 40%. In general, iterations from scheme A are roughly 20 times faster than those of schemes B and C. In addition, the burn-in period seems to be unaffected by changes in the hyperparameters' prior and indicate faster convergence for scheme A (after about 400 iterations) than for schemes B and C (after about 800 iterations). We have opted to allocate the same CPU time to all schemes. Since iterations of scheme A are the fastest, each chain is allowed to run additional iterations, and values stored only when the time limit is reached.

4.1. Simulated dataset

To study the performance of the schemes and the ability of the posterior to estimate the model parameters, a simulated dataset is used. The first difficulty appears because the improper form of (3) prevents direct generation from the model. An approximating alternative inspired by the spatial structure of Anselin (1988) is used. Specifically, the β's are generated from the model

$$\beta_i = \rho\bar{\beta}_{\partial i} + e_i\,, \quad e_i \sim N\left(0, \frac{1}{w_{i+}}\,\Psi^{-1}\right), \tag{11}$$

independently for $i = 1,\ldots,n$.

The independence imposed on the e_i's means that we are approximating (3) by $f(\beta) \propto \prod_{i=1}^{n} f_i(\beta_i;\beta_{-i})$ where each f_i is given by (11). When $\rho = 1$, the f_i's are the full prior conditionals in (5), the prior becomes improper and gives the same approximation suggested in Besag (1975) in the context of observations from a Markov random field (Qian and Titterington, 1991).

The spatial structure used in the simulation is illustrated in Fig. 1. Parameter generation is carried out with $\rho = 0.999$ and completed with observations drawn from model (1) with $n = 558$, $r = 2$, a diagonal matrix $\Psi = diag((0.03)^{-2}, (0.15)^{-2})$, and $\phi = 0.25$.

Table 1 shows the estimated mean value and their Monte Carlo standard errors of $\log \pi(\theta)$, where $\theta = (\phi, \Psi)$ for schemes B and C, and a number of prior value specifications for $v_\phi = v_\psi = v$ and $S_\psi = S_\phi I = SI$. The prior means are, therefore, given by $E(\phi) = E(\Psi_{ii}) = S^{-1}$, for $i = 1,\ldots,r$.

A wide range of prior specifications is used and shows a variety of strengths of prior information ranging from strong to weak, as measured by the number of degrees of freedom, and expected values ranging from small to large for the hyperparameters. The results show that schemes B and C lead to the same posterior values, indicating convergence to the same parameter region. The Monte Carlo errors of scheme C are consistently smaller than those of scheme B, indicating a better mixing of the chains.

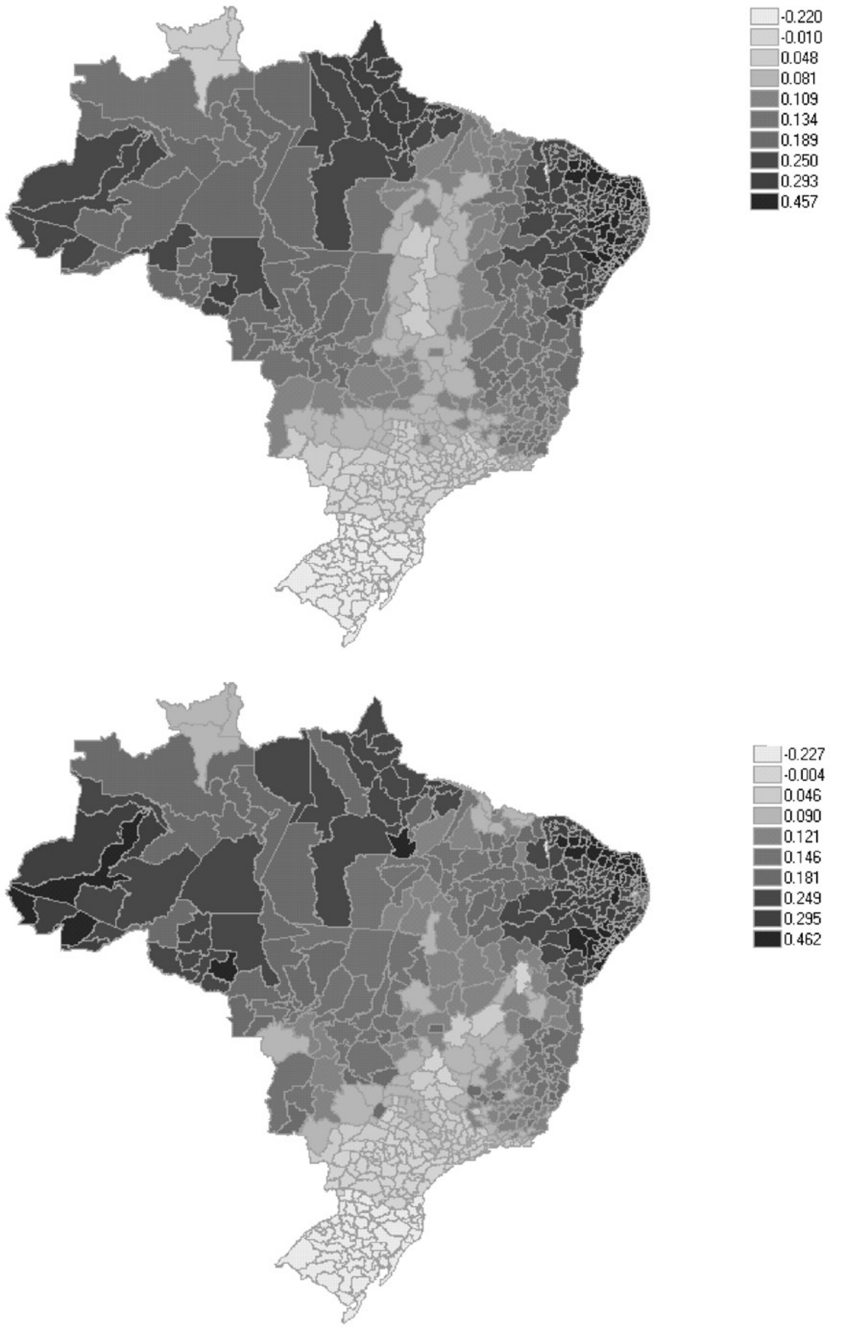
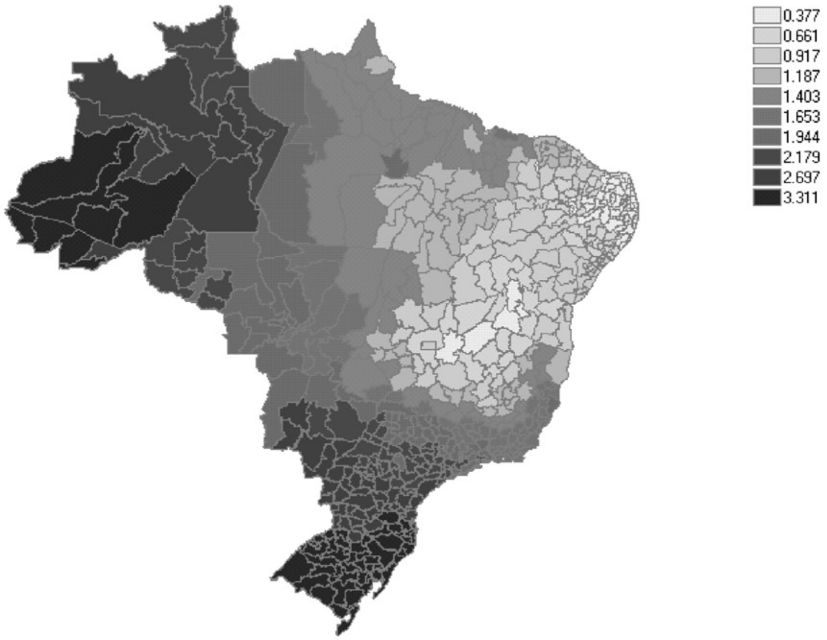


Fig. 1. Spatial summary of estimation of the regression coefficients for the simulated dataset and scheme C. First panel: top—true values of β^1; bottom—posterior mean of β^1; second panel: top—true values of β^2; bottom—posterior mean of β^2; third panel; posterior standard deviation of β^1 (top) and β^2 (bottom).



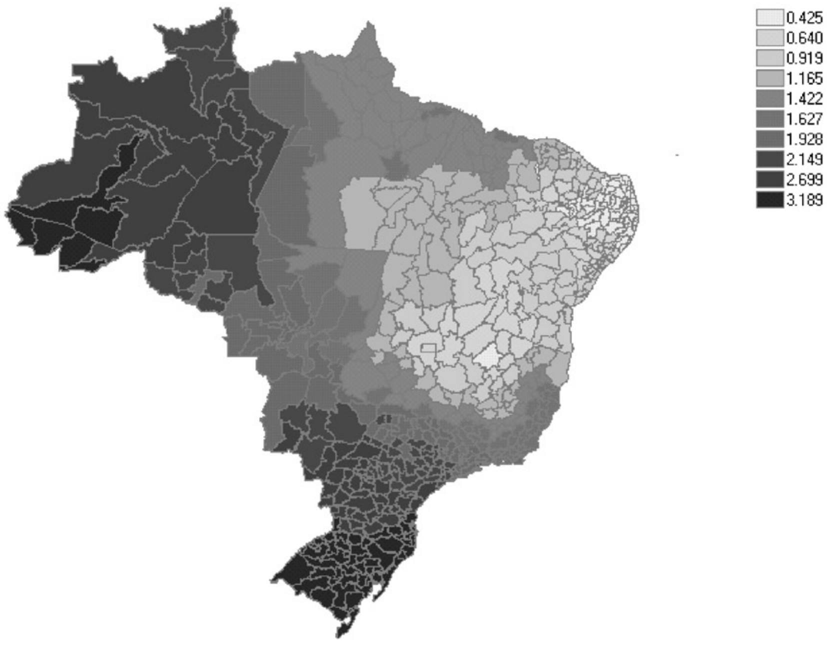


Fig. 1. (*Continued*).

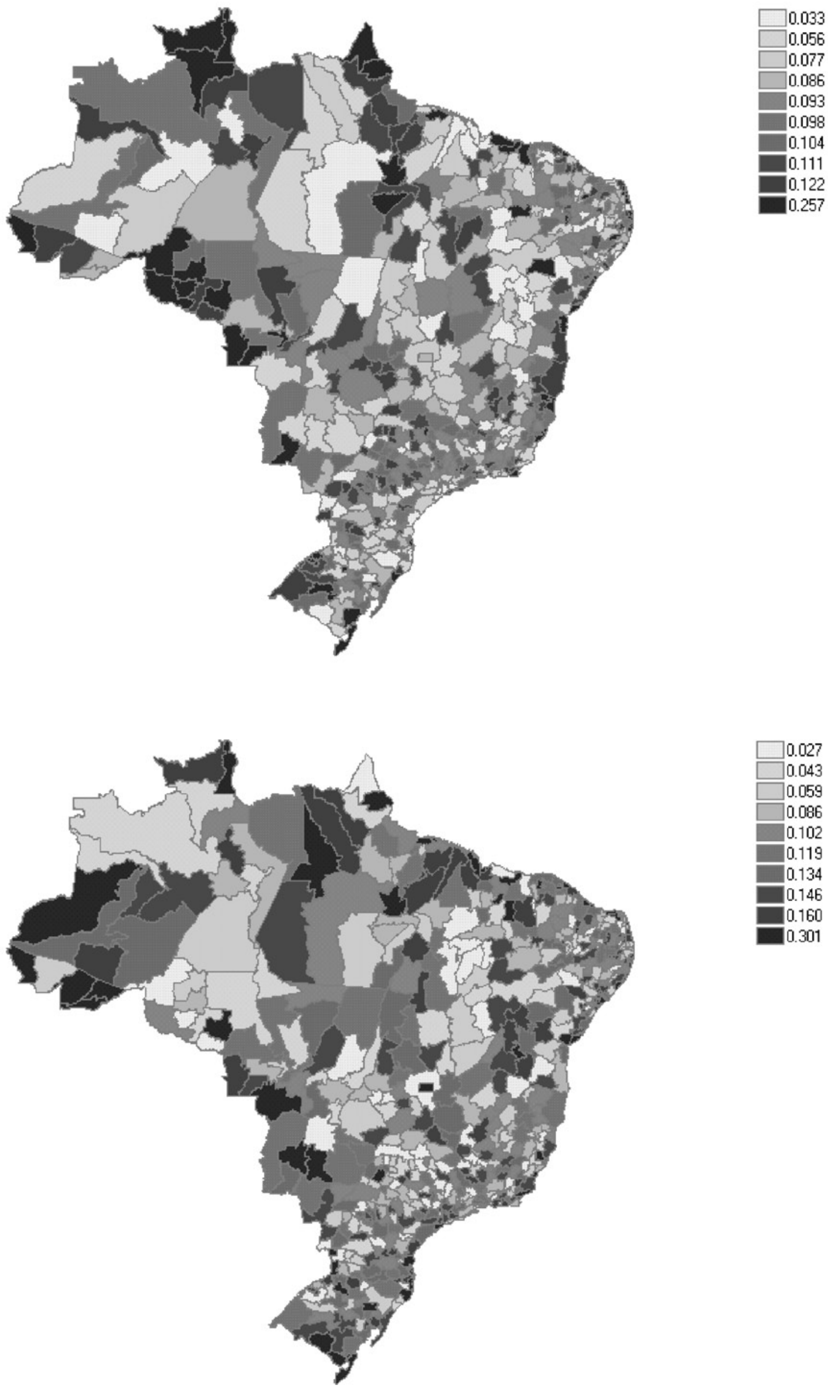


Fig. 1. (*Continued*).

Table 1
Marginal posterior density (in log)

Prior		Scheme	
d.o.f. (v)	Mean ($1/S$)	B	C
10	1	−2297	−2296
		(0.51)	(0.48)
	100	−2103	−2103
		(0.58)	(0.44)
50	1	−2579	−2579
		(0.85)	(0.39)
	100	−1912	−1912
		(0.79)	(0.43)
100	1	−2746	−2746
		(0.63)	(0.39)
	100	−1671	−1669
		(0.76)	(0.43)

MC standard errors are provided in brackets.

Figs. 1–3 show the results of the estimation in space (Fig. 1) and over the line (Figs. 2 and 3) of the spatially varying regression coefficients along with uncertainty bounds for the prior with $v = 10$ and mean 1. The strong spatial pattern of the process is clear from the figures. The estimates appear to reproduce the true spatial pattern well. Squared deviations can be defined as $(\beta_i - \hat{\beta}_i)^2$, where $\hat{\beta}_i$ is the posterior mean of β_i. These quantities can be summed over all regions to provide an overall measure of fit, SSD_i. These figures, provided in Table 2, show no apparent difference between schemes A, B and C.

The SSD figures change very little for all schemes as the parameters of the prior distribution are changed. This indicates that there is strong information in the likelihood for spatially varying regression coefficients; they are reliably estimated, and the results do not change even after substantial changes in the prior. The same is not true for the hyperparameters.

A subset of the results from point estimation of ϕ is presented in Table 3. Posterior estimates do not concentrate around the generated values. They tend to concentrate around prior means, indicating scarcity of information in the likelihood. Credibility intervals are not shown for brevity but indicate similarity between all schemes, and their lengths decrease with increasing prior d.o.f. as expected. Differences between the estimated posterior means with all schemes are well within their Monte Carlo uncertainty, also reported in the table. The Monte Carlo standard errors reflect the better mixing of schemes B and C. Combined with the errors obtained from Table 1, this is an indication of preference for scheme C followed by scheme B and then finally scheme A. Similar results are obtained for the other hyperparameters and other prior specifications.

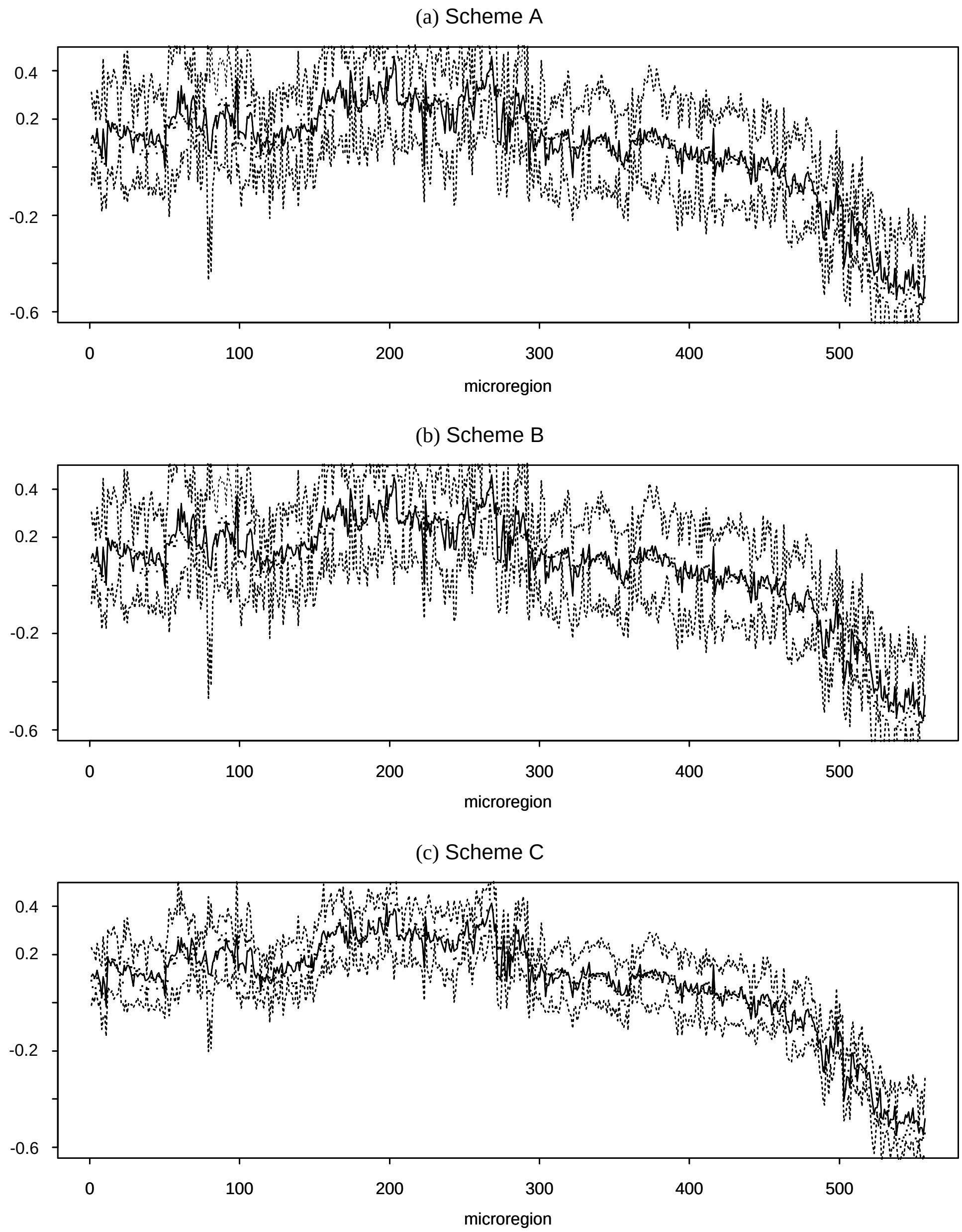


Fig. 2. Linearized summary of estimation of the regression coefficients for the simulated dataset and β^1: (a) Scheme A; (b) Scheme B; (c) Scheme C. True values are represented by dots, posterior means by full lines and 2 posterior standard deviation limits by dashed lines.

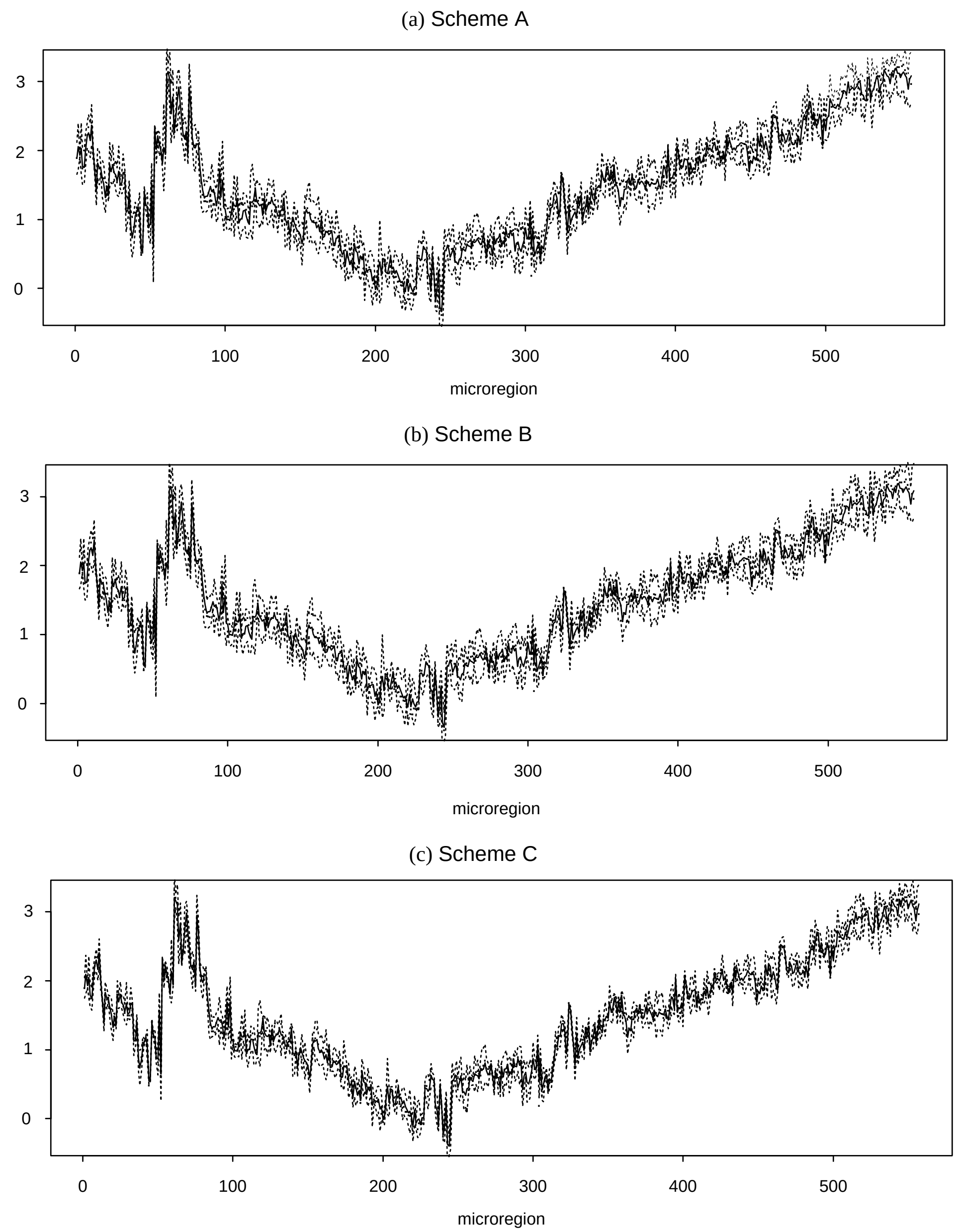


Fig. 3. Linearized summary of estimation of the regression coefficients for the simulated dataset and β^2: (a) Scheme A; (b) Scheme B; (c) Scheme C. True values are represented by dots, posterior means by full lines and 2 posterior standard deviation limits by dashed lines.

Table 2
SSD values for simulated dataset

	Prior		Scheme		
	d.o.f. (v)	Mean ($1/S$)	A	B	C
	10	1	0.042	0.041	0.041
		100	0.039	0.039	0.039
β_1	50	1	0.052	0.052	0.052
		100	0.039	0.039	0.039
	100	1	0.055	0.055	0.055
		100	0.039	0.039	0.039
	10	1	0.073	0.073	0.073
		100	0.071	0.071	0.071
β_2	50	1	0.081	0.081	0.081
		100	0.071	0.071	0.070
	100	1	0.083	0.083	0.083
		100	0.071	0.071	0.071

Table 3
Point estimation of ϕ

	Prior		Scheme		
	d.o.f. (v)	Mean ($1/S$)	A	B	C
ϕ		1	1.93	1.66	1.49
(true = 0.25)	10		(0.31)	(0.27)	(0.28)
		100	160.3	120.8	121.4
			(3.2)	(2.8)	(2.3)

MC standard errors are provided in brackets.

4.2. Real dataset

The Amazon region is a vast area in the North of Brazil that recently is becoming an agricultural frontier. A question of interest is to determine the pace with which land use is changing. Andersen et al. (1997) (AGR, hereafter) proposes a vector autoregressive (VAR) model for land usage that enables projection into the future. Denoting by y_{jkt} the proportion of land in county j used for purpose k at time t and by x_{jt} the proportion of land in county j still covered by forest ($k=1$), the model proposed is

$$y_{jkt} = \beta_{jk0}\,\Delta y_{j1t} + \sum_{m=2}^{K} \beta_{jkm}\, y_{jm,t-1} + e_{jkt},$$

where $j=1,\ldots,J$, $k=1,\ldots,K$, $t=1,\ldots,T$ and Δ is the time lag operator. The temporal dependence is assumed to be captured by the VAR structure implying that

Table 4
Estimated values of β

Coefficient	AGR	Fixed	Spatial
β_0	0.2919	0.1384	0.1420
β_1	−0.0440	−0.0176	0.2760
β_2	1.1543	1.2771	1.0230
β_3	0.1114	0.0545	0.0730

the errors are independent across time and the likelihood is a product of likelihoods (1) over time. The authors acknowledge the presence of heterogeneity between counties by adding an extra layer to the model so that the regression coefficients β_{jkm} vary across counties in an unstructured hierarchical form. Spatially determined variation seems a more reasonable assumption to account for the similarity between neighboring counties due to the stage of land occupation by settlers or proximity to large cities.

In this application, the spatial structure is determined by the $J = 228$ administrative counties in the Amazon region of Brazil, depicted in Fig. 4, with $K = 4$ land usages: forest, crop land, pasture and fallow land, collected in years 1970, 1975, 1980 and 1985 ($T = 4$). Variation of coefficients across counties is modelled with the spatial prior (3) and only results for pasture ($k = 2$) are reported here. The data contains multiple observations per county and, therefore, the likelihood consist of a product of likelihoods (1) over time.

The main results of the analysis are reported in Table 4 and Figs. 4 and 5. Table 4 presents a comparative analysis of the results obtained by AGR with those from the models with fixed and spatially varying vector of regression coefficients. For the AGR and spatial models, where regression coefficients vary across counties, the mean value of the average is reported in the table. There is reasonable agreement between these averages across models, even though the data used by AGR are different because of minor adjustments. We have also compared the fixed and spatial models using the predictive fit statistic D proposed in Gelfand and Ghosh (1998). Their statistic measures goodness of fit but combines it with a penalty for model complexity using decision-theoretic justifications. The values obtained indicate $D_{\text{static}} = 0.035 > 0.016 = D_{\text{spatial}}$, showing a preference for the more complex and more realistic model with spatial variation.

Fig. 4 shows the spatial distribution of the posterior mean of the β_{j21}'s. These coefficients measure the effect of changing land use from agriculture to pasture in each county. The results show a clear pattern of change as one moves from North to South. The effect is more intense in the South of the region. This area is known to be the main access of humans to the region. Human occupation in the region is typically characterized by low technology agriculture which quickly exhausts land resources and leads to pasture use. Our findings confirm this.

Fig. 5 shows the spatial variation arranged over a line for all regression coefficients along with corresponding credibility limits. The coefficients β_{j20}'s are marginally significant and all the other coefficients are clearly significant with the β_{j21}'s and β_{j23}'s showing a similar spatial behavior.

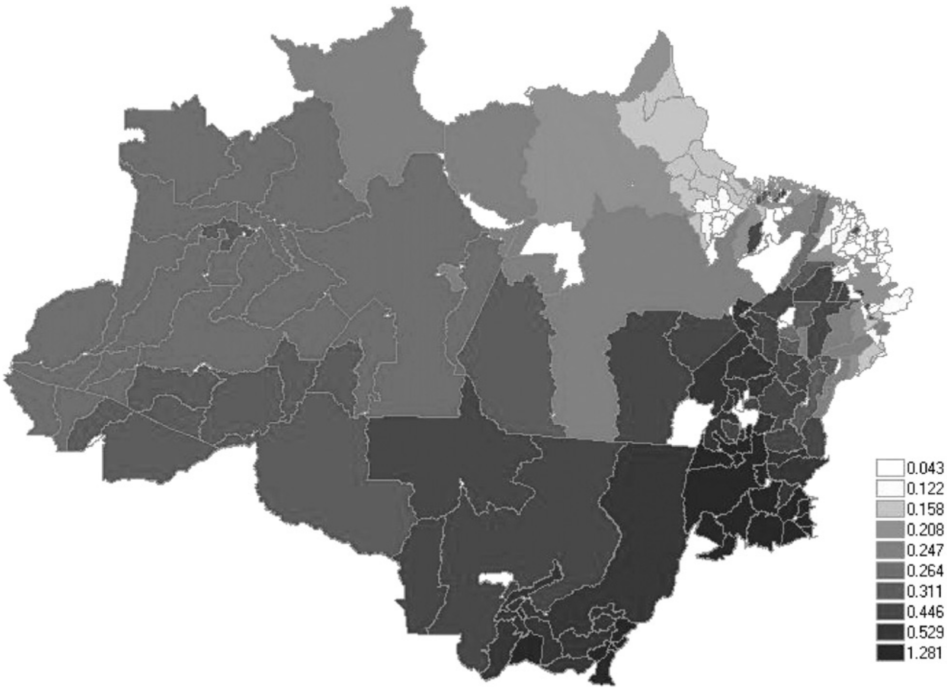


Fig. 4. Spatial summary of the estimation of the coefficient of lagged crop land for the real application.

5. Extensions

There are many interesting extensions to the basic model described above. In this section, we consider some of them, namely, mixed SVRM, other forms of prior specification for β and special forms for the hyperparameter Ψ.

5.1. Mixed SVRM

In their pure form, SVRM have all their regression coefficients subject to a spatial structure. This may be unrealistic and in practice there may be effects that do not vary with space. In general, these effects can be influenced from a variety of sources with an unstructured hierarchical form or a temporal element to them. Attention here is restricted to spatial effects and static effects.

Therefore, the mean responses of the observations are now given by $\mu_i = z_i'\mu + x_i'\beta_i$. The β_i's are still related by the spatial structure (3) but the static regression coefficient μ is not. There are no restrictions on the variables that enter the vectors of the explanatory variables z_i and x_i. Identifiability conditions require that, whenever a covariate enters both z and x, then the associated component of the β_i's must have a fixed sum. Usually, one imposes $\sum_{i=1}^n \beta_{ij} = 0$ so that the corresponding component of μ can be interpreted as an overall basic effect.

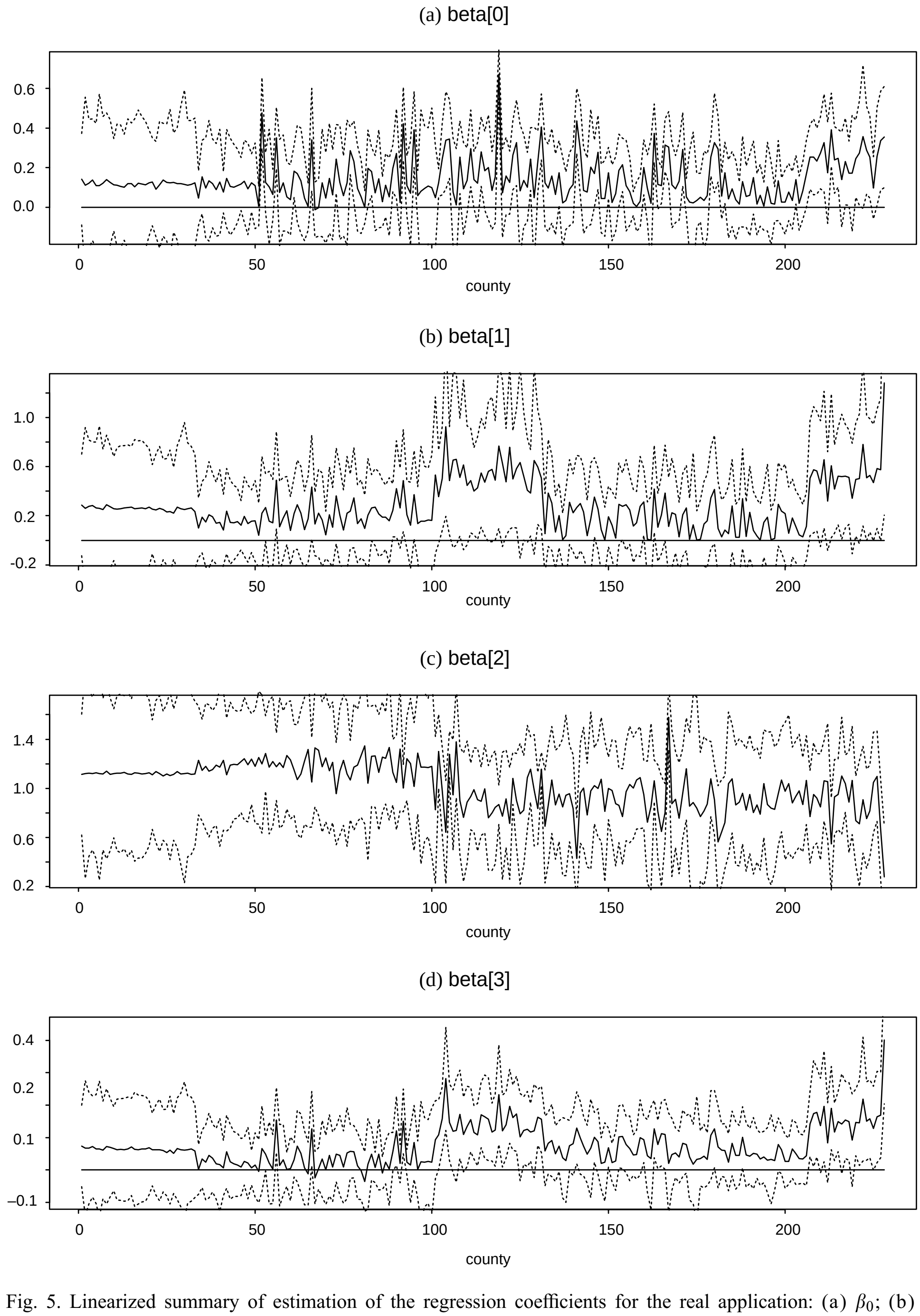


Fig. 5. Linearized summary of estimation of the regression coefficients for the real application: (a) β_0; (b) β_1; (c) β_2; (d) β_3. Posterior means are represented by full lines and 2 posterior standard deviation limits by dashed lines.

Typically, but not necessarily, the model is completed with an independent prior $\mu|\phi \sim N(m, \phi^{-1}M)$ for some vector m and positive definite matrix M. MCMC-based inference proceeds in straightforward fashion. Details may be found in Gamerman et al. (2001).

5.2. Proper prior specifications

One of the theoretical disadvantages of (3) is that it is improper. This stems from the fact that the prior is only informative about distances between β_i's and not about the β_i's themselves. One simple way to correct this is to pin down the β_i's to some point in their space of variation. A simple extension is provided by the prior

$$f(\beta|\phi, \Psi, \lambda) \propto \phi^{nr/2} |\Psi|^{n/2} \times \exp\left\{-\frac{\phi}{2}\left[Q(\beta) + \lambda \sum_{i=1}^{n} (\beta_i - b_i)' \Psi (\beta_i - b_i)\right]\right\}, \tag{12}$$

where $Q(\beta)$ is as in (3) and the additional parameter λ controls the relative weight attached to the 'pinning down' part of the prior. The proportionality constant is now finite and can be obtained analytically. Useful choices for the b_i's are 0 or the MLE obtained in the static (or some local) regression. A special case of this is used as a prior for spatial effects by Fernandez and Green (2000).

This prior is equivalent to $(\beta|\phi, \Psi, \lambda) \sim N[C^{-1}(\lambda I \otimes \Psi)\mu, C]$ where $\mu = (\mu_1', \ldots, \mu_n')'$ and $C^{-1} = (\lambda I + W) \otimes \Psi$. As a consequence, the full prior conditional of β_i is

$$\beta_i|\beta_{-i}, \phi, \Psi, \lambda \sim \beta_i|\beta_{\partial i}, \phi, \Psi, \lambda \sim N\left(w_i \bar{\beta}_{\partial i} + (1 - w_i)\mu_i, \frac{1}{\phi(w_{i+} + \lambda)} \Psi^{-1}\right),$$

where $w_i = w_{i+}/(w_{i+} + \lambda)$, $i = 1, \ldots, n$

5.3. Special forms for the hyperparameter

One special case considers Ψ a diagonal form with diagonal entries $\psi_1, \ldots, \psi_r$. Then, $Q(\beta) = \sum_{j=1}^{r} Q_j(\beta^j)$, where

$$Q_j(\beta^j) = \sum_{i,l=1}^{n} w_{ij} \psi_j (\beta_{ij} - \beta_{lj})^2 = \psi_j \beta^{j\prime} W \beta^j. \tag{13}$$

Sampling schemes for β remain unchanged, although sampling the β^j's may have beneficial mixing properties due to lack of prior correlation. Assumption (13) ensures that their joint full conditional distributions consist of a product of independent densities.

Similar comments apply when the weights w_{ij}'s in the expression of (3) depend on further hyperparameters. An example is given by continuous, geostatistical methods where the weights typically depend on the distance between the regions (i.e., $w_{ij} = \exp(-bd_{ij})$, where d_{ij} is the distance between sites i and j). A prior for the distance attenuation parameter b must be specified and a sampling step for b must be included in the MCMC scheme. Introduction of a further unknown quantity only implies an extra step in each MCMC iteration, the steps associated with the other parameters remaining the same.

6. Concluding remarks

We have considered analyses of different sampling schemes with a spatial regression model. Our simulation results must be interpreted with care since they are based on empirical evidence. Given that we have emulated a few different prior forms we are reasonably confident the results hold for a wider class of spatial processes. The analyses are conducted with software freely available at http://www.ipea.gov.br. This software provides a useful tool for analyses of regression models with the presence of a spatial pattern influencing the effect of the explanatory variables.

The empirical results confirm the general rule anticipated that C > B > A. This is not entirely surprising but the difference in the performance of the schemes is not as sharp as that obtained for state space models by Fruhwirth-Schnatter (1994). Simulation results also show very good estimation of state parameters and a poor result in terms of the hyperparameters. This is probably due to lack of information for these parameters from the likelihood and show the need for informative priors.

All simulations and models considered here assume normal observations. An obvious next step is to consider these models under different observational sampling schemes. There are many instances where regional observations arise in the form of counts or proportions. Unfortunately, scheme C is no longer available under such circumstances and schemes A and B must be adapted to incorporate the Metropolis acceptance steps. This poses another problem of practical relevance. But the main issue in spatial models is clearly the dependence between model parameters and this remains equally important irrespective of the observational model assumed.

Acknowledgements

The authors thank the Special Editor David Belsley for many suggestions and the referees for many comments and criticisms that led to an improved presentation of the paper.

References

Andersen, L., Granger, C., Reis, E., 1997. A random coefficient VAR transition model of the changes in land use in the Brazilian Amazon. Brazilian J. Econometrics 17, 1–13.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht.

Assunção, R.M, Assunção, J.J., Lemos, M., 1998. Induced technical change: a Bayesian spatial varying parameter model, Proceedings of the XX Meeting of the Brazilian Econometric Society.

Assunção, J.J., Gamerman, D., Assunção, R.M., 1999. Regional differences in factor productivities of Brazilian agriculture: a space varying parameter approach. Proceedings of the XV Latin American Meeting of the Econometric Society.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M., 1995. Bayesian analysis of space–time variation in disease risk. Statist. Med. 14, 2433–2443.

Besag, J., 1975. Statistical analysis of non-lattice data. Statistician 24, 179–195.

Besag, J., Higdon, D., 1999. Bayesian analysis of agricultural field experiments (with discussion). J. Roy. Statist. Soc. B 61, 691–746.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math. 43, 1–59.

Cressie, N., 1991. Statistics for Spatial Data. Wiley, New York.

Fernandez, C., Green, P.J., 2000. Modelling spatially correlated data via mixtures: a Bayesian approach. Mimeo. Research Report, University of St. Andrews.

Fruhwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. J. Time Ser. Anal. 15, 183–202.

Gamerman, D., 1997. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman & Hall, London.

Gamerman, D., Moreira, A.R.B., 2002. Bayesian analysis of econometric time series models using hybrid integration rules. Comm. Statist. Theory Meth. 31, 49–72.

Gamerman, D., Moreira, A.R.B., Rue, H., 2001. Space-varying regression models: specifications and simulation. Mimeo. Technical Report, Universidade Federal do Rio de Janeiro.

Gelfand, A., Ghosh, S., 1998. Model choice: a minimum posterior predictive loss approach. Biometrika 85, 1–11.

Gelfand, A.E., Vounatsou, P., 2001. Proper multivariate conditional autoregressive models for spatial data analysis, Mimeo. Research Report 301-15, University of Connecticut.

Gelman, A., Rubin, D.R., 1992. Inference from iterative simulation using multiple sequences (with discussion). Statist. Sci. 7, 457–511.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In: Bernardo, J.M., et al. (Eds.), Bayesian Statistics, Vol. 4. Oxford University Press, Oxford, pp. 169–193.

Kim, H., Sun, D., Tsutakawa, R.K., 2001. A bivariate Bayes method for improving mortality rates with 2-fold CAR model. J. Amer. Statist. Assoc. 96, 1506–1521.

Knorr-Held, L., Rue, H., 2002. On block updating in Markov random field models for disease mapping, Scand. J. Statist., to appear.

Lindley, D.V., Smith, A.F.M., 1972. Bayes estimates for the linear model (with discussion). J. Roy. Statist. Soc. B 34, 1–41.

Mardia, K.V., 1988. Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. J. Multivariate Anal. 24, 265–284.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machine. J. Chem. Phys. 21, 1087–1091.

Moreira, A.R.B., Migon, H.S., 1999. Spatial heterogeneity of the agricultural productivity: Brazil—1970–1996. Proceedings of the XXI Brazilian Meeting on Econometrics, Brazilian Econometrics Society, Rio de Janeiro.

Qian, W., Titterington, D.M., 1991. Estimation of parameters in hidden Markov models. Philos. Trans. Roy. Soc. London A 337, 407–428.

Rubin, D.B., 1988. Using the SIR algorithm to simulate posterior distributions (with discussion). In: Bernardo, J.M., et al. (Eds.), Bayesian Statistics, Vol. 3. Oxford University Press, Oxford, pp. 395–402.

Rue, H., 2001. Fast sampling of Gaussian Markov random fields. J. Roy. Statist. Soc. B 63, 325–338.

Schmidt, A.M., Gamerman, D., Moreira, A.R.B., 1999. An adaptive resampling scheme for cycle estimation. J. Appl. Statist. 26, 619–641.

Shephard, N., 1994. Partial non-Gaussian state space. Biometrika 81, 115–131.

West, M., Harrison, P.J., 1997. Bayesian Forecasting and Dynamic Models, 2nd Edition. Springer, New York.