Journal of Multivariate Analysis 109 (2012) 88–102

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Bayesian spatial models with a mixture neighborhood structure

E.C. Rodrigues *, R. Assunção

Universidade Federal de Minas Gerais 31270-901 Belo Horizonte, MG, Brazil

ARTICLE INFO

Article history:
Received 9 May 2011
Available online 15 March 2012

AMS subject classification codes:
62H11
62M40
62H20
62J12

Keywords:
Disease mapping
Markov random field
Spatial hierarchical models

ABSTRACT

In Bayesian disease mapping, one needs to specify a neighborhood structure to make inference about the underlying geographical relative risks. We propose a model in which the neighborhood structure is part of the parameter space. We retain the Markov property of the typical Bayesian spatial models: given the neighborhood graph, disease rates follow a conditional autoregressive model. However, the neighborhood graph itself is a parameter that also needs to be estimated. We investigate the theoretical properties of our model. In particular, we investigate carefully the prior and posterior covariance matrix induced by this random neighborhood structure, providing interpretation for each element of these matrices.

1. Introduction

In disease mapping, the Bayesian model proposed by Besag et al. [6], and denoted by BYM, is the most popular choice to estimate relative risks in small areas or to evaluate the effects of covariates acting as exposure measurement surrogates. Originally, BYM was introduced to model a cross-section of counts collected in a set of disjoint geographical areas composing a partitioned map. Since then, BYM has been extended into several directions to include space–time generalized linear models [26,28,20,34,30], spatial survival models [9,19], spatially-varying parameters models [3,1,12], and generalized additive models [22]. Multivariate extensions incorporating two correlated sets of spatial effects have also been proposed in recent years [19,13,16,15]. Many of these models can be fit using freely available software such as WinBUGS [25] and BayesX [8].

BYM is based on a conditional autoregressive (CAR) model for the spatial random effects. In the CAR model, spatial dependence is expressed conditionally by requiring that the random effect in a given area, given the values in all other areas, depends only on a small set of neighboring values. More specifically, the random effect b_i associated with the i-th area is the sum $\phi_i + \theta_i$ of two components, where ϕ_i is a spatially structured random effect to which we assigned an improper CAR prior distribution and θ_i is a second set of i.i.d. zero-mean normally distributed unstructured random effects. This is termed a convolution prior [6] because the density of b_i's will be the convolution of the joint densities of the ϕ_i and θ_i vectors.

An essential aspect of the BYM model and its extensions is the specification of the neighborhood structure for the areas. Although this is quite flexible and can be arbitrarily defined, in practice it is typically based only on adjacency relationships. There are few justifications for this practice other than its easy calculation by means of GIS (Geographic information system) routines. A related problem with the BYM model is that the neighborhood structure determines the smoothing degree used in relative risk estimation. Some authors noticed its tendency to oversmooth the risks when the usual adjacency neighborhood structure is used. Therefore, it would be very useful to have a model that allows for multiple neighborhood structure and automatically adapts itself according to the observed data.

* Corresponding author.
E-mail addresses: ericaa_casti@yahoo.com.br (E.C. Rodrigues), assuncao@dcc.ufmg.br (R. Assunção).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.02.017

Despite its crucial role in spatial Bayesian models, very few studies have considered different neighborhood structures for disease mapping problems. One notable exception is MacNab and Dean [26] where the authors considered a model for disease rates with spatial effects structured at two geographical levels. They used infant mortality data over the period 1985–1994 from the province of British Columbia (BC) in Canada. The areas were organized in 21 health units (HUs) that were further subdivided into 79 local health areas (LHAs). Health units (HUs) are administrative health divisions overseeing the functioning of the health sub-units, the local health areas (LHAs). Therefore, it was natural to expect that LHAs within the same HU should share many health service and care characteristics beyond those determined by factors that vary smoothly in space. Hence, they assumed a random effect shared by all LHAs within the same HU. They also considered a neighborhood structure in which two LHAs are considered neighbors if they share boundaries or if there is a third LHA sharing boundaries with both local health areas. This second-order neighborhood structure is less common and it recalls the higher autoregressive order models in the time series setting.

A more recent reference is White and Ghosh [36], who introduced a stochastic neighborhood CAR model where the neighborhood selection depends on unknown parameters. They estimate neighborhood sizes by assuming that there is an unknown cutoff distance. Within this distance proximity weights are equal and sum to one, and beyond it they decline exponentially with distance, reaching zero at the edge of the map. In contrast with most of published applied papers in disease mapping, they base their model on the proper CAR specification rather than BYM. Most people prefer to use BYM, implying in an improper CAR model to deal with the spatial random effects, because the proper CAR model induces little marginal correlation between neighboring areas (see Banergee et al. [4, p. 81] and Assunção and Krainski [2]).

These studies consider only locally larger neighborhoods than the first order neighborhood implied by using simple adjacency. Although in some situations a local neighborhood will be enough to deal with the spatial effects, we feel that spatial models should span a larger range of possibilities. Fundamentally, BYM and its variants consider random effects composed of either unstructured overdispersion or small-scale spatial conditional variation. These are two extreme models and allowing for intermediate situations will be useful in some applications. We will show examples where the typical adjacency neighborhood structure is not sufficient to estimate the underlying risks, providing less smooth estimates than what should be inferred from the data. Our purpose is to introduce spatial effects with that extend beyond the immediate geographical neighborhood. This is likely to be especially useful in situations where the underlying risk changes so smoothly over larger regions as to be considered indistinguishable from a random constant value for all areas within it.

In this work, we investigate more flexible neighborhood structures for spatial conditional autoregressive models. We propose a model in which the neighborhood structure is part of the parameter space. We retain the Markov properties of most Bayesian spatial models. That is, the disease rates follow a conditional autoregressive model, given the neighborhood graph. However, the neighborhood graph itself is a parameter that also needs to be estimated. The methodology described herein permits arbitrary neighborhood extension for incorporating spatial random effects. It provides a simple mechanism for identifying the geographical extent of the conditional influence of neighboring areas.

The manuscript is organized as follows. In Section 2, we introduce the notation and present some models that were proposed previously. In Section 3, we present the definition our model. In Section 4, we investigate the theoretical properties of the model. In particular, we carefully study the prior and posterior covariance matrix induced by this random neighborhood structure, providing interpretation for each element of these matrices. We also present a specific, simple case of our model, allowing for a more thorough understanding of the covariance structure. In Section 6, we illustrate the use of our model for disease mapping. In this section, we also present a simulation study to compare our method with alternative proposals. We end in Section 8 with the main conclusions.

2. Disease mapping

A Bayesian hierarchical model is one of the main tools for making inferences about the underlying relative risks of a disease observed on disjoint geographical areas of a map. Suppose that we have N geographic areas and each has a relative risk ψ_i for $i = 1, \ldots, N$ that needs to be estimated. Bayesian inference is based on the posterior distribution of $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_N)$ given by $f(\boldsymbol{\psi}|y_1, \ldots, y_N) \propto l(y_1, \ldots, y_N|\boldsymbol{\psi})f(\boldsymbol{\psi})$, where $l(y_1, \ldots, y_N|\boldsymbol{\psi})$ is the likelihood function and $f(\boldsymbol{\psi})$ is the prior distribution of the parameter vector $\boldsymbol{\psi}$. Conditional on the values $\psi_1, \ldots, \psi_N$, the values $Y_1, \ldots, Y_N$ are assumed to be independent with a Poisson distribution with mean $\psi_i E_i$, where E_i is the expected value of cases under the hypotheses of constant relative risk over the areas. Modeling the prior distribution $f(\boldsymbol{\psi})$ allows the introduction of spatial dependence between the risks such that close regions tend to have similar relative risks. This dependence appears as a Markovian structure in which the value ψ_i of one area, conditional on all other areas' values, depends only upon the ψ_j values of its neighbors.

More specifically, the relative risk ψ_i is written as

$$\log(\psi_i) = \mu + b_i \tag{1}$$

where μ is the general level of the relative risk and b_i is the random effect for the i-th area. One simple possibility is to assume that the random effects b_i are independent and identically distributed with a normal distribution $N(0, \sigma^2)$. In this case, there will be no spatial effects imposed on the relative risks and the posterior distribution of ψ will reflect this independence. However, one typically expects spatial dependence between the relative risks due to environmental and genetic similarities between neighboring areas. The most popular prior distribution for modeling spatial structure was introduced by Besag

et al. [6]. They decomposed the random effect b_i into two parts, a non-spatially structured component and a spatially structured component:

$$\log(\psi_i) = \mu + \theta_i + \phi_i$$

where $\theta_1, \ldots, \theta_n$ are the non-structured errors, independently and identically distributed according to a normal distribution. The random effects ϕ_i have a spatially structured prior distribution with intrinsic CAR (ICAR) distribution. The ICAR prior distribution is an improper prior with a Markovian structure. The distribution of ϕ_i, conditional on all the other values ϕ_j for $j \neq i$, is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim N\left(\bar{\phi}_i, \frac{\sigma^2}{n_i}\right) \tag{2}$$

where $\bar{\phi}_i$ is the mean of the i-th area neighboring values ϕ_j.

This model presents some identifiability problems for the spatial and non-spatial effects, as noticed by Eberly and Carlin [11]. To fix these problems, Leroux et al. [24] presented an alternative, including a parameter λ which is able to measure the effect of each component. This parameter measures the level of spatial correlation among the areas. In addition to this, it includes a parameter σ^2 to measure the random effect variance. They proposed a multivariate normal distribution for the random effects $\mathbf{b} = (b_1, \ldots, b_N)$ in (1) with the following precision matrix

$$\mathbf{Q} = (\sigma^2)^{-1}\left((1-\lambda)\mathbf{I} + \lambda\mathbf{R}\right) \tag{3}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{R}$ is the precision matrix of the ICAR model, which means that $\mathbf{R}_{ij}$ is equal to n_i, -1, and 0, if $i = j$, $i \sim j$, and otherwise, respectively, where n_i is the number of neighbors of site i and $i \sim j$ means i neighbor of j. For this model, the parameter λ assumes values in the interval [0, 1], so that, the precision matrix $\mathbf{Q}$ is a weighted sum of the $\mathbf{I}$ and $\mathbf{R}$ matrices.

The BYM and Leroux models represent a mixing of two extreme situations. One situation considers a conditional dependence only on the immediate neighbors represented by the single neighborhood structure while the other situation represents the complete independence between the random effects. Both models assume that, if we have information on the immediate neighbors, no additional information about the other areas is necessary to make inference about the random effects. We think that in many practical situations this is too restrictive. Consider, for example, another extreme but possible situation in which the distribution of b_i (and hence, of ψ_i) in a given area, conditional on the rest of the map, should depend upon all the other sites, not only on the immediate neighbors. In this case, all areas are neighboring areas of all other areas. This can be a reasonable model when the region under study is small enough such that economic, social and environmental characteristics are approximately constant over the entire region. This implies on exchangeability between the areas and therefore an all-inclusive dependence between the areas' pairs. Every area gives incremental additional information on a fixed area value, even if conditioning on all the other areas.

3. Model definition

We propose a model that expands the BYM and Leroux models beyond single-neighbor dependence of BYM and Leroux models to a larger class that has geographically increasing orders of neighborhood extension. Through Bayesian updating, we can make inference about the more appropriate neighborhood structure underlying the observed data. More specifically, we extend the weighted sum precision matrix (3) by including matrices that represent neighborhoods of all possible orders in the simple adjacency graph.

Let each area i be a node or site of a graph and connect two nodes by one edge if they share boundaries. Let $\mathbf{A}$ be the $n \times n$ binary adjacency matrix where $\mathbf{A}_{ij} = 1$ if i and j are connected by one edge, and $\mathbf{A}_{ij} = 0$ otherwise. We say that area i is an l-th order neighbor of area j if the (i, j)-th element of the power matrix $\mathbf{A}^l$ is greater than zero and $\mathbf{A}^s_{ij} = 0$, for $s < l$ and $l \geq 1$. The maximum neighborhood order is given by the diameter of the graph, which is the longest path among all the shortest paths that connect two sites. In other words, the diameter counts the minimum number of steps necessary to leave a site and go to any other site in the graph.

In our model, the vector $\mathbf{b} = (b_1, \ldots, b_N)$ in (1) has a multivariate normal distribution with mean zero and precision matrix given by

$$\mathbf{Q} = (\sigma^2)^{-1}\left(\lambda_1\mathbf{I} + \lambda_2\mathbf{R}^{(1)} + \lambda_3\mathbf{R}^{(2)} + \cdots + \lambda_{k+1}\mathbf{R}^{(k)}\right)$$

where $\lambda_1 + \lambda_2 + \cdots + \lambda_{k+1} = 1$ and $\lambda_i \geq 0$ for all i. The integer k is the diameter of the graph and $\mathbf{R}^{(l)}$ is the graph Laplacian that includes neighborhoods up to order l. That is,

$$\mathbf{R}^{(l)}_{ij} = \begin{cases} n_i^{(l)} & \text{if } i = j \\ -1 & \text{if } j \in \partial_i^{(l)} \\ 0 & \text{otherwise} \end{cases}$$

where $n_i^{(l)}$ is the number of neighbors of site i up to order l and $\partial_i^{(l)}$ is the set of neighbors of area i, from order 1 up to order l. Notice that, we are considering that the neighborhood relationship is symmetric, that is, $j \in \partial_i^{(l)}$ if and only if, $i \in \partial_j^{(l)}$. These matrices are linearly independent, ensuring the parameter's identifiability.

This matrix is positive definite if $\lambda_1 > 0$, as it satisfies the sufficient condition of being diagonally dominant. That is, for all $i = 1, \ldots, n$, we have $\mathbf{Q}_{ii} > \sum_{j=1}^{N} |\mathbf{Q}_{ij}|$ because

$$\mathbf{Q}_{ii} = \lambda_1 + \lambda_2 n_i^{(2)} + \lambda_3 n_i^{(3)} + \cdots + \lambda_{k+1} n_i^{(k)} = \lambda_1 + \sum_{j=1}^{N} |\mathbf{Q}_{ij}| > \sum_{j=1}^{N} |\mathbf{Q}_{ij}|.$$

From the precision matrix, it is possible to obtain the conditional distribution $b_i|\mathbf{b}_{-i}$ of each area given the vector $\mathbf{b}_{-i} = (b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n)$. It is a normal distribution with mean $f(\mathbf{b}, \lambda)$ and variance $g(\mathbf{b}, \lambda)$ given by

$$f(\mathbf{b}, \lambda) = \frac{\lambda_2 n_i^{(1)} \bar{b}_i^{(1)} + \lambda_3 n_i^{(2)} \bar{b}_i^{(2)} + \cdots + \lambda_{k+1} n_i^{(k)} \bar{b}_i^{(k)}}{\lambda_1 + \lambda_2 n_i^{(1)} + \lambda_3 n_i^{(2)} + \cdots + \lambda_{k+1} n_i^{(k)}}$$

and

$$g(\mathbf{b}, \lambda) = \frac{\sigma^2}{\lambda_1 + \lambda_2 n_i^{(1)} + \lambda_3 n_i^{(2)} + \cdots + \lambda_{k+1} n_i^{(k)}}$$

where $\bar{b}_i^{(l)}$ is the mean of neighbors of site i up to order l. The conditional expectation is a convex linear combination of the means of its neighbors of all possible orders and the conditional variance is inversely proportional to the number of neighbors of each of these orders multiplied by their respective weight λ_l.

Let $\mathbf{b}_{-ij}$ be the $(n-2)$-dimensional vector obtained by omitting the i-th and j-th coordinates from $\mathbf{b}$. It can be shown that the conditional correlation $\mathrm{Corr}(b_i, b_j|\mathbf{b}_{-ij})$ is given by

$$\mathrm{Corr}(b_i, b_j|\mathbf{b}_{-ij}) \propto \begin{cases} \lambda_2 + \lambda_3 + \cdots + \lambda_k & \text{if } j \in \partial_i^{(1)} \\ \lambda_3 + \cdots + \lambda_k & \text{if } j \in \partial_i^{(2)} - \partial_i^{(1)} \\ \cdot & \\ \cdot & \\ \lambda_k & \text{if } j \in \partial_i^{(k)} - \bigcup_{l=1}^{k-1} \partial_i^{(l)} \end{cases}$$

with the proportionality constant given by the inverse of the square root of

$$\sum_{l=1}^{k} \lambda_l n_i^{(l-1)} \sum_{l=1}^{k} \lambda_l n_j^{(l-1)}$$

and with $n_i^{(0)} \equiv 1$ by definition, for all $i = 1, \ldots, N$. This shows that the conditional correlation between the areas decreases with the neighborhood order l. For example, if a pair of sites are third order neighbors, the conditional correlation between them will be smaller than that between two first order neighbors. Notice also that, if all the λ_l are positive, then the conditional correlation between any pair of areas is different from zero.

We can also write the joint distribution in a more interpretable way:

$$\begin{aligned} f(\mathbf{b}) \propto \exp\Bigg\{ -\frac{1}{2\sigma^2} \Bigg[& \sum_i b_i^2 (\lambda_1 + \cdots + \lambda_{k+1} n_i^{(k)}) - \lambda_2 \sum_i \sum_{j:j\in\partial_i^{(1)}} b_i b_j \\ & - \lambda_3 \sum_i \sum_{j:j\in\partial_i^{(2)}} b_i b_j - \cdots - \lambda_{k+1} \sum_i \sum_{j:j\in\partial_i^{(k)}} b_i b_j \Bigg] \Bigg\} \\ = \exp\Bigg\{ -\frac{1}{2\sigma^2} \Bigg[& \sum_i \Bigg(\lambda_1 b_i^2 + \frac{\lambda_2}{2} \sum_{j:j\in\partial_i^{(1)}} (b_i - b_j)^2 + \cdots + \frac{\lambda_{k+1}}{2} \sum_{j:j\in\partial_i^{(k)}} (b_i - b_j)^2 \Bigg) \Bigg] \Bigg\}. \end{aligned}$$

If $\lambda_l = 0$ for all $l > 1$, we are in the case of independent normal distributions. We can interpret the term associated with λ_l as a penalization for configurations showing too much variation among l-th order neighbors. The larger the value of λ_l, the

smoother is the spatial pattern up to neighborhood order l. This distribution can also be written as

$$f(\mathbf{b}) \propto \left(\exp\left\{-\frac{1}{2\sigma^2}\sum_i b_i^2\right\}\right)^{\lambda_1} \prod_{j=2}^{k}\left(\exp\left\{-\frac{1}{4\sigma^2}\sum_i\sum_{j:j\in\partial_i^{(l)}}(b_i-b_j)^2\right\}\right)^{\lambda_j},$$

which is a geometric mixture of normal distributions.

To complete the model specification, one needs to adopt prior distributions for the weights $(\lambda_1,\ldots,\lambda_k)$ and for the hyperparameter σ^2. In our applications, we assumed an inverse Gamma prior distribution for σ^2 and a uniform distribution on the k-dimensional simplex with the restriction that the $\lambda_l > 0$ and that they add to 1. A more general possibility is to adopt a Dirichlet distribution in this simplex.

To represent the k-th order neighborhood, our model uses the cumulative neighboring areas up to order k. As a referee suggested, an alternative way to define our model is to use only the neighbors that are exactly at k steps away from each area. That is, consider the following precision matrix:

$$\mathbf{Q}' = \frac{1}{\sigma^2}\left(\lambda_1^*\mathbf{I} + \lambda_2^*\mathbf{W}^{(1)} + \lambda_3^*\mathbf{W}^{(2)} + \cdots + \lambda_{k+1}^*\mathbf{W}^{(k)}\right). \tag{4}$$

In this formulation, the neighborhood matrix has the following definition

$$\mathbf{W}_{ij}^{(l)} = \begin{cases} (n^*)_i^{(l)}, & \text{if } i=j \\ -1, & \text{if } j\in(\partial^*)_i^{(l)} \\ 0, & \text{otherwise} \end{cases}$$

where $(n^*)_i^{(l)}$ is the number of neighbors of site i of order l and $(\partial^*)_i^{(l)}$ is the set of neighbors of area i of order l. We need to add the restriction $\lambda_1^* > \lambda_2^* > \cdots > \lambda_{k+1}^*$ to guarantee that the partial correlations decrease with the neighborhood order. This condition implies that there exists non-negative $\lambda_2,\ldots,\lambda_{k+1}$ such that, for $j=2,\ldots,k+1$, we have $\lambda_j^* = \lambda_j + \cdots + \lambda_{(k+1)}$. Substituting these values in the $\mathbf{Q}'$ precision matrix, we have

$$\begin{aligned}\mathbf{Q}' &= \left(\lambda_1^*\mathbf{I} + \lambda_2^*\mathbf{W}^{(1)} + \lambda_3^*\mathbf{W}^{(2)} + \cdots + \lambda_{k+1}^*\mathbf{W}^{(k)}\right) \\ &= \left(\lambda_1^*\mathbf{I} + (\lambda_2+\cdots+\lambda_{(k+1)})\mathbf{W}^{(1)} + (\lambda_3+\cdots+\lambda_{k+1})\mathbf{W}^{(2)} + \cdots + \lambda_{k+1}\mathbf{W}^{(k)}\right) \\ &= \lambda_1^*\mathbf{I} + \lambda_2\mathbf{W}^{(1)} + \lambda_3\left(\mathbf{W}^{(1)}+\mathbf{W}^{(2)}\right) + \cdots + \lambda_{k+1}(\mathbf{W}^{(1)}+\mathbf{W}^{(2)}+\cdots+\mathbf{W}^{(k)}).\end{aligned}$$

Therefore, the two models would be equivalent only if

$$\mathbf{W}^{(1)} + \mathbf{W}^{(2)} + \cdots + \mathbf{W}^{(l)} = \mathbf{R}^{(j)} \quad \text{for } l = 1, 2, \ldots, k.$$

But this is not true for $l \geq 2$. To see this, consider the simplest case, with $l=2$. We have that

$$\left[\mathbf{W}^{(1)} + \mathbf{W}^{(2)}\right]_{ij} = \begin{cases} (n^*)_i^{(1)} + (n^*)_i^{(2)}, & \text{if } i=j \\ -1, & \text{if } j\in(\partial^*)_i^{(1)} \\ -2, & \text{if } j\in(\partial^*)_i^{(2)} \\ 0, & \text{otherwise,} \end{cases}$$

which is different from $\mathbf{R}_{ij}^{(2)}$, defined previously. We will see next that our definition allows us to derive several important properties that help to understand the model. Such developments would not be possible if we had defined the precision matrix as in (4).

4. Model properties

To gain a better understanding of the prior and posterior distribution properties, we obtain its marginal covariance matrix in addition to the conditional correlation given earlier. To avoid a cumbersome notation and long formulas, we will consider the model that includes three different values for λ_l, one corresponding to λ_1 (associated with the individual areas and the independent case), another corresponding to λ_2 (associated with pairs of adjacent areas), and the third one, λ_3, corresponding to the highest possible order k, associated with a complete graph, where every area is a neighbor of every other area. The extension to the general case is straightforward.

Considering only three components, our precision matrix reduces to

$$\mathbf{Q} = (\sigma^2)^{-1}\left(\lambda_1\mathbf{I} + \lambda_2\mathbf{R}^{(1)} + \lambda_3\mathbf{R}^{(k)}\right) \tag{5}$$

where $\mathbf{R}^{(1)}$ is the precision matrix of the ICAR model and $\mathbf{R}^{(k)} = \text{diag}(\mathbf{N}) - \mathbf{1}\mathbf{1}^T$, with $\mathbf{N} = N\mathbf{1}$ and $\mathbf{1} = (1,\ldots,1)$. The precision matrix in (5) can be rewritten as

$$\mathbf{Q} = (\sigma^2)^{-1}\left(\lambda_1\mathbf{I} + \lambda_2\text{diag}(\mathbf{n}) + \lambda_3\text{diag}(\mathbf{N}) - \lambda_2\mathbf{A} - \lambda_3\mathbf{1}\mathbf{1}^T\right)$$

where $\mathbf{A}$ is the binary adjacency matrix and $\mathbf{A}\mathbf{1} = \mathbf{n} = (n_1,\ldots,n_N)$ is the vector which has the number of adjacent neighbors of each area. The following theorem shows what is the inverse of this precision matrix.

Theorem 1. *The inverse of the precision matrix* $\mathbf{Q}$ *is given by*

$$\mathbf{Q}^{-1} = \sigma^2\mathbf{M}^{-1} + \frac{\sigma^2\lambda_3}{1-\lambda_3\sum_{ij} m_{ij}}\left[S_{1+}\, S_{2+}\cdots S_{N+}\right]^T\left[S_{1+}\, S_{2+}\ \cdots\ S_{N+}\right] \tag{6}$$

where $S_{l+} = \sum_j m_{lj} = \sum_i m_{il}$ *and* $\mathbf{M} = \lambda_1\mathbf{I} + \lambda_2\,\mathrm{diag}(\mathbf{n}) + \lambda_3\,\mathrm{diag}(\mathbf{N}) - \lambda_2\mathbf{A}$.

Proof. From matrix algebra, we know that

$$\left(\mathbf{P}+\mathbf{u}\mathbf{v}^T\right)^{-1} = \mathbf{P}^{-1} - \frac{\mathbf{P}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{P}^{-1}}{1+\mathbf{v}^T\mathbf{P}^{-1}\mathbf{u}}, \tag{7}$$

if $\mathbf{P}$ is an invertible matrix and $\mathbf{u}$ and $\mathbf{v}$ are vectors with the same dimension. Let $\mathbf{M} = \lambda_1\mathbf{I} + \lambda_2\mathrm{diag}(\mathbf{n}) + \lambda_3\mathrm{diag}(\mathbf{N}) - \lambda_2\mathbf{A}$ and denote by m_{ij} the ij-th element of $\mathbf{M}^{-1}$. Using result (7), we have that the covariance matrix $\mathbf{Q}^{-1}$ is given by

$$\begin{aligned}\mathbf{Q}^{-1} &= \sigma^2\left(\mathbf{M}^{-1} + \lambda_3\frac{\mathbf{M}^{-1}\mathbf{1}\mathbf{1}^T\mathbf{M}^{-1}}{1-\lambda_3\mathbf{1}\mathbf{M}^{-1}\mathbf{1}^T}\right)\\ &= \sigma^2\mathbf{M}^{-1} + \frac{\sigma^2\lambda_3}{1-\lambda_3\sum_{i,j} m_{ij}}\begin{bmatrix}\sum_j m_{1j}\sum_i m_{i1} & \cdots & \sum_j m_{1j}\sum_i m_{iN}\\ & \vdots & \\ \sum_j m_{Nj}\sum_i m_{i1} & \cdots & \sum_j m_{Nj}\sum_i m_{iN}\end{bmatrix}.\end{aligned}$$

As the matrix $\mathbf{M}$ is symmetric, $\mathbf{M}^{-1}$ is also symmetric and therefore, for all $l = 1,\ldots,N$, we have $\sum_j m_{lj} = \sum_i m_{il}$. Let $S_{l+} = \sum_j m_{lj} = \sum_i m_{il}$. We can write the covariance matrix as

$$\mathbf{Q}^{-1} = \sigma^2\mathbf{M}^{-1} + \frac{\sigma^2\lambda_3}{1-\lambda_3\sum_{ij} m_{ij}}\left[S_{1+}\, S_{2+}\cdots S_{N+}\right]^T\left[S_{1+}\, S_{2+}\ \cdots\ S_{N+}\right]. \quad \square \tag{8}$$

A better understanding of this covariance matrix structure can be obtained by initially considering the matrix $\mathbf{M}^{-1}$. Following the analytical approach adopted by Assunção and Krainski [2], we write

$$\begin{aligned}\mathbf{M}^{-1} &= \mathbf{M}^{-1}\left[\lambda_1\mathbf{I} + \lambda_2\mathrm{diag}(\mathbf{n}) + \lambda_3\mathrm{diag}(\mathbf{N})\right]\left[\lambda_1\mathbf{I} + \lambda_2\mathrm{diag}(\mathbf{n}) + \lambda_3\mathrm{diag}(\mathbf{N})\right]^{-1}\\ &= \left[\mathbf{I} - \lambda_2\mathbf{T}\mathbf{A}\right]^{-1}\mathbf{T}\end{aligned}$$

where

$$\mathbf{T} = \mathrm{diag}\left\{\frac{1}{\lambda_1+\lambda_2 n_1+\lambda_3 N},\ldots,\frac{1}{\lambda_1+\lambda_2 n_N+\lambda_3 N}\right\}.$$

Theorem 2. *The inverse matrix of* $\mathbf{I}-\lambda_2\mathbf{T}\mathbf{A}$ *can be written as*

$$\left[\mathbf{I}-\lambda_2\mathbf{T}\mathbf{A}\right]^{-1} = \left[\mathbf{I} + \lambda_2(\mathbf{T}\mathbf{A}) + \lambda_2^2(\mathbf{T}\mathbf{A})^2 + \lambda_2^3(\mathbf{T}\mathbf{A})^3 + \cdots\right]\mathbf{T}.$$

Proof. A well known linear algebra result [18, p. 45] states that, if $\mathbf{P}$ is a square matrix and each of the terms of the power matrix $\mathbf{P}^k$ tends to zero as k increases, then the inverse $(\mathbf{I}-\mathbf{P})^{-1}$ exists and it is given by $(\mathbf{I}-\mathbf{P})^{-1} = \mathbf{I}+\mathbf{P}+\mathbf{P}^2+\mathbf{P}^3+\cdots$. To use this result with the matrix $[\mathbf{I}-\lambda_2\mathbf{T}\mathbf{A}]^{-1}$, we need to show that the terms $\lambda_2^l[(\mathbf{T}\mathbf{A})^l]_{ij}$ of the power matrix approximate to zero when the power l increases. This will be done finding an upper bound. Consider initially $l=2$. We see that

$$\begin{aligned}\lambda_2^2\left[(\mathbf{T}\mathbf{A})^2\right]_{ij} &= \lambda_2^2\sum_{k=1}^{N}\frac{a_{ik}a_{kj}}{(\lambda_1+\lambda_2 n_i+\lambda_3 N)(\lambda_1+\lambda_2 n_k+\lambda_3 N)}\\ &= \lambda_2^2\sum_{k=1}^{N}\frac{a_{ik}a_{kj}/(n_i n_k)}{(\lambda_1/n_i+\lambda_2+\lambda_3 N/n_i)(\lambda_1/n_k+\lambda_2+\lambda_3 N/n_k)}\\ &< \frac{\lambda_2^2}{(\lambda_1/N+\lambda_2+\lambda_3)^2}\sum_{k=1}^{N}\sum_{k=1}^{N}\frac{a_{ik}}{n_i}\frac{a_{kj}}{n_k},\end{aligned}$$

since $n_i \leq N$. As $\text{diag}(1/\mathbf{n})\mathbf{A}$ is a stochastic matrix, it can be seen as a transition matrix of a random walk on the map with equal probabilities of jumping from a given area to any of its first-order neighbors. In this way, the second term in the multiplication is the probability that a random walk leaves site i and reaches site j in two steps and will be denoted by $p_{ij}^{(2)}$.

For an arbitrary $l \geq 2$, we have

$$\lambda_2^l \left[(\mathbf{TA})^l\right]_{ij} < \left(\frac{\lambda_2}{\lambda_1/N + \lambda_2 + \lambda_3}\right)^l p_{ij}^{(l)}$$

where $p_{ij}^{(l)}$ denotes the probability that the random walk goes from i to j in l steps. Therefore, $p_{ij}^{(l)} \in [0, 1]$ and since $\lambda_2/(\lambda_1/N + \lambda_2 + \lambda_3) < 1$, we have that

$$0 \leq \lim_{l\to\infty} \lambda_2^l \left[(\mathbf{TA})^l\right]_{ij} < \lim_{l\to\infty} \left(\frac{\lambda_2}{\lambda_1/N + \lambda_2 + \lambda_3}\right)^l p_{ij}^{(l)} = 0.$$

This shows that the terms of the matrix $\lambda_2^l \left[(\mathbf{TA})^l\right]$ tends to zero as l goes to infinity and the matrix expansion is valid. □

The elements $\left[(\mathbf{TA})^l\mathbf{T}\right]_{ij}$ of the l-th matrix in this expansion are weighted sums of all possible paths of length l starting at the i-th site and ending at the j-th site. For example, the three first matrices have elements equal to

$$[(\mathbf{TA})\mathbf{T}]_{ij} = \frac{a_{ij}}{(\lambda_1 + \lambda_2 n_i + \lambda_3 N)(\lambda_1 + \lambda_2 n_j + \lambda_3 N)}$$

$$\left[(\mathbf{TA})^2\mathbf{T}\right]_{ij} = \sum_{k=1}^{N} \frac{a_{ik}a_{kj}}{(\lambda_1 + \lambda_2 n_i + \lambda_3 N)(\lambda_1 + \lambda_2 n_k + \lambda_3 N)(\lambda_1 + \lambda_2 n_j + \lambda_3 N)}$$

$$\left[(\mathbf{TA})^3\mathbf{T}\right]_{ij} = \sum_{l=1}^{N}\sum_{k=1}^{N} \frac{a_{ik}a_{kl}a_{lj}}{(\lambda_1 + \lambda_2 n_i + \lambda_3 N)(\lambda_1 + \lambda_2 n_k + \lambda_3 N)(\lambda_1 + \lambda_2 n_l + \lambda_3 N)(\lambda_1 + \lambda_2 n_j + \lambda_3 N)}.$$

Considering the second matrix for illustration, the element $\left[(\mathbf{TA})^2\mathbf{T}\right]_{ij}$ counts all paths $i \to k \to j$ giving a weight inversely proportional to the number of immediate neighbors n_i, n_k, and n_j of these areas. Going from i to j through a highly connected area contributes less to $\mathbf{M}_{ij}^{-1}$ than if the path goes through a poorly connected intermediate area. This shows that two areas in a region of the map with highly connected areas will tend to be less correlated than two areas in a region where the areas have few immediate neighbors.

To complete the understanding of the covariance matrix $\mathbf{Q}^{-1}$ in (8), we consider now the value S_{i+}. We have

$$S_{i+} = \sum_{j=1}^{N} m_{ij} = \sum_{j=1}^{N}\sum_{k=0}^{\infty} \lambda_2^k \left[(\mathbf{TA})^k\mathbf{T}\right]_{ij} = \sum_{k=0}^{\infty} \lambda_2^k \sum_{j=1}^{N} \left[(\mathbf{TA})^k\mathbf{T}\right]_{ij}$$

where we interchange the order of the terms because the sum is absolutely convergent. This quantity is a weighted sum of all paths leaving site i, the weight decreasing with the path length k. Hence, it is inversely related to the average degree of connectivity that the area i has with the other areas in the graph. Note that S_{i+} is a value associated with the i-th area, and not with pairs of areas.

In summary, the covariance $\text{Cov}(b_i, b_j) = \left[\mathbf{Q}^{-1}\right]_{ij}$ is the sum of two components. The first one is $\left[\mathbf{M}^{-1}\right]_{ij}$ and represents a weighted sum of all paths from i to j with weights inversely related to their length and to the connectivity of the areas in the path. The second component is given by the product of $S_{i+}S_{+j}$ where S_{i+} is a score associated with the average connectivity of area i to all the other areas in the map. The first component is influenced by the neighborhood structure through a weighted counting of each path from i to j. The second component is also influenced by the neighborhood structure but it considers only a marginal structure. Its presence in the covariance matrix position (i, j) is by means of the product of these marginal values associated with the areas i and j.

We can write $S_{i+}S_{+j}$ in a different way in order to see how they reflect the structure of a complete graph. Let $\left[\mathbf{A}^k\right]_{ij} = a_{ij}^{(k)}$. Ignoring the weights that multiply the terms of the adjacency matrix, we can approximate S_{i+} by

$$S_{i+} \approx \sum_{j=1}^{N} m_{ij} = \left(\sum_{k=0}^{\infty} a_{i1}^{(k)}\right) + \left(\sum_{k=0}^{\infty} a_{i2}^{(k)}\right) + \cdots + \left(\sum_{k=0}^{\infty} a_{iN}^{(k)}\right)$$

and therefore $S_{i+}S_{+j}$ is approximately equal to

$$\underbrace{\left(\sum_{k=0}^{\infty} a_{i1}^{(k)}\right)\left(\sum_{k=0}^{\infty} a_{1j}^{(k)}\right) + \cdots + \left(\sum_{k=0}^{\infty} a_{iN}^{(k)}\right)\left(\sum_{k=0}^{\infty} a_{Nj}^{(k)}\right)}_{A} + \underbrace{\sum_{l\neq m}\left(\sum_{k=0}^{\infty} a_{il}^{(k)}\right)\left(\sum_{k=0}^{\infty} a_{mj}^{(k)}\right)}_{B}.$$

Reordering the terms in A, if we take the terms whose exponent sum up to k, we will have the following terms

$$\begin{aligned}
&a_{i1}^{(0)} a_{1i}^{(k)} + \cdots + a_{iN}^{(0)} a_{2N}^{(k)} \\
&a_{i1}^{(1)} a_{1i}^{(k-1)} + \cdots + a_{iN}^{(1)} a_{2N}^{(k-1)} \\
&\vdots \\
&a_{i1}^{(k)} a_{1i}^{(0)} + \cdots + a_{iN}^{(k)} a_{2N}^{(0)}.
\end{aligned}$$

All these terms count the number of paths from i to j in k steps. This means that A can be written as

$$N + \sum_{k=1}^{\infty} (\text{number of paths from } i \text{ to } j \text{ in } k \text{ steps})\ (k+1).$$

Considering B, we rearrange the terms aggregating those with exponents adding up to k, with $k = 1, 2, \ldots$. That is,

$$B = \sum_{l \neq m} \sum_{k=1}^{\infty} \sum_{p=0}^{k} a_{il}^{(p)} a_{mj}^{(k-p)}.$$

The term $a_{il}^{(p)} a_{mj}^{(k-p)}$ counts the number of $k+1$ steps paths from i to j and passing through an edge connecting areas l and m. It takes p steps to reach l and $k-p$ additional steps to reach j from m. This edge $l \to m$ can indeed exist in the original adjacency graph, in which case we are counting a truly existing path. If it does not exist, we are counting paths on the original graph with the additional edge $l \to m$. Therefore, the term B can be written as

$$\sum_{l \neq m} \sum_{k=1}^{\infty} (k+1)\ (\text{number of } k+1 \text{ steps paths from } i \text{ to } j \text{ passing through an edge } l \to m)\ .$$

This means that it counts all possible paths in the original graph, possibly adding one additional edge.

5. Posterior covariance matrix

More relevant to the Bayesian data analysis than the prior covariance matrix is the posterior covariance implied by our prior spatial model. To obtain analytical expressions, assume that y_i can be approximated by a normal distribution with variance $1/\tau_y$. The posterior precision matrix is given by

$$\mathbf{Q}^* = \tau_y \mathbf{I} + \mathbf{Q} = \tau_y + \left(\sigma^2\right)^{-1} \left[\lambda_1 \mathbf{I} + \lambda_2 \text{diag}(\mathbf{n}) + \lambda_3 \text{diag}(\mathbf{N}) - \lambda_2 \mathbf{A} - \lambda_3 \mathbf{1}\mathbf{1}^T\right]$$

and therefore, the covariance matrix is

$$\mathbf{Q}^{*-1} = \mathbf{M}^{*-1} + \frac{\left(\sigma^{-2}\lambda_3\right) \left(\mathbf{M}^*\right)^{-1} \left(\mathbf{1}\mathbf{1}^T\right) \left(\mathbf{M}^*\right)^{-1}}{1 - \left(\sigma^{-2}\lambda_3\right) \mathbf{1}^T \left(\mathbf{M}^*\right)^{-1} \mathbf{1}}$$

where

$$\mathbf{M}^* = \left(\tau_y + \frac{\lambda_1}{\sigma^2}\right) \mathbf{I} + \frac{\lambda_2}{\sigma^2} \text{diag}(\mathbf{n}) + \frac{\lambda_3}{\sigma^2} \text{diag}(\mathbf{N}) - \frac{\lambda_2}{\sigma^2} \mathbf{A}.$$

It is rather surprising that it is possible to interpret each one of the two component matrices of the covariance $\mathbf{Q}^{*-1}$. Considering initially $\mathbf{M}^{*-1}$, after some algebraic manipulations analogous to those carried out earlier for the prior covariance matrix, we have that

$$\mathbf{M}^{*-1} = \left[\mathbf{I} - \left(\tau_y \lambda_3\right) \mathbf{T}^* \mathbf{A}\right]^{-1} \mathbf{T}^*$$

where

$$\mathbf{T}^* = \text{diag}\left\{ \frac{1}{\tau_y + \sigma^{-2}\left(\lambda_1 + \lambda_2 n_1 + \lambda_3 N\right)}, \ldots, \frac{1}{\tau_y + \sigma^{-2}\left(\lambda_1 + \lambda_2 n_N + \lambda_3 N\right)} \right\}.$$

The elements of this diagonal matrix involve the data precision τ_y and the weights of the prior covariance $\sigma^{-2}\left(\lambda_1 + \lambda_2 n_i + \lambda_3 N\right)$. The relevance of each of these parts for the posterior covariance will depend on the ratio between the likelihood variance and the prior variance.

The same matrix expansion that was used earlier can be applied here:

$$\mathbf{M}^{*-1} = \mathbf{T}^* + \left(\sigma^{-2}\lambda_2\right) \mathbf{T}^* \mathbf{A} \mathbf{T}^* + \left(\sigma^{-2}\lambda_2\right)^2 \left(\mathbf{T}^* \mathbf{A}\right)^2 \mathbf{T}^* + \left(\sigma^{-2}\lambda_2\right)^3 \left(\mathbf{T}^* \mathbf{A}\right)^3 \mathbf{T}^* + \cdots.$$

As a result, the posterior covariance matrix $\mathbf{Q}^{*-1}$ has the same structure as the prior covariance matrix, being written as the sum of two matrices:

$$\frac{\sigma^{-2}\lambda_3}{1-\sigma^{-2}\lambda_3\sum_{i,j}S^*_{ij}}\begin{bmatrix}\left(S^*_{1+}\right)^2 & S^*_{1+}S^*_{2+} & \cdots & S^*_{1+}S^*_{N+}\\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ S^*_{N+}S^*_{1+} & S^*_{N+}S^*_{2+} & \cdots & \left(S^*_{N+}\right)^2\end{bmatrix}$$

and

$$\begin{bmatrix}m^*_{11} & m^*_{12} & \cdots & m^*_{1N}\\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ m^*_{N1} & m^*_{N2} & \cdots & m^*_{NN}\end{bmatrix}$$

where m^*_{ij} is the (i, j)-th element of the matrix $\mathbf{M}^{*-1}$ and $S^*_{l+} = \sum_j m^*_{lj} = \sum_i m^*_{il}$.

Therefore the posterior covariance matrix can be interpreted in the same way as the prior covariance matrix. The main difference between the two are the weights appearing in the counts of the possible paths between pairs of areas. While they were equal to $(\lambda_1 + \lambda_2 n_i + \lambda_3 N)^{-1}$ in the case of the prior covariance, they are now equal to $\sigma^2/\left(\tau_y + \sigma^2\left(\lambda_1 + \lambda_2 n_i + \lambda_3 N\right)\right)$. This means that, as the prior covariance, the posterior covariance can be decomposed into two components reflecting different aspects of the neighborhood graph. One component is a weighted average of all paths connecting areas i and j, longer paths having smaller weights than shorter ones. Additionally, the paths are weighted according to the connection degree of the intervening areas in the path, more connected paths having less weights. The other component of $[\mathbf{Q}^{*-1}]_{ij}$ reflects intrinsic aspects of the pair of areas i and j. It does not matter where they are located with respect to each other; this covariance component is simply a product of scores specific to each area and, in this sense, has less spatial content than the first component.

5.1. The special case of two components

We consider briefly a specific case in which the inversion of the prior and posterior covariance matrices are feasible and allow an easier interpretation of the covariance matrix. Suppose that, *a priori*, the area-specific values b_i follow a multivariate normal distribution with mean zero and precision matrix

$$\mathbf{Q} = \frac{1}{\sigma^2}\left((1-\lambda)\mathbf{I} + \lambda\left(N\mathbf{I} - \mathbf{1}\mathbf{1}^T\right)\right)$$

where $\lambda \in [0, 1)$. Compared to the model in (3), this model exchanges the first order neighborhood matrix $\mathbf{R}$ of the Leroux model by the matrix associated with the exchangeable risks model of Bernardinelli and Montomoli [5].

Using (7), we can calculate the covariance matrix:

$$\mathbf{Q}^{-1} = \frac{\sigma^2}{1-\lambda+\lambda N}\left[\mathbf{I} + \frac{\lambda}{1-\lambda}\mathbf{1}\mathbf{1}^T\right]$$

and the correlation $\mathrm{Corr}(b_i, b_j) = \lambda$. The correlation approaches 1 as the weight of the exchangeable model increases.

We can also find the posterior covariance matrix, if we assume that the data are normally distributed with variance $(\tau_y)^{-1}$. In this case, the posterior correlation of the random effects of areas i and j is given by

$$\mathrm{Corr}(b_i, b_j|\mathbf{y}) = \frac{\lambda}{\tau_y + (\sigma^2)^{-1}(1-\lambda)}.$$

This correlation is close to zero if λ is also close to zero. In the opposite direction, to get correlation close to 1, we need to have both, λ and $\tau_y\sigma^2$, close to 1. That is, we need an exchangeable component with large relative weight and, at the same time, the underlying risks should have a variation similar to the likelihood variance.

6. Illustrative application

In this section, we analyze the spatial incidence of sudden infant death syndrome (SIDS) in the 100 counties of the North Carolina state for the period 1999–2006. This spatial pattern in the period from 1974 to 1984 was analyzed previously by Symons et al. [35], Cressie [10, p. 386], Kulldorff [21], Lawson and Clark [23], and this early data set is part of many spatial statistics software manuals. There have been found spatial variation of the relative risk with an increasing trend from west to east in the whole USA. According to the National Center for Health Statistics, the US SIDS incidence rate (per thousand live births) has been decreasing steadily from 1.53 in 1980 to 0.51 in 2005. The southern region presents the highest rates and, in the period 1999–2006, the North Carolina rate was 0.73 cases per thousand live births. One of the main aims of the

Table 1
DIC and logarithm score criteria for North Carolina data base using Gamma(0.5, 0.0005) (first row) and Gamma(0.01, 0.01) (second row). The models compared are BYM, Leroux and our model with all and with three components. The logarithm score criterium is evaluated with the importance weights and the importance resampling methods.

Prior	All comp	3 comp	Leroux	BYM
	DIC			
Gamma(0.5, 0.0005)	438.63	438.59	439.27	439.28
Gamma(0.01, 0.01)	438.74	438.66	439.63	441.85
Logarithm score using the importance weights method				
Gamma(0.5, 0.0005)	2.22	2.24	7.87	2.48
Gamma(0.01, 0.01)	2.25	2.22	7.86	2.62
Logarithm score using the importance resampling method				
Gamma(0.5, 0.0005)	2.05	2.04	2.82	2.40
Gamma(0.01, 0.01)	2.03	2.04	2.67	2.27

spatial analysis of the SIDS underlying risk is to find hints for the identification of unknown risk factors. We show how our model can be used in this problem considering the effect of known risk factors.

We fitted all models using the software WinBUGS [25] to obtain the posterior distribution of the relative risks. Taking all possible neighborhood matrices $\mathbf{R}^{(l)}$, we have l varying from 1 to 19, where the maximum is determined by the graph diameter, as defined in Section 3. We also considered the particular three-components model, which uses only the identity matrix, the first order neighborhood matrix, and the matrix $\mathbf{1}\mathbf{1}^T$. We adopted a gamma distribution with parameters equal to either 0.5 and 0.0005 or 0.01 and 0.01 for all inverse variance parameters and a uniform distribution on the l-dimensional simplex for the weights $(\lambda_1, \ldots, \lambda_l)$. We ran the Markov chain Monte Carlo (MCMC) chains for 30,000 iterations, with 15,000 iterations as burn-in, and convergence was assessed by a variety of methods, including graphical diagnostics. The posterior inference was based on a thinned sample of 1000 elements, resulting from retaining every 15-th simulated parameter vector. In order to compare the different models, we calculated the deviance information criterion (*DIC*) proposed by Spiegelhalter et al. [31].

The *DIC* values are presented in the first row of Table 1. The model proposed by Leroux has the poorest fit followed by the model with all neighborhood components and the BYM model. Although they have similar values, it is clear that the model with three components is the best one for these data. In order to check the model sensitivity with respect to the choice of the prior distribution for the variance parameters, we fit the model considering a Gamma distribution with parameters 0.01 and 0.01 for these precision parameters. The values of the *DIC* criterion are shown in the second row of Table 1. The results are almost the same as before. Again, the best model is that with three components, while the BYM and Leroux models had the worst fit.

The *DIC* has been criticized as an inadequate measure to evaluate models and it should be considered cautiously [29]. Therefore, in addition to this global measure, we also calculated a cross-validation posterior predictive distribution check proposed by Stern and Cressie [32]. We computed the approximated conditional probability ordinate using their importance weights and the importance resampling methods. The basic idea of posterior predictive checking is to assess the fitness of the model in a given area in a two step procedure. In the first one, we obtain a predictive distribution for the i-th area without using the observed count in the area in question. In the second one, we compare the truly observed disease count in that area with the predictive distribution evaluating how extreme it is.

More specifically, let $\boldsymbol{\theta}$ be the vector of all parameters in a given Bayesian model and $\mathbf{Y}_{-i}$ denote the data vector without the i-th area count. Let $p(\boldsymbol{\theta}|\mathbf{Y}_{-i})$ denote the posterior distribution of $\boldsymbol{\theta}$ computed without the observation in the i-th region. We define a cross-validation posterior predictive distribution of $Y^{\text{rep}}_{i,-i}$ as

$$CPO_i = p\left(Y^{\text{rep}}_{i,-i}|\mathbf{Y}_{-i}\right) = \int p(Y^{\text{rep}}_{i,-i}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{Y}_{-i})d\boldsymbol{\theta}$$

where $Y^{\text{rep}}_{i,-i}$ is a predicted value for the count in region i based on the given model and data $\mathbf{Y}_{-i}$. This measure is also called conditional predictive ordinate (CPO). A small value of the CPO_i indicates that the i-th observation is very unlikely under the model and the remaining observations.

As it is very costly to refit the model without each observation in turn, Stern and Cressie [32] avoid the refitting of the model using two different methods. They propose the use of importance weighting and importance resampling to approximate the posterior distribution that would be obtained if the analysis were repeated without a small area. In order to compare the observed CPO's, we used a summary measure known as logarithmic score [14]. This is a scoring rule providing an evaluation of a model forecasting performance based on the posterior predictive distribution. This measure is calculated as

$$LS = -\frac{\sum_{i=1}^{N} \log(CPO_i)}{N}. \tag{9}$$

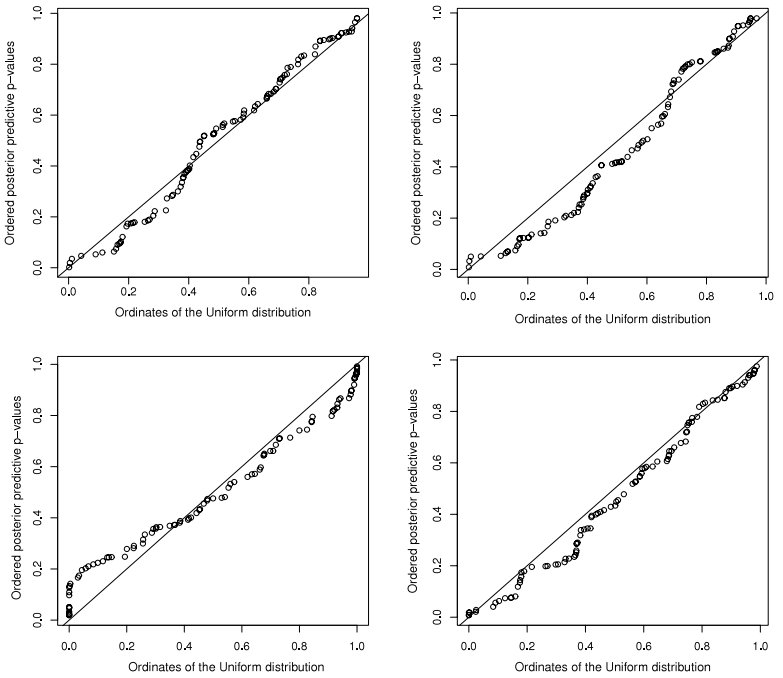

Fig. 1. *QQ*-plot of *p*-values using Gamma(0.5, 0.0005) for all components, three components, Leroux and Bym models. The *p*-values were calculated using the cross-validation proposal of Marshall and Spiegelhalter [27].

The lower this value, the better the model. According to Stone [33], this logarithm score is asymptotically equivalent to the Akaike Information Criterion if the observations are independent.

Table 1 shows the values computed for these measures for the two priors considered before. Considering the resampling weight method, we note that the models with all the components and the one with three components presented the best performance with respect to this criterion, since they have lower values. The model proposed by Leroux had the poorest performance among the four. Table 1 also shows the results using the method of importance resampling using the same priors. Once again, our two models, with all and with three components, were better than the others. It is also noticeable that Leroux model had a poor performance in all the cases.

One additional cross-validation measure, proposed by Marshall and Spiegelhalter [27], can be used to evaluate the goodness of fit of the models. This method is based on the simulation of both, replicate random effects and data, and it is simpler to apply than the methods from Stern and Cressie. The simplicity comes from the embedding of the leave-one-out predictive distributions replications within the MCMC simulations. The Bayesian *p*-value is defined as the minimum between $P(Y^{\text{rep}}_{i,-i} < y_i|y_{-i}) + \frac{1}{2}P(Y^{\text{rep}}_{i,-i} = y_i|y_{-i})$ and $P(Y^{\text{rep}}_{i,-i} > y_i|y_{-i}) + \frac{1}{2}P(Y^{\text{rep}}_{i,-i} = y_i|y_{-i})$. These *p*-values should be approximately uniformly distributed if the model is correct and Marshall and Spiegelhalter [27] suggested a *QQ*-plot as a diagnostic tool for model checking. Fig. 1 shows the *QQ*-plot for each of the models using a Gamma(0.5, 0.0005) as a hyperprior while Fig. 2 shows the same *QQ*-plot using a Gamma(0.01, 0.01) as a hyperprior. The model proposed by Leroux is not adequate as the points clearly depart from the straight line, while the other three models have their *p*-values equally well fitted by the uniform distribution.

6.1. Including covariates

Most epidemiologic studies involve risk factors. The spatial analysis of disease rates should always take into account known or suspected risk factors. The random effects modeled with Bayesian spatial models stand for unknown risk factors and their estimation through the posterior distribution could help on spotting underlying causes for these as yet unknown risks. There is not much knowledge of the syndrome's biological cause or potential causes but some epidemiological studies have found an ecological correlation between SIDS rates and social-economic conditions (see [17]). Black, American Indian or Eskimo infants have a larger incidence of SIDS, as well as those under maternal risks such as being a teenage mother, being a smoker, drug or alcohol user, and having inadequate prenatal care. Therefore, we included the following covariates in our model: the average proportion of Black and American Indians in the county population from 2001 to 2009 (see http://www.census.gov/popest/counties/asrh/CC-EST2009-RACE5.html), the proportion of mothers who had prenatal care and the proportion of mothers who were smokers from 2005 to 2009 (see http://www.epi.state.nc.us/SCHS/data/databook/).

We centered all three covariates and fitted the all components model with the three covariates simultaneously present in the model. We obtained the posterior densities in Fig. 3. Fitting our model with three components gave virtually the same

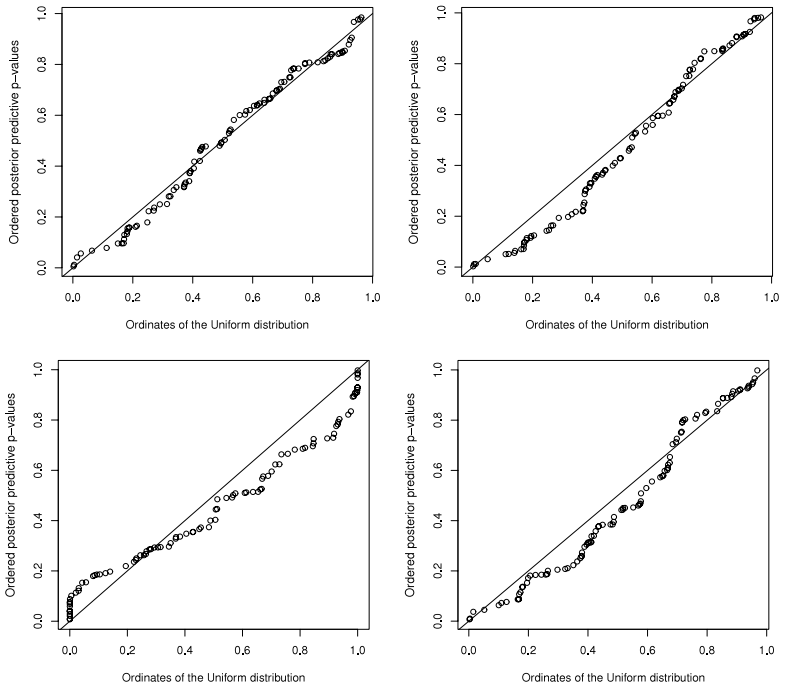

Fig. 2. *QQ*-plot of *p*-values using Gamma(0.01, 0.01) for all components, three components, Leroux and Bym models. The *p*-values were calculated using the cross-validation proposal of Marshall and Spiegelhalter [27].

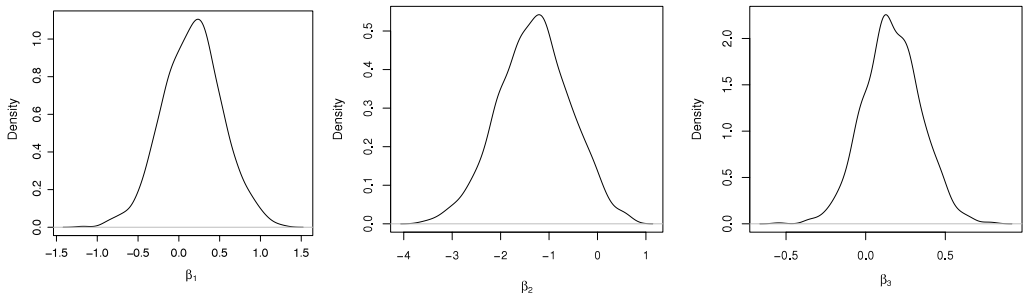

Fig. 3. Posterior density of the β coefficients for three covariates using our model with all components for the spatial random effect. The first plot refers to the proportion of Black and American Indians in the population, the second plot refers to proportion of mothers who had prenatal care and the third plot refers to the proportion of mothers who were smokers.

result. We find evidence of covariate effects only for the proportion of mothers who had prenatal care (second plot), since zero is on the border of the 95% highest density interval (given by $(-2.719, 0.139)$) and the posterior probability that the covariate coefficient is less than zero is given by 0.963. We refitted the model three times, each time with a single covariate and the only significant covariate was again the proportion of mothers who had prenatal care. Focusing on the model with this single covariate, we obtain a posterior mean equal to -1.419. This means that a 1% increase in prenatal care leads to an average reduction in the SIDS risk of $\exp(-1.419 * 0.01) = 0.987\%$ or 1.41% reduction.

7. Simulation study

In this section, we present a simulation study that helps to understand our formulation better and that shows clearly its advantages and benefits with respect to the other two main approaches available to spatial statisticians, the Leroux and the BYM models. We used the North Carolina counties with the observed live births in the period 1999–2006. The precision coefficient σ^2 is fixed and equal to 5 in all simulations. The simulated SIDS counts were generated according to six different scenarios, as we explain next.

The first model for the SIDS counts assumed an extreme situation, in which we have a constant underlying rate equal to the observed NC SIDS rate (0.73 per thousand live births). That is, each y_i is generated independently from a Poisson distribution with mean $E_i = 0.73m_i/1000$, where m_i is the observed number of live births in the i-th county. This implies that the relative risk ψ_i is equal to 1 for all areas. The other five scenarios were less extreme and had spatially varying relative

risks. In these other cases, y_i was simulated independently from a Poisson with mean $E_i\psi_i$ where $\psi_i = \exp(b_i)$. In the second scenario, $\psi_i \approx 1$ for all i, implying that the precision matrix is composed basically by the neighborhood matrix full of 1's. More specifically, b_i follows our model with $\lambda_{20} = 0.979$ and $\lambda_1 = \cdots = \lambda_{19} = (1 - \lambda_{20})/19 = 0.001$.

In the third scenario, we used our component model with four heavily weighted neighborhood matrices in the precision matrix: the identity matrix, the first and the second neighborhood order matrices, and the matrix full of 1's, with $\lambda_1 = 0.007$, $\lambda_2 = 0.421$, $\lambda_3 = 0.351$, and $\lambda_{20} = 0.210$. All the other λ's are small and equal to 0.001. The fourth scenario had a high weight associated with the second neighborhood order matrix, and moderate weights associated with the identity and the matrix full of 1's. That is, $\lambda_1 = 0.108$, $\lambda_2 = 0.011$, $\lambda_3 = 0.540$, $\lambda_{20} = 0.324$. All the other λ's are equal to 0.001. The fifth scenario follows the Leroux model with $\lambda_1 = \lambda_2 = 0.5$ and all the other λ's equal to zero. The sixth scenario follows a CAR model with $\rho = 0.99$ to mimic the behavior of the improper ICAR prior.

We fitted four different models to each simulated dataset: our model with 20 increasing neighborhood orders, our model with three components (described in Section 5.1), and the BYM and Leroux models. The prior distribution for the precision parameter was taken as a Gamma(0.5, 0.0005) in all models. In all cases, we ran the MCMC for 3000 iterations with 1500 as a burn-in period.

Let

$$MSE_i = \frac{1}{B}\sum_{j=1}^{B}(\psi_i^{(j)} - \exp(b_i))^2$$

where $\psi_i^{(j)}$ is the j-th simulated value of the relative risk ψ_i, $\exp(b_i)$ is the realized relative risk under each one of the scenarios, and B is the number of simulations retained after burn-in. Note that $b_i = 0$ in the first scenario. Denote by $\overline{MSE}$ the average of the MSE_i values, $i = 1, \ldots, 100$. We considered four summary statistics to evaluate the fitted models: the average $\overline{MSE}$, the *DIC*, and the two logarithm scores, based on importance weights and on importance resampling. The measure $\overline{MSE}$ is our preferred criterium to select the best model since we compare the estimated with the true relative risks in each model. Of course, this is only possible in simulations, not in real data analysis. We simulated 10 independent copies of each scenario. The results shown below are the averages of the summary statistics in these 10 independent replications.

Table 2 shows the values of these evaluation measures for each one of the four possible models in each scenario. Considering the $\overline{MSE}$ criterium, our model is always the best one, either the three components model or the 20 components model. This is rather surprising considering that at least in one case (Scenario 5), we are fitting a model (Leroux) to data generated according to this same model. It is also clear that the BYM model is the worst model in all scenarios. In all of them the three component model had almost the same $\overline{MSE}$ as the 20 component model. The third column shows how careful one must be when using the *DIC* measure. In all scenarios, the difference between the *DIC* measures is very small. Furthermore, these numbers are averages and, in some replications, the *DIC* did not select the best model. For example, in 4 of the 10 replications of scenario 1, *DIC* selected Leroux or BYM although $\overline{MSE}$ indicated clearly that our model was better. The *CPO* measures are not very sensitive either, with differences showing up in the third decimal place in most cases. In the Web-based supplementary material we show the estimated posterior densities of the differences between ψ_j and the true values of the relative risks realized in one particular and typical simulation.

8. Conclusions

In our model, we considered a precision matrix equal to a weighted average of increasing neighborhood matrices. One possibility we have not explored in this paper is to define a continuous version of this model. Let $\lambda(t)$ be a probability density function defined for $t \in [0, 1]$ and $\mathbf{R}^{(t)}$ be a continuously defined precision matrix. Assume that $\mathbf{R}^{(t)}$ as a function of t is an injective function. The precision matrix of the mixture model is given then by

$$\mathbf{Q} = \frac{1}{\sigma^2}\int_0^1 \lambda(t)\mathbf{R}^{(t)}dt.$$

This model would allow different degrees of neighborhood and could be more flexible to adapt to empirical data.

Another possible extension of the model is to include other kinds of neighborhood structure in the mixture of matrices that compose the precision. For example, we can include a matrix which has neighborhood criteria based on the size of the cities. It is also possible to treat space–time data including matrices that represent time relationship.

The BYM model is very popular but one problem with it is to find the appropriate spatial smoothing degree to estimate the relative risks. In fact, other authors have noticed its tendency to oversmooth the estimates in some cases [7]. The model we treat in this paper allows for the multiple definition of a smoothing neighborhood. In our model, the λ_j parameters control this smoothing automatically. The model can be specially useful in the situation where the underlying risk is practically constant. However, our simulation study shows that in many other spatial underlying structures our models were able to fit the data better than current spatial alternative models. In particular, the three components model is a very good option as it has a small number of parameters and it is able to estimate the true relative risk much better than other models with almost the same number of parameters.

One important outcome of this paper is to provide an interpretation for the posterior distributions involved in our model. We were able to show how the correlation between neighbors depends on the vector of λ_j values and on the graph structure.

Table 2
$\overline{MSE}$, *DIC*, and *LS* for the simulation of the six scenarios. Summary statistics are the average of 10 independent replications of each model.

Model	$\overline{MSE}$	*DIC*	Logarithm-score (importance weights)	Logarithm-score (importance resampling)
Scenario 1				
All comp	0.0024	400.7232	2.0029	1.9925
Three comp	0.0024	400.8519	2.0034	1.9928
Leroux	0.0042	401.0085	2.0038	1.991
BYM	0.0209	403.0931	2.0114	1.9812
Scenario 2				
All comp	0.0019	407.8364	2.0396	2.0282
Three comp	0.0023	407.9339	2.0398	2.0279
Leroux	0.0051	408.0303	2.0403	2.024
BYM	0.0235	408.6293	2.0417	2.0057
Scenario 3				
All comp	0.0022	407.9054	2.0393	2.0287
Three comp	0.0024	407.9057	2.0391	2.0280
Leroux	0.0047	408.0214	2.0398	2.0250
BYM	0.0218	409.5973	2.0455	2.0113
Scenario 4				
All comp	0.0027	407.1517	2.0354	2.0238
Three comp	0.0021	407.0670	2.0349	2.0245
Leroux	0.0048	407.2710	2.0357	2.0214
BYM	0.0233	408.6199	2.0408	2.0052
Scenario 5				
All comp	0.0101	411.4990	2.0578	2.0431
Three comp	0.0098	411.5712	2.0581	2.0444
Leroux	0.0115	411.5052	2.0572	2.0409
BYM	0.0295	412.1749	2.0606	2.0221
Scenario 6				
All comp	0.0106	412.7021	2.0637	2.0509
Three comp	0.0104	412.6090	2.0631	2.0508
Leroux	0.0131	412.5638	2.0629	2.0454
BYM	0.0299	412.5616	2.0618	2.0229

We view our model as an additional tool the statistician has available to made inference about the relative risks of disease mapping problems. However, the model can also be applied to other type of spatial data that requires the specification of neighborhood structures such as space–time problem or spatial survival data analysis.

Appendix. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2012.02.017.

References

[1] R.M. Assunção, Space varying coefficient models for small area data, Environmetrics 14 (2003) 453–473.
[2] R.M. Assunção, E.T. Krainski, Neighborhood dependence in Bayesian spatial models, Biometrical Journal 51 (2009) 851–869.
[3] R.M. Assunção, J.E. Potter, S. Cavenaghi, A Bayesian space varying parameter model applied to estimating fertility schedules, Statistics in Medicine 14 (2002) 2057–2075.
[4] S. Banerjee, B.P. Carlin, A.E. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, in: Monographs on Statistics & Applied Probability, Chapman & Hall/CRC, 2004.
[5] L. Bernardinelli, C. Montomoli, Empirical bayes versus fully Bayesian analysis of geographical variation in disease risk, Statistics in Medicine 11 (1992) 983–1007.
[6] J. Besag, J. York, A. Mollié, Bayesian image restoration, with two applications in spatial statistics, Annals of the Institute of Statistical Mathematics 43 (1991) 1–20.
[7] N. Best, S. Richardson, A. Thomson, A comparison of Bayesian spatial models for disease mapping, Statistical Methods in Medical Research 14 (2005) 35–39.
[8] A. Brezger, T. Kneib, S. Lang, Bayesx-software for Bayesian inference based on Markov chain Monte Carlo simulation techniques, 2003.
[9] B. Carlin, S. Banerjee, Hierarchical multivariate car models for spatio-temporally correlated survival data, in: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (Eds.), Bayesian Statistics, vol. 7, Oxford University Press, 2003, pp. 45–63.
[10] N. Cressie, Statistics for Spatial Data, John Wiley & Sons, 1991.
[11] L.E. Eberly, B.P. Carlin, Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models, Statistics in Medicine 19 (2000) 2279–2294.
[12] A.E. Gelfand, K. Hyon-Jung, C. Sirmans, S. Banerjee, Spatial modeling with spatially varying coefficient processes, Journal of the American Statistical Association 98 (2003) 2057–2075.

[13] A.E. Gelfand, P. Vounatsou, Proper multivariate conditional autoregressive models for spatial data analysis, Biostatistics 4 (2003) 11–25.
[14] T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association 102 (2007) 359–378.
[15] L. Held, G. Graziano, C. Frank, H. Rue, Joint spatial analysis of gastrointestinal infectious diseases, Statistical Methods in Medical Research 15 (2006) 465–480.
[16] L. Held, I. Natario, S. Fenton, H. Rue, N. Becker, Towards joint disease mapping, Statistical Methods in Medical Research 14 (2005) 61–82.
[17] C. Hunt, F. Hauck, Sudden infant death syndrome, Nelson Textbook of Pediatrics 174 (2007) 1736–1742.
[18] M. Iosifescu, Finite Markov Processes and Their Applications, John Wiley and Sons, Chichester, 1989.
[19] X. Jin, B.P. Carlin, Multivariate parametric spatiotemporal models for county level breast cancer survival data, Lifetime Data Analysis 11 (2005) 5–27.
[20] L. Knorr-Held, N. Best, A shared component model for detecting joint and selective clustering of two diseases, Journal of the Royal Statistical Society, Series A 164 (2001) 73–85.
[21] M. Kulldorff, A spatial scan statistic, Communications in Statistics Theory and Methods 26 (1997) 1481–1496.
[22] S. Lang, L. Fahrmeir, Bayesian generalized additive mixed models, a simulation study, Journal of the Royal Statistical Society, Series C (Applied Statistics) 50 (2001) 201–220.
[23] A.B. Lawson, A. Clark, Spatial mixture relative risk models applied to disease mapping, Statististics in Medicine 21 (2002) 359–370.
[24] B.G. Leroux, X. Lei, N. Breslow, Estimation of disease rates in small areas: a new mixed model for spatial dependence, Statistical Models in Epidemiology; the Environment and Clinical Trials (1999) 179–192.
[25] D. Lunn, A. Thomas, N. Best, D. Spiegelhalter, Winbugs—a Bayesian modelling framework: concepts, structure, and extensibility, Statistics and Computing 10 (2000) 325–337.
[26] C. MacNab, C.B. Dean, Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models, Statistics in Medicine 19 (2000) 2421–2435.
[27] E. Marshall, D. Spiegelhalter, Approximate cross-validatory predictive checks in disease mapping models, Statistics in Medicine 22 (2003) 1649–1660.
[28] M.A. Martínez-Beneito, A. López-Quilez, P. Botella-Rocamora, An autoregressive approach to spatio-temporal disease mapping, Statistics in Medicine 27 (2008) 2874–2889.
[29] M. Plummer, Penalized loss functions for Bayesian model comparison, Biostatistics 9 (2008) 523–539.
[30] G.L. Silva, C.B. Dean, T. Niyonsenga, A. Vanasse, Hierarchical Bayesian spatiotemporal analysis of revascularization odds using smoothing splines, Statistics in Medicine 27 (2008) 2381–2401.
[31] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. Van der Linde, Bayesian measures of model complexity and fit (with discussion), Journal of the Royal Statistical Society, Series B 64 (2002) 583–639.
[32] H.S. Stern, N. Cressie, Posterior predictive model checks for disease mapping models, Statistics in Medicine 19 (2000) 2377–2397.
[33] M. Stone, An asymptotic equivalence of choice of model by cross-validation and akaikes criterion, Journal of the Royal Statistical Society, Series B (Methodological) 39 (1997) 44–47.
[34] D.C. Sun, R.K. Tsutakawa, H. Kim, Z.Q. He, Spatio-temporal interaction with disease mapping, Statistics in Medicine 19 (2000) 2015–2035.
[35] M.J. Symons, R.C. Grimson, Y.C. Yuan, Clustering of rare events, Biometrics 39 (1983) 193–205.
[36] G. White, S.K. Ghosh, A stochastic neighborhood conditional autoregressive model for spatial data, Computational Statistics and Data Analysis 53 (2009) 3033–3046.