BAYESIAN SPATIAL MODELS WITH MIXTURE NEIGHBORHOOD STRUCTURE

Erica Castilho Rodrigues[1], **Renato Martins Assunção**[1]

Abstract: *In Bayesian disease mapping, one needs to specify a neighborhood structure to make inference on 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 usual Bayesian spatial models: given the neighborhood graph, the 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.*

Palavras-chave: Disease mapping, Markov Random Field, Spatial Hierarchical Models.

Introduction

In disease mapping, the Bayesian model proposed by Besag, York and Mollié [7], 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 measurements surrogates. Originally, BYM was introduced to model a cross-section of counts collected in a set disjoint geographical areas composing a partitioned map. Since then, BYM has been extended into several directions to include space-time generalized linear models [26, 27, 19, 33, 29], spatial survival models [10, 21], spatially-varying parameters models [1, 2, 12], and generalized additive models [22]. Multivariate extensions incorporating two correlated sets of spatial effects have also been proposed in recent years [21, 13, 15, 16]. 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 θ_i associated with the i-th area is the sum $\phi_i + \psi_i$ of two components where ϕ_i is a spatially structured random effect assigned an improper CAR prior distribution, and ψ_i is a second set of i.i.d. zero-mean normally distributed unstructured random effects. This is termed a convolution prior [7] since the density of θ_i's will be the convolution of the joint densities of ϕ_i vector and the ψ_i vector.

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 conveniently 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 the relative risks estimation. Some authors

[1]Departamento de Estatística, UFMG,
ericaa_casti@yahoo.com.br, assuncao@est.ufmg.br

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 evidence.

Notwithstanding its crucial role in the spatial Bayesian models, very few studies have considered different neighborhood structures for disease mapping problems. One notable exception is [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 and their sub-divisions, 79 local health areas. 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 HUs 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 usual and it reminds the higher autoregressive order models in the time series setting.

A more recent reference is [35], who introduced a stochastic neighborhood CAR model where the selection of the neighborhood depends on unknown parameters. They estimate how far should the neighborhood of the areas be assuming proximity weights that stay constant and equal to 1 up to a certain distance and thereafter decreases exponentially towards zero. In contrast with most of the published applied papers in disease mapping, they base their model in the proper CAR model 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 [4] and [3]).

These studies consider only locally larger neighborhoods than the first order neighborhood provided by the simple adjacency between the areas. 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 variations consider the random effects being composed of either unstructured over-dispersion or small range spatial conditional variation. These are two extremes 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 longer range than the immediate geographical neighborhood. This is likely to be useful specially 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 spatial conditional autoregressive models in terms of the neighborhood structure. We propose a model in which the neighborhood structure is part of the parameter space. We retain the same Markov property that is part 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 , we introduce the notation and present some models that were proposed previously. In section present the definition our model. In Section , 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, in section , a specific and simpler case of our model allowing for a more thorough understanding of the covariance structure. In Section , we illustrate the use of our model for disease mapping. We end in Section

presenting the main conclusions.

Disease Mapping

A Bayesian hierarchical model is one of the main tools to make inference on the underlying relative risks of a disease observed into disjoint geographical areas of a map. These models can be described in this way: suppose that we have N geographic areas and each of them has a relative risk ψ_i for $i = 1, ..., N$, that needs to be estimated. The Bayesian inference is based on the posterior distribution of $\boldsymbol{\psi} = (\psi_1, ..., \psi_N)$ given by $\boldsymbol{f}(\boldsymbol{\psi}|y_1, ..., y_N) = l(y_1, ..., y_N|\boldsymbol{\psi})f(\boldsymbol{\psi})$, where $l(y_1, ..., y_N|\boldsymbol{\psi})$ is the likelihood function and $\boldsymbol{f}(\boldsymbol{\psi})$ is the prior distribution of the parameters vector $\boldsymbol{\psi}$. Conditionally on the values $\psi_1, ..., \psi_N$, the values $Y_1, ..., Y_N$ are supposed 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. The modeling of the prior distribution $\boldsymbol{f}(\boldsymbol{\psi})$ 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, conditionally on all other areas values, depends only upon the ψ_j's 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 of 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 anticipates a spatial dependence between the relative risks due to environmental and genetic similarities of neighboring areas. The most popular distribution to reflect the spatial structure of the data in the prior distribution was introduced by [7]. 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 of the spatial and non-spatial effects, as noticed by [9]. To fix this problem, [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 is included a parameter σ^2 to measure the random effect variance. He proposed a multivariate normal distribution for the random effects $\mathbf{b} = (b_1, \ldots, b_N)$ in (1) with the following precision matrix

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

where $\mathbf{I}$ is the identity matrix and $\mathbf{R}$ is the precision matrix of the ICAR model, which means that

$$\mathbf{R}_{ij} = \begin{cases} n_i & \text{if } i = j \\ -1 & \text{if } j \sim j \\ 0 & \text{otherwise} \end{cases}$$

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 $[0,1]$, and hence, the precision matrix $\mathbf{Q}$ is a weighted sum of the $\mathbf{I}$ and $\mathbf{R}$ matrices.

Another spatial mixture model was proposed by [23]. Spatial discontinuities in the risk surface can be over smoothed by the BYM and Leroux models and hence they suggested the addition of another spatial effects component $(w_1, \ldots, w_N)$. This additional component uses the absolute differences $|w_i - w_j|$, rather than squared differences $(w_i - w_j)^2$, between neighboring areas in the prior density:

$$\pi(w_1, \ldots, w_N | \lambda) \propto \frac{1}{\sqrt{\lambda}} \exp\left(-\frac{1}{\sqrt{\lambda}} \sum_{i \sim j} |w_i - w_j|\right). \quad (4)$$

The relative risks are the sum of the unstructured component θ_i and a mixture of the usual spatial components ϕ_i given in (2) and the prior in (4):

$$\log(\psi_i) = \mu + \theta_i + p_i \phi_i + (1 - p_i) w_i$$

where $p_1, \ldots, p_N$ are i.i.d with Beta(α, α) distribution.

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 consider that, if we have information on the immediate neighbors, no additional information about the other areas is necessary to make inference on 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 the economic, social and environmental characteristics are approximately constant over the entire region. This implies a certain 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.

Model definition

We propose a model that expands the single neighborhood structure 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 i is a l-th order neighbor of 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, it 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:

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

where $\lambda_1 + \lambda_2 + ... + \lambda_{k+1} = 1$ 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,

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

where $n_i^{(l)}$ is the number of neighbors of site i 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, $i \in \partial_j^{(l)}$. It is important to point out that these matrices are linearly independent, ensuring the parameters identifiability.

This matrix is positive definite, as it satisfies the sufficient condition of being diagonal dominant. That is, for all $i = 1, \ldots, n$

$$\mathbf{Q}_{ii} > \sum_{j=1}^{N} |\mathbf{Q}_{ij}|$$

because

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

as $\lambda_1 \in (0, 1)$, and, therefore, $\mathbf{Q}$ can be a precision matrix.

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

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

and

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

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 the $(n-2)$ dimensional $\mathbf{b}_{-ij}$ be the vector $\mathbf{b}$ without its $i-th$ and $j-th$ coordinates. It can be shown that the conditional correlation between the areas, $\text{Corr}(b_i, b_j | \mathbf{b}_{-ij})$ is given by

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

with the proportionality constant given by the inverse of

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

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:

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

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

$$
\begin{aligned}
f(\mathbf{b}) \propto & \left(\exp\left\{-\frac{1}{2\sigma^2}\sum_i b_i^2\right\}\right)^{\lambda_1} \\
& \left(\exp\left\{-\frac{1}{4\sigma^2}\sum_i \sum_{j:j\in\partial_i^{(l)}} (b_i - b_j)^2\right\}\right)^{\lambda_2} \\
& \ldots \\
& \left(\exp\left\{-\frac{1}{4\sigma^2}\sum_i \sum_{j:j\in\partial_i^{(l)}} (b_i - b_j)^2\right\}\right)^{\lambda_k}
\end{aligned}
$$

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.

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 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 $\mathbf{R}^{(k)}$ is given by

$$\mathbf{R}^{(k)} = \text{diag}(\mathbf{N}) - \mathbf{1}\mathbf{1}^T$$

where $\mathbf{N} = N\mathbf{1}$ with N being the total number of areas in the map and $\mathbf{1}$ denotes a N-dimensional vector of ones. 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 $\mathbf{A1} = \mathbf{n} = (n_1, \ldots, n_N)$ 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.

Using the definitions of $\mathbf{A}$, $\mathbf{N}$ and $\mathbf{n}$, the inverse of the precision matrix

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

is given by

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

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

Proof.:

From matrix algebra, we know that

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

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

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

As the matrix $\mathbf{M}$ is symmetric, $\mathbf{M}^{-1}$ is also symmetric and therefore, for all $l = 1, \ldots, N$,

$$\sum_j m_{lj} = \sum_i m_{il}\,.$$

Let $S_{l+} = \sum_j m_{lj} = \sum_i m_{il}$. We can write the covariance matrix as

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

To understand this covariance matrix, we consider initially the matrix $\mathbf{M}^{-1}$ by following the analytical approach adopted by [3]. We can write

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

where

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

The following theorem shows the matrix $[\mathbf{I} - \lambda_2 \mathbf{TA}]^{-1}$ presents an interpretable format.

Using the definitions of $\mathbf{T}$, $\mathbf{A}$ and λ_2, consider the matrix

$$[\mathbf{I} - \lambda_2 \mathbf{TA}]^{-1}$$

this inverse matrix can be written as a series as follows:

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

Proof¿:

A well known linear algebra result [18] 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

$$(\mathbf{I} - \mathbf{P})^{-1} = \mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \mathbf{P}^3 + \ldots$$

To use this result with the matrix $[\mathbf{I} - \lambda_2 \mathbf{TA}]^{-1}$, we need to show that the terms $\lambda_2^l [(\mathbf{TA})^l]_{ij}$ of the power matrix approximate zero when the power l increases. This will be done finding an upper bound. Consider initially $l = 2$. We see that

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

since $n_i \leq N$. As $\text{diag}(1/\mathbf{n})\mathbf{A}$ is an 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 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 the areas have. Going from i to j through a highly connected area has a smaller contribution to $\mathbf{M}^{-1}_{ij}$ 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 has few immediate neighbors.

To complete the understanding of the covariance matrix $\mathbf{Q}^{-1}$ in (6), we consider now the value S_{i+}. We have

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

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. It is inversely related to the average degree of connectivity that area i has with the other areas in the graph. It is a value associated with the individual area, not with specific pairs of areas.

In summary, the covariance $\mathrm{Cov}(b_i, b_j) = \mathbf{Q}^{-1}_{ij}$ is the sum of two components. The first one is $\mathbf{M}^{-1}_{ij}$ 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 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 individual areas. Tho following interpretation of these terms show how they reflect the kind of structure that was defined in the model.

We can write $S_{i+}S_{+j}$ in a different way, in order to see how they reflect the structure of a complete graph. To make this interpretation easier to be understood, we will leave out the weights the multiply the terms of the adjacency matrix. We will consider just the terms a_{ij}, which are one if i and j are neighbors and zero otherwise. Therefore we have an approximation for S_{i+} that is given by

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

So we see that

$$S_{i+}S_{+j} \approx \left[\left(\sum_{k=0} a_{i1}^{(k)}\right) + \cdots + \left(\sum_{k=0} a_{iN}^{(k)}\right)\right]\left[\left(\sum_{k=0} a_{1j}^{(k)}\right) + \cdots + \left(\sum_{k=0} a_{Nj}^{(k)}\right)\right]$$

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

Lets analyze first term A. We can see that we will have in this sum the following term

$$a_{i1}^{(0)}\,a_{1i}^{(0)} + \cdots + a_{iN}^{(0)}\,a_{2N}^{(0)}$$

and this components are all equal to 1, so their sum is equal to N.
We will also have terms of this kind

$$a_{i1}^{(0)}\,a_{1i}^{(1)} + \cdots + a_{iN}^{(0)}\,a_{2N}^{(1)}$$
$$a_{i1}^{(1)}\,a_{1i}^{(0)} + \cdots + a_{iN}^{(1)}\,a_{2N}^{(0)}$$

and both of them count the possible ways of going from i to j in one step. If we consider the terms whose exponents sum up two, we have

$$a_{i1}^{(0)}\,a_{1i}^{(2)} + \cdots + a_{iN}^{(0)}\,a_{2N}^{(2)}$$
$$a_{i1}^{(1)}\,a_{1i}^{(1)} + \cdots + a_{iN}^{(1)}\,a_{2N}^{(1)}$$
$$a_{i1}^{(2)}\,a_{1i}^{(0)} + \cdots + a_{iN}^{(2)}\,a_{2N}^{(0)}$$

Each of these three terms counts the number of paths that go from i to j in 2 steps. In general, what we see is that if we take the terms whose exponent sum up k, we will have terms of the kind

$$a_{i1}^{(0)}\,a_{1i}^{(k)} + \cdots + a_{iN}^{(0)}\,a_{2N}^{(k)}$$
$$a_{i1}^{(1)}\,a_{1i}^{(k-1)} + \cdots + a_{iN}^{(1)}\,a_{2N}^{(k-1)}$$
$$\vdots$$
$$a_{i1}^{(k)}\,a_{1i}^{(0)} + \cdots + a_{iN}^{(k)}\,a_{2N}^{(0)}$$

This means that the paths that go from i to j in k steps are counted $k+1$ times. Therefore the term A can be written as

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

Now we analyze the format of term B. We can see that

$$\sum_{l=1}^{N}\sum_{m\neq l}\left(\sum_{n=0}^{\infty} a_{il}^{(n)}\right)\left(\sum_{n=0}^{\infty} a_{mj}^{(n)}\right) =$$
$$\left(\sum_{n=0}^{\infty} a_{i1}^{(n)}\right)\left(\sum_{n=0}^{\infty} a_{2j}^{(n)}\right) + \left(\sum_{n=0}^{\infty} a_{i1}^{(n)}\right)\left(\sum_{n=0}^{\infty} a_{3j}^{(n)}\right) + \cdots + \left(\sum_{n=0}^{\infty} a_{i1}^{(n)}\right)\left(\sum_{n=0}^{\infty} a_{Nj}^{(n)}\right)$$
$$+ \left(\sum_{n=0}^{\infty} a^{(n)}\right)\left(\sum_{n=0}^{\infty} a^{(n)}\right) + \left(\sum_{n=0}^{\infty} a^{(n)}\right)\left(\sum_{n=0}^{\infty} a^{(n)}\right) + \cdots + \left(\sum_{n=0}^{\infty} a^{(n)}\right)\left(\sum_{n=0}^{\infty} a^{(n)}\right)$$

$$+ \left(\sum_{n=0}^{\infty} a_{iN}^{(n)}\right)\left(\sum_{n=0}^{\infty} a_{1j}^{(n)}\right) + \left(\sum_{n=0}^{\infty} a_{iN}^{(n)}\right)\left(\sum_{n=0}^{\infty} a_{2j}^{(n)}\right) + \cdots + \left(\sum_{n=0}^{\infty} a_{iN}^{(n)}\right)\left(\sum_{n=0}^{\infty} a_{(N-1)j}^{(n)}\right)$$

To see what this sum is counting lets consider first the terms for which $l = 1$ and $m = 2$. In the same way that we did for the term *A*, we will analyze first the multiplication whose exponents sum up one.

$$\left[a_{i1}^{(0)} + a_{2j}^{(1)}\right] + \left[a_{i1}^{(1)} + a_{2j}^{(0)}\right]$$

These represent twice the paths that go from *i* to *j* in two steps and have to pass in the edge 1-2. If this edge already existis in the graph, this is just one of its possible paths. However, if this edge does not exist, it is like that we are including a new one in the graph and counting other paths that were not considered before.

Lets know consider the multiplications whose exponents sum up two.

$$\left[a_{i1}^{(0)} + a_{2j}^{(2)}\right] + \left[a_{i1}^{(1)} + a_{2j}^{(1)}\right] + \left[a_{i1}^{(2)} + a_{2j}^{(0)}\right]$$

We see that they represent three times the number of paths of three steps that go from *i* to *j* and pass in the edge 1-2.

In general, for the terms whose exponent sum up *k*, we will count *k* times the number of possible paths of *k* steps from *i* to *j* and that have to pass in edge 1-2.

We just considered here the pair of sites 1 and 2, but the sum is over all possible pairs of sites. Therefore, the term *B* can be written in this way

$$\sum\sum_{k\neq l}^{\infty}{}_{n=0}(k+1)\,(\text{number of paths of n steps from i to j that pass in edge k-l})\,.$$

This means that it counts all possible paths in the graph considering that all possible edges in the graph, in fact, exist. In other words, it counts all possible paths in a complete graph.

Posterior Covariance Matrix

More relevant to the Bayesian data analysis 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

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

and therefore, the covariance matrix is

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

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

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

where

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

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

The same matrix expansion that was used earlier can be applied here. Thus we have

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

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

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

and

$$\begin{bmatrix} m^*_{11} & m^*_{12} & \ldots & m^*_{1N} \\ \vdots & & & \\ m^*_{N1} & m^*_{N2} & \ldots & m^*_{NN} \end{bmatrix}.$$

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

Therefore the posterior covariance matrix can be interpreted in the same way as the prior covariance matrix. The only difference here are the weights appearing in the counts of the possible paths between pairs of areas. While they were equal to $(\lambda_1 + \lambda_2 n_i + \lambda_3 N)^{-1}$ for the prior, they are now equal to $\sigma^2/(\tau_y + \sigma^2(\lambda_1 + \lambda_2 n_i + \lambda_3 N))$. 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.

The specific 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 follows a multivariate normal distribution with mean zero and precision matrix

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

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

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

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

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

Tabela 1: DIC criterion for North Carolina data base using Gamma(0.5, 0.0005) (first row) and Gamma(0.01, 0.01) (second row).

Prior	All comp	Three components	Leroux	BYM
Gamma(0.5, 0.0005)	439.51	438.92	439.94	439.28
Gamma(0.01, 0.01)	438.74	438.58	439.25	440.52

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

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

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 λ close to 1 and also the ratio $(\sigma^2)\tau_y$ between prior and likelihood variances close to one. 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.

Illustrative application

In this section we will consider the analysis of the spatial incidence of sudden infant death syndrome (SIDS) in the 100 counties of the North Carolina state for the period 1999-2006. The spatial pattern of this disease in the period 1974 to 1984 has been previously analyzed by [34, 11, 20, 23], and this early data set is part of many spatial statistics software manuals. Spatial analysis of the underlying risk could hints on unknown risk factors. Indeed there have been found spatial variation of the relative risk with an increasing trend from west to east in the whole U.S.A. 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 to 2006, the North Carolina rate was 0.73 cases per thousand live births.

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. We considered also the particular three-components model, which uses only the identity matrix, the first order neighborhood matrix, and the the matrix $\mathbf{1}\mathbf{1}^T$. We adopted a gamma distribution with parameters equal to 0.5 and 0.0005 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 [30]. 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 BYM model. Although they have similar values, from these results, it is clear that the model with three components is the one that best fits the 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 Table reftab:dicsnc01. The results remained almost the same as before. Again, the best model is the model with three components, while the BYM and Leroux models had the

Tabela 2: Logarithm Score for North Carolina data base using the resampling weight method for Gamma(0.5, 0.0005) (first row) and Gamma(0.01, 0.01) (second row).

Prior	All comp	Three components	Leroux	BYM
Gamma(0.5, 0.0005)	2.24	2.24	6.89	2.48
Gamma(0.01, 0.01)	2.24	2.22	7.56	2.63

worst fits.

The DIC has been criticized as an inadequate measure to evaluate models and it should be considered cautiously ([28]). Therefore, in addition to this global measure, we also calculated a cross-validation posterior predictive distribution check proposed by [31].

We computed the approximated conditional probability ordinate using importance weighting and importance resampling, as it was proposed by [31]. The basic idea of posterior predictive checking is to assess the model's fit to the count in a given area by 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, [31] avoid the refitting of the model using importance weighting and importance resampling to approximate the posterior distribution that would be obtained if the analysis were repeated without the small area. We used their proposal in this paper.

In order to compare the observed CPO's we used a summary measure known as Logarithmic Score [14]. This is a scoring rule. This means that it provides an evaluation of the model regarding forecasts performance, based on the posterior predictive distribution. This measure is calculated as

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

The lower this value, the better the model is fitted. According to [32], that score is asymptotically equivalent to Akaike Information Criterion if the observations are independent.

Table 2 shows the values computed for these measures using the resampling weight method for the two priors considered before. 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. Tables 3 show the results using the method of importance resampling using the same priors. We notice here that again our models, the one with all components and the one with three, were well above the others. It is also noticeable that Leroux model had a poor performance in all the cases.

Tabela 3: Logarithm Score for North Carolina data base using the importance resampling method for Gamma(0.5, 0.005) (first row) and Gamma(0.01, 0.01) (second row).

Prior	All comp	Three components	Leroux	BYM
Gamma(0.5, 0.0005)	2.05	2.05	2.97	2.40
Gamma(0.01, 0.01)	2.05	2.04	2.73	2.38

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 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 [6]. The model we treat in this paper allows for the multiple definition of a smoothing neighborhood. In our model, the λ_j parameters control automatically this smoothing. The model can be specially useful in the situation where the underlying risk is practically constant.

We were able to interpret the conditional distributions involved in our model. The correlation between neighbors depend on the vector of λ_j values and on the graph structure.

We view our model as an additional tool the staticician has available to made inference on 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.

Referências

[1] Assunção, R.M., Potter, J.E., A bayesian space varying parameter model applied to estimating fertility schedules, *Statistics in Medicine*, **14**, 2057-2075, 2002.

[2] Assunção, R.M., Space varying coefficient models for small area data, *Environmetrics*, **14**, 453-473, 2003.

[3] Assunção, R. M., Krainski, E. T., Neighborhood Dependence in Bayesian Spatial Models. *Environmetrics*, 851-869, 2009.

[4] Banerjee, S., Carlin, B. P., Gelfand, A. E., Hierarchical Modeling and Analysis for Spatial Data. *Chapman & Hall/CRC Monographs on Statistics & Applied Probability*, 2004.

[5] Bernardinelli, L., Montomoli, C., Empirical bayes versus fully bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 983-1007, 1992.

[6] Best, N., Richardson, S., Thomson, A., A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research,* 35-39, 2005.

[7] Besag, J., Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics,* **43**, 1-20, 1991.

[8] Brezger, A., Kneib, T., Lang, S., Sudden infant death syndrome, 325-337, 2003.

[9] Eberly, L. E., Carlin, B. P., Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine,* 2279-2294, 2000.

[10] Carlin, B., Banerjee, S., Hierarchical multivariate car models for spatio-temporally correlated survival data, *Bayesian Statistics,* **7**, 45-63, 2003.

[11] Cressie, N., Statistics for spatial data. *New York: John Wiley & Sons,* 1991.

[12] Gelfand, A.E., Hyon-Jung, K., Sirmans, C., Sudipt, B., Spatial modeling with spatially varying coefficient processes, *Bayesian generalized additive mixed models. A simulation study,* **98**, 2057-2075, 2001.

[13] Gelfand, A.E., Vounatsou, P.,Proper multivariate conditional autoregressive models for spatial data analysis, *Biostatistics,* **4**, 11-25, 2003.

[14] Gneiting, T., Raftery, A. E., Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association,* 359-378, 2007.

[15] Held, L., Natario, I., Fenton, S., Rue, H., Becker, N., Towards joint disease mapping, *Statistical Methods in Medical Research,* **14**, 61-82, 2005.

[16] Held, L., Graziano, G., Frank, C., Rue, H., Joint spatial analysis of gastrointestinal infectious diseases, *Statistical Methods in Medical Research,* **15**, 465-480, 2006.

[17] Hunt, C., Hauck, F., Sudden infant death syndrome, *Nelson Textbook of Pediatrics,* 1736-1742, 2007.

[18] Iosifescu M., Finite Markov Processes and Their Applications. *Chichester: John Wiley and Sons,* 1980.

[19] Knorr-Held, L., Best, N., A shared component model for detecting joint and selective clustering of two diseases, *Journal of the Royal Statistical Society, Series A,* **164**, 73-85, 2000.

[20] Kulldorff, M., A spatial scan statistic. *Communications in Statistics-Theory and methods,* 1481-1496, 1997.

[21] Jin, X., Carlin, B.P., Multivariate parametric spatiotemporal models for county level breast cancer survival data, *Lifetime Data Analysis,* **11**, 5-27, 2005.

[22] N, T. (2017). Mau nanya dong dok: Kanker Payudara: Informasi, Penyebab, Gejala, stadium dan Pengobatan. [online] Mau nanya dong dok. Available at: https://nanyadongdok.blogspot.com/2017/06/kanker-payudara-informasi-penyebab.html [Accessed 2 Jul. 2015].

[23] Lawson, A. B., Clark, A., Spatial mixture relative risk models applied to disease mapping. *Statististics in Medicine,* 359–370, 2002.

[24] Leroux, B. G., Lei, X., Breslow, N., Estimation of disease rates in small areas: A new mixed model for spatial dependence. *In Statistical Models in Epidemiology; the Environment and Clinical Trials,* 179-192,1999.

[25] Lunn, D., Thomas, A., Best, N., Spiegelhalter, D., Winbugs - a bayesian modelling framework: concepts, structure, and extensibility, *Statistics and Computing,* **10**, 325-337, 2000.

[26] MacNab, C., Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models, *Statistics in Medicine,* **19**, 2421-2435, 2000.

[27] Martínez-Beneito, M.A., López-Quilez, A., Botella-Rocamora, P., An autoregressive approach to spatio-temporal disease mapping *Statistics in Medicine,* **27**, 2874-2889, 2008.

[28] Plummer, M., Penalized loss functions for Bayesian model comparison. *Biostatistics,* 523-539, 2008.

[29] Silva, L.G, Dean, C.B., Niyonsenga, T., Vanasse, A., Hierarchical bayesian spatio-temporal analysis of revascularization odds using smoothing splines, *Statistics in Medicine,* **164**, 2381-2401, 2008.

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

[31] Stern, H. S., Cressie, N., Posterior predictive model checks for disease mapping models. *Statistics in Medicine,* 2377-2397, 2000.

[32] Stone, M., An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological),* 44-47, 1977.

[33] Sun, D.C., Tsutakawa, R.K., Kim, H., He, Z.Q., Spatio-temporal interaction with disease mapping, *Statistics in Medicine,* **19**, 2015-2035, 2000.

[34] Symons, M. J., Grimson, R. C., Yuan, Y. C., Clustering of rare events. *Biometrics,* 193-205, 1983.

[35] White, G., Ghosh, S. K., A stochastic neighborhood conditional autoregressive model for spatial data. *Computational Statistics and Data Analysis,* 3033-3046, 2009.