ERICA CASTILHO RODRIGUES

ESTRUTURA DE COVARIÂNCIA EM MODELOS BAYESIANOS ESPACIAIS

Belo Horizonte
07 de dezembro de 2012

ERICA CASTILHO RODRIGUES
Orientador: Renato Martins Assunção

ESTRUTURA DE COVARIÂNCIA EM MODELOS BAYESIANOS ESPACIAIS

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Estatística.

ERICA CASTILHO RODRIGUES

Belo Horizonte
07 de dezembro de 2012

ERICA CASTILHO RODRIGUES
Advisor: Renato Martins Assunção

ESTRUTURA DE COVARIÂNCIA EM MODELOS BAYESIANOS ESPACIAIS

Thesis presented to the Graduate Program in Estatística of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Estatística.

ERICA CASTILHO RODRIGUES

Belo Horizonte
December 7, 2012

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Estrutura de Covariância em Modelos Bayesianos Espaciais

ERICA CASTILHO RODRIGUES

Tese defendida e aprovada pela banca examinadora constituída por:

Ph. D. Renato Martins Assunção – Orientador
Universidade Federal de Minas Gerais

Ph. D. Rosângela Helena Loschi
Universidade Federal de Minas Gerais

Ph. D. Marcos Oliveira Prates
Universidade Federal de Minas Gerais

Ph. D. Alexandre Loureiros Rodrigues
Universidade Federal do Espírito Santo

Ph. D. Thaís Cristina Oliveira da Fonseca
Universidade Federal do Rio de Janeiro

Belo Horizonte, 07 de dezembro de 2012

Acknowledgments

Agradeço primeiramente a Deus, por me permitir terminar mais essa etapa.

Não poderia deixar de agradecer em primeiro lugar aos meus pais, Ricardo e Marcia. À minha mãe, pela paciência em me ouvir todos os dias, por participar em todos os momento, sempre me confortando e me incentivando. Ao meu pai, por sempre acreditar em mim, por sempre me ajudar em tudo que fosse possível. À Gabriela, pela compreensão e pela torcida sempre constante. Gostaria de agradecer também ao Taciano, pelo amor e incentivo incondicionais. Por sempre me ouvir, me dar conselhos e por sempre compartilhar comigo cada momento dessa jornada. Não poderia deixar de agradecer também a todos de sua família, Mariana, Arnaldo, Huascar e Glenda, muito obrigada por tudo. Agradeço às minhas grandes amigas, Poliana, Natália e Sarah, pelos ótimos momentos compartilhados e por sempre torcerem por mim.

Devo agradecer também a todos os amigos da UFMG. À Vanessa, com seu jeito meigo e generoso, quantos momentos de sufoco não passamos juntas. Às meninas do Leste, Aline, Letícia, Thaís, Márcia, Paola pelas horas de trabalho tão agradáveis ao lado de vocês. Pelos muitos momentos de descontração e por fazerem meu dia bem mais divertido. Ao João Vitor, pelo pouco tempo que trabalhamos juntos. Ao Marquinhos por toda a ajuda acadêmica e pelo bom humor a todo momento. Gostaria também de agradecer a todos os mestres do Departamento de Estatística. Em especial à Rosângela e Denise, pelo incentivo sempre e pelo tanto que me ensinaram. Não posso deixar de agradecer também aos mestres do Departamento de Matemática, Bernardo e Remy, que acreditaram em mim bem no comecinho e me iniciaram na vida acadêmica. Gostaria também de agradecer ao Renato pela orientação, sempre firme, segura, por me mostrar com é importante amar o que se faz. Pela paciência em entender minhas dificuldades e por me motivar a superá-las a todo momento. Gostaria de agradecer também aos professores Alexandre Loureiros, Marcos Oliveira, Thaís Fonseca e Wagner Meira pelos comentários valiosos para melhoria deste trabalho.

Resumo

No mapeamento de doenças, é necessário especificar uma estrutura de vizinhança para fazer inferências sobre a distribuição geográfica dos riscos relativos. Essa estrutura pode ser usada para modelar a dependência espacial dos dados. Um ponto importante é como modelar essa dependência, qual tipo de covariância será definida entre os pares de áreas. Neste trabalho é feita uma análise da estrutura de covariâncias para dados de áreas. Em partes desse trabalho novos modelos são propostos e, em outras, modelos presentes na literatura são analisados cuidadosamente. Em um contexto um pouco diferente, a dependência entre os dados pode ser utilizada para recuperar outros tipos de informação, como, por exemplo, a localização dos eventos. Resolvemos esse tipo de problema para o caso específico de uma rede social, o *Twitter*. As arestas do grafo agora não representam mais vizinhança geográfica, mas sim relações de amizades entre os usuários. Mostramos como essa informação associada ao tipo de publicação de cada usuário pode ser utilizada para inferir sua localização.

Palavras-chaves: Mapeamento de doenças; Campos Aleatórios de Markov; Modelos Hierárquicos Espaciais; Naive Bayes; Twitter.

Abstract

In disease mapping, one must specify a neighborhood structure to make inferences about the geographic distribution of the relative risks. This structure can be used for modeling spatial dependence of the data. An important point is how to model this dependence, which type of covariance is defined between pairs of areas. This work presents an analysis of covariance structure for area data. In parts of this work new models are proposed and in other parts some models proposed in the literature are analyzed carefully. In a somewhat different context, the dependency between data can be used to retrieve other kind of information, such as the location of events. We solve this kind of problem for the specific case of a social network, *Twitter*. The edges of the graph no longer represent geographical neighborhood, but friendships relationships among users. We show how this type of information associated with the publication of each user can be used to infer their location.

Key words: Disease Mapping ; Markov Random Fields, Hierarchical Spatial Models; Naive Bayes; Twitter.

Resumo Estendido

O mapeamento das taxas de doenças é de grande importância em estudos de epidemiologia. Essa técnica permite analizar como determinada doença se espalha na região em estudo, levando em conta características ambientais, socio-econômicas, genéticas e fatores como o risco de contágio. Dessa maneira, torna possível a identificação de locais com risco elevado e que, portanto, devem sofrer intervenção por agentes de saúde. Elliott e Wartenberg (2004) [26] fazem uma revisão sobre o assunto, apresentam alguns problemas relacionados a esse tipo de análise, bem como técnincas que vem sendo desenvolvidas.

O procedimento que é adotado usualmente consiste em mapear o estimador de máxima verossimilhança dessa quantidade, que é denominado SMR (*Standardized Mortality Ratio*), sobre as diferentes regiões geográficas. Esse estimador é dado pela razão entre o número de casos da doença observados naquela área, pelo número que seria esperado caso não houvesse nenhuma estrutura espacial, ou seja, se o número de casos dependesse apenas da população sob risco.

Porém, quando a doença é rara e a população é muito pequena, essas estimativas são muito instáveis e acabam causando distorções na apresentação do mapa. Isso torna inviável a visualização de padrões de distribuição da doença. A solução encontrada para esse problema é tentar suavizar essa estimativa. Essa suavização pode ser feita utilizando-se Modelos Bayesianos Hierárquicos. Nesse tipo de modelagem, supõe-se que as contagens da doença seguem uma distribuição de Poisson com uma média que é igual ao risco relativo da área multiplicado pelo número de casos esperados na região sob a hipótese de risco constante. O logaritmo do risco relativo é modelado por uma normal cuja média pode depender de algumas covariáveis e soma-se a essa média um erro aleatório que apresenta uma estrutura espacial. Isso é feito colocando uma distribuição a *priori* que reflita tal variabilidade. Essa estrutura deve comportar aspectos que são semelhantes entre áreas próximas no mapa, como características ambientais, econômicas e sociais. Tais distribuições a *priori* podem ser definidas como Campos Markovianos. Isso siginifica que a distribuição condicional de uma área dadas as suas vizinhas não depende do resto do mapa. Em outras palavras, áreas que não são vizinhas são condicionalmente independentes. No caso em que as variáveis aleatórias têm distribuição normal, essa estrutura de indepedência condicional é definida através da inversa da matriz de precisão, ou seja, a inversa da matriz de covariâncias da distribuição normal multivariada. Esses tipos de modelos são conhecidos como Campos Markovianos Gaussianos e um detalhado estudo sobre esse assunto pode ser encontrado em Rue et at. (2005) [63]. Os termos não nulos

da matriz de precisão correspondem a arestas que ligam áreas condicionalmente dependentes e as entradas da diagonal fornecem as variâncias condicionais. Nesse contexto, um ponto importante é a especificação da estrutura de vizinhança. Essa estrutura é responsável por induzir certos tipos de relação de dependência entre as áreas.

Neste trabalho analisamos esse tipo de dependência em determinados contextos e verificamos como ela se comporta para alguns modelos específicos. Apresentamos uma maneira alternativa de se definir a estrutura de vizinhança espacial e mostramos quais as consequências dessa definição para a matriz de covariâncias induzida. Mostramos ainda que determinados tipos de modelos podem apresentar interpretações até o momento não evidentes, mas que podem ajudar na análise de resultados em aplicações práticas. Analisamos ainda casos em que a definição desse tipo de modelo leva a resultados não intuitivos que devem ser analisados com cuidado para evitar que modelos sem sentido prático sejam ajustados a diversas bases de dados sem muita cautela.

Lidamos ainda com um tipo de problema em que a variável para a qual se deseja se fazer inferência é a própria localização. Apenas parte da base de dados possui essa informação disponível e queremos realizar uma imputação para os casos em que ela não está presente. Essa imputação é feita considerando a estrutura de dependência entre os dados, que é modelada utilizando-se um grafo. Mais uma vez, um ponto importante é como incorporar no modelo essa dependência entre os dados e utilizar essa informação de maneira adequada para realizarmos nossas estimações. Uma complicação adicional neste segundo tipo de problema é o grande volume de dados que estamos tratando. Os dados se referem a usuários de redes sociais, o que significa que o número de nós do grafo possui uma ordem de grandeza muito superior àquela quando estamos trabalhando com mapeamento de doenças. Dessa maneira, os métodos propostos devem ser simples o suficiente a fim de não tornar o processo inferencial inviável computacionalmente.

Neste trabalho apresentamos soluções para os problemas acima descritos e mostramos através de resultados práticos e teóricos quais as vantagens e os pontos fracos das metodologias propostas. O texto está dividido em quatro artigos independentes entre si que serão brevemente descritos a seguir.

O primeiro deles é intitulado *Bayesian spatial models with a mixture neighborhood structure* e foi publicado em Agosto de 2012 no *Journal of Multivariate Analysis*. Nesse artigo apresentamos um modelo para dados espaciais no qual a estrutura de vizinhança é parte do espaço paramétrico. Mantemos a propriedade markoviana dos modelos Bayesianos espaciais: dado o grafo de vizinhança, a taxa de incidência de uma doença segue um modelo condicional autoregressivo. Definimos a matriz de precisão do modelo de uma maneira mais flexível. Isso permite que o modelo seja capaz de ajustar diversos tipos de estruturas espaciais presentes nos dados. Analisamos as propriedades teóricas do modelo proposto e mostramos que ele apresenta resultados superiores aos demais tanto para casos simulados como para exemplos reais.

O segundo artigo, *Covariance decomposition in multivariate spatial models*, consiste em uma análise cuidadosa de um modelo espacial para dados multivariados. Nesse caso, a resposta

observada em cada área consiste em um vetor de atributos. Define-se então um modelo que seja capaz de captar a dependência espacial entre as áreas, a dependência entre os atributos de cada área e a depenência entre variáveis observadas em áreas distintas. Esse modelo é definido a partir de um conjunto de distribuições condicionais que apresentam um formato intuitivo. Porém as covariâncias marginais induzidas apresentam um formato complexo. Para fins de interpretação de modelos ajustados, é de grande importância entendermos como as correlações marginais se comportam. Nós mostramos neste trabalho que as matrizes de covariâncias a *priori* e a *posteriori* podem ser escritas como uma soma infinita de matrizes. Obtemos interpretações para as componentes que aparecem nessa soma e mostramos através de um exemplo real como tais interpretações podem auxiliar na análise de problemas práticos.

O terceiro artigo, cujo título é *Analisando o modelo espacial de decaimento exponencial*, consiste em uma análise de um modelo espacial econométrico. Esse modelo foi proposto por [49] e consiste em uma tentativa de resolver problemas numéricos adjacentes ao processo de inferência de modelos espaciais para dados de área. Apesar de tornar o processo inferencial mais eficiente computacionalmente, esse modelo possui uma estrutura de correlações marginais pouco intuitiva. Mostramos que ele apresenta comportamentos não intuitivos para diversos tipos de grafos de vizinhança, desde os mais simples até os mais complexos. Em especial, mostramos que para esse modelo podemos ter situações em que as correlações marginal e condicional entre duas áreas podem ter sinais opostos. Esse tipo de comportamento não é observado para modelos clássicos de séries temporais, como os modelos Autoregressivos, e nem para modelos bastante difundidos em estatística espacial, como os Condicionais Autoregressivos. Mostramos esses resultados através de exemplos experimentais e demonstrações analíticas.

No quarto artigo, *Inferindo a localização de usuários do Twitter*, analisamos um outro tipo de dado. Nosso grafo agora é formado por relações de amizades entre usuários de redes sociais, em particular do *Twitter*. O nosso objetivo nesse caso é inferir a localização dos usuários. Alguns trabalhos apresentados na literatura, como Davis et al. (2004) [24], utilizam a rede de amizades para fazer essa inferência. Outros trabalhos , como de Cheng et al. (2012) [19], coletam informação do texto para alcançar esse objetivo. Vamos incorporar as duas informações e usá-las simultâneamente para fazer inferência sobre a localização. Essa inferência é feita a nível de cidade. Portanto vamos atribuir um rótulo a cada usuário do grafo, que corresponde à cidade na qual ele mora. Consideramos que esse rótulos dispostos no grafo seguem um Campo Markoviano. Nesse caso, como a variável resposta é categórica, não é viável utilizar um Campo Gaussiano. Utilizamos, então, o Modelo de Potts que é muito utilizado na área de processamento de imagem e consiste em um Campo Markoviano para o qual a variável resposta corresponde a uma classificação dos sítios. Nesse modelo, a suposição principal é de que áreas conectadas diretamente tenderão a pertencer à mesma classe. Para recuperarmos a informação do texto utilizaremos o método *Naive Bayes*. Esse método classifica objetos com base em um conjunto de caracteríticas. No nosso casso, esse conjunto de características são as palavras postadas pelos usuários. Uma suposição subjacente ao método é de que, dadas as classes dos objetos, tais características são independentes entre si. Unindo essas

duas ferramentas, modelo de *Potts* e *Naive Bayes*, fomos capazes de incorporar informações sobre o texto e a rede no processo de inferência. Mostramos por meio de experimentos que a metodologia proposta atinge taxas de acertos muito superiores aos trabalhos apresentados na literatura.

Contents

List of Figures

List of Tables

Chapter 1

Bayesian spatial models with mixture neighborhood structure

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.

Keywords: Disease mapping; Markov Random Field; Spatial Hierarchical Models.

1.1 Introduction

In disease mapping, the Bayesian model proposed by [12], 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 of disjoint geographical areas composing a partitioned map. Since then, BYM has been extended into several directions to include space-time generalized linear models [52, 58, 43, 71, 67], spatial survival models [17, 39], spatially-varying parameters models [5, 1, 28], and generalized additive models [45]. Multivariate extensions incorporating two correlated sets of spatial effects have also been proposed in recent years [39, 30, 35, 34]. Many of these models can be fit using freely available software such as WinBUGS [51] and BayesX [16].

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 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 [12] 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.

Despite its crucial role in spatial Bayesian models, very few studies have considered different neighborhood structures for disease mapping problems. One notable exception is [52] 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 [74], 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 speficication 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 [7], page 81) and [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 1.2, we introduce the notation and present some models that were proposed previously. In section 3.2 present the definition our model. In Section 1.4, 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 2.6, 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 1.8 with the main conclusions.

1.2 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 $i = 1, ..., N$ that needs to be estimated. Bayesian inference is based on the posterior distribution of $\boldsymbol{\psi} = (\psi_1, ..., \psi_N)$ given by $f(\boldsymbol{\psi}|y_1, ..., y_N) \propto l(y_1, ..., y_N|\boldsymbol{\psi})f(\boldsymbol{\psi})$, where $l(y_1, ..., y_N|\boldsymbol{\psi})$ is the likelihood function and $f(\boldsymbol{\psi})$ is the prior distribution of the parameter vector $\boldsymbol{\psi}$. Conditional on the values $\psi_1, ..., \psi_N$, the values $Y_1, ..., Y_N$ 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

$$\log(\psi_i) = \mu + b_i \tag{1.1}$$

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 [12]. 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

$$\phi_i | \boldsymbol{\phi}_{-i} \sim N\left(\bar{\phi}_i, \frac{\sigma^2}{n_i}\right) \tag{1.2}$$

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 [25]. To fix these problem, [47] 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 (1.1) with the following precision matrix

$$\mathbf{Q} = (\sigma^2)^{-1} \left((1 - \lambda)\mathbf{I} + \lambda \mathbf{R}\right) \tag{1.3}$$

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.

1.3 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 (1.3) 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 (1.1) has a multivariate normal distribution with mean zero and precision matrix given by:

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

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,

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

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 parameters 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

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

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

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

and

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

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 $\text{Corr}(b_i, b_j|\mathbf{b}_{-ij})$ is given by

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

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:

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

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

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

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:

$$\mathbf{Q}' = \frac{1}{\sigma^2}\left(\lambda_1^*\mathbf{I} + \lambda_2^*\mathbf{W}^{(1)} + \lambda_3^*\mathbf{W}^{(2)} + ... + \lambda_{k+1}^*\mathbf{W}^{(k)}\right). \tag{1.4}$$

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

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

Therefore, the two models would be equivalent only if

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

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 (1.4).

1.4 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 neighbor of every other area. The extension to the general case is straightforward.

Considering only three components, our precision matrix reduces to

$$\mathbf{Q} = (\sigma^2)^{-1}\left(\lambda_1\mathbf{I} + \lambda_2\mathbf{R}^{(1)} + \lambda_3\mathbf{R}^{(k)}\right) \tag{1.5}$$

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 (1.5) 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*

$$\mathbf{Q}^{-1} = \sigma^2\mathbf{M}^{-1} + \frac{\sigma^2\lambda_3}{1-\lambda_3\sum_{ij} m_{ij}} [S_{1+}\ S_{2+} \ldots S_{N+}]^T [S_{1+}\ S_{2+} \quad ... \quad S_{N+}] \tag{1.6}$$

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

Proof. From matrix algebra, we know that

$$\left(\mathbf{P} + \mathbf{u}\mathbf{v}^T\right)^{-1} = \mathbf{P}^{-1} - \frac{\mathbf{P}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{P}^{-1}}{1+\mathbf{v}^T\mathbf{P}^{-1}\mathbf{u}}\,, \tag{1.7}$$

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 (1.7), we have that the covariance matrix $\mathbf{Q}^{-1}$ is given by

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

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

$$\mathbf{Q}^{-1} = \sigma^2\mathbf{M}^{-1} + \frac{\sigma^2\lambda_3}{1-\lambda_3\sum_{ij} m_{ij}} [S_{1+}\ S_{2+} \ldots S_{N+}]^T [S_{1+}\ S_{2+} \quad ... \quad S_{N+}]\ . \tag{1.8}$$

$\Diamond$

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 [2], we write

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

where

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

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

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

Proof. A well known linear algebra result ([38], page 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 $(\mathbf{I}-\mathbf{P})^{-1} = \mathbf{I}+\mathbf{P}+\mathbf{P}^2+\mathbf{P}^3+\ldots$. 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

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

since $n_i \leq N$. As $\mathrm{diag}(1/\mathbf{n})\mathbf{A}$ 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 $\lambda_2^l\left[(\mathbf{TA})^{\mathbf{l}}\right]$ tends to zero as l goes to infinity and the matrix expansion is valid.

$\Diamond$

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 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 (1.8), 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 $\mathrm{Cov}(b_i, b_j) = \left[\mathbf{Q}^{-1}\right]_{ij}$ 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 exists 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.

1.5 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

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

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

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

where

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

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:

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

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

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

and

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

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

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 $\left(\lambda_1 + \lambda_2 n_i + \lambda_3 N\right)^{-1}$ 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.

1.5.1 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 (1.3), this model exchanges the first order neighborhood matrix $\mathbf{R}$ of Leroux model by the matrix associated with the exchangeable risks model of [8].

Using (1.7), 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.

1.6 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 [72], [22] (page 386), [44], [46], 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 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. One of the main aims of the 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 [51] 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 3.2. We considered also 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 [68].

The DIC values are presented in the first row of Table 1.1. 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 1.1. 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 [62]. Therefore, in addition to this global measure, we also calculated a cross-validation posterior predictive distribution check proposed by [69]. 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-

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

Table 1.1: 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.

validation posterior predictive distribution of $Y_{i,-i}^{\text{rep}}$ as

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

where $Y_{i,-i}^{\text{rep}}$ 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, [69] 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 [32]. This is a scoring rule providing an evaluation of a model forecasting performance based on the posterior predictive distribution. This measure is calculated as

$$LS = -\frac{\sum_{i=1}^{N}\log(CPO_i)}{N}\,. \tag{1.9}$$

The lower this value, the better the model. According to [70], this logarithm score is asymptotically equivalent to the Akaike Information Criterion if the observations are independent.

Table 1.1 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 1.1 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 [57], 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 [57] suggested that a QQ-plot as a diagnostic tool for model checking. Figure 1.1 shows the QQ-plot for each of the models using a Gamma$(0.5, 0.0005)$ as a hyperprior while Figure 1.2 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.

1.6.1 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 stands 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 [36]). 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 Figure 1.3. Fitting our model with three components gave virtually the same 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.

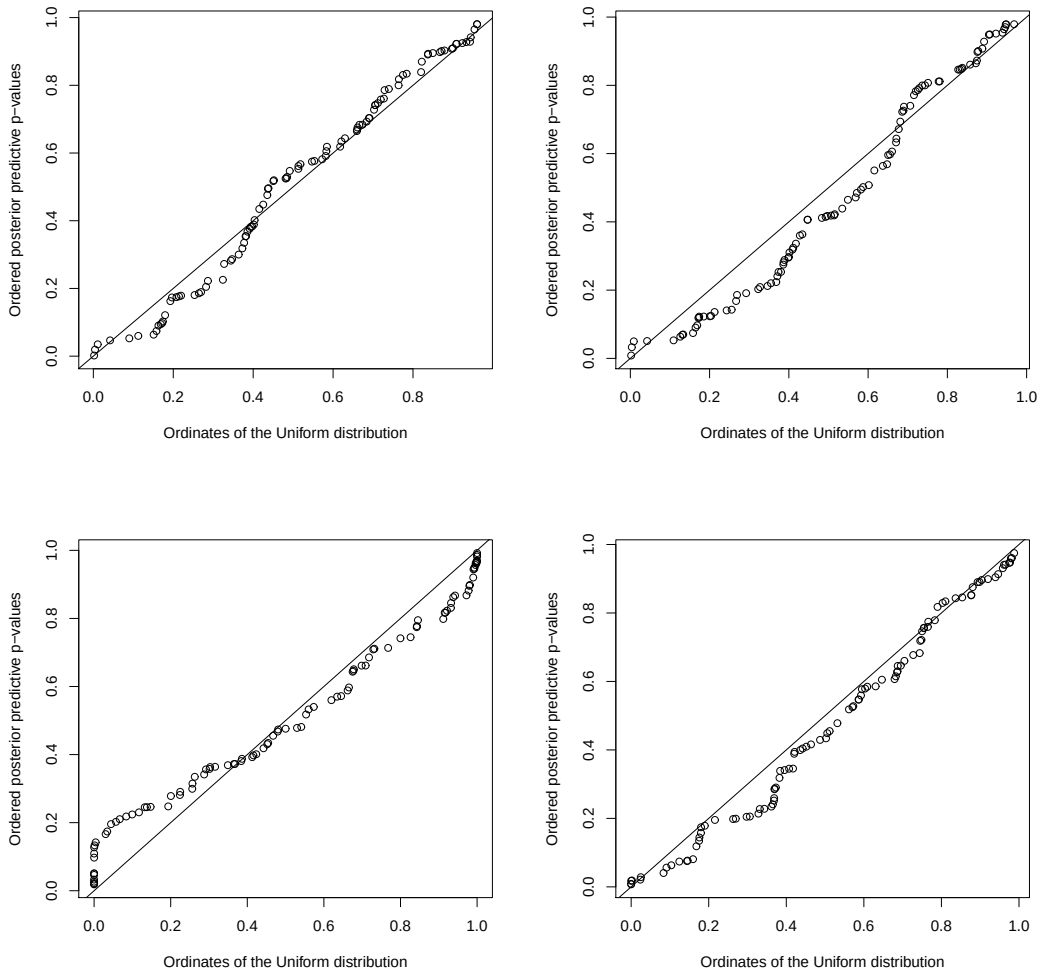

Figure 1.1: QQ-plot of p-values using Gamma(0.5, 0.0005) for all components, three components, Leroux and Bym models. The p-values were calculated using the cross-validation proposal of [57].

1.7 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

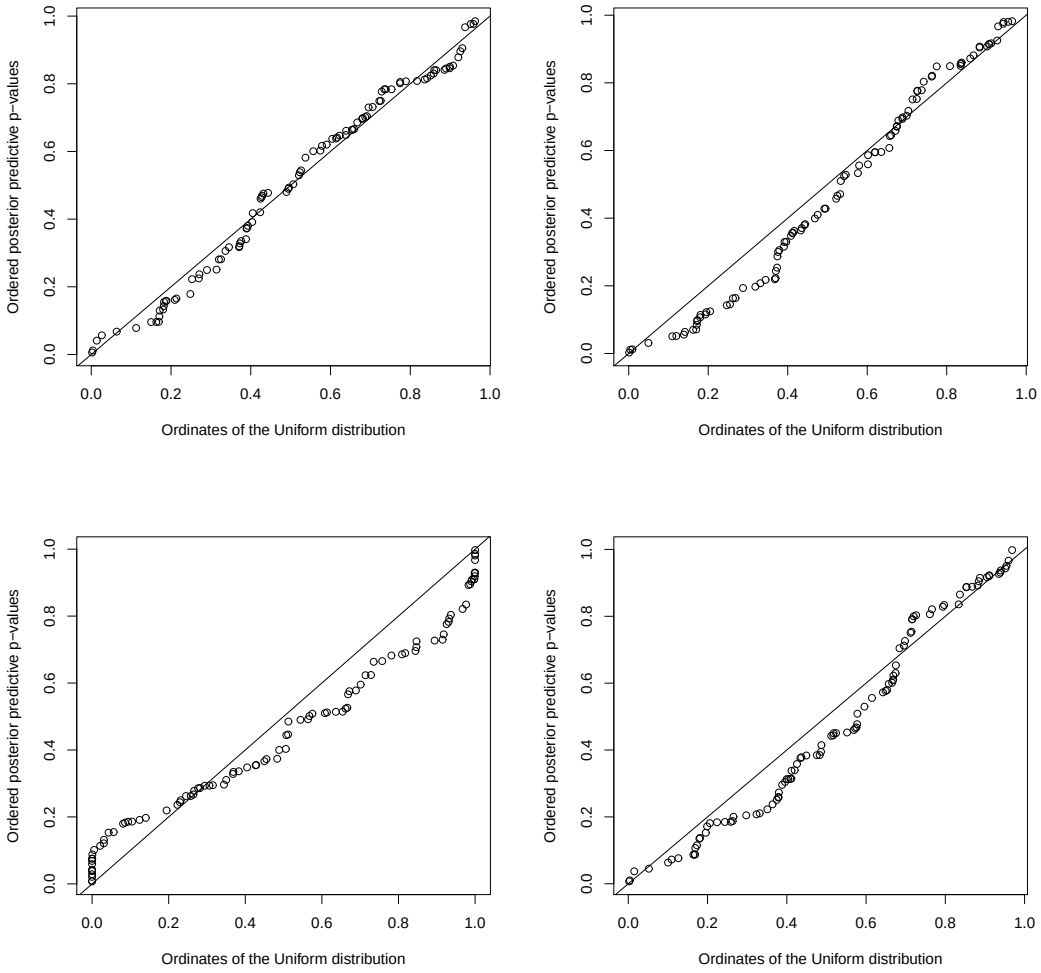

Figure 1.2: QQ-plot of p-values using Gamma(0.01, 0.01) for all components, three components, Leroux and Bym models. The p-values were calculated using the cross-validation proposal of [57].

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 $\lambda_1 = \ldots = \lambda_{19} = (1 - \lambda_{20})/19 = 0.001$.

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

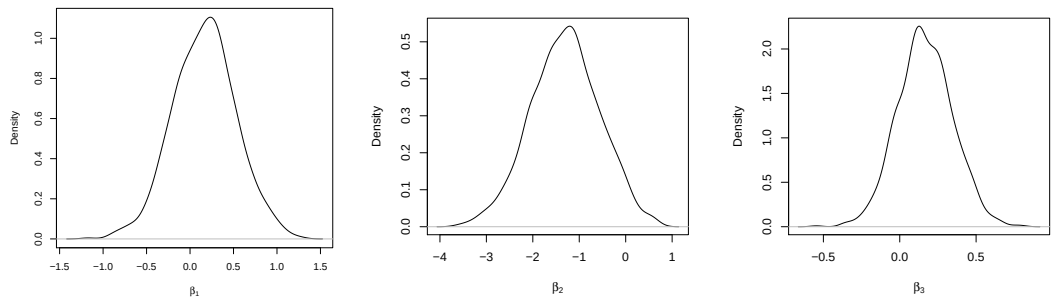

Figure 1.3: Posterior density of the β coefficients for three covariates using our model with all components for the spatial random effect. The first plot refers to the proportion of Black and American Indians in the population, the second plot refers to proportion of mothers who had prenatal care and the third plot refers to the proportion of mothers who were smokers.

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 1.5.1), 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 1.2 shows the values of these evaluation measures for each one of the four possible models in each scenario. Considering the $\overline{MSE}$ criterium, our models are 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

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

Table 1.2: $\overline{MSE}$, DIC, and LS for the simulation of the six scenarios. Summary statistics are the average of 10 independent replications of each model.

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.

1.8 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 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 [13]. 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. 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.

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.

Chapter 2

Covariance decomposition in multivariate spatial models

Abstract

In this paper, we deal with a Bayesian model for multivariate spatial data which specifies the joint prior distribution for the graph-structured random effects through conditional distributions. The conditional distributions are based on an intuitively simple spatial structure but they entail non-intuitive structures for the prior and posterior marginal covariance matrix which are not amenable to closed-form solutions. We derive explicit expressions in the form of a matrix series for the prior and posterior covariance matrices of the underlying parameters. The terms in this infinite matrix sum are associated with weighted paths in a three-dimensional graph connecting geographical areas and variables. Our analytical expressions decompose the covariance into factors that depend either on the spatial structure or the cross-variable partial correlations. We provide intuitive interpretation for the components appearing in these analytical formulas. We illustrate how this decomposition may be useful to understand the correlations between spatial multivariate variables with an analysis based on Florida counties.

Keywords: Markov random field; Conditional autoregressive model; Bayesian hierarchical model; Lattice data; Conditional independence.

2.1 Introduction

The development of spatial statistical methods had great advances in the last two decades. Major innovations were developed for disease rates mapping, geostatistics, and spatial point process data [6, 23, 27]. The applications are as diverse as in epidemiologic, economic, demographic and environmental problems [59, 48, 4, 21]. In this paper, the initial motivation is this last type of application. The study of environmental problems requires the analysis of spatial and temporal data in widely different geographical and temporal scales. Accordingly, statisticians have developed several models allowing for complex structures that describe the environmental phenomena.

The purely spatial methods for modeling a single response random variable with covariates and spatial dependence have been intensively studied and many of its statistical problems received good solutions. The model proposed by [11] is the main framework when one deals with area (or lattice) data with spatial dependence. However, in many applications we have multivariate responses in each area. An important case is the space-time situation, where the multivariate response is a time series observed at each location. Some models have been proposed to deal with these types of data, but we have not yet reached a satisfactory model and hence there is not clearly dominating models. An excellent survey of the space-time data analysis is the newly released book [23]. Researchers are actively searching for a model capturing the diverse types of structure that may be present in multivariate and space-time models and, at the same time, simple enough to be interpretable and computable.

[64] introduced a promising model for multivariate spatial response in lattice data. In a follow-up paper, [65] adapted this proposal to analyze Regional Climate Model (RCM) data. The RCM outputs maps of several variables associated with a set of geographical locations. The aim of [65] is to develop a multivariate model for these multivariate RCM outputs that takes into account the spatial dependence in the variables. They use a multivariate Markov Gaussian Random Field (MMGRF) based on a graph whose neighborhood structure reflects not only the geographical proximity, but also the crossed dependence between different variables. A previous attempt to use a MMGRF was presented by [56] but this earlier model has several implementation difficulties except in rather simple cases. [65] introduced a new way to model multivariate spatial data. The main idea of the authors is to represent the multivariate lattice as if it was univariate, but with a more complex structure. The most important implication is that the usual theory for univariate Markov random fields [63] can be applied to their graph structure.

The model proposed by [64] and [65] is very general and an important alternative for modeling multivariate data, including space-time data, especially considering the scarcity of flexible models for these more complex situations. However, this model inherits most of the difficulties associated with the Markov random field (MRF) models. They both define the joint distribution through a set of conditionally specified distributions that depend heavily on the neighborhood structure. As usual in graphical models, the neighborhood graph reflects the conditional independence structure. It allows to reduce substantially the joint distribution complexity by specifying the zeros of the precision matrix in a simple and intuitive way, connected to the spatial dependence. However, the effect that the neighborhood structure and other parameters have on the implied covariance structure is not well understood and many puzzling results have been reported [73, 10, 6, 63, 3].

The MMGRF has the same difficulty: although its partial correlation matrix is sparse and simple, the correlation matrix is dense and it is not intuitive. Many times, the covariance structure is more easily understood by the user and hence it is important to know what are the implied consequences of assuming a certain precision matrix for the spatial multivariate model. The intuitive gap between the simple precision structure and the induced covariance matrix is larger when the MMGRF is used as a model for richly structured unobserved random

effects. Since this the main use for the MMGRF, the hiatus between the conditionally specified distributions and the covariance structure becomes a severe problem for the practice of data analysis.

Recently, there are at least two previous successful attempts to understand the induced covariance matrix of conditionally specified models. [40] showed how to decompose the covariance matrix associated with a graphical model, not necessarily with spatial content, in more intuitive terms. Their approach is more difficult to apply to MMRGF seen as prior models for random effects. In the purely spatial case, [3] showed that the covariance matrix of MRF models is associated with connected paths between areas in the map. Their approach is more amenable to the analysis of the joint posterior covariance and it is the one we follow in this paper.

Our objective in this work is to study the covariance structure induced by multivariate Markov Gaussian Random field models, focusing on the model proposed by [65]. We show that the covariance between two random variables in the multivariate spatial field can be decomposed in terms of the paths between variables in the three-dimensional multivariate spatial neighborhood graph underlying the model. The decomposition is a weighted sum of these paths and these weights are determined by the intervening variables in the three-dimensional graph.

The manuscript is organized as follows. In Section 2.2, we present the model introduced by [65] and introduce changes that are important to the correct interpretation of the general multivariate spatial cases. In the next section, we derive explicit expressions for the prior and posterior covariance structure when the MMGRF model is used as a prior distribution for random effects. In section 2.6, we use the spatial structure of the counties in Florida to illustrate how our decomposition can be useful for data analysis. In section 2.7, we present the conclusions and the main considerations.

2.2 Model Definition

Denote the j-th variable measured at the i-th area by y_{ij}, with $i = 1, \ldots, n$, and $j = 1, \ldots, p$. The MMGF proposed by [65] is a Gaussian model for the np random variables Y_{ij} collected in a single vector as

$$\begin{aligned} \boldsymbol{Y} &= (\boldsymbol{Y}_{1\cdot}, \ldots, \boldsymbol{Y}_{n\cdot}) \\ &= (Y_{11}, \ldots, Y_{1p}, Y_{21}, \ldots, Y_{2p}, \ldots, Y_{n1}, \ldots, Y_{np}) \ . \end{aligned} \tag{2.1}$$

Since $\boldsymbol{Y}$ has a multivariate normal distribution, it suffices to specify the first two multivariate moments. Without loss of generality, the mean can be assumed equal to zero. Let $\boldsymbol{\Sigma}$ be the $np \times np$ covariance matrix with elements $\boldsymbol{\Sigma}_{ij,kl} = \mathrm{Cov}(Y_{ij}, Y_{kl})$. We will denote by $\boldsymbol{Y}_{-S}$ the vector with all random variables in $\boldsymbol{Y}$ except those in the index set S. The Markov Gaussian Random fields are specified through the precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$, which is typically sparse. The reason for imposing a large number of zeros in this precision matrix is the property that

Y_{ij} is conditionally independent of Y_{kl} given all the other values in $\boldsymbol{Y}_{-\{ij,kl\}}$ if, and only if, $\boldsymbol{Q}_{ij,kl} = 0$ [see 63, p. 21]. Indeed, from the elements of $\boldsymbol{Q}$, we can obtain the partial correlation $\text{Corr}(Y_{ij}, Y_{kl}|\boldsymbol{Y}_{-\{ij,kl\}}) = -\boldsymbol{Q}_{ij,kl}/\sqrt{\boldsymbol{Q}_{ij,ij}\boldsymbol{Q}_{kl,kl}}$.

In the univariate case, every Markov random field has an associated graph that can be represented in the plane with each node standing for a random variable measured in a geographical area. Typically, the nodes are placed in the geographical centroids of the corresponding areas. There is an edge connecting two areas if, and only if, their partial correlation is non-null. In the multivariate case, we can imagine a stack of maps, one for each variable. The combination of an area and a variable is called a site in this paper. In Figure 2.1 we represent a spatial dataset with two variables measured in each area. This is represented by two regular lattices, one for each variable. The sites can be seen as nodes of a three-dimensional graph. As usual in Markov random fields, the partial correlation between y_{ij} and y_{kl} is represented by the presence or absence of edges connecting these sites. That is, in the three-dimensional graph, two sites are not connected by an edge if, and only if, they are independent of each other conditioned on the values of the remaining sites. These edges identify the neighborhood structure necessary to specify a MMGF.

[65] included three types of neighborhood between the sites, as illustrated in the Figure 2.1. The first type, called within-variable neighborhood type, represents conditional dependence between measurements of the same variable measured in two geographically close areas. In principle, it is possible to allow within-variable neighborhoods to change between variables. However, in practice, this is taken to be the same for all variables. The left hand side plot shows the edges corresponding to this first neighborhood type. There are no edges connecting sites of different variables (or maps) in this type of neighborhood. These edges will appear in the next two neighborhood types. The second type, called within-location neighborhood type, captures the conditional dependence between two distinct variables in the same area in the map. The center graph illustrates this neighborhood type, with the i-th area in the two different maps connected by an edge. We show also the edges of the first neighborhood type with faint lines. Finally, the third type represents the conditional dependence between two distinct variables in geographically close areas and it is called cross-variable neighborhood type. The right hand side plot shows this third type of neighborhood between y_{ij} in the j-th variable map and its neighbors y_{kl} in the l-th variable map, with $i \neq k$.

The MMGF proposed by [65] specify the conditional expectation in the following way:

$$E(Y_{ij}|\mathbf{Y}_{-ij}) = \mu_{ij} + \sum_{k\neq i} b_{ijkj}(y_{kj} - \mu_{kj}) + \sum_{l\neq j} b_{ijil}(y_{il} - \mu_{il}) + \sum_{k,l\neq i,j} b_{ijkl}(y_{kl} - \mu_{kl})\,. \tag{2.2}$$

where b_{ijkl} are parameters that will be specified latter. In this expression, each sum represents one of the three types of dependence, in the same order in which we introduced them. For the conditional variance, the authors simplify the model by assuming that it is constant in

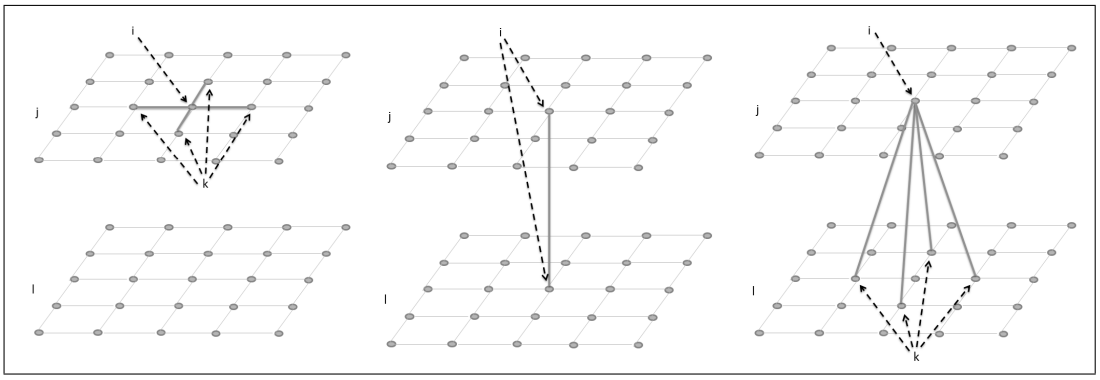

Figure 2.1: Examples of different types of neighborhoods using variables j and l measured in a spatial regular lattice. The left hand side frame shows a within-variable spatial neighborhood, while the middle frame shows a within-location neighborhood. The right hand side frame demonstrates the neighborhood associated with cross-variable connections.

the map of each variable, possibly changing only when we change variables:

$$\mathrm{Var}(Y_{ij}|\mathbf{Y}_{-ij}) = \tau_j^2 \,. \tag{2.3}$$

Additional restrictions must be imposed on the parameters b_{ijkl} and τ_j^2 to yield a symmetric, positive-definite covariance matrix for the joint distribution. [65] assume constraints that make b_{ijkl} independent of i and k. With these restrictions, the conditional distribution specification has some non-intuitive aspects. Firstly, it is not reasonable to assume that the conditional variance is constant in each map. If one tries to predict y_{ij} out of the remaining values $\boldsymbol{y}_{-ij}$ in the graph using the conditional mean (2.2), the prediction variance should depend on the number of neighbors of a given area. The larger the number of neighbors of a given area, the greater the information we have to carry out a prediction. Secondly, in a non-regular grid, a weighted mean of neighboring values must depend on the specific area in consideration. Hence, if we want the conditional expectation (2.2) to represent a weighted mean of the neighboring deviates $y_{kj} - \mu_{kj}$, the coefficients b_{ijkl} must depend on i, which it is not possible in the specification adopted by [65]. This implies that the conditional expectation (2.2) is more like a sum over neighboring areas rather than some kind of weighted mean. To see that this is not reasonable, consider a simple example. Take $\mu_{kj} = 0$ for all k, j, and $b_{ijkl} = 1$ if i and k are geographical neighbors, and equal to zero, otherwise. One area with three other spatial neighboring areas with y values equal to 10 has its conditional expected value equal to 30. If the area has only one neighbor with y value equal to 10, its conditional expected value is equal to 10.

We believe that it is more appropriate to divide both, the conditional expectation and the conditional variance, by the total number of neighboring sites. The case study considered by [65] uses a regular lattice and hence this modification would be unnecessary. However, irregular lattices are very common in applications and our modification is more appropriate in these cases. Let d_i be the number of within-variable spatial neighboring sites of the i-th area in the j-th variable and let p be the number of variable. Then, we define the total number

of neighboring sites of area i as equal to

$$d_i + (p-1) + d_i(p-1) = d_i p + p - 1$$

which is simply the number of edges involving the ij site. In this way, the conditional expectation becomes

$$E(Y_{ij}|\mathbf{Y}_{-ij}) = \frac{1}{d_i p + p - 1}\left(\mu_{ij} + \sum_{k\neq i} b_{ijkj}(y_{kj} - \mu_{kj}) + \sum_{l\neq j} b_{ijil}(y_{il} - \mu_{il}) + \sum_{k,l\neq i,j} b_{ijkl}(y_{kl} - \mu_{kl})\right)$$

In the usual case where b_{ijkl} are binary, this conditional mean can be rewritten as

$$E(Y_{ij}|\mathbf{Y}_{-ij}) = \frac{\mu_{ij}}{d_i p + p - 1} + \frac{d_i \bar{y}_{-i} + (p-1)\bar{y}_{-j} + \bar{y}_{-ij} d_i (p-1)}{d_i + (p-1) + d_i(p-1)}$$

where $\bar{y}_{-i}$ is the mean of $y_{kj} - \mu_{kj}$ for k varying among the d_i within-variable spatial neighbors of the i-th area. The value $\bar{y}_{-j}$ is the mean of $y_{il} - \mu_{il}$ for l varying among the within-location neighboring nodes il of the j-th variable in area i. Finally, $\bar{y}_{-ij}$ is the mean of $y_{kl} - \mu_{kl}$ for the nodes kl that are cross-variable neighbors of node ij. The conditional mean of each site is a convex linear combination of the means of their three types of neighboring sites.

In the same way, the conditional variance is defined as

$$\mathrm{Var}(Y_{ij}|\mathbf{Y}_{-ij}) = \frac{\tau_j^2}{d_i p + p - 1},$$

becoming inversely proportional to the number of neighboring sites.

2.3 Covariance matrix

Let $\boldsymbol{\tau} = (\tau_1^2, \ldots, \tau_p^2)^t$ be a p-dimensional vector and $\mathrm{diag}(\boldsymbol{\tau}) = \mathbf{T}$ a diagonal matrix with the elements of $\boldsymbol{\tau}$. Define also another diagonal matrix with the total number of neighbors of each area, $\mathbf{V} = \mathrm{diag}(\mathbf{v})$ where $\mathbf{v}$ is a vector with i-th element given by $v_i = d_i p + p - 1$. The within-variable neighborhood matrix is defined by

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_{11} & \cdots & \delta_{1n} \\ & \ddots & \\ \delta_{n1} & \cdots & \delta_{nn} \end{bmatrix}.$$

where $\delta_{ik} = 1$, if the areas i and k are geographical neighbors, and $\delta_{ik} = 0$, otherwise. We also set $\delta_{ii} = 0$.

The partial correlation between the within locations random variables Y_{ij} and Y_{il} is given

by

$$\rho_{ijil} = \text{Corr}(Y_{ij}, Y_{il}|\mathbf{Y}_{-ij,il}) = \frac{b_{ijil}(d_i p + p - 1)}{\tau_j^2} \frac{\tau_j \tau_l}{\sqrt{d_i p + p - 1}\sqrt{d_i p + p - 1}} = \frac{b_{ijil}\tau_l}{\tau_j} .$$

To guarantee the symmetry of the precision matrix, [65] assume that this partial correlation does not depend on the area i and denote it simply by $\rho_{ijil} \equiv \rho_{jl}$. This simplification determine the value of b_{ijil} as equal to $\rho_{jl}\tau_j/\tau_l$. This implies that b_{ijil} does not involve the number of neighboring sites and hence (2.2) can not be written as a weighted mean.

Let $\mathbf{A}$ be the $p \times p$ matrix given by

$$\mathbf{A} = \begin{bmatrix} 1 & \cdots & -\rho_{1p} \\ \vdots & \ddots & \vdots \\ -\rho_{p1} & \cdots & 1 \end{bmatrix} .$$

The matrix $\mathbf{A}$ is the same for all areas in the map and it is sometimes called the within-location partial correlation matrix.

Another simplification assumed by [65], sufficient to guarantee the symmetry of the precision matrix $\mathbf{Q}$, is to take $b_{ijkl} = \phi_{jl}\tau_l/\tau_j$ for $i \neq k$ and $\delta_{ik} \neq 0$. For $\delta_{ik} = 0$, we set $b_{ijkl} = 0$. Again, this implies that b_{ijkl} does not depend on the neighboring areas i and k. Let the $p \times p$ matrix containing these terms be defined as

$$\mathbf{B} = \begin{bmatrix} \phi_{11} & \cdots & \phi_{1p} \\ \vdots & \ddots & \vdots \\ \phi_{p1} & \cdots & \phi_{pp} \end{bmatrix} .$$

Our definition of $\mathbf{B}$ differs from [65] by a negative sign, because it simplifies the mathematical expresssions in our paper. [64] and [65] allow $\phi_{jl} \neq \phi_{lj}$ and, therefore, an asymmetric $\mathbf{B}$. However, these authors found posterior estimates for the matrix $\mathbf{B}$ in their applied examples that are essentially symmetric. The right frame of Figure 4 in [65] and the bottom plot of Figure 9 in [64] show kernel density estimates of the posterior marginals for ϕ_{12} and ϕ_{21} with such a large overlap that justifies assuming that they are the same. This symmetric restriction is also present in the models of [14], [42], [61], [18] and [31]. Therefore, from now on, we assume that $\phi_{jl} = \phi_{lj}$.

2.3.1 Blocking by areas

In this section, the precision matrix is the same as that considered by [65]. Its dimension is $np \times np$ and it can be represented by a $n \times n$ block matrix, each block being a $p \times p$ matrix.

The block in position (i,k) refers to areas i and k, where $1 \leq i,k \leq n$ and it is given by:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_1(d_1p+p-1) & -\mathbf{B}_{12}\delta_{12} & \dots & -\mathbf{B}_{1n}\delta_{1n} \\ -\mathbf{B}_{21}\delta_{21} & \mathbf{A}_2(d_2p+p-1) & \dots & -\mathbf{B}_{2n}\delta_{2n} \\ \vdots & \vdots & & \vdots \\ -\mathbf{B}_{n1}\delta_{n1} & -\mathbf{B}_{n2}\delta_{n2} & \dots & \mathbf{A}_n(d_np+p-1) \end{bmatrix} \left[\mathbf{I} \otimes \mathbf{T}^{-1}\right]$$

where

$$\mathbf{A}_i = \begin{bmatrix} 1 & \cdots & -b_{i1ip} \\ & \ddots & \\ -b_{ipi1} & \cdots & 1 \end{bmatrix} \qquad \mathbf{B}_{ik} = \begin{bmatrix} b_{i1k1} & \cdots & b_{i1kp} \\ & \ddots & \\ b_{ipk1} & \cdots & b_{ipkp} \end{bmatrix}$$

We can write

$$(d_ip+p-1)\mathbf{T}^{-1}\mathbf{A}_i = \begin{bmatrix} \frac{d_ip+p-1}{\tau_1^2} & & -\frac{b_{ijil}(d_ip+p-1)}{\tau_1^2} \\ & \ddots & \\ -\frac{b_{ilij}(d_ip+p-1)}{\tau_p^2} & & -\frac{b_{ilij}}{\tau_p^2} \end{bmatrix}.$$

$$= (d_ip+p-1)\left(\mathbf{T}^{-1/2}\right)\mathbf{A}\left(\mathbf{T}^{-1/2}\right)$$

where $\mathbf{A}$ was previously defined.

For the matrices $\mathbf{B}_{ij}$, we have that

$$\mathbf{T}^{-1}\mathbf{B}_{ik} = \mathbf{T}^{-1/2}\mathbf{B}\mathbf{T}^{-1/2}$$

Therefore, the precision matrix is symmetric and it can be written as

$$\begin{bmatrix} (d_1p+p-1)\mathbf{T}^{-1/2}\mathbf{A}\mathbf{T}^{-1/2} & \cdots & -\delta_{1n}\mathbf{T}^{-1/2}\mathbf{B}\mathbf{T}^{-1/2} \\ -\delta_{21}\mathbf{T}^{-1/2}\mathbf{B}^{\mathbf{T}}\mathbf{T}^{-1/2} & \cdots & -\delta_{2n}\mathbf{T}^{-1/2}\mathbf{B}\mathbf{T}^{-1/2} \\ & \vdots & \\ -\delta_{n1}\mathbf{T}^{-1/2}\mathbf{B}^{\mathbf{T}}\mathbf{T}^{-1/2} & \cdots & (d_np+p-1)\mathbf{T}^{-1/2}\mathbf{A}\mathbf{T}^{-1/2} \end{bmatrix}. \tag{2.4}$$

Due to the symmetry of $\mathbf{B}$, we can write the precision matrix as

$$\mathbf{Q} = \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \begin{bmatrix} (d_1p+p-1)\mathbf{A} & & -\delta_{1n}\mathbf{B} \\ & \ddots & \\ -\delta_{1n}\mathbf{B} & & (d_np+p-1)\mathbf{A} \end{bmatrix} \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right)$$

$$\begin{aligned} &= \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right)\left[\mathbf{V} \otimes \mathbf{A} - \boldsymbol{\delta} \otimes \mathbf{B}\right]\left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \\ &= (\mathbf{I} \otimes \mathbf{T}^{-1/2})(\mathbf{V} \otimes \mathbf{A})\left[\mathbf{I} - (\mathbf{V} \otimes \mathbf{A})^{-1}(\boldsymbol{\delta} \otimes \mathbf{B})\right](\mathbf{I} \otimes \mathbf{T}^{-1/2}). \end{aligned}$$

Here we use some results from linear algebra [37, p. 45]. If $\mathbf{P}$ is a square matrix and all elements of the power matrix $\mathbf{P}^k$ tend to zero as k increases, then the inverse $(\mathbf{I}-\mathbf{P})^{-1}$ exists

and it is given by

$$(\mathbf{I}-\mathbf{P})^{-1}=\mathbf{I}+\mathbf{P}+\mathbf{P}^2+\mathbf{P}^3+\dots \tag{2.5}$$

We will apply this matrix expansion to the matrix $\left[\mathbf{I}-(\mathbf{V}\otimes\mathbf{A})^{-1}(\boldsymbol{\delta}\otimes\mathbf{B})\right]^{-1}$. For that, we will verify the necessary conditions to have

$$\lim_{k\to\infty}\left((\mathbf{V}\otimes\mathbf{A})^{-1}(\boldsymbol{\delta}\otimes\mathbf{B})\right)^k=0\,.$$

Given that $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{V})=\mathbf{AC}\otimes\mathbf{BV}$ e $(\mathbf{A}\otimes\mathbf{B})^{-1}=\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$, we can rewrite that matrix as

$$((\mathbf{V}\otimes\mathbf{A})^{-1}(\boldsymbol{\delta}\otimes\mathbf{B}))^k=\left((\mathbf{V}^{-1}\otimes\mathbf{A}^{-1})(\boldsymbol{\delta}\otimes\mathbf{B})\right)^k=(\mathbf{V}^{-1}\boldsymbol{\delta})^k\otimes(\mathbf{A}^{-1}\mathbf{B})^k$$

$$=\begin{bmatrix}\mathbf{0} & & \frac{\delta_{1n}}{d_1p+p-1}\mathbf{A}^{-1}\mathbf{B}\\ & \ddots & \\ \frac{\delta_{n1}}{d_np+p-1}\mathbf{A}^{-1}\mathbf{B} & & \mathbf{0}\end{bmatrix}^k\,.$$

The term $(\mathbf{V}^{-1}\boldsymbol{\delta})^k$ can be rewritten as

$$(\mathbf{V}^{-1}\boldsymbol{\delta})^k=\begin{bmatrix}\frac{\delta_{11}}{d_1} & & \frac{\delta_{1n}}{d_1}\\ & \ddots & \\ \frac{\delta_{n1}}{d_n} & & \frac{\delta_{nn}}{d_n}\end{bmatrix}^k\begin{bmatrix}\frac{d_1}{d_1p+p-1} & & 0\\ & \ddots & \\ 0 & & \frac{d_n}{d_np+p-1}\end{bmatrix}^k$$

The first matrix in the right hand side of this last expression is a Markov matrix and, as a consequence, the elements of its k-th power will be between 0 and 1. When taking the k-th power of the second matrix, it will remain a diagonal matrix with its diagonal elements raised to the k-th power. However, since

$$0<\frac{d_i}{d_ip+p-1}<1 \text{ for all } i$$

then

$$\lim_{k\to\infty}\left(\frac{d_i}{d_ip+p-1}\right)^k=0\,.$$

This means that, if $(\mathbf{A}^{-1}\mathbf{B})^k$ is limited, then $\left((\mathbf{V}\otimes\mathbf{A})^{-1}(\boldsymbol{\delta}\otimes\mathbf{B})\right)^k$ goes to zero and we can use (2.5). Applying this result to the matrix $\left[\mathbf{I}-(\mathbf{V}\otimes\mathbf{A})^{-1}(\boldsymbol{\delta}\otimes\mathbf{B})\right]^{-1}$, we can write it in the following way:

$$\left[\mathbf{I}-\left((\mathbf{V}\otimes\mathbf{A})^{-1}(\boldsymbol{\delta}\otimes\mathbf{B})\right)\right]^{-1}=\sum_{k=0}^{\infty}\mathbf{C}^k$$

where

$$\mathbf{C}=(\mathbf{V}\otimes\mathbf{A})^{-1}(\boldsymbol{\delta}\otimes\mathbf{B}) \tag{2.6}$$

Therefore, the covariance matrix is equal to

$$[\mathbf{I} \otimes \mathbf{T}^{1/2}] \sum_k \mathbf{C}^k [\mathbf{V}^{-1} \otimes \mathbf{A}^{-1}][\mathbf{I} \otimes \mathbf{T}^{1/2}] . \tag{2.7}$$

The matrices multiplying $\sum_k \mathbf{C}^k$ are diagonal matrices or involve only within-location partial correlations. Therefore, the graph structure corresponding to the within-variable and cross-variable neighborhood types is fully concentrated in the matrix $\sum_k \mathbf{C}^k$, to be interpreted in section 2.4.

2.3.2 Blocking by variables

Another way to think about this problem can be analised. We can consider the vector $\boldsymbol{Y}$ ordered in a different way. Let

$$\boldsymbol{Y}^* = (Y_{11}, \ldots, Y_{n1}, Y_{12}, \ldots, Y_{n2}, \ldots, Y_{1p}, \ldots, Y_{np}) . \tag{2.8}$$

In this case, the precision matrix $\boldsymbol{Q}$ is a $p \times p$ block matrix, where each block has dimension $n \times n$. The block in position j, l (where $1 \leq j, l \leq p$) represents the association between variables j and l measured in any pair of neighboring areas. Interpreting the problem by this point of view, we have another way to derive the precision matrix. The results are very similar to those presented in the section 2.3.1, but there is an exchange between the matrices involved in Kronecker product of equation (2.6). Here it tuns out to be $C^* = (\mathbf{A} \otimes \mathbf{V})^{-1}(\mathbf{B} \otimes \boldsymbol{\delta})$.

2.4 Interpreting the decomposition

In this section, we interpret the matrices

$$C = (\mathbf{V} \otimes \mathbf{A})^{-1}(\boldsymbol{\delta} \otimes \mathbf{B}) \qquad C^* = (\mathbf{A} \otimes \mathbf{V})^{-1}(\mathbf{B} \otimes \boldsymbol{\delta})$$

involved in the covariance matrix (2.7) expressed as a power series as well as its equivalent form in the case of variable blocking.

As we saw previously, these terms can be rewritten as

$$C^k = (\mathbf{V}^{-1}\boldsymbol{\delta})^k \otimes (\mathbf{A}^{-1}\mathbf{B})^k \qquad (C^*)^k = (\mathbf{A}^{-1}\mathbf{B})^k \otimes (\mathbf{V}^{-1}\boldsymbol{\delta})^k .$$

Therefore, the matrix expansion is the sum of Kronecker products. In this product, one factor is associated solely with the geographical locations of the areas and it represents the transition matrix of a random walk on the within-variable neighborhood graph. The other factor involved in the product represents association between the p variables, irrespectively of the spatial location. The interaction between the two types of dependence is induced by the Kronecker product of the matrices. To better understand each one of these factors, we write

the general form of the block matrix C^k. The block associated with areas i and h is given by

$$[C^k]_{ih} = [(\mathbf{V}^{-1}\boldsymbol{\delta})^k \otimes (\mathbf{A}^{-1}\mathbf{B})^k]_{ih} = [(\mathbf{V}^{-1}\boldsymbol{\delta})^k]_{ih}(\mathbf{A}^{-1}\mathbf{B})^k \,. \tag{2.9}$$

We deal initially with the term $[(\mathbf{V}^{-1}\boldsymbol{\delta})^k]_{ih}$ of the matrix

$$(\mathbf{V}^{-1}\boldsymbol{\delta})^k = \begin{bmatrix} \frac{\delta_{11}}{d_1} & & \frac{\delta_{1n}}{d_1} \\ & \ddots & \\ \frac{\delta_{n1}}{d_n} & & \frac{\delta_{nn}}{d_n} \end{bmatrix}^k \begin{bmatrix} \frac{d_1}{d_1 p+p-1} & & 0 \\ & \ddots & \\ 0 & & \frac{d_n}{d_n p+p-1} \end{bmatrix}^k . \tag{2.10}$$

Observe that the first matrix in the right hand side is a transition matrix of a random walk. As observed by [3], the terms of the k-th power of this matrix can be seen as weighted sum of all paths from area i to area h in k steps. The paths are those determined by the within-variable neighborhood graph. Each path contributes with a positive term inversely proportional to the product of the number of spatial neighbors of each area comprising the path. Multiplying this matrix by the diagonal matrix in the right hand side of (2.10) simply multiplies each row by a constant. This implies that the block of C^k associated with areas i and h is the product of $(\mathbf{A}^{-1}\mathbf{B})^k$ by a number representing a weighted sum of all possible paths from i to h.

We turn now to the interpretation of the $(\mathbf{A}^{-1}\mathbf{B})^k$ terms of $\mathbf{C}^k$ block associated with areas i and h. Let $\mathbf{A}_i^+$ be the ii-th block of the precision matrix $\mathbf{Q}$ as expressed in (2.4):

$$\mathbf{A}_i^+ = \left(\frac{1}{d_i p + p - 1}\mathbf{T}\right)^{-1/2} \mathbf{A} \left(\frac{1}{d_i p + p - 1}\mathbf{T}\right)^{-1/2} = \left(\mathbf{T}_i^+\right)^{-1/2} \mathbf{A} \left(\mathbf{T}_i^+\right)^{-1/2}$$

and therefore

$$\mathbf{A} = \left(\frac{1}{d_i p + p - 1}\mathbf{T}\right)^{1/2} \mathbf{A}_i^+ \left(\frac{1}{d_i p + p - 1}\mathbf{T}\right)^{1/2}$$

Note that $\mathbf{T}_i^+$ is the $p \times p$ diagonal matrix with partial variances and given by

$$\mathbf{T}_i^+ = \operatorname{diag} \begin{bmatrix} \operatorname{Var}(Y_{i1}|\mathbf{Y}_{-i1}) \\ \vdots \\ \operatorname{Var}(Y_{i1}|\mathbf{Y}_{-ip}) \end{bmatrix} = \frac{1}{d_i p + p - 1} \operatorname{diag} \begin{bmatrix} \tau_1^2 \\ \vdots \\ \tau_p^2 \end{bmatrix}$$

We can also rewrite the matrix $\mathbf{B}$ by pivoting around area i as

$$\mathbf{B} = \left(\mathbf{T}_i^+\right)^{1/2} \mathbf{B}_i^+ \left(\mathbf{T}_i^+\right)^{1/2}$$

where $\left(\mathbf{T}_i^+\right)^{1/2}$ is defined as above and $\mathbf{B}_i^+$ is the i,h-th block of the precision matrix $\mathbf{Q}$ in (2.4) multiplied by $d_i p + p - 1$. Note that $\mathbf{B}_i^+$ is the zero matrix if $\delta_{ih} = 0$.

Therefore we see that

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{B} &= \left(\mathbf{T}_i^+\right)^{-1/2} \left(\mathbf{A}_i^+\right)^{-1} \left(\mathbf{T}_i^+\right)^{-1/2} \left(\mathbf{T}_i^+\right)^{1/2} \mathbf{B}_i^+ \left(\mathbf{T}_i^+\right)^{1/2} \\ &= \left(\mathbf{T}_i^+\right)^{-1/2} \left(\left(\mathbf{A}_i^+\right)^{-1} \mathbf{B}_i^+\right) \left(\mathbf{T}_i^+\right)^{1/2} . \end{aligned}$$

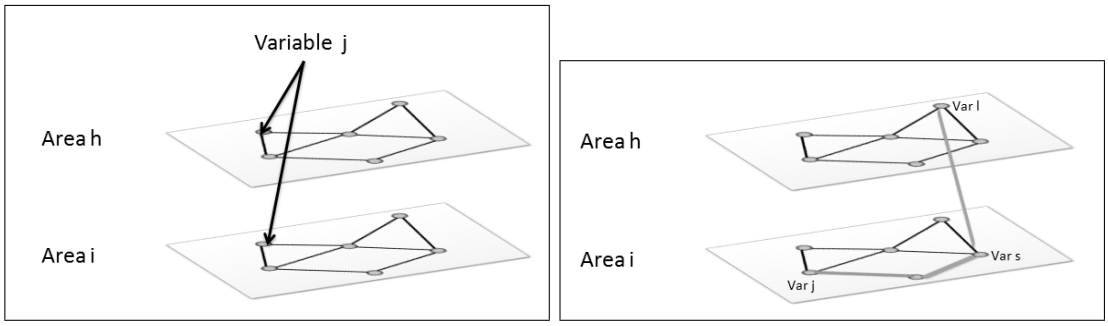

Figure 2.2: Dual graph of areas i and k (right). An example of a possible path between these areas (left).

Raising this matrix to the square we get the following expression

$$\left(\mathbf{A}^{-1}\mathbf{B}\right)^2 = \left(\mathbf{T}_i^+\right)^{-1/2}((\mathbf{A}_i^+)^{-1}\mathbf{B}_i^+)^2\left(\mathbf{T}_i^+\right)^{1/2}$$

and in general

$$\left(\mathbf{A}^{-1}\mathbf{B}\right)^k = \left(\mathbf{T}_i^+\right)^{-1/2}((\mathbf{A}_i^+)^{-1}\mathbf{B}_i^+)^k\left(\mathbf{T}_i^+\right)^{1/2} \ .$$

Now let us look closely to the product $((\mathbf{A}_i^+)^{-1}\mathbf{B}_i^+)$. We have seen that $\mathbf{A}_i^+$ is the precision matrix of the $p \times 1$ vector $\mathbf{Y}_{i.}$, and therefore $(\mathbf{A}_i^+)^{-1}$ is a covariance matrix. [40] showed that the terms j, l of a graphical model covariance matrix can be written as weighted sums of all possible paths between the variables j and l. Using this result, we can interpret the term

$$[(\mathbf{A}_i^+)^{-1}\mathbf{B}_i^+]_{jl} = \sum_{s=1}^{p}\left[(\mathbf{A}_i^+)^{-1}\right]_{js}\left[\mathbf{B}_i^+\right]_{sl} \tag{2.11}$$

as a weighted sum of paths from variable j in area i to variable l in the neighboring area h. In order to facilitate the interpretation, we consider a kind of dual of the original neighborhood graph, where each area has a graph associated with their variables, as illustrated in Figure 2.2 (left). This variable-graph is induced by the matrix $\mathbf{A}$, with edges conecting variables if, and only if, $\rho_{jl} \neq 0$. Therefore, the edge structure of these variable-graphs is the same for all areas in the map.

Consider the possible edges connecting the variable-graph of area i to the nodes of the variable-graph of area h. The random variables Y_{is} and Y_{hl} are linked if, and only if, $\text{Cor}(Y_{is}, Y_{hl}|\mathbf{Y}_{-\{is,hl\}}) \neq 0$ or, equivalently, if, and only if, $\delta_{ih}\phi_{sl} \neq 0$. This is the same as having $[\mathbf{B}_i^+]_{sl} \neq 0$. Therefore, the term $\left[(\mathbf{A}_i^+)^{-1}\right]_{js}\left[\mathbf{B}_i^+\right]_{sl}$ is the product of the weighted sum of all paths from j to s in the variable-graph of area i times a factor that is proportional to $\phi_{sl}\delta_{ih}$. This is the sum of all weigthed paths from Y_{ij} to Y_{hl}, where it is possible to change areas only in the last step. Figure 2.2 (right) shows an example of a possible path.

In conclusion, the product $[(\mathbf{A}_i^+)^{-1}\mathbf{B}_i^+]$ represents the interaction between two variable graphs of neighboring areas. If the areas are not neighbors, the entire matrix is null. When we raise this product to the k-th power we are just counting more paths, but the interpretation is the same.

2.5 Posterior Covariance Matrix

[65] use their multivariate model as a prior distribution for random effects in a hierarchical model to analyze RCM data assuming a normal distribution likelihood. The posterior covariance induced by this prior multivariate model is more relevant to the analyst than the covariance prior distribution. Fortunately, we can obtain explicit forms for the posterior covariance matrix in the normal distribution case, as we explain now. For the first hierarchical level, we will keep similar notation and assumptions as [65]. Let $\mathbf{y}_j$ be independent n-dimensional vectors representing a RCM outputs where j is the j-th variable, $j = 1, \ldots, p$, with the following distribution

$$\mathbf{y}_j | \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \mathbf{h}_j, \sigma_j^2 \sim N(\mathbf{X}_1 \boldsymbol{\alpha}_j + \mathbf{h}_j, \sigma_j^2 \mathbf{I}) \, .$$

In this model, $\mathbf{X}_1 \boldsymbol{\alpha}_j$ is a fixed effect associated with the j-th variable. The spatial effects are included by means of the term $\mathbf{h}_j$. The set of $\mathbf{h}_1, \ldots, \mathbf{h}_p$ follows the multivariate MMGF spatial model defined in section 2.2.

Therefore, conditioned on the vectors $\mathbf{h}_j$, the joint distribution can be written as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_p \end{bmatrix} \sim N\left(m + \mathbf{H}, \mathbf{I} \otimes \operatorname{diag}(\boldsymbol{\sigma^2})\right)$$

where $\mathbf{m} = (\mathbf{X}_1 \boldsymbol{\alpha}_1, \ldots, \mathbf{X}_1 \boldsymbol{\alpha}_p)^t$, $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_p)^t$ and $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_p^2)$.

We can write the posterior precision matrix equal to the sum of the likelihood and the prior precision matrices. Following the same approach of [3], we have

$$\begin{aligned} & \mathbf{I} \otimes \operatorname{diag}^{-1}(\boldsymbol{\sigma^2}) + \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \left[\mathbf{V} \otimes \mathbf{A} - \boldsymbol{\delta} \otimes \mathbf{B}\right] \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \\ = \; & \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \left(\mathbf{I} \otimes \mathbf{T}^{1/2}\right) \left(\mathbf{I} \otimes \operatorname{diag}^{-1}(\boldsymbol{\sigma^2})\right) \left(\mathbf{I} \otimes \mathbf{T}^{1/2}\right) \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \\ & + \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \left[\mathbf{V} \otimes \mathbf{A} - \boldsymbol{\delta} \otimes \mathbf{B}\right] \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \\ = \; & \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \Big[\left(\mathbf{I} \otimes \mathbf{T}^{1/2}\right) \left(\mathbf{I} \otimes \operatorname{diag}^{-1}(\boldsymbol{\sigma^2})\right) \left(\mathbf{I} \otimes \mathbf{T}^{1/2}\right) \\ & + \mathbf{V} \otimes \mathbf{A} - \boldsymbol{\delta} \otimes \mathbf{B}\Big] \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \, . \end{aligned}$$

The matrix $\left(\mathbf{I} \otimes \mathbf{T}^{1/2}\right) \left(\mathbf{I} \otimes \operatorname{diag}^{-1}(\boldsymbol{\sigma^2})\right) \left(\mathbf{I} \otimes \mathbf{T}^{1/2}\right)$ is a diagonal matrix given by

$$\left(\mathbf{I} \otimes \mathbf{T}^{1/2}\right) \left(\mathbf{I} \otimes \operatorname{diag}^{-1}(\boldsymbol{\sigma^2})\right) \left(\mathbf{I} \otimes \mathbf{T}^{1/2}\right) = \mathbf{I} \otimes \operatorname{diag}(\boldsymbol{\sigma^2})^*$$

where $(\boldsymbol{\sigma^2})^* = (1/(\sigma_1^2 \tau_1^2), \ldots, 1/(\sigma_p^2 \tau_p^2)) \, .$

Therefore, the posterior precision matrix becomes

$$\left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \left[\mathbf{I} \otimes \operatorname{diag}(\boldsymbol{\sigma^2})^* + \mathbf{V} \otimes \mathbf{A} - \boldsymbol{\delta} \otimes \mathbf{B}\right] \left(\mathbf{I} \otimes \mathbf{T}^{-1/2}\right) \, .$$

and the posterior covariance matrix turns out be

$$\left(\mathbf{I}\otimes\left(\mathbf{T}^{1/2}\right)\right)\left[\mathbf{I}\otimes \mathrm{diag}(\boldsymbol{\sigma^2})^* + \mathbf{V}\otimes\mathbf{A} - \boldsymbol{\delta}\otimes\mathbf{B}\right]^{-1}\left(\mathbf{I}\otimes\left(\mathbf{T}^{1/2}\right)\right) .$$

Considering the middle matrix

$$\left[\mathbf{I}\otimes \mathrm{diag}(\boldsymbol{\sigma^2})^* + \mathbf{V}\otimes\mathbf{A} - \boldsymbol{\delta}\otimes\mathbf{B}\right]^{-1}$$

we have

$$\left[\mathbf{I}\otimes \mathrm{diag}(\boldsymbol{\sigma^2})^* + \mathbf{V}\otimes\mathbf{A} - \boldsymbol{\delta}\otimes\mathbf{B}\right]^{-1} = \left[\mathbf{I} - \mathbf{U}^{-1}\left(\boldsymbol{\delta}\otimes\mathbf{B}\right)\right]^{-1}\mathbf{U}^{-1}$$

where $\mathbf{U} = \mathbf{I}\otimes \mathrm{diag}(\boldsymbol{\sigma^2})^* + \mathbf{V}\otimes\mathbf{A}$.

If the terms of the matrix $\left(\mathbf{U}^{-1}\left(\boldsymbol{\delta}\otimes\mathbf{B}\right)\right)^k$ tend to zero as k increases, we can use the matrix expansion (2.5) obtaining:

$$\left[\mathbf{I} - \mathbf{U}^{-1}\left(\boldsymbol{\delta}\otimes\mathbf{B}\right)\right]^{-1} = \mathbf{I} + \mathbf{U}^{-1}\left(\boldsymbol{\delta}\right)\otimes\mathbf{B} + \left(\mathbf{U}^{-1}\left(\boldsymbol{\delta}\otimes\mathbf{B}\right)\right)^2 + \ldots$$

Take the $p\times p$ block at the position (i,h) associated with areas i and h of the $np\times np$ matrix $\left[\mathbf{I} - \mathbf{U}^{-1}\left(\boldsymbol{\delta}\otimes\mathbf{B}\right)\right]^{-1}$. This block is weighted sum of matrices associated with the paths from i to h where the weighting factor comes from the matrices $\mathbf{U}^{-1}$ and $\mathbf{B}$. To better understand this interpretation, consider the third term in the expansion when $k=2$:

$$\begin{aligned}
[\left(\mathbf{U}^{-1}\boldsymbol{\delta}\otimes\mathbf{B}\right)^2]_{ih} &= \sum_{l=1}^{n}(\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_i p + p - 1)\mathbf{A})^{-1}\boldsymbol{\delta}_{il}\mathbf{B}(\mathrm{diag}(\boldsymbol{\sigma^2})^* \\
&\quad +(d_l p + p - 1)\mathbf{A})^{-1}\boldsymbol{\delta}_{lh}\mathbf{B} \\
&= \left(\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_i p + p - 1)\mathbf{A}\right)^{-1}\mathbf{B}\left(\sum_{l=1}^{n}\boldsymbol{\delta}_{il}\boldsymbol{\delta}_{lh}\left(\mathrm{diag}(\boldsymbol{\sigma^2})^*\right.\right. \\
&\quad \left.\left.+(d_l p + p - 1)\mathbf{A})^{-1}\right)\mathbf{B}\right)
\end{aligned}$$

The sum in the right hand side of the last equation yields a $p\times p$ matrix that considers all the paths from i to h in two steps. Each of these paths receives a weight given by the $p\times p$ matrix $\left(\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right)^{-1}$ associated with the number d_l of spatial neighbors of the intervening area l. To verify how this weight matrix changes as d_l changes, we take the derivative with respect to $d_l p + p - 1$ considered as a continuous value. Since

$$\frac{\partial\mathbf{P}^{-1}}{\partial x} = -\mathbf{P^{-1}}\frac{\partial\mathbf{P}}{\partial x}\mathbf{P}^{-1} ,$$

we obtain

$$\frac{\partial\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1}}{\partial(d_l p + p - 1)} =$$

$$-\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1}\mathbf{A}\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1} .$$

Observe that the term (u,v) of this matrix is given by

$$\begin{aligned}
&\sum_s \Big[- \left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1} \mathbf{A} \Big]_{us} \Big[\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1} \Big]_{sv} \\
= \; & -\sum_s \sum_r \left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1}_{ur} \mathbf{A}_{ts} \Big[\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1} \Big]_{sv} \\
= \; & -\left(\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1}_{u} \right) \mathbf{A} \left(\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1}_{v} \right)^t
\end{aligned}$$

where $\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]_u$ denotes the u-th row of the matrix. As $\mathbf{A}$ is a principal diagonal block of a definite positive matrix, it is itself a positive definite matrix. Therefore, we know that

$$\left(\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1}_{u} \right)^t \mathbf{A} \left(\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1}_{v} \right) > 0$$

since $\left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]_i \neq 0$. This implies that

$$\frac{\partial \left[\mathrm{diag}(\boldsymbol{\sigma^2})^* + (d_l p + p - 1)\mathbf{A}\right]^{-1}}{\partial (d_l p + p - 1)} < 0 \, .$$

For larger k, the formulas are longer and more complicated but they are analogous to $k = 2$.

In conclusion, the weighting matrix decreases as the number of neighbors $d_l p + p - 1$ increases or, equivalently, the number of within-variable neighbors d_l increases. Therefore, the weighted sum is such that paths going through more densely connected areas receive less weight than paths going through more isolated areas. This also happens in the purely spatial univariate case, as observed by [3].

2.6 An illustrative synthetic example

In this section, we provide an example to show how the decomposition we developed in this paper can be useful in the practice of data analysis. As spatial setting, we took the Florida map partitioned into $n = 68$ counties and $\delta_{ih} = 1$ if, and only if, areas i and h share border (see the map in Figure 2.3). We took $p = 2$ and we adopted

$$\mathbf{A} = \begin{bmatrix} 1 & -\rho_{12} \\ -\rho_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.12 \\ 0.12 & 1 \end{bmatrix} ,$$

for the within-location partial correlation matrix, which is a matrix similar to that estimated by [65] in their studies of the precipitation and temperature variables over the western United States and part of western Canada. For $\mathbf{B}$, we departed from the values estimated by [65]. Their parameters implied on marginal correlations $\mathrm{Cor}(Y_{ij}, Y_{hl})$ between neighboring areas in the same variable no larger than 0.08, a very low value. Hence, to make the example more

interesting, we adopted

$$\mathbf{B} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} = \begin{bmatrix} 0.80 & 0.30 \\ 0.30 & 0.85 \end{bmatrix}.$$

Finally, we fixed $\tau_1 = \tau_2 = 1$. With these choices, we obtained $\mathbf{Q}$ and $\mathbf{\Sigma}$.

Figure 2.3 show the marginal correlations $\text{Cor}(Y_{ij}, Y_{hl})$ among all pairs of random variables at increasing spatial neighborhood order s. That is, $s = 1$ implies that the areas i and h are geographically adjacent, $s = 2$ means that there is one intervening area between i and h, etc. The vertical axis are in the logarithmic scale. Clearly, the correlations decrease linearly (in log-scale) with the spatial neighborhood order, although there is some variability depending on the pair of areas considered. [3] showed how this variation can be understood in the purely spatial case.

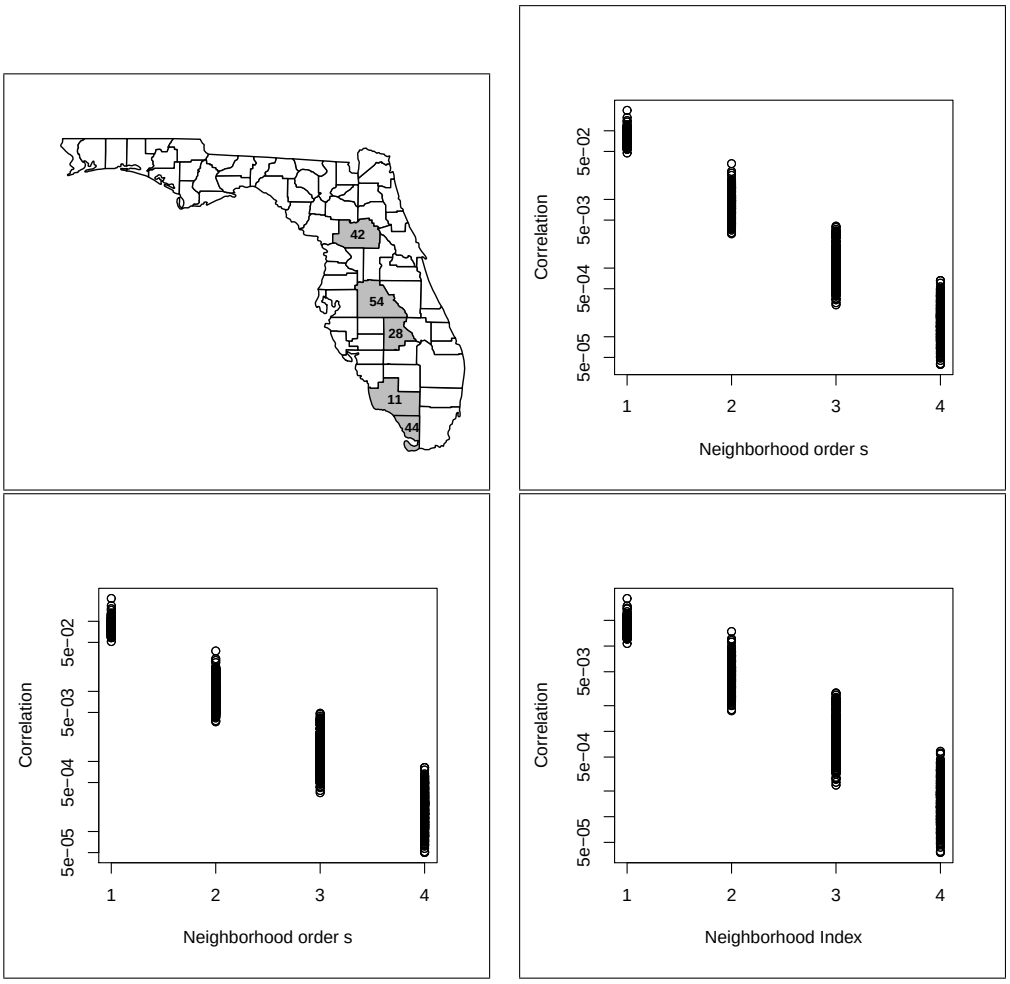

Figure 2.3: Map of Florida counties with some areas identified. The other three plots show the marginal correlations $\text{Cor}(Y_{ij}, Y_{kl})$ (in log-scale) versus the spatial neighborhood order s. In clockwise direction, they represent variables (j, l) equal to $(1, 1)$, $(2, 2)$, and $(1, 2)$.

The usefulness of our covariance decomposition can be appreciated in Figure 2.4. As we saw, the covariance between Y_{ij} and Y_{hl} is determined by the infinite sum $\sum_k [\mathbf{C}^k]_{ih}$ of the i, h-th blocks of the 2×2 matrix $\mathbf{C}^k$ given in (2.9). Each element in this block is the product

of a factor representing the weighted sum of all k-steps paths from area i to area h in the purely geographical neighborhood graph times a factor representing the partial correlation structure between variables without consideration to the spatial structure.

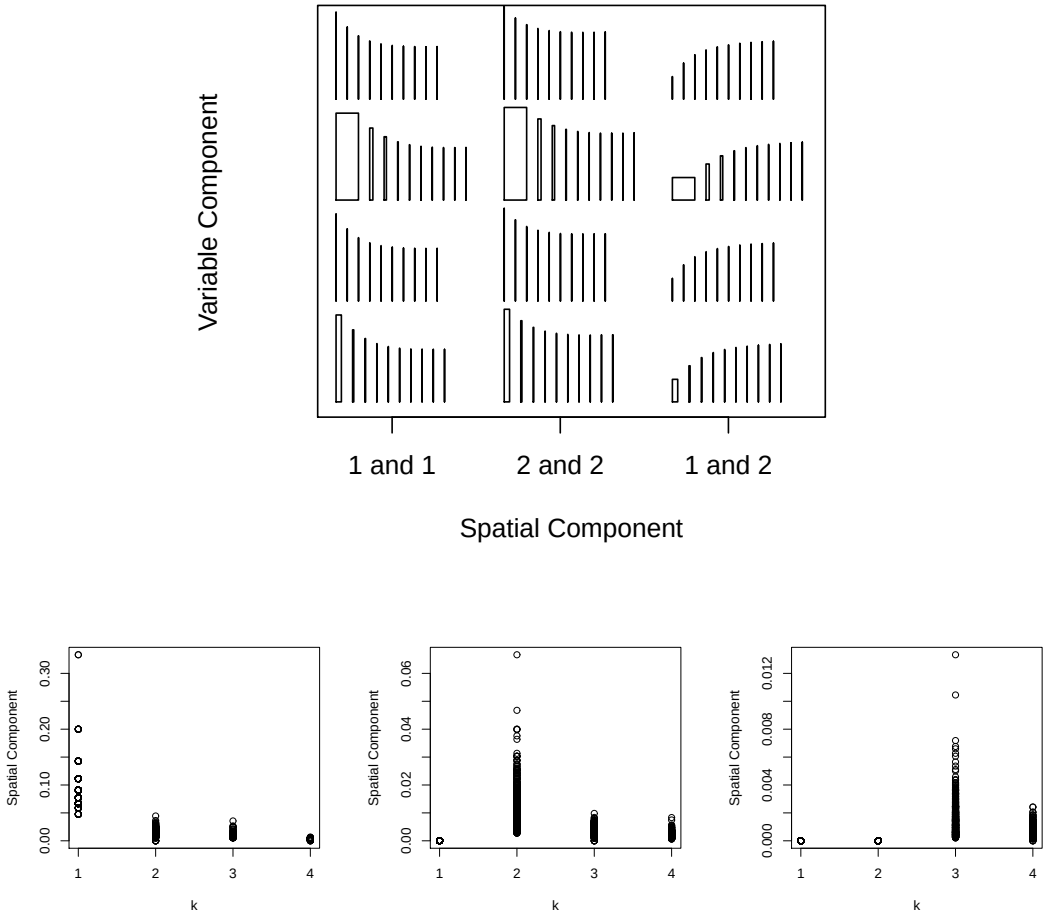

Figure 2.4: Top: Rectangles showing the spatial and variable components of covariance between Y_{ij} and Y_{kl} for increasing neighborhood order k and for different pairs of areas. Bottom: Spatial component as a function of k for all first, second, and third order spatial neighboring areas.

In the top plot of Figure 2.4, we show graphically how these two factors interact to generate the elements in $[\mathbf{C}^k]_{ih}$ as k varies. Each row represents a pair of areas (i, h) and there are three blocks of rectangles representing the possible pair (j, l) of variables. In each block, the successive rectangles represent the increasing power k. Each rectangle has area equal to the elements of $[\mathbf{C}^k]_{ih}$ with the basis length representing $\left[\left(\mathbf{V}^{-1}\delta\right)^k\right]_{ih}$, which is a weighted sum of all spatial paths from i to h. We call this factor the *spatial component* of the covariance. The rectangle's height is the *variable component* and it is given by the (j, l)-th element of the matrix $\left(\mathbf{A}^{-1}\mathbf{B}\right)^k$.

From top to bottom, the four rows of rectangles corresponds to the pair of areas (i, h) equal to $(44, 28)$, $(44, 11)$, $(28, 42)$, and $(28, 54)$, respectively. These areas are highlighted in the Florida counties' map. The first and third pairs are composed by areas that are not contiguous, being separated from each other by three and two intervening areas, respectively. The second and fourth rows use neighboring areas, the former being located at the south corner of the map, with few other neighbors, while the latter are located in the middle of the map and hence possessing more neighbors. The first column of rectangles represent the variables $j = 1$ and $l = 1$, while the second and third columns are $2, 2$ and $1, 2$, respectively. Concerning the first two columns, for all pairs of areas, the rectangles shrink quickly to zero showing visually that the convergence of the matrix series $\sum_k \mathbf{C}^k$ is rather quick. The variable component is dominant in the rectangle areas and it decreases slowly as k increases. The shrinking towards a null area rectangle is mainly due to the spatial component, which decreases quickly to zero in rows two and four, the adjacent pairs of areas. In the case of non-neighboring areas (rows one and three), this decrease is even faster. Indeed, the first rectangles' areas are zero as one is not able to reach area h from area i in less than $k = 3$ and $k = 2$ steps, respectively. For larger k, the rectangles areas are not exactly equal to zero but they are essentially zero as the spatial component is very small. The third column of rectangles show the cross-variable correlation with $j = 1$ and $l = 2$. As in the within-variable case, the rectangles shrink quickly to zero with k, showing that the first terms in the infinite sum $\sum_k [\mathbf{C}^k]_{ih}$ are responsible for the bulk of the values. However, in this case, a different behavior is observed for the variable component as k increases. Although the rectangles areas decrease towards zero, their heights increases as k varies. We conjecture that this behavior could be due to the larger number of possible jumps between the variables 1 and 2 maps as k increases. This allows for more interaction between the variables and hence a larger relative weight for the variable component.

To show that these conclusions are not specific for the pair of areas considered in this first plot, we produced the bottom row of plots. As the variable component is the same for all pair of areas and it has been already shown as the rectangle's heights in the previous plot, we considered only the spatial component in this second row of plots. They show the spatial component $\left[\left(\mathbf{V}^{-1}\delta\right)^k\right]_{ih}$ in the vertical axis versus the number of steps k in the horizontal axis. From left to right, the plots show respectively all pairs of geographically neighboring areas ($s = 1$), pairs of areas with $s = 2$, and $s = 3$. Indeed, the decrease to zero is fast with the first terms largely accounting for the final value.

2.7 Concluding remarks

In this paper, we provided a close look at the induced covariance structure of a flexible spatial model for multivariate lattice data recently introduced in the literature. The model is conditionally specified and it implies on a covariance structure that is not intuitively derived from this model specification. We have been able to find explicit formulas for this covariance structure and to interpret its components. Hence, our work provides a deeper understanding

of the promising [65] model. Given the increasing importance of the multivariate spatial data analysis, including spatial-temporal data, and given the scarcity of flexible models in the literature, this model is likely to reach great popularity in the next few years and a clear interpretation of their structure will be crucial in this eventual success. In fact, if there is any chance that this multivariate model provide a good fit for complex data, it seems prudent to first understand it deeper. It is essential to have a clear idea of which type of correlation structure is implied by the definitions, otherwise we can be simply fitting a model to the data set that does not have any meaningful interpretation.

Chapter 3

Analisando o modelo espacial de decaimento exponencial

Abstract

Neste trabalho a analisamos as estruturas de covariâncias parcial e marginal presentes no modelo proposto por [49]. Mostramos que esse modelo apresenta comportamentos não intuitivos para diversos tipos de grafos de vizinhança, desde os mais simples até os mais complexos. Em especial, mostramos que para esse modelo podemos ter situações em que a correlações marginal e condicional entre duas áreas podem ter sinais opostos. Mostramos esses resultados através de exemplos experimentais e demonstrações analíticas.

3.1 Introdução

Em estudos de epidemiologia, um dos principais interesses dos pesquisadores é investigar como a incidência de uma doença varia geograficamente. Ao analisarmos esse padrão, torna-se possível identificar regiões que apresentam um número de casos acima do esperado e que devem sofrer intervenções por agentes de saúde. A modelagem desse tipo de padrão deve levar em conta a dependência espacial entre as áreas, a fim de reduzir o efeito de ruídos presentes em contagens feitas em pequenas áreas [53]. Os dados coletados podem estar tanto na forma de coordendas geográficas, que indicam a posição exata do caso da doença, ou agregados por unidades tais como condados, municípios e setores censitários. Neste trabalho trataremos com dados coletados da segunda maneira, que são denominados dados de área.

Quando estamos lidando com esse tipo de dado, temos uma região $\mathbf{R}$ particionada em n áreas disjuntas $A_1, A_2, \ldots, A_n$. Tais que $\cup_{i=1}^{n} A_i = \mathbf{R}$. Os dados observados em cada uma dessas áreas são, tipicamente, somas ou médias de variáveis observadas sobre cada uma dessas unidades. A fim de introduzir a dependência espacial entre elas, precisamos definir uma estrutura de vizinhança de acordo com a forma como essas áreas estão dispostas em toda a região. Uma vez definida a estrutura de vizinhança, podemos utilizá-la para definirmos modelos que reflitam a dependência espacial dos dados. Uma abordagem muito recorrente se baseia nas idéias dos modelos autoregressivos de séries temporais. Dois modelos muito populares, que

incluem esse tipo de estrutura, são os modelos CAR (*conditional autoregressive*) e SAR (*simultaneous autoregressive*), que foram propostos originalmente por [9] e [75], respectivamente. Vejamos agora com mais detalhes como o modelo SAR é definido, pois será com base na ideia central desse modelo que o trabalho será desenvolvido.

O modelo SAR é especificado por meio de um conjunto de equações de regressão nas quais a variável resposta é a observação em uma determinada área e as variáveis explicativas são as observações em suas vizinhas. Esse sistema de equações é resolvido de uma forma simultânea e daí surge o nome do modelo. Vejamos sua definição formal.

Seja y_i o valor da variável Y_i observado na área A_i. O modelo SAR é determinado pela solução simultânea do conjunto de equações dado por:

$$Y_i = \mu_i + \sum_{j=1}^{n} s_{ij}(Y_j - \mu_j) + \epsilon_i \quad \text{para} \quad i = 1, \ldots, n \tag{3.1}$$

onde os ϵ_i são normais i.i.d. com variância σ^2, $\mu_i = E(Y_i)$, s_{ij} são constantes, geralmente conhecidas, e $s_{ii} = 0$.

Vamos definir $\mathbf{Y}$ como vetor de dimensão n, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. O conjunto de equações apresentado anteriormente leva a uma distribuição normal multivariada para $\mathbf{Y}$:

$$\mathbf{Y} \sim N_n\left(\boldsymbol{\mu}, \sigma^2(\mathbf{I}_n - \mathbf{S})^{-1}(\mathbf{I}_n - \mathbf{S})\right)$$

onde $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ é o vetor de médias, I_n é a matriz identidade de ordem n e $\mathbf{S}$ é uma matriz $n \times n$ formada pelas constantes s_{ij}.

Devido ao formato de sua distribuição conjunta, a forma mais adequada de se estimar o modelo SAR é usando máxima verossimilhança. Porém, mesmo usando essa abordagem, o processo de estimação ainda não é simples, apesar da estrutura de covariância não apresentar grandes complexidades. Diante desse problema, [49] apresentam um modelo com objetivo de resolver problemas numéricos e tornar o processo de estimação mais eficiente. Eles definem a matriz de covariância de uma forma um pouco diferente. Essa nova matriz de covariância apresenta certas propriedades que simplificam bastante o processo de maximização. Apesar de resolver esse problema, o modelo proposto apresenta alguns aspectos não intuitivos. O ponto principal é que utilizando-se essa abordagem podemos ter correlações marginais e parciais entre pares de áreas que possuem sinais opostos.

O nosso objetivo então é entender melhor o modelo e apontar alguns de seus aspectos não intuitivos, tentando sempre explicitar qual o motivo do comportamento observado. O artigo está organizado da seguinte maneira. Na Seção 3.2 apresentamos o modelo proposto por [49]. Nas Seções 3.3, 3.4 e 3.5 apontamos aspectos não intuitivos do modelo, primeiro para o caso do *lattice* regular, em seguida para uma série temporal e finalmente para mapas reais com estrutura de vizinhança espacial. Na Seção 4.5 apresentamos as conclusões e discussões sobre o trabalho.

3.2 Definição do Modelo

Para que a distribuição normal induzida pelo conjunto de equações simultâneas (3.1) seja própria, a matriz $(\mathbf{I}-\mathbf{S})$ deve ser de posto completo. Uma das possibilidades para se garantir essa propriedade é definir $\mathbf{S} = \rho\mathbf{D}$ e colocar restrições no intervalo onde o parâmetro ρ é definido. A matriz $\mathbf{D}$ pode ser especificada de várias maneiras. As duas abordagens mais comuns são as seguintes. A primeira delas, vamos denotá-la por $\mathbf{W}$, é definir a matriz como uma matriz binária, ou seja, $\mathbf{W}_{ij}$ recebe valor 1 se as áreas i e j forem vizinhas e zero, caso contrário, sendo que $\mathbf{W}_{ii} = 0$. Uma outra alternativa comumente utilizada é definir a matriz $\mathbf{D}$ como uma versão padronizada por linhas da matriz $\mathbf{W}$. Ou seja, $\mathbf{D}_{ij} = \mathbf{W}_{ij}/\mathbf{W}i+$, onde $\mathbf{W}i+ = \sum_j \mathbf{W}_{ij}$. O parâmetro ρ é um parâmetro que mede associação espacial entre as áreas. Pode-se mostrar que $(\mathbf{I} - \rho\mathbf{D})$ será não singular se

$$\rho \in \left(\frac{1}{\lambda_1}, \frac{1}{\lambda_n}\right)$$

onde $\lambda_1 < 0 < \lambda_n$ são, respectivamente, o menor e maior autovalores da matriz $\mathbf{D}$.

Considerando-se a matriz $\mathbf{D}$ padronizada por linhas, temos que $\lambda_n = 1$ e, nesse caso, o intervalo em que ρ está definido fica na forma $\rho \in (1/\lambda_1, 1)$. Sob essas condições, é possível mostrar que a matriz $(\mathbf{I} - \rho\mathbf{D})^{-1}$ pode ser expressa como

$$(\mathbf{I} - \rho\mathbf{D})^{-1} = \mathbf{I} + \rho\mathbf{D} + \rho^2\mathbf{D}^2 + \rho^3\mathbf{D}^3 \ldots \tag{3.2}$$

onde a matriz $\mathbf{D}^k$ representa as vizinhanças de ordem k. Por vizinhança de segunda ordem entre as áreas i e j entendemos que precisamos percorrer no mínimo duas arestas do grafo de vizinhança para saírmos da área i e chegar em j. Vizinhança de terceira ordem significa que precisamos de passar por no mínimos três arestas. O mesmo é válido para matriz $\mathbf{D}^k$ em geral, que irá representar vizinhanças de k-ésima ordem, ou seja, pares de áreas para os quais precisamos de passar por no mínimo k arestas para sair de uma e chega até outra. Notamos, portanto, a partir da equação (3.2), que a influência de vizinhanças mais longínguas cai geometricamente nesse modelo. [3] analisaram como esse decaimento influencia no comportamento das correlações marginais induzidas pelo modelo CAR.

A ideia de [49] é propor um modelo no qual essa influência de vizinhanças mais distantes cai de maneira mais rápida. Uma possibilidade seria substituir, na versão original do modelo, a matriz $(\mathbf{I} - \rho\mathbf{D})^{-1}$ pela matriz

$$\mathbf{I} + \alpha\mathbf{D} + \frac{\alpha^2}{2!}\mathbf{D}^2 + \frac{\alpha^3}{3!}\mathbf{D}^3 \ldots$$

Essa série é denominada matriz exponencial e é denotada por $e^{\alpha\mathbf{D}}$. Ela tem propriedades interessantes que serão listadas a seguir. Considere $\mathbf{X}$ e $\mathbf{Y}$ duas matrizes de dimensão $n \times n$ e sejam a e b dois escalares, então

$$
\begin{aligned}
e^{\mathbf{0}} &= \mathbf{I};\\
e^{a\mathbf{X}}e^{b\mathbf{X}} &= e^{(a+b)\mathbf{X}};\\
\text{se } \mathbf{XY} = \mathbf{YX} &\text{ então } e^{\mathbf{Y}}e^{\mathbf{X}} = e^{\mathbf{X}}e^{\mathbf{Y}} = e^{\mathbf{X}+\mathbf{Y}};\\
\text{se } \mathbf{Y} &\text{ é inversível então } e^{\mathbf{YXY}^{-1}} = \mathbf{Y}e^{\mathbf{X}}\mathbf{Y}^{-1}.
\end{aligned}
$$

No modelo proposto por [49], a matriz de covariâncias é definida como

$$\Sigma_\alpha = \sigma^2\left(e^{-\alpha\mathbf{D}'}e^{-\alpha\mathbf{D}}\right)$$

onde $\mathbf{D}$ é uma matriz de pesos espaciais não-negativa $n \times n$. Como já mencionado anteriormente, tipicamente, define-se $\mathbf{D}_{ij} > 0$ se i e j são vizinhas e $\mathbf{D}_{ij} = 0$ caso contrário.

Uma abordagem semelhante a essa foi proposta por [20] para modelar matrizes de covariância em geral. Os autores definem essa matriz com uma função da matriz de delineamento. Uma grande vantagem dessa definição é que ela garante que a matriz de covariâncias será sempre positiva definida. Dessa maneira, torna-se dispensável colocar restrições nos parâmetros durante o processo de estimação para garantir que tal propriedade seja satisfeita. Uma outra vantagem é a facilidade de se obter a inversa da matriz exponencial.

O modelo proposto por [20] é mais genérico e pode ser usado em diversas situações. Porém, em muitas delas, os aspectos não intuitivos identificados para o modelo proposto por [49] também estarão presentes. Uma de suas principais aplicações são para dados longitudinais. Nesse caso também não é razoável que as correlações parciais e marginais tenham sinais opostos. Isso significaria, por exemplo, que se a medida do tempo t de um indivíduo tende a estar acima (abaixo) da média, a medida do tempo $t+1$ também tenderia a ficar acima (abaixo) da média. Porém, condicionando-se na informação obtida em todos os outros instantes de tempo, essa relação assume uma direção oposta. É possível verificar que para determinadas matrizes de delineamento, isso de fato ocorre. Observe porém que esse tipo de comportamento pouco intuitivo não é observado para modelos clássicos de séries temporais, como os modelos autoregressivos, por exemplo. Se seus parâmetros são todos positivos, a função de autocorrelação e a função de autocorrelação parcial terão ambas sinal positivo para qualquer ordem de vizinhança considerada.

Veremos agora vários exemplos que mostram que essa definição proposta por [49] pode induzir correlações marginais e condicionais com sinais opostos. Em outras palavras, podemos ter um par de áreas i e j cuja correlação marginal é positiva, porém quando condicionamos nas demais áreas do mapa, a correlação condicional passa a ser negativa. Esse tipo de problema já foi amplamente estudado quando estamos lidando com estrutura de dependência entre variáveis e é conhecido como "Paradoxo de Simpson" ([15]). Esse paradoxo diz que o sinal da correlação entre duas variáveis pode se modificar de acordo com os grupos que consideramos dentro da população. Ou seja, a correlação marginal, sem considerar o fator grupo, tem um sinal diferente da correlação obtida quando se condiciona no grupo ao qual o indivíduo

pertence.

Esse tipo de comportamento não intuitivo para modelos espaciais foi apontado por [6] (veja página 169) que falam que para os modelos SAR e CAR *the transformation to (the covariance matrix)* Σ_Y *is very complicated and very non-linear. Positive conditional association can become negative unconditional association.* Ocorre porém, que para os modelos SAR e CAR esse tipo de comportamento é observado apenas em situações de pouco interesse prático, que são aquelas em que existe uma associação espacial negativa, ou seja, uma repulsão entre as áreas. Uma explicação bem simples para esse fato pode ser obtida a partir de um resultado mostrado por [41]. Tal resultado garante que se todas as correlações parciais têm sinal positivo, então a correlação marginal também será positiva. Na especificação dos modelo SAR e CAR, para valores positivos do parâmetro de associação espacial, as correlações parcial entre quaisquer pares de áreas são sempre positivas. Dessa maneira, a correlação marginal também terá sinal positivo. Já para o modelo de [49], como veremos ao longo deste trabalho, essa troca de sinais ocorre mesmo quando existe uma associação positiva entre as áreas.

3.3 *Lattice* Regular

O primeiro caso a ser analisado é uma situação simples, em que estamos em um *lattice* regular. Isso significa que o número de vizinhos (n_i) de cada área é fixo e igual a 4. Vamos assumir que a matrix $\mathbf{D}$ é uma matriz padronizada, ou seja, $[\mathbf{D}]_{ij} = 1/n_i$ se as áreas i e j são vizinhas e 0 caso contrário. Vamos considerar ainda que $\sigma^2 = 1$. Dessa forma, a matriz de covariância do modelo se reduz a

$$\Sigma_\alpha = e^{-\alpha\mathbf{D}'}e^{-\alpha\mathbf{D}} .$$

Como $\mathbf{D}' = \mathbf{D}$ e $\mathbf{D}'\mathbf{D} = \mathbf{D}\mathbf{D}'$ temos que

$$\Sigma_\alpha = e^{-\alpha\mathbf{D}'}e^{-\alpha\mathbf{D}} = e^{-\alpha\mathbf{D}'-\alpha\mathbf{D}} = e^{-2\alpha\mathbf{D}} .$$

Sabemos que, pela definição da matriz exponencial

$$e^{-2\alpha\mathbf{D}} = \mathbf{I} - (2\alpha)\mathbf{D} + \frac{(2\alpha)^2}{2!}\mathbf{D}^2 - \frac{(2\alpha)^3}{3!}\mathbf{D}^3 + \frac{(2\alpha)^4}{4!}\mathbf{D}^4 - \frac{(2\alpha)^5}{5!}\mathbf{D}^5 + \ldots$$

O termo ij da matriz $\mathbf{D}^k$ nos fornece a probabilidade de um passeio aleatório que percorre o grafo de vizinhança sair da área i em chegar em j em k passos. Sabemos que para um *lattice* regular só é possível ir para um vizinho de ordem ímpar (de primeira, terceira, quinta, etc) em um número ímpar de passos. E para um vizinho de ordem par (segunda, quarta, sexta, etc) em um número par de passos. Isso significa que se k é impar

$$\mathbf{D}^k_{ij} \begin{cases} > 0 & \text{se } i \text{ e } j \text{ são vizinhos de ordem ímpar} \\ = 0 & \text{caso contrário} \end{cases}$$

se k é par

$$\mathbf{D}_{ij}^{k} \begin{cases} > 0 & \text{se } i \text{ e } j \text{ são vizinhos de ordem par} \\ = 0 & \text{caso contrário .} \end{cases}$$

Portanto se i e j são vizinhos de primeira ordem, por exemplo, o termo ij da matriz de covariância fica na forma

$$[e^{-2\alpha\mathbf{D}}]_{ij} = -(2\alpha)[\mathbf{D}]_{ij} - \frac{(2\alpha)^3}{3!}[\mathbf{D}^3]_{ij} - \frac{(2\alpha)^5}{5!}[\mathbf{D}^5]_{ij} + \dots \tag{3.3}$$

ou seja, a correlação entre as áreas i e j terá o sinal oposto ao sinal do parâmetro α.

Sabemos ainda que a matriz de precisão (inverso da matriz de covariância) nesse caso será dada por

$$\left(e^{-\alpha\mathbf{D}}\right)^{-1} = e^{\alpha\mathbf{D}} = \mathbf{I} + (2\alpha)\mathbf{D} + \frac{(2\alpha)^2}{2!}\mathbf{D}^2 + \frac{(2\alpha)^3}{3!}\mathbf{D}^3 + \frac{(2\alpha)^4}{4!}\mathbf{D}^4 + \frac{(2\alpha)^5}{5!}\mathbf{D}^5 + \dots$$

Como o sinal da correlação condicional é dado pelo oposto do sinal do termo da matriz de precisão, veremos que, para o *lattice* regular, todas as correlações condicionais entre quaisquer pares de sítios terão sinal oposto ao sinal de α.

Por exemplo, se i e j são vizinhos de primeira ordem o termo ij dessa matriz fica

$$[e^{-2\alpha\mathbf{D}}]_{ij} = (2\alpha)[\mathbf{D}]_{ij} + \frac{(2\alpha)^3}{3!}[\mathbf{D}^3]_{ij} + \frac{(2\alpha)^5}{5!}[\mathbf{D}^5]_{ij} + \dots$$

Comparando essa expressão com aquela apresentada em (3.3) percebemos que as correlações parciais e marginais entre i e j apresentam o mesmo sinal.

Vamos olhar agora para um par de vizinhos de segunda ordem, i e j. Nesse caso temos que a o termo ij da matriz de covariância será dado por

$$[e^{-2\alpha\mathbf{D}}]_{ij} = \frac{(2\alpha)^2}{2!}[\mathbf{D}^2]_{ij} + \frac{(2\alpha)^4}{4!}[\mathbf{D}^4]_{ij} + \frac{(2\alpha)^6}{6!}[\mathbf{D}^6]_{ij} + \dots \tag{3.4}$$

Já o termo ij da matriz de precisão fica

$$[e^{2\alpha\mathbf{D}}]_{ij} = \frac{(2\alpha)^2}{2!}[\mathbf{D}^2]_{ij} + \frac{(2\alpha)^4}{4!}[\mathbf{D}^4]_{ij} + \frac{(2\alpha)^6}{6!}[\mathbf{D}^6]_{ij} + \dots$$

comparando essa expressão com (3.4) percebemos que a correlação marginal entre i e j será sempre positiva, porém a correlação parcial será sempre negativa. Esse mesmo comportamento pode ser observado para todos vizinhos de ordem par. Ou seja, para quaisquer pares de vizinhos de ordem par, a correlação marginal entre eles será sempre positiva, porém quando condicionamos nas demais áreas do mapa essa correlação passa a ter valor negativo. Esse não é um resultado razoável para um modelo de dependência espacial, visto que uma correlação marginal positiva significa que se a área i apresenta um valor acima da média, a área j tenderá também a apresentar um valor acima do esperado. Porém, se consideramos conhecidos os valores de todas as áreas do mapa, o fato da área i ter um valor acima da média, induz a área

j a ter um valor abaixo da média.

Para os vizinhos de ordem par observamos ainda outra restrição forte do modelo. Ele só permite que existam correlações marginais positivas entre vizinhos desse tipo. Ou seja, vizinhos de segunda, quarta, sexta ordem só poderão estar positivamente correlacionados quando utilizamos esse modelo, independentemente do valor de α.

Vejamos agora quais são as consequências desses resultados para as esperanças condicionais, visto que a esperança condicional tem uma interpretação um pouco mais direta do que a correlação condicional. Considere Campo Gaussiano Markoviano cuja matriz de precisão é denotada por $\mathbf{Q}$ com entradas Q_{ij} e cujo vetor de médias $\boldsymbol{\mu}$ é igual a zero. Sabe-se que a esperança condicional da variável aleatória em um sítio i, dado o resto do mapa ($\mathbf{Y}_{-i}$), pode ser obtida a partir da seguinte expressão

$$E(Y_i|\mathbf{y}_{-i}) = -\frac{1}{Q_i i}\sum_{j:i\sim j} Q_{ij}y_j \tag{3.5}$$

onde $i \sim j$ significa i vizinho de j. Vimos que para o caso do *lattice* regular se i e j são vizinhos de ordem ímpar, $Q_{ij} < 0$ e se i e j são vizinhos de ordem par, $Q_{ij} > 0$. Notamos então, a partir da equação (3.5), que se quisermos prever o valor da variável $\mathbf{Y}$ usando valores observados de seus vizinhos, os valores observados para os vizinhos de ordem ímpar darão uma contribuição positiva, já os valores observados para os vizinhos de ordem par entrarão um peso negativo para esperança. Isso não parece razoável, visto que a correlação marginal da área i com todos os seus vizinhos tem sinal positivo, de maneira que todos eles deveriam entrar com peso positivo na média condicional.

Observe que, apesar do caso do *lattice* regular ser um caso muito específico de estrutura de vizinhança, ele está presente em muita situações práticas, como, por exemplo, no processamento de imagens. Veremos mais a frente que esse tipo de comportamento se repete mesmo para casos em que não consideramos um *lattice* regular.

A fim de entendermos melhor qual o significado desta troca de sinal entre as as correlações marginais e condicionais, apresentaremos na seção a seguir um breve estudo de simulação para o caso unidimensional, ou seja, quando os dados são observados ao longo do tempo.

3.4 *Lattice* unidimensional

O objetivo dessa seção é mostrar visualmente os tipos de dados que são gerados quando se considera a estrutura de covariância proposta por [49]. Para facilitar o entendimento e a apresentação gráfica vamos considerar o caso em que os dados são observados ao longo do tempo. Fixamos os parâmetros $\sigma^2 = 1$ e $\alpha = -1$ e geramos séries temporais com 51 observações. A estrutura de vizinhança é uma Cadeia de Markov de ordem um. A série gerada é apresentada na Figura 3.1 à esquerda. Duas das observações foram apagadas e seus valores foram gerados novamente a partir das distribuições condicionais da variável nesses dois pontos no tempo, dado todo o resto da série. O gráfico à direita da Figura 3.1 mostra a série

com as duas observações faltantes. Em outra palavras, temos um conjunto de observações

$$x_1, x_2, \ldots, x_{51} \, .$$

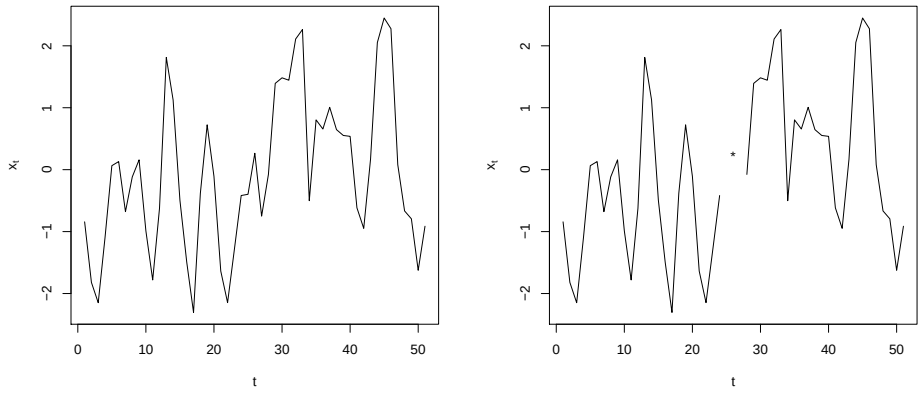

Figure 3.1: Série temporal com 51 observações simuladas (esquerda) e série simulada retirando-se as observações de ordem 25 e 27 (direita).

Descartamos os valores x_{25} e x_{27} e consideramos essas posições como se fossem observações perdidas do nosso banco de dados. Sabemos que a distribuição condicional do vetor (X_{25}, X_{27}), dado todo o resto da série, é uma normal bivariada. Geramos então vários pares de observações com tal distribuição a fim de verificar o tipo de relação que aparece entre as duas observações. Na Figura 3.2, o gráfico da esquerda mostra em linha cheia as observações da série original e a linha tracejada mostram os valores gerados a partir da distribuição condicional. Os ponto marcados em preto representam os valores das esperanças condicionais. Observa-se que, como já era esperado, se os valores de um dos dois pontos selecionados está abaixo da esperança condicional, o outro tende a estar acima de sua esperança condicional e vice versa. A Figura 3.2 apresenta ainda o gráfico de dispersão entre várias observações geradas para essas duas variáveis. Notamos claramente que existe uma correlação negativa entre as duas variáveis, o que já era de se esperar visto que os dois sítios considerados são vizinhos de segunda ordem. Além disso, para quatro dentre os cinco pontos gerados, se a observação para um sítio está abaixo (acima) de sua média condicional, a observação do outro estará acima (abaixo).

A tendência geral desse tipo de comportamento pode ser observada gerando-se mais observações da distribuição condicional de (X_{25}, X_{27}) dado o resto da série. Os mesmos gráficos apresentados na Figura 3.2 são mostrados na Figura 3.3, mas agora com 20 pares observações gerados no primeiro gráfico e 200 gerados no segundo. Nota-se que existe claramente uma correlação negativa entre as observações nos dois pontos que estão sendo analisados.

A fim de compararmos o comportamento da série gerada a partir do modelo analisado com o comportamento de um modelo ARIMA clássico, geramos 51 observações de um processo autoregressivo de ordem 2. Os parâmetros do modelo foram fixados em 0,3 e 0,4 e a variância foi fixada em $\sigma^2 = 1$. Os dados foram gerados da mesma forma como feito no caso anterior.

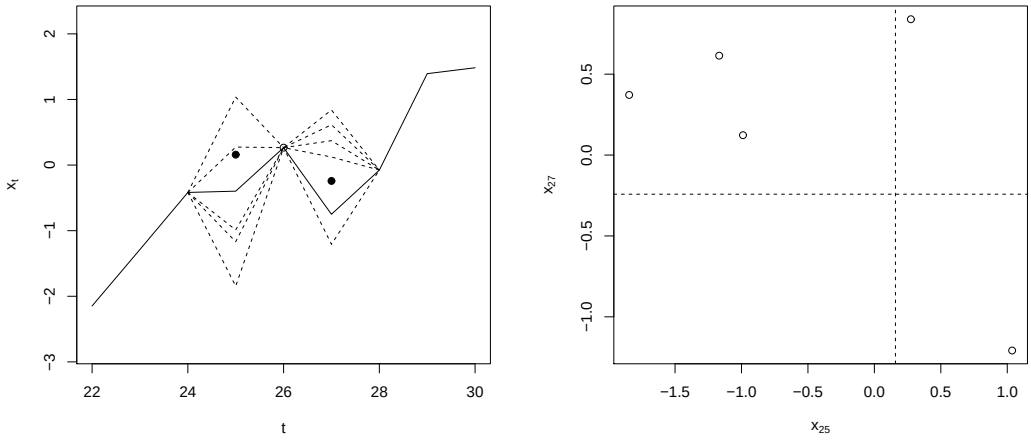

Figure 3.2: Série temporal gerada, linhas tracejadas representam as cinco observações geradas a partir das distribuições condicionais e os pontos marcam as esperanças condicionais (esquerda). Gráfico de dispersão das cinco observações geradas a partir da distribuição condicional (direita).

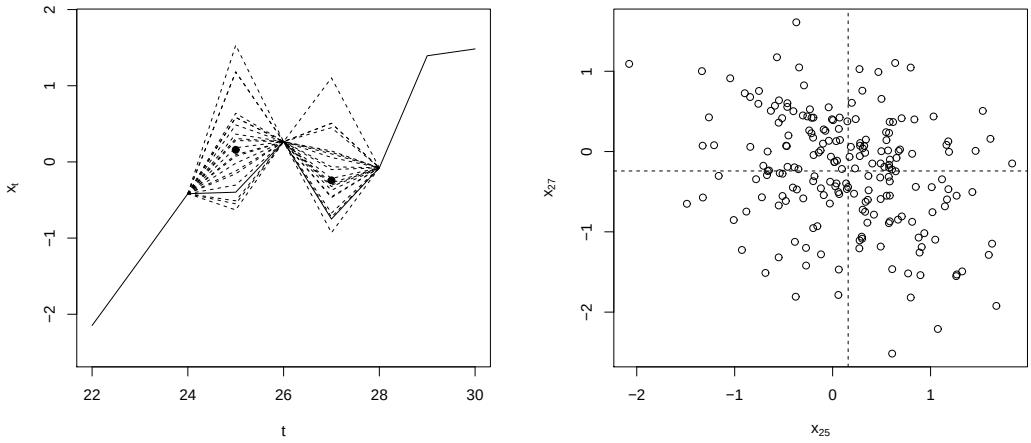

Figure 3.3: Série temporal gerada, linhas tracejadas representam as 20 observações geradas a partir das distribuições condicionais e os pontos marcam as esperanças condicionais (esquerda). Gráfico de dispersão das 200 observações geradas a partir da distribuição condicional (direita).

Os gráficos da Figura 3.4 mostram, à esquerda, a série original em linha cheia e as observações geradas através das distribuições condicionais em linha tracejada. Os pontos em preto marcam as esperanças condicionais. À direita da Figura 3.4 é apresentado o gráfico de dispersão das observações geradas a partir da distribuição condicional. Observa-se que, ao contrário do que ocorre no caso anterior, existe claramente uma associação positiva entre as observações nos dois pontos.

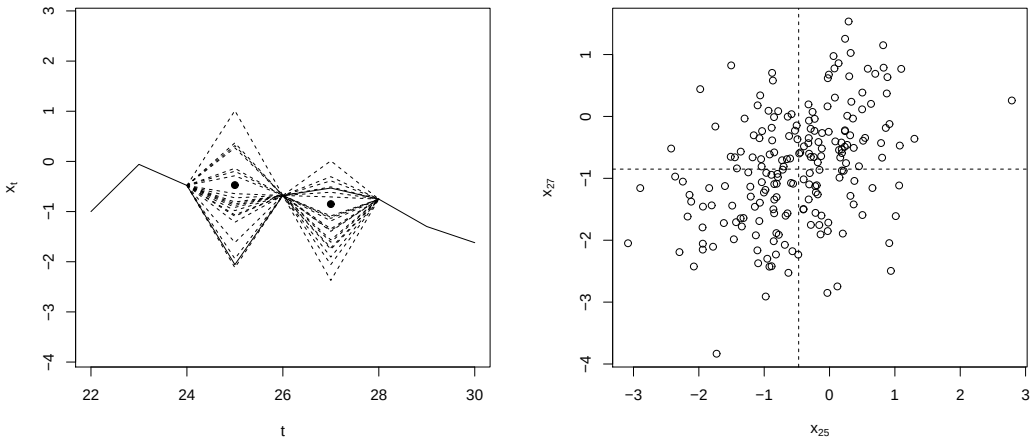

Figure 3.4: Série temporal gerada de um processo autoregressivo de ordem 2. Linhas tracejadas representam as 20 observações geradas a partir das distribuições condicionais e os pontos marcam as esperanças condicionais (esquerda). Gráfico de dispersão das 200 observações geradas a partir da distribuição condicional (direita).

3.5 *Lattice* irregular espacial

No caso de mapas reais, o *lattice* será tipicamente irregular. Os fenômenos apontados anteriormente ainda se repetem nessa situação. Consideraremos primeiramente um caso bem próximo do *lattice* regular, que apresenta resultados contraintuitivos adicionais. Observa-se que a troca de sinal ocorre até mesmo entre vizinhos de primeira ordem. Essa situação será ilustrada com o mapa de Iowa.

O mapa analisado é apresentado na Figura 3.5. O critério de vizinhança utilizado foi o de adjacência, ou seja, duas áreas são consideradas vizinhas se, e somente se, dividem fronteira e a matriz de vizinhança é especificada como uma matriz binária, que assume apenas valores zero e um. Fixamos $\alpha = -1$ e $\sigma^2 = 1$. De maneira que a matriz de covariância fica na forma

$$\Sigma_\alpha = e^{\mathbf{W}'} e^{\mathbf{W}}$$

onde $\mathbf{W}$ é uma matriz binária tal que $[\mathbf{W}]_{ij} = 1$, se i e j são vizinhas, e zero, caso contrário.

Iremos verificar agora para quais pares de vizinhos as correlações marginais e condicionais possuem sinais opostos. O primeiro mapa da Figura 3.5 mostra as ligações entre vizinhos de primeira ordem. O segundo mapa dessa mesma figura apresenta as ligações entre áreas vizinhas que possuem correlação marginal positiva. Nota-se que todos os pares destacados no mapa anterior satisfazem essa condição. O mapa logo abaixo à esquerda mostra ligações entre áreas vizinhas que possuem correlações condicionais positivas. Portanto, para todas as ligações apresentadas nessa figura não há troca de sinais. Já o mapa à direita mostra ligações entre áreas vizinhas que possuem correlação condicional negativa. Isso significa que para tais ligações ocorre a troca de sinal. Notamos, então, que para esse exemplo mesmo vizinhos de

primeira ordem possuem correlações marginais e condicionais com sinais opostos. Apesar da disposição das áreas no mapa se parecer muito com um *lattice* regular, para esse caso temos um comportamento ainda menos intuitivo.

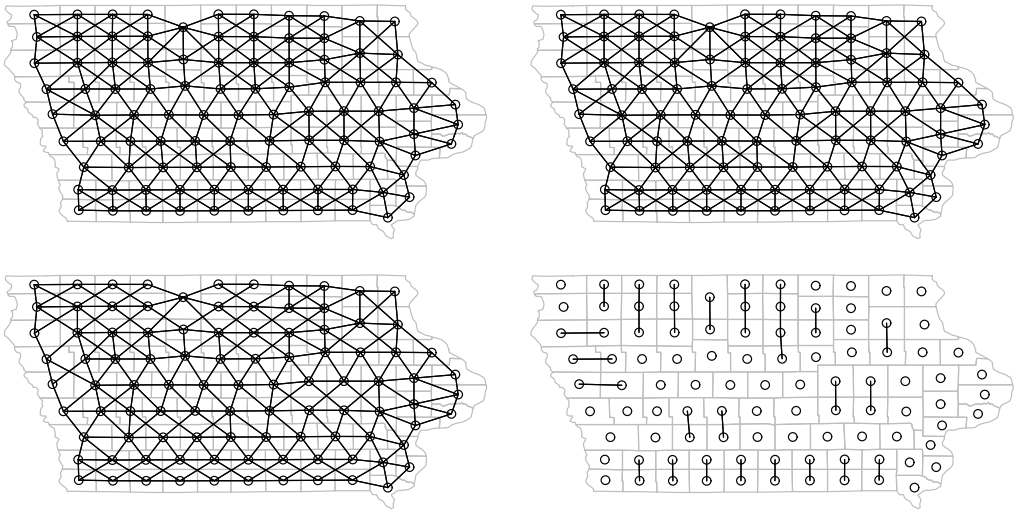

Figure 3.5: Mapa Iowa representando correlações marginais e condicionais entre pares de áreas vizinhas. A primeira figura mostra o mapa de Iowa com as ligações representando vizinhos de primeira ordem. A segunda mostra ligações entre vizinhos de primeira ordem que possuem correlação marginal positiva. A terceira figura mostra ligações entre pares de áreas vizinhas que possuem correlação condicional positiva. A quarta figura mostra ligações entre pares de áreas que possuem correlações condicionais negativas.

Mapas reais costumam apresentar áreas de dimensões muito diferentes, induzindo grafos de vizinhanças muito irregulares. Essa situação é mais difícil de ser analisada pois a topologia do grafo torna-se mais complexa.

Para ilustrar esse tipo de situação, vamos considerar o mapa utilizado na aplicação apresentada por [49]. Os autores analisam a participação dos americanos nas eleições presidenciais no ano de 2000. Os dados são agregados em nível de condados e todo o estado do Texas é excluído da análisde devido a problemas de irregularidades nas eleições. O mapa utilizado é apresentado na Figura 3.6. Os autores consideram dois tipos de vizinhança: adjacência e k vizinhos mais próximos. Por motivos de simplificação, consideraremos aqui apenas o critério de adjacência. As ligações entre algumas áreas vizinhas de acordo com esse critério são mostradas no primeiro gráfico da Figura 3.6. Observe que estamos ampliando apenas uma região dos Estados Unidos a fim de facilitar a visualização. O valor estimado para α na aplicação foi de -0.74 e esse será o valor utilizado aqui.

O segundo gráfico da Figura 3.6 alguns dos pares de vizinhos que possuem correlação marginal positiva. Todos vizinhos de primeira ordem apresentam essa propriedade. Isso já era esperado pois para $\alpha < 0$ todos os termos da matriz de covariância são positivos. O primeiro gráfico da segunda linha nessa mesma figura mostra as ligações entre vizinhos de primeira ordem que possuem correlação parcial positiva. Observe que são todos vizinhos de primeira ordem, ou seja, nesse caso não existe troca de sinal. O último gráfico dessa figura

apresenta ligações entre vizinhos de primeira ordem que possuem correlação parcial negativa, ou seja, nenhum par de vizinhos, o que confirma a observação feita anteriormente.

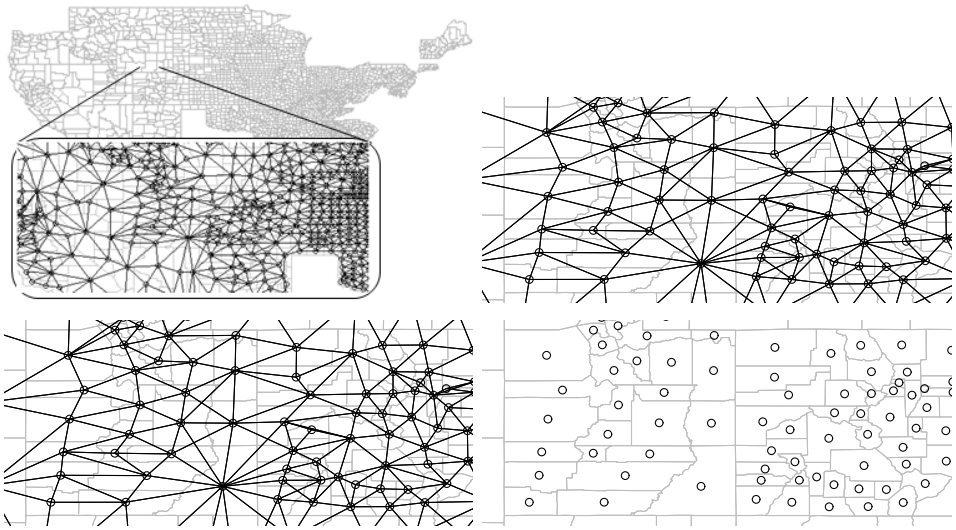

Figure 3.6: Vizinhos de primeira ordem, vizinhos de primeia ordem com correlação marginal positiva, vizinhos de primeira ordem com correlação parcial positiva e vizinhos de primeira ordem com correlação parcial negativa.

O primeiro gráfico da Figura 3.7 mostra as ligações entre vizinhos de segunda ordem. O segundo gráfico dessa figura mostra os pares de vizinhos de segunda ordem que possuem correlação marginal positva. Assim como ocorre para vizinhança de primeira ordem, nesse caso também a correlação entre quaisquer pares de áreas é positiva. O primeiro gráfico da segunda linha da Figura 3.7 mostra as ligações entre pares de vizinhos de segunda ordem que possuem correlação parcial positiva, ou seja, nenhum par de vizinhos. Já o gráfico seguinte mostra as ligações entre vizinhos de segunda ordem que possuem correlação parcial negativa. Observamos então que ao considerarmos vizinhos de segunda ordem, observamos troca de sinal para todos os pares. Esse comportamento se repete em geral para vizinhos de ordem par.

Notamos então que as correlações para esse mapa se comportam da mesma maneira que o gráfico regular. Analisando-se outros mapas, com configurações completamente distintas, para esse valor de α, o mesmo comportamento foi observado. Veremos que isso ocorre pois, como o valor de α é pequeno, as primeiras ordens de vizinhança dominam a matriz de covariância e de precisão.

Vamos mostrar agora porque esse comportamento ocorrerá para qualquer tipo de mapa. Vimos que a matriz de covariâncias do modelo proposto é dada por

$$\sigma^2 e^{-\alpha \mathbf{D}'} e^{-\alpha \mathbf{D}} .$$

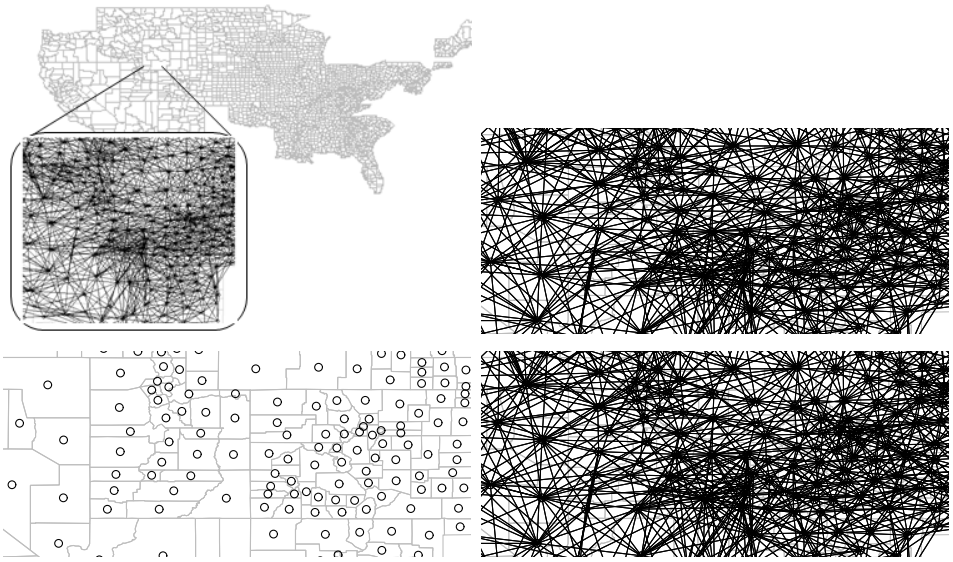

Figure 3.7: Vizinhos de segunda ordem, vizinhos de segunda ordem com correlação marginal positiva, vizinhos de segunda ordem com correlação parcial positiva e vizinhos de segunda ordem com correlação parcial negativa.

Ignorando o termo de variância σ^2 e usando a definição de $e^{-\alpha\mathbf{D}}$ esse produto fica na forma

$$\left(\mathbf{I} - \alpha\mathbf{D}' + \frac{\alpha^2}{2!}\mathbf{D}'^2 - \frac{\alpha^3}{3!}\mathbf{D}'^3 + \dots\right)\left(\mathbf{I} - \alpha\mathbf{D} + \frac{\alpha^2}{2!}\mathbf{D}^2 - \frac{\alpha^3}{3!}\mathbf{D}^3 + \dots\right)$$

Percebemos então que se $\alpha < 0$ todos os termos da matriz de covariâncias serão positivos. Ou seja, a correlação entre quaisquer pares de áreas será positiva. Vejamos agora qual o formato da matriz de precisão. Temos que essa matriz é dada por

$$(\sigma^2)^{-1}e^{\alpha\mathbf{D}}e^{\alpha\mathbf{D}'}\ .$$

Novamente, usando a definição de $e^{\alpha\mathbf{D}}$ ficamos com

$$\left(\mathbf{I} + \alpha\mathbf{D} + \frac{\alpha^2}{2!}\mathbf{D}^2 + \frac{\alpha^3}{3!}\mathbf{D}^3 + \dots\right)\left(\mathbf{I} + \alpha\mathbf{D}' + \frac{\alpha^2}{2!}\mathbf{D}'^2 + \frac{\alpha^3}{3!}\mathbf{D}'^3 + \dots\right)\ .$$

Nesse caso, se α é negativo teremos o produto de dois somatórios que envolvem termos positivos e negativos. Dessa forma, os termos da matriz de precisão podem ser positivos ou negativos, dependendo de quais matrizes dominem a soma. Podemos reescrever o produto na forma

$$\mathbf{I} + \alpha\mathbf{D}' + \alpha\mathbf{D} + \frac{\alpha^2}{2!}\mathbf{D}'^2 + \frac{\alpha^2}{2!}\mathbf{D}^2 + \alpha^2\mathbf{D}'\mathbf{D} + \dots \tag{3.6}$$

Notamos ainda que como α é pequeno e $\mathbf{D}$ está limitada entre $[0, 1]$ à medida que aumentamos k os termos

$$\frac{\alpha^k}{k!}\mathbf{D}^k \quad \text{e} \quad \frac{\alpha^k}{k!}\mathbf{D}'^k$$

ficam cada vez mais próximos de zero. Portanto, a soma apresenta em (3.6) será dominada apenas pelos expoentes mais baixos. Vamos considerar primeiramente o caso em que duas áreas i e j são vizinhas de primeira ordem. Temos então que os termos $[\mathbf{D}]_{ij}$ e $[\mathbf{D}']_{ij}$ são ambos maiores que zero. Dessa maneira a soma em (3.6) fica dominada pelos termos

$$\alpha\mathbf{D}' + \alpha\mathbf{D} .$$

Como estamos considerando $\alpha < 0$ essa soma fica menor que zero, o que implica em uma correlação condicional positiva. Ou seja, se i e j são vizinhos de primeira ordem, as correlações marginal e condicional possuem o mesmo sinal, como já tínhamos observado no exemplo.

Consideremos agora o caso em que i e j são vizinhos de segunda ordem. Nesse caso temos que $[\mathbf{D}]_{ij} = [\mathbf{D}']_{ij} = 0$, porém $[\mathbf{D}^2]_{ij}$ e $[\mathbf{D}'^2]_{ij}$ são ambos maiores que zero. Portanto a soma em (3.6) fica dominada pelos termos

$$\frac{\alpha^2}{2!}\mathbf{D}'^2 + \frac{\alpha^2}{2!}\mathbf{D}^2 + \alpha^2\mathbf{D}'\mathbf{D} .$$

Como todas as matrizes envolvidas na soma acima são positivas o resultado será um número positivo. Isso implica em uma correlação condicional negativa. Ou seja, se i e j são vizinhos de segunda ordem, as correlações marginal e condicional possuem sinais trocados. Mostramos assim porque o comportamento observado no exemplo anterior não é devido à estrutura de vizinhança utilizada. Para um valor pequeno de α, seja qual for o mapa considerado, sempre haverá troca de sinais nas correlações de vizinhos de segunda ordem.

3.5.1 Como a correlação parcial varia com α

A forma como as correlações parciais variam com α também apresenta aspectos pouco intuitivos.

Vejamos primeiramente qual seria o tipo de relação esperada entre o parâmetro α e o grau de associação entre as áreas ρ. De acordo com [49] a relação entre o parâmetro α do modelo proposto e o parâmetro ρ do modelo SAR é dada por

$$\alpha \approx \ln(1-\rho) .$$

Como os valores de ρ que possuem interesse prático são $\rho > 0$, observamos dois pontos importantes a partir dessa relação. O primeiro deles é que os valores de α que terão interesse prático serão $\alpha < 0$. Além disso, observe que

$$\frac{d\ln(1-\rho)}{d\rho} = -\frac{\rho}{1-\rho}$$

e como $\rho < 1$, a derivada possui sinal negativo. Portanto, à medida que ρ aumenta, α diminui.

Tendo isso em vista, seria razoável que as correlações marginais e condicionais tivessem um comportamento decrescente com relação ao parâmetro α. Veremos aqui que para o caso das correlações condicionais esse fato não é sempre verdade.

Considerando novamente o mesmo mapa utilizado por [49] vamos encontrar os valores das correlações condicionais para diversos valores do parâmetro α.

Verificaremos inicialmente como se dá essa relação para os pares de sítios que são vizinhos de primeira ordem. A Figura 3.8 apresenta valores das correlações entre vizinhos de primeira ordem para α variando entre $[-1, 0]$. Notamos, portanto, que a correlação condicional decresce com o valor de α. E, como já explicado anteriormente, esse é um comportamento razoável, pois significa que à medida que a associação espacial entre as áreas fica mais forte, a correlação condicional entre determinados pares de áreas irá crescer. Por exemplo, $\alpha = 0$ equivale a $\rho = 0$, ou seja, significaria a independência entre as áreas. Nesse caso, percebemos pelo gráfico que a correlação parcial entre os vizinhos de primeira ordem é zero. Para $\alpha = -1$ temos que $\rho \approx 0.632$, ou seja, deveria haver um aumento na correlação parcial entre as áreas. Pelo gráfico apresentado na Figura 3.8 notamos que é isso que ocorre.

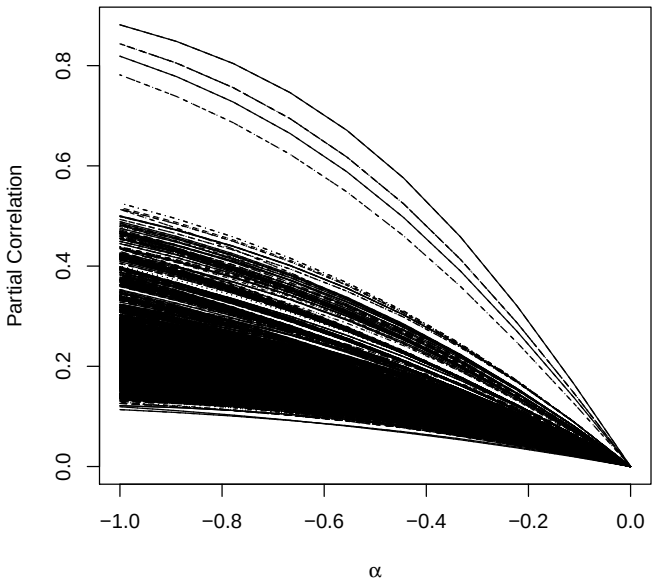

Figure 3.8: Correlação entre vizinhos de primeira ordem para diferentes valores de α.

Vejamos agora o que ocorre para os pares de sítios que são vizinhos de segunda ordem. A Figura 3.9 apresenta valores das correlações entre vizinhos de segunda ordem para α variando entre $[-1, 0]$. Notamos, portanto, que a correlação condicional cresce com o valor de α. Percebe-se que nesse caso, à medida que o valor de α diminui, associação espacial fica mais forte, a correlação parcial entre vizinhos de segunda ordem aumenta, porém ela tem sinal negativo. Esse não é um comportamento razoável, pois o que seria esperado era que à medida que a associação espacial aumentasse, a correlação parcial entre as áreas deveria aumentar, porém no sentido positivo e não negativo, como ocorre.

Mostraremos agora o motivo pelo qual esse comportamento não intuitivo ocorre para vizinhos de segunda ordem. Relembre que a matriz de precisão é dada por

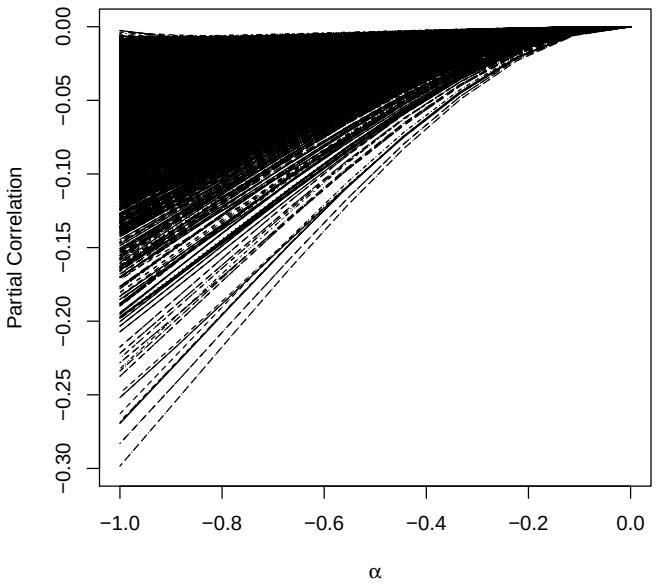

Figure 3.9: Correlação entre vizinhos de segunda ordem para diferentes valores de α.

$$\sigma^2 e^{\alpha \mathbf{D}} e^{\alpha \mathbf{D}'} .$$

Iremos agora derivar esse produto de matrizes em relação ao parâmetro α a fim de verificar seu comportamento. A derivada do produto de duas matrizes pode ser obtida de forma análoga à derivada do produto de duas funções, usando a seguinte propriedade. Sejam $\mathbf{X}$ e $\mathbf{Y}$ duas matrizes quaisquer, que dependem de α, então

$$\frac{d(\mathbf{XY})}{d\alpha} = \frac{d\mathbf{X}}{d\alpha}\mathbf{Y} + \mathbf{X}\frac{d\mathbf{Y}}{d\alpha}$$

Aplicando essa propriedade à matriz de precisão (3.6) temos que

$$\frac{de^{\alpha \mathbf{D}} e^{\alpha \mathbf{D}'}}{d\alpha} = \frac{de^{\alpha \mathbf{D}}}{d\alpha} e^{\alpha \mathbf{D}'} + e^{\alpha \mathbf{D}} \frac{de^{\alpha} \mathbf{D}'}{d\alpha} .$$

Vamos usar agora o fato de que

$$\frac{de^{\alpha \mathbf{D}}}{d\alpha} = \mathbf{D} e^{\alpha \mathbf{D}} .$$

Portanto

$$\frac{de^{\alpha \mathbf{D}} e^{\alpha \mathbf{D}'}}{\alpha} = \mathbf{D} e^{\alpha \mathbf{D}} e^{\alpha \mathbf{D}'} + e^{\alpha \mathbf{D}} \mathbf{D}' e^{\alpha \mathbf{D}'} .$$

Usando a definição de $e^{\alpha \mathbf{D}}$, esse produto fica

$$\mathbf{D}\left(I + \alpha \mathbf{D} + \alpha^2 \frac{\mathbf{D}^2}{2!} + \ldots\right)\left(I + \alpha \mathbf{D}' + \alpha^2 \frac{\mathbf{D}'^2}{2!} + \ldots\right)$$

$$+\left(I+\alpha\mathbf{D}+\alpha^2\frac{\mathbf{D}^2}{2!}+\dots\right)\mathbf{D}'\left(I+\alpha\mathbf{D}'+\alpha^2\frac{\mathbf{D}'^2}{2!}+\dots\right)$$

que pode ser reescrito como

$$\mathbf{D}+\alpha\mathbf{D}\mathbf{D}'+\alpha\mathbf{D}^2+\alpha^2\mathbf{D}^2\mathbf{D}'+\alpha^2\frac{\mathbf{D}^2}{2!}+\dots$$
$$+\mathbf{D}'+\alpha\mathbf{D}\mathbf{D}'+\alpha\mathbf{D}'^2+\alpha^2\mathbf{D}\mathbf{D}'^2+\alpha^2\frac{\mathbf{D}'^2}{2!}+\dots$$

Como essa soma converge absolutamente, podemos mudar a ordem do somatório. Supondo que α é pequeno, podemos ignorar os termos com potência maior ou igual a três e ficamos com

$$(\mathbf{D}+\mathbf{D}')+\alpha(\mathbf{D}\mathbf{D}'+\mathbf{D}'\mathbf{D}+\mathbf{D}^2+\mathbf{D}'^2)+\alpha^2\left(\mathbf{D}^2\mathbf{D}'+\mathbf{D}\mathbf{D}'^2+\frac{\mathbf{D}^2}{2!}+\frac{\mathbf{D}'^2}{2!}+\dots\right)$$

Analisaremos agora qual o sinal dessa derivada quando o par de áreas i e j são vizinhos de primeira ou segunda ordem. Quando são vizinhos de primeira ordem, o primeiro termo da soma domina e a derivada é positiva. Ou seja, a correlação condicional aumenta quando o parâmetro α diminui. Lembre-se que a correlação condicional tem o sinal oposto ao sinal de cada um dos termos da matriz de precisão. Portanto, à medida que a associação espacial entre as áreas aumenta ($\alpha \to -\infty$), a correlação condicional entre todos vizinhos de primeira ordem aumenta. Por outro lado, se i e j são vizinhos de segunda ordem, o primeiro termo da soma é zero e o segundo termo é diferente de zero e, portanto, domina a soma. Como, em geral, estamos interessados no caso em que $\alpha < 0$ essa soma assume valor negativo. Isso significa que, se duas áreas são vizinhas de segunda ordem o valor da correlação condicional entre elas diminui à medida que o parâmetro α diminui. Ou seja, à medida que a associação espacial entre as áreas aumenta, a correlação condicional entre vizinhos de segunda ordem fica mais forte. Porém fica mais forte negativamente, o que não é intuitivo, visto que o aumento da associação espacial está causando uma maior repulsão entre as áreas.

3.6 Conclusão

Mostramos nesse trabalho que o modelo proposto por [49], apesar de apresentar vantagens, principalmente no que se refere à eficiência computacional, apresenta alguns aspectos pouco intuitivos. Esse tipo de comportamento pode ser observado para outros modelo amplamente utilizados, como por exemplo os modelos CAR. Porém no caso desse último modelo isso só ocorre quando o parâmetro de dependência espacial tem sinal negativo, ou seja, um caso com pouca aplicabilidade em problemas reais. Já para o modelo de [49], esse comportamento é observado mesmo em situações de grande interesse prático. Portanto, a aplicação desse modelo deve ser feita com cautela, visto que os resultados apresentados podem estar refletindo um tipo de dependência espacial que não possui menor sentido prático.

Chapter 4

Inferindo a localização de usuários do *Twitter*

Abstract

As redes sociais como o *Twitter* e o *Facebook* são fontes valiosas para monitoramento de eventos em tempo real, como terremotos, epidemias, etc. Para esse tipo de vigilância uma informação essencial é a localização do usuário. Grande parte desse conteúdo não possui uma informação geográfica associada, visto que nem todos usuários optam por divulgar esse tipo de informação. Entretanto, características do comportamento dos usuários, como os amigos aos quais ele se associa e os tipos de mensagens que ele publica podem nos dar dicas de sua localização espacial. Neste trabalho apresentamos um método para inferir a localização espacial de usuários do *Twitter*. Ao contrário das abordagens apresentadas até o momento, nós incorporamos dois tipo de informação no processo de inferência, o texto postado pelos usuários e a sua rede de amigos. O método é avaliado e as taxas de acerto ficam em torno de 90%.

4.1 Introdução

A Internet 2.0 têm se tornado uma fonte inesgotável e barata de dados a serem analisados. Os usuários são motivados continuamente a publicarem suas opiniões, características pessoais e registrar fatos que ocorrem a sua volta. Isso tudo de uma maneira bem simples e rápida. Serviços de microblog, como o *Twitter*, permitem a disseminação da informação em uma velocidade sem precedentes. Criado em 2006, o *Twitter* é um microblog através do qual os usuários podem publicar mensagens de até 140 caracteres, que são chamadas de *tweets*.

Os usuários desse serviço podem funcionar como uma espécie de radar que provém informação em tempo real sobre eventos como terremotos [66], epidemias [33], etc. Para fazer esse tipo de vigilância, uma informação essencial é a localização do usuário. Ela pode ser informada de três maneiras distintas, cada uma delas com graus diferentes de precisão e acurácia. O usuário pode informar em seu perfil o local onde mora. Como esse campo pode ser preenchido livremente, um grande volume de localizações inválidas, como "Marte", ou com baixa precisão,

como "Brasil", são informadas pelos usuários. A segunda maneira consiste em obter a localização geográfica a partir do endereço de IP da máquina. Esse tipo de georeferenciamento não é muito confiável e deve ser atualizado continuamente. No Brasil, por exemplo, esse serviço localiza corretamente 72% dos IP's dentro de um raio de 40 quilômetros. A terceira forma é obtida a partir das coordenadas do GPS de aparelhos celulares. Esse terceiro tipo é o que tem maior precisão e confiabilidade, porém, como está restrito aos casos em que o usuário posta a mensagem de um aparelho celular com GPS e ainda permite que essa informação seja divulgada, esse tipo de informação geográfica está presente apenas em uma pequena fração dos *tweets*. Em alguns países como o Brasil, essa proporção não passa de 1%.

Apesar da informação geográfica não ser explícita em grande parte dos casos, alguns aspectos sobre o comportamento do usuário podem nos dar dicas sobre sua localização. Por exemplo, o conjunto de *tweets* publicados por um usuário podem nos fornecer informação sobre onde ele reside. Alguns trabalhos têm sido desenvolvidos nesse sentido. [19] estimam a localização do usuário identificando palavras que caracterizam determinadas localizações, com por exemplo o termo "rockets" que está associado à cidade de Houston. Os autores definem que esse tipo de palavra deve ter uma alta frequência em um determinado ponto do espaço e essa frequência deve cair rapidamente quando nos afastamos desse ponto. [55] inferem a localização do usuário, também com base no texto, utilizando algoritmos de classificação. Esses últimos são capazes de inferir a localização para diferentes níveis de granularidade: cidade, estado e zona temporal.

Além do texto, as relações de seguidor/seguido entre os usuários também podem nos trazer informação geográfica. Sabe-se que, principalmente em países onde a língua falada não é o inglês, as relações de amizade no *Twitter* tendem a refletir a proximidade geográfica entre usuários. Tendo isso em vista, a rede de amizades pode ser usada como fonte de informação para o processo de inferência. [24] propõem um método de estimação segundo o qual localização de um usuário será aquela mais frequente entre seus amigos. Ao determinar as relações de amizade consideram que dois usuários são amigos apenas se eles se seguem mutuamente. Isso evita que páginas institucionais ou perfis de celebridades atrapalhem o processo de inferência. Ao longo deste trabalho consideraremos essa mesma definição de amizade entre os usuários. Alguns dos problemas encontrados por esses autores se referem ao reduzido número de amigos que alguns usuários possuem, o que dificulta bastante o processo de inferência. Além disso, usuários com muitos amigos também são fonte de erros, visto que tais relações de amizades, muito provavelmente, não refletem a proximidade geográfica.

A nossa principal contribuição neste trabalho é apresentar um método que seja capaz de incorporar os dois tipos de informação, o texto e a rede de amizades, para fazer inferência sobre a localização do usuário. Este manuscrito está organizado da seguinte maneira. Na Seção 4.2 apresentamos com mais detalhes os principais métodos apresentados até o momento para se resolver o problema de inferência de localização de usuários em redes sociais. Na Seção 4.3 descrevemos a metodologia proposta. Em seguida, na Seção 4.4, apresentamos os resultados experimentais obtidos até o momento. E, finalmente, na Seção 4.5, apresentamos as principais conclusões do trabalho e os trabalhos futuros a serem realizados.

4.2 Trabalhos Relacionados

O trabalho de [24] têm como objetivo aumentar o número de *tweets* geolocalizados. Dessa maneira são capazes de aprimorar o uso do *Twitter* como ferramenta de monitoramento em tempo real, capturando de maneira adequada tendências espaço-temporais das mensagens. A motivação original desse trabalho se iniciou a partir de uma tentativa de monitoramento da dengue apresentada por [33]. A base de dados utilizada pelos autores corresponde aos *tweets* relevantes para o monitoramento dessa doença. Ocorre, porém, que se forem coletados apenas os usuários que publicam mensagens sobre a dengue, o grafo ficaria muito esparso e pouco informação estaria disponível para o processo de inferência. A solução encontrada foi utilizar esses usuários inicialmente coletados como sementes e em seguida fazer uma busca em profundidade no grafo. Essa coleta corresponde a um tipo de amostragem "bola de neve". Ao realizar essa coleta, eles iniciam com 5 usuários sementes e fazem a busca em profundidade terminando com 61.400 usuários. Desses usuários, apenas 24.767 tinham uma informação válida sobre localização em seu perfil. Nos casos em que o usuário tinha mais de uma de uma fonte de informação, GPS e IP, por exemplo, consideravam aquela mais confiável. Para as informações obtidas através do GPS, como elas se referem cada *tweet* postado pelo usuário, eles consideravam localização mais frequente entre as 100 últimas postagens. Observaram, porém, que as 10 últimas publicações já eram suficiente para atingir os objetivos desejados.

A inferência sobre a localização de cada usuário é feita através de um esquema de votação. A localização de um usuário é definida como aquela mais frequente entre seus amigos. Entretanto, esse processo de inferência apresenta uma série de complicações. Primeiramente, para usuários com poucos vizinhos, a inferência não é confiável. Usuários com muitos vizinhos também apresentam o mesmo tipo de problema, visto que muito provavelmente suas relações de amizade não estarão refletindo proximidade geográfica. Um outro problema encontrado é que esse esquema de votação pode levar a empates. A solução encontrada pelos autores para resolver alguns desses problemas consiste em definir um número mínimo e máximo de amigos necessários para se inferir a localização. Eles mostram experimentalmente que esses dois parâmetros têm grande influência sobre os resultados obtidos. A performance da metodologia é avaliada através de duas medidas: sensibilidade e precisão. A sensibilidade é definida como a porcentagem de usuários para os quais inferiu-se corretamente a localização. A precisão é definida como a proporção de localizações inferidas corretamente. Avaliam os resultados para o número mínimo de amigos iguais a 1, 5, 10, 15 e 20; e o número máximo igual a 50, 100 e 200. Ao definir esses limiares podem determinar que apenas o usuário para o qual a inferência está sendo realizada deve ter o número de amigos dentro do intervalo ou que seus amigos também devem satisfazer essa condição. Os autores observam que, ao exigir que tanto o usuário quanto seus amigos satisfaçam os critérios estabelecidos, os resultados obtidos são melhores. Para resolver o problema do empate, determinam um número mínimo de votos necessários para se inferir a localização do usuário. Testam os valores 2, 3, 5 e 10. Observam que o valor 2 leva a melhores resultados. Selecionando-se os melhores critérios conseguem uma taxa de acerto próxima a 80%.

O método proposto por [24] é bem simples, mas necessita da calibração de vários parâmetros, como número mínimo e máximo de vizinhos considerados; frequência mínima de amigos residentes em um determinado local. Outros métodos de classificação para dados dispostos em grafos já foram apresentados na literatura e uma revisão ampla sobre o assunto pode ser obtida em [54]

[50] apresentam uma metodologia que é capaz de classificar os sítios de uma maneira semi-supervisionada. Isso significa que apenas alguns vértices precisarão estar rotulados. Os autores mostram que essa nova proposta apresenta bons resultados se comparada com aqueles obtidos nos demais trabalhos. A ideia deles consiste em utilizar o algoritmo *PageRank* [60] para inferir os rótulos dos vértices. O *PageRank* é um algoritmo de classificação de páginas da internet utilizado pelo *Google*. Ele associa um peso a cada página, que mensura qual a sua importância em toda a rede. Para obtenção desses pesos, considera-se um modelo segundo o qual o usuário percorre as páginas da internet seguindo um passeio aleatório. Em cada passo ele escolhe, com igual probabilidade, uma dentre as páginas referenciadas por aquela na qual ele está atualmente. Os pesos associados à páginas serão dados pela distribuição estacionária dessa Cadeia de Markov. Para evitar que essa cadeia entre em um estado absorvente, a cada passo, atribui-se uma probabilidade p de que o usuário escolha aleatoriamente entre todas as páginas disponíveis na internet. Essa propriedade é denominada teleportação. A forma como usuário escolhe uma dessas páginas pode ser definida de várias maneiras. A mais simples delas é considerar todas as probabilidade iguais. Pode-se a atribuir ainda probabilidades que dependam de fatores externos à rede. [50] modificam essas probabilidades de teleportação a fim de inferir a classificação dos nós do grafo. A ideia dos autores é obter o *PageRank* das páginas um número k de vezes, onde k é o número de categorias às quais cada nó pode pertencer. Para cada uma, eles consideram que o passeio pode teleportar apenas para os sítios pertencentes àquela categoria. Por exemplo, a primeira vez corresponde a categoria de número 1. Então a teleportação só será permitida para aquelas nós rotulados com a categoria 1. Após calcular esses pesos k vezes, cada nó terá a sua importância medida em cada um dos k cenários. A categoria de cada nó não rotulado será dada pela categoria para a qual esse nó possui maior peso. A intuição por trás do método é que se um determinado nó tem grande importância no grafo quando a teleportação só pode ser feita para nós da categoria i, com alta probabilidade esse nó também deve ser da categoria i. Aplicamos essa técnica para inferência das localizações dos usuários do *Twitter* e os resultados obtidos são apresentados na Seção 4.4.

Como mencionado anteriormente, uma outra fonte de informação que pode ser utilizada para se fazer inferência sobre a localização do usuário são as mensagens que e ele publica. [19] fazem uso do fato de que os *tweets* postados pelos usuários podem conter alguma informação sobre sua posição geográfica. Essa informação pode vir de nomes específicos ou expressões que tenham maior probabilidade de estarem associadas a um determinado local. Por exemplo, a expressão *Howdy*, que em português significa "olá", é tipicamente utilizada no estado do Texas. Ao se utilizar esse tipo de informação no processo de inferência, aparecem uma série de complicações. Algumas delas são apontadas pelos autores. As mensagens postadas

apresentam muito ruído. Elas abordam os mais diversos assuntos: comida, esportes, diálogos pessoais, etc. Uma pequena fração possui, de fato, conteúdo espacial. Um outro problema é a presença recorrente de gírias e expressões informais. Além disso, existe o problema da mobilidade do usuário. Ele pode morar em local, mas postar mensagens sobre outro. Por exemplo, habitantes de Nova York podem postar mensagens sobre o terremoto no Haiti. O usuário pode ainda ter mais de uma localização válida. Ele pode, por exemplo, estar viajando ou se mudar de um local para o outro. Todos esses aspectos dificultam o processo de inferência e alguns deles são abordados pelos autores do trabalho.

[19] utilizam como base dados uma amostra aleatória de 1% de todos os *tweets* postados em um determinado período. Dentro dessa amostra, descartam todas mensagens que não possuem uma localização válida associada. Observam que essa amostra representa bem a distribuição espacial da população americana. O processo de inferência é realizado estimando-se a probabilidade do usuário pertencer a determinada cidade, de acordo com as palavras por ele publicadas. Denotam por $S_{tweets}(u)$ conjunto de *tweets* postados pelo usuário u e por $p(i|S_{tweets}(u))$ probabilidade do usuário u estar localizado na cidade i dado que ele postou um conjunto de *tweets* $S_{tweets}(u)$. O objetivo é estimar $p_u(i)$, ou seja, a probabilidade do usuário u pertencer à cidade i. Primeiramente é feito um pré-processamento do texto. Removem-se caracteres especiais, palavras com frequência muito baixa e as *stoping words*, que são artigos, preposições, etc. O conjunto filtrado de palavras é denotado por $S_{words}(u) \subset S_{tweets}(u)$. A probabilidade de interesse $p_u(i)$ é estimada por

$$p_u(i) = \sum_{w \in S_{words}(u)} p(i|w)p(w)$$

onde w representa cada uma das palavras postadas pelo usuário, $p(w)$ é a proporção de vezes que a palavra w aparece em todo conjunto de dados e $p(i|w)$ proporção de vezes em que a cidade i aparece dentre as vezes que w é citada.

Os autores observam, porém, que essa maneira de estimar não gera bons resultados. Conseguem localizar apenas 10,21% dos usuários a menos de 100 milhas de sua localização verdadeira. Identificam, então, duas possíveis fontes de problemas. A primeira delas é que muitas palavras são postadas em diversos locais, ou seja, não têm informação local, como, por exemplo, as palavras *peace* e *world*. Essas palavras não trazem nenhum informação relevante e adicionam ruído aos dados. O segundo problema é que muitas cidades, principalmente aquelas com população pequena, têm um conjunto muito esparso de palavras. Para resolver o primeiro problema, os autores tentam identificar as palavras que devem ser classificadas como "palavras locais". Elas devem ser caracterizadas por apresentarem uma alta frequência em um ponto central e sua frequência deve cair rapidamente quando nos afastamos desse ponto. Esses pontos são definidos como centróides das cidades. Consideram um ponto fixo. Denotam por d a distância do centróide de onde a palavra foi postada até o ponto considerado. A probabilidade de uma palavra a w ser postada à uma distância d do centro é dada por

$$Cd^{-\alpha}$$

onde C identifica a frequência no centro e α controla a velocidade com que a frequência cai quando d aumenta. Seja S um conjunto de ocorrências da palavra w. Tal palavra pode ou não ser mencionada na cidade i. A probabilidade de ser mencionada é

$$Cd_i^{-\alpha}$$

e de não ser mencionada é

$$1 - Cd_i^{-\alpha} .$$

Portanto log-verossimilhança do modelo é dada por

$$f(C, \alpha) = \sum_{i \in S} \log Cd_i^{-\alpha} + \sum_{i \notin S} \log(1 - Cd_i^{-\alpha}) .$$

Estimam C e α para cada uma das palavras usando máxima verossimilhança. Se a palavra w é local, os valores estimados de C e α devem ser altos. Portanto, uma forma simples de se definir quais são as palavras locais seria definindo-se limiares para os parâmetros C e α. Essa estratégia, porém, não apresenta bons resultados. Os autores, então, optam por utilizar um método de classificação denominado *SimpleCart*.

O segundo problema, de esparsidade das palavras, é resolvido aplicando-se um método de suavização. Isso é feito de três maneiras distintas. A primeira delas é utilizar o método de Laplace, que consiste em somar um em todas as contagens. Uma segunda possibilidade é realizar uma suavização geográfica, que consiste em utilizar a informações dos vizinhos espaciais de cada cidade. A terceira tentativa dos autores foi utilizar o mesmo modelo que identifica as palavras locais, ou seja, a probabilidade de um usuário pertencer a cidade i, dado que publicou a palavra w, será definida por $p'(i|w) = C(w)d_i^{-\alpha(w)}$.

Os resultados experimentais mostram que a metodologia mais simples, sem filtrar as palavras ou fazer suavização, atinge uma taxa de acertos de apenas 10,21%. Ao selecionarem apenas as palavras locais, essa proporção sobre para 49%. Incorporando a suavização na metodologia, a porcentagem de acerto aumenta para 51%. Verifica-se, portanto, que o grande impacto nos resultados é obtido quando se classifica as palavras em locais e não locais. O ganho marginal ao se incorporar a suavização é bem reduzido, o que indica que talvez ela seja até dispensável. Os autores apontam duas possíveis direções de trabalhos futuros. A primeira delas desenvolver um modelo capaz de inferir a movimentação do usuário ao longo do tempo. A segunda consiste associar a informação do texto com a informação da rede para prever a localização. Esse último problema que é o foco principal deste trabalho.

Até o presente momento, nenhum trabalho publicado apresentou uma abordagem que integra as duas informações: o grafo de amizades e o texto do *tweet*. Na próxima seção apresentamos um método de inferência que leva em conta as duas fontes de informação. Para tanto, o modelo considerado para o grafo será um Campo de Markov e a informação do texto será extraída através do método *Naive Bayes*.

4.3 Metodologia

Vamos denotar por θ_i a localização do usuário i. A informação geográfica será a cidade usuário. Chamaremos de $\boldsymbol{\theta}_{-i}$ o vetor com as localizações de todos usuários, com exceção do i-ésimo. O vetor $\mathbf{w}_i$ será formado por todas as palavras dos *tweets* postados pelos usuários nos últimos tempos. O nosso objeto então é, para um usuário i cuja localização é desconhecida, encontrar o valor mais provável de θ_i. Para encontrarmos esse valor, precisamos da distribuição de probabilidade de todo o vetor $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$, onde N é o número total de usuários analisados. Essa distribuição será obtida através de um amostrador de *Gibbs* [29]. Vamos gerar os valores dos θ_i desconhecidos a partir da seguinte distribuição condicional:

$$P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i)\,. \tag{4.1}$$

Sabe-se que a amostra do vetor de θ's gerados dessa maneira terá uma distribuição que aproxima a distriuição conjunta $P(\boldsymbol{\theta})$.

Para encontrarmos a expressão da distribuição condicional completa em (4.1) , podemos fazer algumas simplificações. Sabemos que

$$P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i) \propto P(\theta_i|\boldsymbol{\theta}_{-i})P(\mathbf{w}_i|\theta_i, \boldsymbol{\theta}_{-i})\,.$$

Vamos analisar primeiramente o termo $P(\theta_i|\boldsymbol{\theta}_{-i})$. É razoável supor que as localizações dos usuários no grafo seguem um Campo de Markov. Dessa maneira $P(\theta_i|\boldsymbol{\theta}_{-i})$ se simplifica em

$$P(\theta_i|\boldsymbol{\theta}_{-i}) = P(\theta_i|\boldsymbol{\theta}_{\partial i})$$

onde o vetor $\boldsymbol{\theta}_{\partial i}$ contém a localização de todos os vizinhos de i. Essa é uma suposição razoável, visto que se fornecemos a informação sobre a localização dos amigos de um usuário, toda a informação contida no resto da rede é dispensável. Vamos supor ainda que a distribuição das localizações dos usuários pode ser modelada por um campo markoviano denominado Modelo de Potts [76]. Esse modelo é uma generalização de um modelo mais simples, o Modelo de Ising. No modelo de Ising cada sítio pode pertencer a duas classes, o que seria equivalente a termos apenas duas localizações. Já no modelo de Potts podemos ter um número arbitrário de classes ou de localizações. De acordo com esse modelo, probabilidade de um sítio pertencer a uma determinada classe será uma função crescente do número de vizinhos desse sítio pertencentes à essa classe, ou seja

$$P(\theta_i|\boldsymbol{\theta}_{\partial i}) \propto \exp\left(\beta \sum_{j:j\in\partial_i} \sigma_{ij}\right)$$

onde σ_{ij} é uma função indicadora, que recebe valor 1 se i e j pertencem à mesma classe e zero, caso contrário. O parâmetro β é conhecido como a temperatura do modelo e mede o grau de interação entre os sítios. Para $\beta > 0$ temos um modelo atrativo, ou seja, sítios vizinhos tenderão pertencer à mesma classe.

Vejamos agora como podemos simplificar o termo $P(\mathbf{w}_i|\theta_i, \boldsymbol{\theta}_{-i})$. Primeiramente, observe que se queremos predizer o texto de um usuário, dado que sabemos sua localização, a informação geográfica sobre seus amigos é desnecessária. Dessa maneira, essa probabilidade se simplifica a

$$P(\mathbf{w}_i|\theta_i, \boldsymbol{\theta}_{-i}) = P(\mathbf{w}_i|\theta_i) \, .$$

Para encontrarmos o valor de $P(\mathbf{w}_i|\theta_i)$ utilizaremos o método *Naive Bayes*. Vamos considerar que as palavras postadas pelo usuário são independentes entre si, ou seja

$$P(\mathbf{w}_i|\theta_i) = \prod_j P(w_{ij}|\theta_i)$$

onde w_{ij} denota a j-ésima palavra publicada pelo i-ésimo usuário. Essa suposição é aparentemente pouco razoável, porém método *Naive Bayes* têm mostrados ótimos resultados apesar de sua simplicidade [77]. Cada uma das probabilidade $P(w_{ij}|\theta_i)$ pode ser estimada como a proporção de vezes que a palavra w_{ij} aparece dentre todas as palavras publicadas por usuários que residem na localização θ_i.

Temos então que a probabilidade condicional $P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i)$ fica na forma

$$P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i) \propto P(\theta_i|\boldsymbol{\theta}_{\partial i}) \prod_j P(w_{ij}|\theta_i) \, .$$

E os valores de θ_i para aqueles usuários cuja localização é desconhecida podem ser atualizados através do amostrador de *Gibbs*.

4.4 Resultados Experimentais

A fim de verificar a adequação do método proposto consideramos um conjunto de 8477 usuários do *Twitter* residentes nas cidades de Belo Horizonte, Rio de Janeiro e São Paulo. O número total de usuários coletados dessas três cidades são respectivamente 1402, 4061 e 3014. Consideramos a princípio apenas essas três localidades, pois em cidades muito pequenas a informação geográfica disponível não é suficiente para fazermos inferência. Esses 8477 usuários formam um grafo composto por 140715 arestas, o qual é apresentado na Figura 4.1. Notamos que existe uma componente fortemente conectada nesse grafo e vários usuários isolados, com menos de dois amigos.

Desses 8477, escondemos a localidade de 1000 e vamos tentar prever o seu valor correto. A inferência sobre a localização do usuário foi feita em um primeiro momento usando apenas o grafo. O método utilizado foi aquele proposto por [50]. Essa metodologia foi implementada para os dados do *Twitter* utilizando o pacote *igraph*, inserido no software R. A probabilidade de teleportação foi fixada em 0,5. Separamos 1000 usuários para fazer a inferência e a taxa de acerto obtida foi de 58,4%. Para os usuários residentes em Belo Horizonte, a proporção de acertos foi de 62,89%. Dentre os usuários residentes no Rio de Janeiro e São Paulo as proporções de acerto foram de respectivamente 67,98% e 43,22%. Notamos que o método

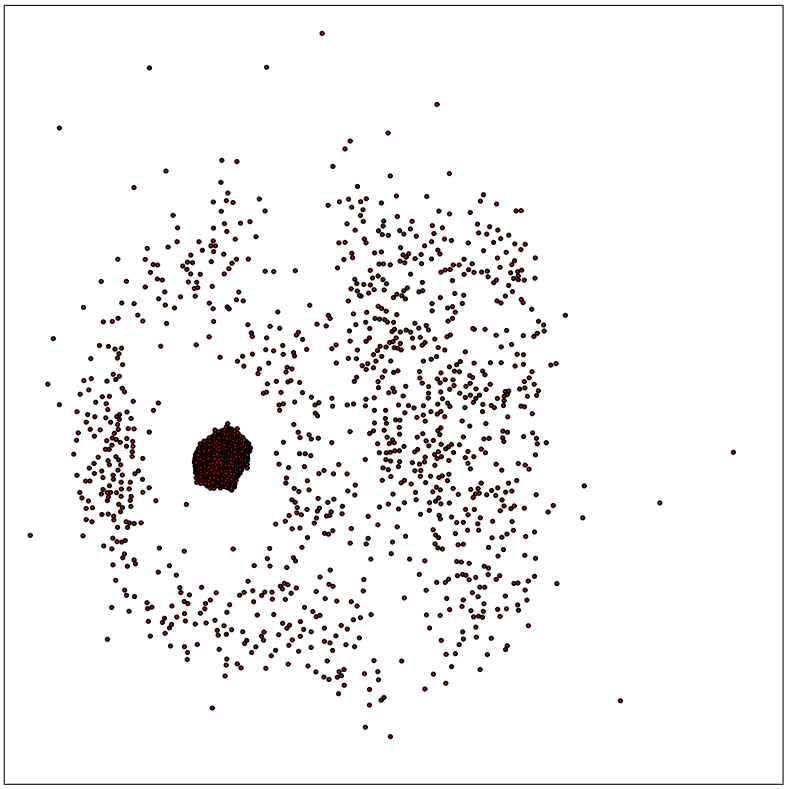

Figure 4.1: Grafo de relações amizades dos 8477 usuários coletados.

apresentou uma pior performance para os usuários residentes na cidade de São Paulo. Notamos que essa abordagem, apesar de apresentar uma taxa de acerto próxima a alguns trabalhos anteriores, como [19], não apresenta ainda resultados satisfatórios. Vamos aplicar então o segundo método, que incorpora informações da rede e do texto simultaneamente.

Para a implementação da segunda metodologia precisamos, além do grafo de amizades dos usuários, dos *tweets* por eles publicados. Como até o momento não foi possível finalizar a coleta de texto dos usuários, os mesmos foram simulados. Por motivos de simplificação, consideramos um dicionário com apenas 20 palavras. A distribuição de palavras postadas por cada usuário irá depender do local onde ele reside. A distribuição de palavras para cada cidade segue uma Multinomial de vetor de parâmetros $\boldsymbol{\alpha}$. Os parâmetros $\boldsymbol{\alpha}$ são gerados de uma Dirichlet com todos parâmetros iguais a um. O número de palavras que cada usuário posta segue uma distribuição Poisson de parâmetro 7.

De posse dos dados simulados, fixamos a temperatura em $\beta = 1$ e atualizamos o amostrador de *Gibbs* 1000 vezes. Consideramos como a localidade do usuário i aquela que tivesse maior probabilidade, ou seja, aquela que foi mais frequente no nosso processo de amostragem. Em seguida, contamos dentre os usuários para os quais estávamos estimando a localização, em quantos deles conseguimos acertar. Dentre os 1000 usuários, a localização foi predita corretamente para 922, o que representa um taxa de acerto de 92,2%. Dentre os usuários residentes em Belo Horizonte, a proporção de acerto foi de 93,91%. Para os que moram no Rio de Janeiro e São Paulo as proporções de acerto foram de 91,68% e 92,1% respectivamente. Isso mostra

que o método conseguiu uma boa taxa de acerto para as três cidades.

A fim de avaliarmos a qualidade do método proposto executamos uma validação cruzada com fator 10. Ou seja, dividimos o conjunto de dados em 10 partes e realizamos o processo inferência para cada uma dessas partes utilizando o restante dos dados. Em seguida, tiramos a média das taxas de acerto obtidas em cada uma dessas 10 vezes. A taxa de acerto média obtida para todos os usuários foi de 91,16%. Para os usuários residentes no Rio de Janeiro e São Paulo, a taxa média de acerto foi de respectivamente 91,39% e 91,42%. Dentre os usuários que residem na cidade de Belo Horizonte, essa proporção foi 72,45%. A média de acerto para essa cidade foi baixa, pois em uma das realizações a taxa de acerto foi de apenas 50%. Não se conseguiu até o momento nenhuma explicação para esse resultado atípico. Percebe-se, portanto, que em média o método apresenta resultados bons, com taxas de acerto que ultrapassam aquelas obtidas em trabalhos anteriores, que não passam de 80%. A Figura 4.2 apresenta as matrizes de confusão dos resultados obtidos pelo método proposto. A matriz à esquerda mostra os valores da medida de sensibilidade, ou seja, a probabilidade do usuário ser classificado na cidade i, dado que ele é de fato da cidade i. Por exemplo, um usuário de Belo Horizonte, tem uma probabilidade de 72,45% de ser classificado como sendo de Belo Horizonte; 6,39% de chance de ser classificado como do Rio de Janeiro e 21,17% de chance de ser classificado como de São Paulo. A matriz à direita apresenta os valores da precisão do método, a probabilidade de um usuário ser da cidade i, dado que ele foi classificado como sendo dessa cidade. Por exemplo, se um usuário é classificado como sendo de Belo Horizonte, ele tem 77,39% de ser de fato de Belo Horizonte, 7,63% de ser do Rio de Janeiro e 14,98% de chance de ser de São Paulo. Nota-se então, que para a cidade de Belo Horizonte grande parte dos erros consistem em classificar o usuário como sendo de São Paulo. Isso pode ocorrer devido à uma intereção maior entre os usuários dessas duas cidades. De um modo geral, notamos que o método apresenta bons resultados e as probabilidades de acertos são claramente altas.

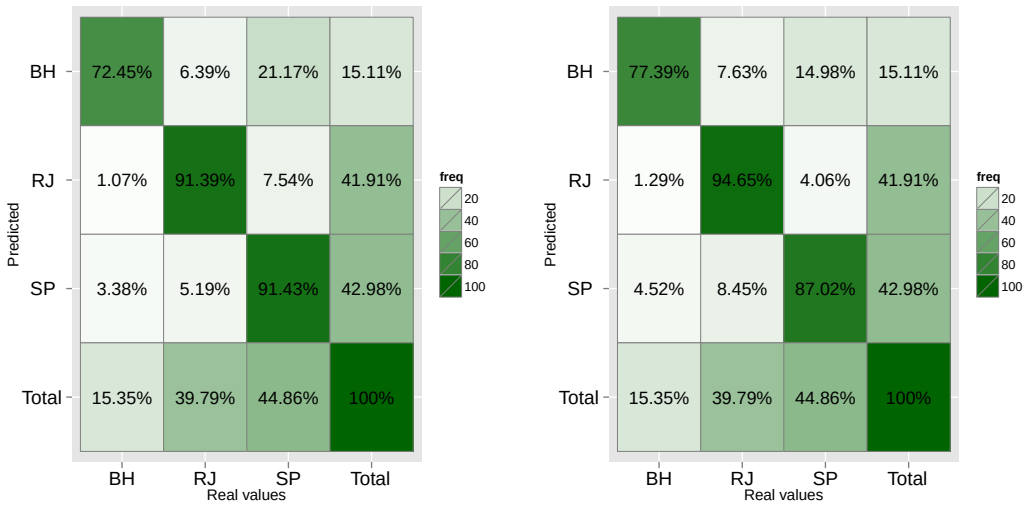

Figure 4.2: Matrizes de confusão dos resultados obtidos aplicando-se a metodologia proposta. A matriz à esquerda apresenta medidas de sensibilidade e à direita as medidas de precisão do método.

4.5 Conclusão

Neste trabalho apresentamos uma metodologia para inferir a localização de usuários do *Twitter*. Notamos que ao integrarmos as informações dos textos do usuário e de sua rede de amizade, fomos capazes de inferir sua localização com uma alta precisão.

Como trabalhos futuros, pretendemos utilizar a base de textos real publicada pelos usuários do *Twitter*. Essa base de texto está em processamento e alguns problemas adicionais já estão sendo identificados ao lidarmos com o texto real. Algumas dessas complicações já foram apontadas por [19]. Como por exemplo, o fato de muitas palavras postadas por usuários não conterem informação espacial alguma. Muito provavelmente precisaremos filtrar essas palavras de alguma maneira. Pretende-se ainda aplicar o método para uma base de dados maior, com um maior número de cidades e usuários. A metodologia utilizada neste trabalho pode ser estendida para analisar dados de outra redes sociais, como *Facebook*, *Flickr*, *Instagram*, etc.

Bibliography

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

[2] R. M. Assunção and E. T. Krainski. Neighborhood dependence in bayesian spatial models. *Biometrical Journal*, 51:851–869, 2009.

[3] R. M. Assunção and E. T. Krainski. Neighborhood dependence in bayesian spatial models. *Biometrical Journal*, 51:851–869, 2009.

[4] R. M. Assunção, J. Potter, and S. Cavenaghi. A bayesian space varying parameter model applied to estimating fertility schedules. *Statistics in Medicine*, 21:2057–2075, 2003.

[5] R. M. Assunção, J. E. Potter, and S. Cavenaghi. A bayesian space varying parameter model applied to estimating fertility schedules. *Statistics in Medicine*, 14:2057–2075, 2002.

[6] S. Banerjee, B. Carlin, and A. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2003.

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

[8] L. Bernardinelli and C. Montomoli. Empirical bayes versus fully bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 11:983–1007, 1992.

[9] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

[10] J. Besag and C. Koopeberg. On conditinal and intrinsic autoregressions. *Biometrika*, 82:733–46, 1995.

[11] J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–20, 1991.

[12] J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–20, 1991.

[13] N. Best, S. Richardson, and A. Thomson. A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–39, 2005.

[14] D. Billheimer, T. Cardoso, E. Freeman, P. Guttorp, H. Ko, and M. Silkey. Natural variability of benthic species composition in the delaware bay. environmental and ecological statistics. *Environmental and Ecological Statistics*, 4:95–115, 1997.

[15] C. R. Blyth. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67:364–366, 1972.

[16] A. Brezger, T. Kneib, and S. Lang. Bayesx-software for bayesian inference based on markov chain monte carlo simulation techniques, 2003.

[17] B. Carlin and S. Banerjee. Hierarchical multivariate car models for spatio-temporally correlated survival data. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 7*, pages 45–63. Oxford University Press, 2003.

[18] B.P. Carlin and S. Banerjee. Hierarchical multivariate car models for spatio-temporally correlated data. In *Bayesian Statistics, vol. 7*, pages 45–63. Bernardo, J. and Bayarri, M. and Berger, J. and Dawid, A. and Heckerman, D. and Smith, A. and West, M., 2003.

[19] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768. ACM, 2010.

[20] Y. M. T. Chiu, T. Leonard, and KW. Tsui. The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):pp. 198–210, 1996.

[21] D. Cooley and S.R. Sain. Spatial hierarchical modeling of precipitation extremes from a regional climate model. *Journal of Agricultural, Biological and Environmental Statistics*, 15:381–402, 2010.

[22] N. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1991.

[23] N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley and Sons, 2011.

[24] C. A. Davis Jr., G. L. Pappa, D. R. R. de Oliveira, and F. L. Arcanjo. Inferring the Location of Twitter Messages Based on User Relationships. Transactions in GIS. In Valerie Barr and Zdravko Markov, editors, *FLAIRS Conference*, pages 735–751. AAAI Press, 2004.

[25] L. E. Eberly and B. P. Carlin. Identifiability and convergence issues for markov chain monte carlo fitting of spatial models. *Statistics in Medicine*, 19:2279–2294, 2000.

[26] P. Elliot and Wartenberg D. Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspect*, 112(9):998–1006, 2004.

[27] A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of Spatial Statistics*. CRC Press, 2010.

[28] A. E. Gelfand, K. Hyon-Jung, C.F. Sirmans, and S. Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98:2057–2075, 2003.

[29] A. E. Gelfand and A. F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

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

[31] A.E. Gelfand and P. Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4:11–15, 2003.

[32] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[33] J. Gomide, A. Veloso, W. Meira, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the ACM*, WebSci'11, pages 1–8, 2011.

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

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

[36] C.E. Hunt and F.R. Hauck. Sudden infant death syndrome. *Nelson Textbook of Pediatrics*, 174:1736–1742, 2007.

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

[38] M. Iosifescu. *Finite Markov Processes and Their Applications*. Chichester: John Wiley and Sons, 1989.

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

[40] B. Jones and M. West. Covariance decomposition in undirected gaussian graphical models. *Biometrika*, 92:779–786, 2005.

[41] B. Jones and M. West. Covariance decomposition in undirected gaussian graphical models. *Biometrika*, 92(4):779–786, 2005.

[42] H. Kim, D. Sun, and R.K. Tsutakawa. A bivariate bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, 96:1506–1521, 2001.

[43] L. Knorr-Held and N. Best. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, Series A*, 164:73–85, 2001.

[44] M. Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and methods*, 26:1481–1496, 1997.

[45] S. Lang and L. Fahrmeir. Bayesian generalized additive mixed models. a simulation study. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50:201–220, 2001.

[46] A. B. Lawson and A. Clark. Spatial mixture relative risk models applied to disease mapping. *Statististics in Medicine*, 21:359–370, 2002.

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

[48] J. LeSage and R. K. Pace. *Introduction to Spatial Econometrics.* Chapman and Hall/CRC, 2009.

[49] J. P. LeSage and R. K. Pace. A matrix exponential spatial specification. *Journal of Econometrics*, 140(1):190–214, September 2007.

[50] F. Lin and W. Cohen. Semi-supervised classification of network data using very few labels. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '10, pages 192–199, Washington, DC, USA, 2010. IEEE Computer Society.

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

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

[53] Y.C. MacNab. On gaussian markov random fields and bayesian disease mapping. *Statistical Methods in Medical Research*, 20:49–68, 2011.

[54] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.*, 8:935–983, 2007.

[55] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? inferring home locations of twitter users. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[56] K. Mardia. Multidimensional multivariate gaussian markov random fields with applications to image processing. *Journal of Multivariate Analysis*, 24:265–284, 1988.

[57] E.C. Marshall and D.J. Spiegelhalter. Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, 22:1649–1660, 2003.

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

[59] J. Norton and X. Niu. Intrinsically autoregressive spatiotemporal models with application to aggregated birth outcomes. *Journal of the American Statistical Association*, 104:638–649, 2009.

[60] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[61] A.N. Pettitt, I.S. Weir, and A.G. Hart. A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modeling large sets of binary data. *Statistics and Computing*, 12:353–367, 2002.

[62] M. Plummer. Penalized loss functions for bayesian model comparison. *Biostatistics*, 9:523–539, 2008.

[63] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, 2005.

[64] S. R. Sain and N. Cressie. A spatial model for multivariate lattice data. *Journal of Econometrics*, 140:226–259, 2007.

[65] S. R. Sain, R. Furrer, and N. Cressie. A spatial analysis of multivariate output from regional climate models. *The Annals of Applied Statistics*, 5:150–175, 2011.

[66] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860. ACM, 2010.

[67] G. L. Silva, C. B. Dean, T. Niyonsenga, and A. Vanasse. Hierarchical bayesian spatiotemporal analysis of revascularization odds using smoothing splines. *Statistics in Medicine*, 27:2381–2401, 2008.

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

[69] H. S. Stern and N. Cressie. Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19:2377–2397, 2000.

[70] M. Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:44–47, 1997.

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

[72] M. J. Symons, R. C. Grimson, and Y. C. Yuan. Clustering of rare events. *Biometrics*, 39:193–205, 1983.

[73] M. Wall. A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference*, 121:311–324, 2004.

[74] G. White and S. K. Ghosh. A stochastic neighborhood conditional autoregressive model for spatial data. *Computational Statistics and Data Analysis*, 53:3033–3046, 2009.

[75] P. Whitle. On stationary process in the plane. *Biometrika*, 41(3-4):434–449, 1954.

[76] F. Y. Wu. The potts model. *Rev. Mod. Phys.*, 54:235–268, 1982.

[77] H. Zhang. The Optimality of Naive Bayes. In Valerie Barr and Zdravko Markov, editors, *FLAIRS Conference*. AAAI Press, 2004.