Renato Couto Rampaso

Análise Bayesiana de Dados Espaciais Explorando Diferentes Estruturas de Variância

Presidente Prudente

2014

Renato Couto Rampaso

Análise Bayesiana de Dados Espaciais Explorando Diferentes Estruturas de Variância

Dissertação apresentada ao Programa de Pós-Graduação em Matemática Aplicada e Computacional da Faculdade de Ciências e Tecnologia da UNESP para obtenção do título de Mestre em Matemática Aplicada e Computacional

Universidade Estadual Paulista – UNESP

Faculdade de Ciências e Tecnologia

Programa de Pós-Graduação em Matemática Aplicada e Computacional

Orientador: Aparecida Doniseti Pires de Souza

Coorientador: Edilson Ferreira Flores

Presidente Prudente

2014

Agradecimentos

Muitas pessoas foram essenciais para a realização deste trabalho e merecem meus mais sinceros agradecimentos.

À toda minha família: pai, mãe, irmãs e sobrinhos por todo carinho, suporte e incentivo aos estudos que me fizeram chegar até aqui.

Aos amigos de mestrado, Ana Paula, Suelen, Leandro, Pedro e Taciana pelas várias horas de estudos juntos e desafios superados durante o percorrer do curso.

À minha orientadora, professora Aparecida e ao meu coorientador, professor Edilson, pelo apoio e por toda a sabedoria e conhecimentos compartilhados que nortearam todo o meu trabalho, além de despertarem a minha paixão à Estatística.

Aos membros da banca, professora Vilma e professor Ricardo, pelas sugestões que contribuíram de maneira significativa para o a aprimoramento do trabalho.

Agradeço também à minha noiva, Débora, por todo seu amor e companheirismo. Em todos os momentos sempre esteve ao meu lado, dando-me o apoio necessário nos momentos difíceis que me conduziram até aqui.

Resumo

No mapeamento de doenças, o objetivo geral é estudar a incidência ou risco de mortalidade causado por uma determinada doença em um conjunto de regiões geográficas. É comum assumir que a variável resposta, geralmente uma contagem, segue uma distribuição de Poisson, cuja taxa média pode ser explicada por um grupo de covariáveis e um efeito aleatório. Para este efeito aleatório, considera-se modelos autorregressivos condicionais (CAR) que carregam informação sobre a relação de vizinhança entre as regiões. Tais relações de vizinhança são expressas por meio da matriz de variâncias presente nestes modelos. Cada modelo CAR possui características distintas que atendem a diferentes propósitos a serem considerados pelo pesquisador. O foco do trabalho foi o estudo e comparação de alguns modelos autorregressivos condicionais propostos na literatura. Para a melhor compreensão das características de cada modelo, duas aplicações com dados epidemiológicos foram conduzidas para modelar o risco de óbito por Doença de Crohn e Colite Ulcerativa e por Câncer de traqueia, brônquios e pulmões no Estado de São Paulo, no período de 2008 a 2012, utilizando cinco diferentes distribuições a priori CAR para os efeitos aleatórios. Estudos desta natureza envolvendo a Doença de Crohn e a Colite Ulcerativa são escassos, o que motiva a melhor compreensão da distribuição espacial dessas doenças. Adicionalmente, o Câncer de Pulmões é a causa de morte por câncer mais comum do mundo. Áreas com elevado risco foram identificadas. Por fim, um estudo de simulação foi feito para reforçar os resultados obtidos e averiguar o desempenho dos modelos na presença de diferentes níveis de dependência espacial.

Palavras-chaves: Modelos autorregressivos condicionais. Mapeamento de doenças. Inferência bayesiana espacial.

Abstract

In disease mapping, the overall goal is to study the incidence or risk of mortality caused by a specific disease in a number of geographical regions. It is common to assume that the response variable, generally a count, follows a Poisson distribution, whose average rate can be explained by a group of covariates and a random effect. For this random effect, it is considered conditional autoregressive models (CAR), which carry information about the neighborhood relationship between the regions. Such neighborhood relations are expressed by the variance matrix present in the models. Each CAR model has distinct characteristics that serve different purposes to be considered by the researcher. The focus of this dissertation was the study and comparison of some conditional autoregressive models proposed in the literature. For better understanding of the characteristics of each model, two applications with epidemiological data were conducted to model the risk of death due to Crohn's Disease and Ulcerative Colitis, and due to trachea, bronchus and lung cancer in the State of São Paulo, in the period of 2008-2012, using different CAR prior distributions for the random effects. Studies of this nature involving Crohn's Disease and Ulcerative Colitis are scarce, which motivates a better understanding of the spatial distribution of these diseases. In addition, the Cancer Lungs is the most common cause of cancer deaths in the world. Areas with high risk were identified. Finally, a simulation study was done to strengthen the results and assess the performance of the models in the presence of various levels of spatial dependence.

Key-words: Conditional autoregressive models. Disease mapping. Spatial bayesian inference.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Compreender a distribuição espacial de dados oriundos de fenômenos aleatórios consiste em um grande desafio para o esclarecimento de questões importantes em diversas áreas do conhecimento. Do ponto de vista da estatística, o interesse está em desenvolver modelos probabilísticos que representem adequadamente a distribuição espacial do fenômeno em estudo. Exemplos de aplicação envolvem o mapeamento de doenças e da criminalidade em uma região de interesse, a modelagem de poluentes do ar em um centro urbano, a modelagem da quantidade de chuva em uma determinada região, a análise de imagens de satélites, experimentação agrícola, entre outras.

No mapeamento de doenças, o objetivo geral é estudar a incidência ou risco de mortalidade causado por uma determinada doença em um conjunto de regiões geográficas. Tal área tem relevante importância em estudos epidemiológicos e na definição de políticas públicas, identificando locais que apresentem taxas elevadas e possíveis fatores de riscos associados. Stern e Cressie (2000) afirmaram que mapas que apresentam taxas simples, como a própria incidência ou a razão do número de óbitos da doença e a população da área em estudo têm sido criticados por não serem confiáveis, devido à variância não constante associada com a heterogeneidade nos tamanhos populacionais.

Uma alternativa para a análise de dados de área é considerar modelos que suavizem as taxas estudadas. Como a variável aleatória de interesse é em geral uma contagem que representa a ocorrência de um determinado fenômeno aleatório, é comum assumir que a variável resposta segue uma distribuição de Poisson, cuja taxa média pode ser explicada por um grupo de covariáveis e um efeito aleatório. Para esse efeito aleatório, considera-se uma distribuição a priori que carrega a informação sobre a relação de vizinhança entre as áreas da região de interesse.

Uma classe de modelos bastante popular utilizada para representar os efeitos aleatórios espaciais é a autorregressiva condicional (CAR). A especificação desses modelos está diretamente ligada à sua matriz de variâncias. Matriz essa de extrema importância por permitir a incorporação da estrutura espacial no modelo. Diferentes especificações para essa matriz resultam em modelos CAR distintos. Uma imposição necessária para definir um modelo válido é que tal matriz seja simétrica e definida positiva, o que requer certas restrições.

Nesse contexto, Lee (2011) apresentou um estudo envolvendo quatro diferentes modelos da classe CAR: o modelo intrínseco (ICAR), de convolução, de Cressie e de Leroux. O objetivo desta dissertação é a comparação desses diferentes modelos CAR através de aplicações e de um estudo de simulação como em Lee (2011), e complementar

com outras abordagens para a representação da estrutura de vizinhança espacial presentes na literatura, como, por exemplo, o modelo de Lu, que permite estimar componentes da matriz de adjacências.

A escassez de estudos nesse contexto, que compara os diferentes modelos CAR e suas estruturas de variância para os efeitos aleatórios justifica a realização desse trabalho. Aqui, são listadas as particularidades de cada modelo, vantagens e desvantagens que permitirão ao pesquisador embasamento suficiente para escolher o mais apropriado e que mais atenda às suas necessidades.

Este relatório está organizado como se segue. No Capítulo 2 é apresentada uma breve revisão da literatura, apontando os acontecimentos históricos mais relevantes e algumas aplicações, utilizando os modelos autorregressivos condicionais. No Capítulo 3, os modelos são apresentados, com a definição de suas formulações, características e particularidades. No Capítulo 4, duas aplicações com dados reais utilizando os modelos CAR são conduzidas. A primeira refere-se a dados sobre o número de óbitos por local de residência causados pela Doença de Crohn e Colite Ulcerativa no Estado de São Paulo, no período de 2008 a 2012. A segunda é referente ao número de óbitos causados por Câncer de traqueia, brônquios e pulmões para o mesmo local e período. Para reforçar os resultados obtidos e com o interesse de se averiguar o desempenho dos modelos na presença de diversos níveis de dependência espacial, um estudo de simulação é apresentado no Capítulo 5. Finalmente, no Capítulo 6 um panorama geral sobre os resultados obtidos e considerações finais são levantadas.

2 Revisão da Literatura

Neste capítulo é feita uma breve revisão da literatura para os acontecimentos históricos e estudos mais relevantes acerca da classe de modelos autorregressivos condicionais e suas principais aplicações.

Um artigo considerado inovador foi publicado por Besag (1974), que introduziu a noção de modelos autorregressivos com abordagem condicional para dados de área. Tal artigo foi marcante e bastante elogiado por diversos pesquisadores renomados da época, tais como D. R. Cox, A. G. Hawkes, P. Clifford e P. Whittle. Besag concluiu que a abordagem condicional para a especificação e análise de interação espacial é mais atrativa que a abordagem conjunta.

Um dos problemas que os modelos espaciais dessa natureza lidavam era a dificuldade computacional para a estimação. O grande avanço no uso prático desses modelos veio, porém, na década de 90, juntamente com a revolução causada pelo uso de métodos computacionais de Monte Carlo via Cadeias de Markov (MCMC). Tais métodos foram popularizados entre os bayesianos pelo artigo de Gelfand e Smith (1990). Essa nova forma de abordagem para a inferência tornou a estimação dos efeitos aleatórios mais simples e funcional.

Besag, York e Mollié (1991) trazem, novamente, outro trabalho relevante para a área. Nesse artigo, os autores fazem uso do modelo ICAR, aplicado na área de restauração bayesiana de imagens e trazem pela primeira vez o modelo de convolução para o mapeamento de doenças.

Ideias iniciais de modelos CAR próprios são apresentadas em Cressie (1993) e Haining (1993). Stern e Cressie (2000) apresentaram uma aplicação de um modelo CAR próprio. Nesse trabalho foram propostos dois métodos para aproximar uma validação cruzada da distribuição preditiva a posteriori, com o intuito de se verificar se valores extremos são potenciais descobertas epidemiológicas ou apenas evidências de um mal ajuste do modelo. Nesse mesmo ano, Leroux, Lei e Breslow (2000) propuseram um novo modelo CAR próprio considerado mais geral, que acomoda diversos outros modelos de acordo com a especificação de seus parâmetros.

Também é importante ressaltar a contribuição de Thomas et al. (2004), que desenvolveram o GeoBUGS, módulo existente no software OpenBUGS (*Bayesian inference Using Gibbs Sampling*) para produção de mapas referentes a dados de mapeamentos de doenças e outros modelos espaciais. Nesse módulo encontram-se implementados os modelos intrínseco e de convolução, além de exemplos já bastante explorados na literatura, como o referente aos dados de mortalidade por câncer labial na Escócia, presentes em Clayton e

Kaldor (1987) e Breslow e Clayton (1993).

Uma nova abordagem para modelos CAR foi apresentada por Lu e Carlin (2005) e Lu et al. (2007). Em tais modelos, o foco é conduzir uma análise de fronteira que permita identificar "barreiras" entre regiões vizinhas. A matriz de adjacências proposta é aleatória e estima-se a probabilidade de que duas regiões com fronteira em comum sejam consideradas vizinhas, utilizando para tal um modelo de regressão logística.

Além desses artigos já citados, Assunção e Krainski (2009) exploraram diversos aspectos da matriz de covariância a posteriori do modelo ICAR, ajudando a melhor entender alguns resultados que antes eram contra-intuitivos e justificando o porquê de modelos CAR serem uma escolha eficiente para acomodar os efeitos aleatórios espaciais. Outro trabalho relevante foi apresentado por Rodrigues e Assunção (2012), no qual propuseram um modelo que considera ordens mais altas de vizinhança espacial. Isto é, a estrutura de vizinhança é parte do espaço paramétrico e pode ser estimada.

Lee (2011) apresentou um estudo de simulação e o desenvolvimento de uma aplicação comparando os modelos ICAR, de convolução, de Cressie e de Leroux, envolvendo o ajuste de dados relacionados à incidência de câncer em *Greater Glasgow*, na Escócia, no período de 2001 a 2005, através desses modelos. Nessa aplicação, foi utilizado um conjunto de covariáveis, como o aumento no consumo do tabaco, poluição do ar, além de algumas variáveis sócio-econômicas. Como já mencionado, tal artigo serviu de base para o desenvolvimento desta dissertação.

Modelos autorregressivos condicionais também são utilizados para dados longitudinais. Achcar et al. (2011) aplicaram modelos de regressão de Poisson espaço-temporais na área epidemiológica, estudando a contagem de casos de malária na Floresta Amazônica brasileira, no período de 1999 a 2008. Covariáveis que podem ter importância na predição de casos de malária foram incluídas no modelo, como por exemplo o nível de desmatamento.

A aplicação de modelos CAR é feita, em sua maioria, no mapeamento de doenças. Entretanto, os modelos espaciais especificados condicionalmente provaram sua utilidade também em outras áreas, como em análise e processamento de imagens (MOLINA; KATSAGGELOS; MATEOS, 1999), estudos de associação geográficas (LEE; FERGUSON; MITCHELL, 2009), experimentação agrícola (MARTIN, 1990), classificação de dados de satélite (WIMBERLY et al., 2009), entre outras.

3 Modelos Autorregressivos Condicionais

Neste capítulo será apresentada a formulação geral da classe de modelos Bayesianos Hierárquicos para mapeamento de doenças. Além disso, alguns modelos CAR propostos na literatura serão explorados com detalhes. Antes, porém, será traçado um breve paralelo com outras propostas existentes para lidar com dados com estrutura espacial.

Basicamente, existem duas diferentes maneiras de se especificar um modelo espacialmente estruturado. Uma delas é a abordagem geoestatística para o processo espacial. Nessa abordagem, é comum assumir que os dados são relacionados a pontos, observados no centro ou no centroide de cada região e as distâncias entre esses pontos são utilizadas para representar a estrutura espacial através de uma função de covariância. Tais funções de covariância devem respeitar algumas condições matemáticas para admitir formas permissíveis (ver Ripley (1981, p.10-13) e Christakos (1984) para um estudo mais aprofundado sobre a permissibilidade de funções de covariância). Haining (1993, p.90-94) cita alguns exemplos de funções de covariância, entre elas a função de correlação triangular, o modelo esférico e a função exponencial. Wall (2004) cita um problema comum em se utilizar a abordagem geoestatística, que é a arbitrariedade de se escolher o centroide e atribuir a ele toda a informação da região. Por outro lado, uma vantagem dessa abordagem é que a função de covariância espacial é modelada diretamente e sua estrutura possui um entendimento mais simples e direto.

A outra maneira de se especificar um modelo espacial leva em conta uma estrutura de vizinhança de índices discretos. Aqui, cada observação não está atrelada a somente um ponto, mas sim a toda uma região. Nessa abordagem, ao invés de se definir um centroide para se trabalhar com distâncias entre as regiões, utiliza-se uma matriz que informa quais são os vizinhos de cada região. Uma vez que tal estrutura de vizinhança está definida, são adotados modelos autorregressivos similares a modelos de séries temporais discretas para representar os efeitos aleatórios espaciais. Um dos mais populares está na classe de modelos autorregressivos condicionais (CAR). Esses modelos se mostraram bastante úteis no contexto da Inferência Bayesiana, utilizados como distribuições a priori para efeitos aleatórios em modelos hierárquicos. Nas seções subsequentes serão definidos diferentes modelos CAR conhecidos da literatura.

3.1 Mapeamento de Doenças e a Abordagem Bayesiana

Em mapeamento de doenças, o interesse é investigar a incidência ou o risco de determinada doença em n áreas contíguas. Tais riscos são frequentemente dispostos em mapas, que ajudam a visualizar a distribuição da incidência da doença, detectar áreas

com riscos elevados e possíveis indícios de associação espacial. Por exemplo, pode-se estar interessado em visualizar o número de óbitos ou incidência de determinado tipo de câncer ao longo das regiões de um estado.

De acordo com Gelfand e Fuentes (2010), o mapeamento de doenças refere-se a uma coleção de métodos que estendem a estimação em pequenas áreas para utilizar diretamente a configuração espacial e assumir correlação espacial positiva entre observações, tomando essencialmente mais informação de áreas vizinhas do que áreas distantes, além de suavizar taxas locais para valores locais entre vizinhos. O termo mapeamento de doenças surgiu em Clayton e Kaldor (1987), que definiram métodos Bayesianos Empíricos construídos a partir de uma regressão de Poisson com interceptos aleatórios definidos com correlação espacial.

Para a construção de modelos de regressão de Poisson, seja $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ o vetor de valores observados de uma determinada doença. Sabe-se que a ocorrência de determinadas doenças está ligada a fatores de risco conhecidos, como a distribuição etária e por sexo da população. Sendo assim, para retirar o efeito que essas diferenças podem causar na incidência da doença, os valores esperados, denotados por $\mathbf{E} = (E_1, E_2, ..., E_n)$, são calculados, considerando a estrutura demográfica do local. O vetor $\mathbf{E}$ pode ser visto como os valores esperados sob a hipótese de risco constante entre as regiões.

Devido ao importante papel desempenhado pelo vetor de valores esperados $\mathbf{E}$, é de interesse uma breve elucidação do procedimento para sua obtenção. Suponha que existam K diferentes estratos demográficos. Neste trabalho, foram considerados diferentes grupos de acordo com a faixa etária da população (detalhes adicionais sobre os dados coletados são discutidos no Capítulo 4). Considere P_{ik} o total populacional da região i, no estrato k, $i = 1, ..., n$ e $k = 1, ..., K$. Além disso, seja r_k a proporção da população que se espera desenvolver a doença, no estrato k. Então, o valor esperado para a região i é

$$E_i = \sum_{k=1}^{K} P_{ik} r_k.$$

Nesse caso, o valor de r_k precisa ser estimado. Cressie (1993) cita duas diferentes maneiras para se realizar tal estimação. Se as estimativas estiverem disponíveis em uma fonte externa do conjunto de dados (por exemplo, em um registro de dados internacionais), então os dados serão padronizados externamente. Entretanto, a abordagem mais comum é adotar uma padronização interna, utilizando o próprio vetor de valores observados da doença $\mathbf{Y}$. Um estimador natural é a proporção da população no estrato k que desenvolveu a doença. Formalmente, seja Y_{ik} o número de casos observados na i-ésima região, no estrato k. Note que $Y_i = \sum_k Y_{ik}$. Então, o estimador de r_k por padronização interna é

$$\hat{r}_k = \frac{\sum_{i=1}^{n} Y_{ik}}{\sum_{i=1}^{n} P_{ik}}.$$

A medida mais simples para mensurar o risco de uma doença é através da razão de mortalidade padronizada — *SMR* (*standardised mortality ratio*, em inglês) —, calculada, para uma determinada área i, por $SMR_i = Y_i/E_i$. Valores superiores a um implicam um risco de doença acima do esperado, enquanto que abaixo de um indicam um risco abaixo do esperado para a respectiva área. Entretanto, um baixo valor de E_i pode acontecer se a população de determinado local é muito baixa ou se a doença em estudo é rara, o que implica um alto risco não verossímil para a região em questão.

Para contornar tal problema, modelos espaciais bayesianos podem ser adotados. Tais modelos permitem utilizar covariáveis que podem fornecer informação sobre o risco de mortalidade da doença, além de um conjunto de efeitos aleatórios que capturam a dependência entre regiões vizinhas. Uma formulação geral dessa classe de modelos é dada por

$$\begin{aligned} Y_i|E_i, R_i &\sim \text{Poisson}(E_i R_i) \\ \log(R_i) &= \alpha + \boldsymbol{x}_i^t \boldsymbol{\beta} + \phi_i \; i = 1, 2, ...n. \end{aligned} \tag{3.1}$$

Aqui, R_i denota o risco da doença na área i, que é estimado considerando um intercepto α comum a todas as regiões, um conjunto de p covariáveis $x_i^t = (x_{i1}, ..., x_{ip})$ e um efeito aleatório ϕ_i. A distribuição a priori utilizada para o parâmetro α é uma *flat*, o que corresponde a uma uniforme imprópria U($-\infty$,$+\infty$) em toda a reta real. Mais detalhes que justificam a escolha dessa priori serão fornecidos na Subseção (3.2.1). Para os parâmetros do vetor de coeficientes $\boldsymbol{\beta}$, são assumidas em geral distribuições a priori normais com baixa precisão. A precisão é definida como o inverso da variância.

Na distribuição de Poisson, $\text{Var}(Y_i) = \text{E}(Y_i)$, muito embora em casos práticos ocorre que $\text{Var}(Y_i) > \text{E}(Y_i)$. Esse cenário é conhecido como superdispersão. O efeito aleatório ϕ_i busca capturar a dependência espacial e/ou alguma superdispersão presente nos dados. Uma superdispersão pode ocorrer devido a existência de fatores de risco não mensurados.

3.2 Especificação de um Modelo CAR

Por ser de fácil manipulação e bastante flexível, a distribuição normal multivariada é utilizada para representar a distribuição conjunta dos efeitos aleatórios. Pode-se assumir então,

$$\boldsymbol{\phi} \sim \text{NM}\left(\boldsymbol{\mu}, \Sigma(\boldsymbol{\theta})\right), \tag{3.2}$$

em que $\boldsymbol{\mu}$ é um vetor de médias e $\Sigma(\boldsymbol{\theta})$ é a matriz de variâncias que considera a estrutura espacial através de um conjunto de parâmetros $\boldsymbol{\theta}$.

Seja n o número total de áreas em estudo. De um modo geral, considera-se que

$$\Sigma(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{\Phi}; \qquad \boldsymbol{\Phi} = (\mathbf{I} - \rho\, \mathbf{C})^{-1} \mathbf{M}, \tag{3.3}$$

em que σ^2 é um parâmetro geral de variância, $\mathbf{I}$ é uma matriz identidade $n \times n$, ρ é um parâmetro que mensura a dependência espacial, $\mathbf{C} = C_{ij}$ é uma matriz de associação espacial de zeros em sua diagonal e $\mathbf{M} = M_{ii}$ é uma matriz diagonal.

Os modelos para $\boldsymbol{\phi}$ em (3.2) são geralmente especificados através de n distribuições condicionais completas $p(\phi_i|\boldsymbol{\phi}_{-i})$, em que $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$. Para definir um modelo válido, deve-se especificar as matrizes C_{ij} e M_{ii} de tal modo que a matriz $\boldsymbol{\Phi} = (\mathbf{I} - \rho\,\mathbf{C})^{-1}\mathbf{M}$ seja simétrica e definida positiva.

Inspecionando $\boldsymbol{\Phi}^{-1}$, nota-se que $\boldsymbol{\Phi}$ é simétrica se $m_{jj}c_{ij} = m_{ii}c_{ji}$. Além disso, é fácil perceber que $\boldsymbol{\Phi} = \mathbf{M}^{\frac{1}{2}}\left(\mathbf{I} - \rho\,\mathbf{M}^{-\frac{1}{2}}\mathbf{C}\mathbf{M}^{\frac{1}{2}}\right)^{-1}\mathbf{M}^{\frac{1}{2}}$. Logo, a matriz de variâncias é definida positiva para $\rho \in (\rho_{min}, \rho_{max})$, em que $1/\rho_{min}$ e $1/\rho_{max}$ são o menor e maior autovalor da matriz $\mathbf{M}^{-\frac{1}{2}}\mathbf{C}\mathbf{M}^{\frac{1}{2}}$, respectivamente. Na prática, espera-se uma dependência espacial positiva. Assim, pode-se limitar o intervalo de ρ em $(0, \rho_{max})$. Um valor de $\rho = 0$ implica independência espacial.

Seja $j \sim i$ a notação que denota a presença de vizinhança entre as áreas j e i. Duas áreas são ditas vizinhas se partilham fronteira em comum, embora outras maneiras de definir vizinhanças possam ser adotadas (ver, por exemplo, Cressie (1993, p.384-385)). As distribuições condicionais completas de um modelo CAR geral podem ser definidas como

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \mathrm{N}\left(\mu_i + \rho\sum_{j\sim i} c_{ij}(\phi_j - \mu_j), \sigma^2 m_{ii}\right). \tag{3.4}$$

Pode-se mostrar que a especificação condicional resulta em uma distribuição conjunta normal multivariada, já estabelecida em (3.2). Isso é feito utilizando-se o Teorema da Fatoração.

Teorema da Fatoração (BESAG, 1974). Suponha que se $x_1, ..., x_n$ podem ocorrer individualmente nas áreas $1, ..., n$, respectivamente, então elas podem ocorrer conjuntamente. Formalmente, se $P(x_i) > 0$, para todo i, então $P(x_1, ..., x_n) > 0$. Essa é a chamada condição de *positividade*, de Hammersley e Clifford (1971), assumida ao longo deste teorema. Na prática, essa condição é usualmente satisfeita. Seja o espaço amostral Ω o conjunto de todas as possíveis realizações de $\boldsymbol{x} = (x_1, ..., x_n)$. Isto é, $\Omega = \{\boldsymbol{x} : P(\boldsymbol{x} > 0)\}$. Seja $P(.)$ a função de probabilidade conjunta. Então segue que, para duas realizações $\boldsymbol{x}$ e $\boldsymbol{y}$ de Ω,

$$\frac{P(\boldsymbol{x})}{P(\boldsymbol{y})} = \prod_{i=1}^{n} \frac{P(x_i|x_1, ..., x_{i-1}, y_{i+1}, ..., y_n)}{P(y_i|x_1, ..., x_{i-1}, y_{i+1}, ..., y_n)}. \tag{3.5}$$

Prova: Utilizando probabilidade condicional, pode-se escrever

$$P(\boldsymbol{x}) = P(x_n|x_1, ..., x_{n-1})P(x_1, ..., x_{n-1}); \tag{3.6}$$

aqui, $P(x_1, ..., x_{n-1})$ não pode ser fatorada em uma forma útil, já que, por exemplo, $P(x_{n-1}|x_1, ..., x_{n-2})$ não é facilmente obtida das distribuições condicionais dadas. Entre-

tanto, com a introdução de y_n, tem-se

$$P(\boldsymbol{x}) = \frac{P(x_n|x_1, ..., x_{n-1})}{P(y_n|x_1, ..., x_{n-1})} P(x_1, ..., x_{n-1}, y_n). \tag{3.7}$$

Sob a condição de positividade, o denominador em (3.7) é positivo. Operando agora com x_{n-1} em $P(x_1, ..., x_{n-1}, y_n)$, chega-se em

$$P(x_1, ..., x_{n-1}, y_n) = \frac{P(x_{n-1}|x_1, ..., x_{n-2}, y_n)}{P(y_{n-1}|x_1, ..., x_{n-2}, y_n)} P(x_1, ..., x_{n-2}, x_{n-1}, y_{n-1}, y_n)$$

após similar introdução de y_{n-1}. Novamente a condição de positividade é utilizada. Continuando esse processo, chega-se em (3.5) e o teorema está provado.

O teorema da fatoração tem uma versão para dados contínuos, em que $P(.)$ é substituído por uma função densidade de probabilidade $p(.)$. Então, a condição de normalidade é $\int_\Omega p(\boldsymbol{y})d\boldsymbol{y} = 1$ (CRESSIE, 1993).

Proposição: A especificação condicional em (3.4) implica que

$$\boldsymbol{\phi} \sim \text{NM}\left(\boldsymbol{\mu}, \sigma^2(\mathbf{I} - \rho\,\mathbf{C})^{-1}\mathbf{M}\right), \tag{3.8}$$

desde que $(\mathbf{I} - \rho\,\mathbf{C})$ seja invertível e $(\mathbf{I} - \rho\,\mathbf{C})^{-1}\mathbf{M}$ seja simétrica e definida positiva.

Prova: Utilizando o teorema da fatoração para densidades de probabilidade e tomando $\boldsymbol{x} = \boldsymbol{\phi}$ e $\boldsymbol{y} = \boldsymbol{\mu}$ em (3.5), tem-se que

$$\begin{aligned}
log\left(\frac{p(\boldsymbol{\phi})}{p(\boldsymbol{\mu})}\right) &= -\frac{1}{2\sigma_i^2}\sum_{i=1}^{n}\left(\phi_i - \mu_i - \rho\sum_{j=1}^{i-1} c_{ij}(\phi_j - \mu_j)\right)^2 \\
&\quad + \frac{1}{2\sigma_i^2}\sum_{i=1}^{n}\left(\rho\sum_{j=1}^{i-1} c_{ij}(\phi_j - \mu_j)\right)^2 \\
&= -\frac{1}{2\sigma_i^2}\sum_{i=1}^{n}(\phi_i - \mu_i)^2 \\
&\quad + \frac{1}{\sigma_i^2}\rho\sum_{i=1}^{n}\sum_{j=1}^{i-1} c_{ij}\,(\phi_i - \mu_i)\,(\phi_j - \mu_j) \\
&= -(1/2)(\boldsymbol{\phi} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{I} - \rho\mathbf{C})(\boldsymbol{\phi} - \boldsymbol{\mu}).
\end{aligned} \tag{3.9}$$

O lado direito em (3.9) é o expoente de uma distribuição normal multivariada n-dimensional com vetor de médias $\boldsymbol{\mu}$ e matriz de variâncias $(\mathbf{I} - \rho\,\mathbf{C})^{-1}\mathbf{M}$.

A verossimilhança do modelo CAR em (3.8) pode ser expressa como

$$(2\pi)^{-n/2}|\mathbf{M}|^{-1/2}|\mathbf{I} - \mathbf{C}|^{1/2}\exp\left\{-(1/2)(\boldsymbol{\phi} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{I} - \rho\mathbf{C})(\boldsymbol{\phi} - \boldsymbol{\mu})\right\}. \tag{3.10}$$

3.2.1 Modelo Intrínseco

Um dos modelos CAR mais simples é o intrínseco (ICAR, abreviando-se), proposto por Besag, York e Mollié (1991), em que a matriz de variâncias $\boldsymbol{\Phi}$ não é definida positiva.

Por esse motivo, o modelo é impróprio e somente pode ser usado como distribuição a priori e não como verossimilhança para os dados. Esse modelo é obtido da Equação (3.4) fazendo $c_{ij} = 1/n_i$ se as áreas i e j forem adjacentes e 0 caso contrário, $m_{ii} = 1/n_i$, sendo n_i o número de vizinhos da área i e $\rho = \rho_{max}$, que com tais especificações de c_{ij} e m_{ii} implica em $\rho = 1$. Aqui, a matriz $\mathbf{C}$ é equivalente à clássica matriz de pesos normalizados. Sem perda de generalidade, também é assumido $\mu_i = 0$, $i = 1, ..., n$. Sua distribuição condicional completa pode ser expressa como

$$\phi_i|\phi_{-i} \sim \mathrm{N}\left(\frac{1}{n_i}\sum_{j\sim i}\phi_j, \frac{\sigma^2}{n_i}\right). \tag{3.11}$$

A esperança condicional do efeito aleatório ϕ_i é a média dos efeitos de seus vizinhos. Sua estrutura de variância nos induz que quanto maior o número de vizinhos da área i, n_i, maior a precisão condicional de ϕ_i.

Por se tratar de um modelo impróprio, alguns cuidados são necessários. Em geral, é conveniente assumir que os efeitos têm média zero. Além disso, uma restrição é necessária para que os efeitos aleatórios somem zero ($\sum_{i=1}^{n} \phi_i = 0$). Normalmente, essa restrição é feita numericamente recentralizando as amostras de ϕ_i em torno de sua própria média (BEST; RICHARDSON; THOMSON, 2005). Besag e Kooperberg (1995) mostraram que restringir os efeitos para somar zero e especificar um intercepto separado com distribuição a priori invariante U($-\infty$,$+\infty$) (*flat*) é equivalente ao modelo sem restrição e sem intercepto. Logo, incluindo-se o intercepto no modelo, é necessária a utilização de uma distribuição a priori *flat* U($-\infty$,$+\infty$) para o mesmo (THOMAS et al., 2004). Impostas tais restrições, a distribuição conjunta do vetor de efeitos aleatórios $\boldsymbol{\phi}$ é normal multivariada, com vetor de médias $\mathbf{0}$ e matriz de variâncias (singular) $\sigma^2\mathbf{D}^-$ (em que $\mathbf{D}^-$ representa a inversa generalizada da matriz $\mathbf{D}$), sendo o *ij-ésimo* elemento dessa matriz $\mathbf{D}$ definido como

$$d_{ij} = \begin{cases} n_i, & \text{se } j = i \\ -1, & \text{se } j \sim i \\ 0, & \text{caso contrário.} \end{cases} \tag{3.12}$$

O modelo ICAR aqui considerado apresenta algumas desvantagens. A força da dependência espacial entre os efeitos aleatórios é sempre considerada máxima (justamente por considerar $\rho = 1$), sendo o modelo adequado na presença de forte correlação espacial. Outro ponto é que o parâmetro de variância σ^2 é utilizado tanto para captar superdispersão quanto dependência espacial (LEROUX; LEI; BRESLOW, 2000).

3.2.2 Modelo de Convolução

Outro modelo bastante popular para efeitos aleatórios é o modelo de convolução. Também proposto por Besag, York e Mollié (1991), combina-se o modelo intrínseco (estruturado espacialmente) com um efeito aleatório adicional (não estruturado espacialmente).

Esse modelo é dado por

$$\begin{aligned} \phi_i &= \theta_i + \psi_i, \\ \theta_i &\sim \mathrm{N}(0, \sigma_\theta^2), \\ \boldsymbol{\psi} = (\psi_1, ..., \psi_n) &\sim \mathrm{ICAR}(W, \sigma_\psi^2). \end{aligned} \tag{3.13}$$

Aqui, o termo $\boldsymbol{\psi}$ tem distribuição a priori ICAR, descrita na Equação (3.11). Uma justificativa para a inclusão de θ_i é feita a partir do estudo empírico de Breslow (1984), que observou uma variação extra adicional no modelo log-linear de Poisson. Isto é, o efeito aleatório $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ — independente entre as áreas — é adicionado com o objetivo de absorver essa variação adicional não capturada pelo efeito aleatório espacial intrínseco. Logo, variando o tamanho relativo desses dois componentes aleatórios, espera-se captar diferentes níveis de correlação espacial. O nome convolução se dá porque a densidade dos efeitos aleatórios $\phi_i's$ será a convolução das funções densidades de probabilidade conjuntas dos vetores $\boldsymbol{\theta}$ e $\boldsymbol{\psi}$ (RODRIGUES; ASSUNÇÃO, 2012).

Na estimação, pode ocorrer que um dos dois termos seja o dominante. Caso isso ocorra para o termo $\boldsymbol{\psi}$, então os riscos são estimados considerando uma maior estrutura espacial. Por outro lado, se o efeito aleatório $\boldsymbol{\theta}$ tiver maior peso, então a estimação dos riscos será concentrada em torno da média geral μ.

Algumas críticas são apontadas para o modelo de convolução. Cada observação é representada por dois efeitos aleatórios, mas apenas a soma $\theta_i + \psi_i$ é identificável (LEE, 2011). Alem disso, Eberly, Carlin et al. (2000) apontam alguns problemas de convergência na estimação, que pode ser afetada pela escolha das distribuições a priori para os parâmetros e até mesmo pelos valores iniciais escolhidos para a cadeia.

3.2.3 Modelo de Cressie

O modelo de Cressie, mais lembrado como CAR próprio, é uma alternativa para se captar diferentes níveis de dependência espacial. Suas ideias iniciais podem ser encontradas em Cressie (1993), com detalhes adicionais em Stern e Cressie (2000). A ideia aqui é introduzir ρ, de forma a ser um parâmetro a ser estimado pelo modelo.

Para as matrizes $\mathbf{C}$ e $\mathbf{M}$, será considerada aqui a mesma especificação imposta no modelo ICAR, isto é, $c_{ij} = 1/n_i$ se as áreas i e j forem adjacentes e 0 caso contrário e $m_{ii} = 1/n_i$. Com essas configurações, o limite superior para ρ é 1. Como o traço da matriz $\mathbf{M}^{-\frac{1}{2}}\mathbf{C}\mathbf{M}^{\frac{1}{2}}$ é igual a zero, então $\rho_{min} < 0 < \rho_{max}$ (pois $\mathbf{M}^{-\frac{1}{2}}\mathbf{C}\mathbf{M}^{\frac{1}{2}}$ é uma matriz quadrada com zeros em sua diagonal e $tr(\mathbf{M}^{-\frac{1}{2}}\mathbf{C}\mathbf{M}^{\frac{1}{2}}) = \sum_{i=1}^{n} \delta_i = 0$, sendo δ_i o i-ésimo autovalor da matriz $\mathbf{M}^{-\frac{1}{2}}\mathbf{C}\mathbf{M}^{\frac{1}{2}}$). Logo, um intervalo plausível que faz desse um modelo próprio é $0 \leq \rho < 1$, sendo $\rho = 1$ o modelo ICAR. Assim, o conjunto de efeitos aleatórios $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ tem uma distribuição normal multivariada com matriz de precisão $\sigma^2\,\mathbf{Q}^{-1}$,

isto é,

$$\boldsymbol{\phi} \sim \text{NM}\left(\boldsymbol{\mu}, \sigma^2\,\mathbf{Q}^{-1}\right),\ \ 0 \leq\ \rho < 1, \tag{3.14}$$

em que o *ij-ésimo* elemento da matriz $\mathbf{Q}$ é definido como

$$q_{ij} = \begin{cases} n_i, & \text{se } j = i \\ -\rho, & \text{se } j \sim i \\ 0, & \text{caso contrário.} \end{cases}$$

A distribuição condicional completa univariada para os efeitos ϕ_i é então

$$\phi_i|\phi_{-i} \sim \text{N}\left(\rho\,\frac{1}{n_i}\sum_{j\sim i}\phi_j, \frac{\sigma^2}{n_i}\right). \tag{3.15}$$

A variância condicional é a mesma do modelo intrínseco. A esperança condicional desse modelo pode ser vista como uma média ponderada da média local dos efeitos aleatórios (com peso ρ) e da média geral zero (com peso 1-ρ):

$$E(\phi_i|\phi_{-i}) = (1-\rho)\times 0 + \rho \times \frac{1}{n_i}\sum_{j\sim i}\phi_j.$$

Para $\rho = 0$ considera-se independência, já que os valores irão se concentrar em torno da mesma média zero. Por outro lado, valores de ρ próximos a um indicam forte correlação espacial. Quando $\rho = 1$, tem-se o modelo intrínseco.

Ressalta-se, contudo, que existe certa flexibilidade na especificação das matrizes $\mathbf{C}$ e $\mathbf{M}$. Stern e Cressie (2000), por exemplo, consideraram $C_{ij} = \sqrt{E_j/E_i}$ se as áreas i e j forem adjacentes e 0 caso contrário e $M_{ii} = 1/E_i$. Essa especificação, entretanto, altera o intervalo permissível de variação de ρ, sendo necessário o cuidado de se obter um novo intervalo para conseguir um modelo próprio.

Uma crítica feita a esse modelo se deve ao fato de que na ausência de dependência espacial entre os efeitos aleatórios (ρ próximo de zero), a variância condicional é invariante e continua dependendo do número de vizinhos n_i.

3.2.4 Modelo de Leroux

Outro modelo, considerado uma forma mais geral que o de Cressie, foi proposto por Leroux, Lei e Breslow (2000). O conjunto de efeitos $\boldsymbol{\phi}$ são representados por uma distribuição normal multivariada

$$\boldsymbol{\phi} \sim \text{NM}\left(\mathbf{0}, \sigma^2[\rho\,\mathbf{D} + (1-\rho)\mathbf{I}]^{-1}\right). \tag{3.16}$$

Nesse caso, a matriz de precisão é $\mathbf{L} = \rho\mathbf{D} + (1-\rho)\mathbf{I}$, em que $\mathbf{I}$ denota uma matriz identidade de ordem n e a matriz $\mathbf{D}$ é a mesma já definida na Equação (3.12). É fácil perceber que para $\rho = 0$, $\mathbf{L} = \mathbf{I}$ e tem-se um modelo com efeitos aleatórios independentes.

Já quando $\rho = 1$, $\mathbf{L} = \mathbf{D}$, ou seja, obtém-se o modelo intrínseco. Se $0 \leq \rho < 1$, a distribuição conjunta (3.16) é própria. A distribuição condicional completa univariada é então dada por

$$\phi_i|\phi_{-i} \sim \mathrm{N}\left(\frac{\rho}{n_i\rho + 1 - \rho}\sum_{j\sim i}\phi_j, \frac{\sigma^2}{n_i\rho + 1 - \rho}\right). \tag{3.17}$$

Agora, a variância condicional tem uma forma mais atrativa que no modelo (3.15). Para ρ próximo de 1, a variância condicional se aproxima de σ^2/n_i. Já quando $\rho = 0$, a variância condicional se resume a σ^2, não mais dependendo do número de vizinhos n_i. A esperança condicional pode ser vista como uma média ponderada entre a média local dos efeitos aleatórios (com peso $n_i\rho$) e da média geral 0 (com peso $1 - \rho$), isto é,

$$E(\phi_i|\phi_{-i}) = \frac{1-\rho}{n_i\rho + 1 - \rho} \times 0 + \frac{n_i\rho}{n_i\rho + 1 - \rho} \times \frac{1}{n_i}\sum_{j\sim i}\phi_j.$$

Além disso, a variância condicional também pode ser vista como uma média ponderada entre a variância local do modelo intrínseco (com peso $1 - \rho$) e a variância de um modelo independente (com peso $n_i\rho$)

$$V(\phi_i|\phi_{-i}) = \frac{1-\rho}{n_i\rho + 1 - \rho} \times \sigma^2 + \frac{n_i\rho}{n_i\rho + 1 - \rho} \times \frac{\sigma^2}{n_i}.$$

Observe que o parâmetro ρ na Equação (3.17) é fixo para todas as áreas. Uma abordagem alternativa para o modelo de Leroux é variar esse parâmetro entre as regiões. Com isso, o nível de correlação espacial pode variar ao longo de todo o mapa. A distribuição condicional para os efeitos aleatórios pode ser expressa como

$$\phi_i|\phi_{-i} \sim \mathrm{N}\left(\frac{\rho_i}{n_i\rho_i + 1 - \rho_i}\sum_{j\sim i}\phi_j, \frac{\sigma^2}{n_i\rho_i + 1 - \rho_i}\right). \tag{3.18}$$

Ressalta-se que em geral, essa especificação pode resultar em uma "super-parametrização", um indício de que os dados não fornecem informação suficiente para a inferência dos parâmetros à posteriori (MACNAB et al., 2006).

Rodrigues (2011) e Banerjee, Gelfand e Carlin (2003) apontam que modelos CAR próprios podem induzir pouca correlação entre áreas vizinhas. Essa desvantagem pode ser um dos motivos que faz dos modelos intrínseco e de convolução escolhas mais populares para acomodar os efeitos aleatórios espaciais.

3.2.5 Modelo de Lu

Lu et al. (2007) desenvolveram um modelo em que a matriz de adjacências não é mais determinística, mas sim aleatória. Em outras palavras, a estrutura de variância/vizinhança do modelo é agora estimada. Isso é feito modelando os pesos da matriz de vizinhanças como

$$w_{ij}|p_{ij} \sim \text{Bernoulli}\left(p_{ij}\right), \tag{3.19}$$

em que

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mathbf{z}'_{ij}\boldsymbol{\gamma}.$$

Aqui, $\mathbf{z}_{ij}$ é um conjunto de características conhecidas inerentes às regiões i e j, com correspondente vetor de parâmetros $\boldsymbol{\gamma}$. Tais características podem ser, por exemplo, a distância entre os centroides das duas regiões geográficas. Agora, as áreas i e j são consideradas vizinhas com probabilidade p_{ij}, dado que elas partilham fronteira em comum. O modelo de Lu permite que a estrutura de vizinhança seja determinada pelo valor do processo em cada área e por variáveis que determinam o quão similar duas regiões são. Assumindo o caso mais simples — e utilizado no decorrer deste trabalho —, com a presença de apenas uma covariável, o modelo é então definido como

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_0 + \gamma_1 z_{ij}. \tag{3.20}$$

Modelos dessa natureza são especialmente úteis em análise de fronteiras espaciais. Nesse tipo de análise, são determinadas fronteiras entre áreas que alternam valores altos e baixos para o risco de mortalidade. Tal método é frequentemente chamado de *wombling*, referência ao artigo precursor na área de Womble (1951).

A especificação para o vetor de efeitos aleatórios remete à Equação (3.8), isto é,

$$\boldsymbol{\phi} \sim \mathrm{NM}\left(\mathbf{0}, \sigma^2(\mathbf{I}-\mathbf{C})^{-1}\mathbf{M}\right),$$

em que $c_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$ para i e j adjacentes e 0 caso contrário e $m_{ii} = \frac{1}{\sum_j w_{ij}}$, em que $\mathbf{W}$ é a matriz de adjacências. Contudo, o determinante da matriz de precisão $\mathbf{M}^{-1}(\mathbf{I}-\mathbf{C})/\sigma^2 = (\mathbf{M}-\mathbf{W})/\sigma^2$ é zero para os casos em que uma área não possua vizinhos. Nesse modelo, tal cenário pode acontecer mesmo no caso em que não existam "ilhas" no conjunto de regiões em estudo, já que a natureza aleatória para a matriz de adjacências permite que $w_{ij} = 0$, ainda que i e j partilhem fronteira em comum. Para contornar tal problema, Lu et al. (2007) propuseram uma aproximação que resulta em um cálculo adequado para o determinante da matriz de precisão. Isso é feito substituindo $\mathbf{M}$ por $\mathbf{M}^*$, adicionando um pequeno valor positivo ϵ aos elementos diagonais de $\mathbf{M}$. Dessa forma, a matriz $(\mathbf{M}^*-\mathbf{W})/\sigma^2$ é diagonal dominante, simétrica e definida positiva.

Uma crítica apontada para esse modelo é que o vetor de parâmetros $\boldsymbol{\gamma}$, efeitos das covariáveis no modelo de regressão logística, nem sempre identifica satisfatoriamente os dados. Em alguns casos, distribuições a priori com teor informativo são necessárias para assegurar o efeito desejável pelas covariáveis $\mathbf{z}_{ij}$ na análise de fronteira.

3.3 Especificação Completa dos Modelos e Procedimentos de Inferência

Considerando os modelos propostos em (3.1) e (3.2), com diferentes estruturas de variância para a distribuição a priori especificada para o vetor de efeitos aleatórios $\boldsymbol{\phi}$, é preciso ainda especificar as distribuições a priori para ρ (modelo de Cressie e Leroux), $\boldsymbol{\gamma}$ (modelo de Lu) e σ (todos os modelos). Seguindo o trabalho de Gelman (2006), para o desvio padrão, foi especificada uma distribuição a priori não informativa com distribuição uniforme, isto é,

$$p(\sigma) \sim U(0, T).$$

Alguns estudos de sensibilidade para o valor de T já foram realizados (LEE, 2011), apontando que, desde que seja um valor suficientemente grande, essa escolha não altera as estimativas produzidas pelas distribuições a posteriori. Logo, aqui optou-se por fixar $T = 10$.

Para o parâmetro ρ, dos modelos de Cressie e de Leroux, definiu-se o intervalo de variação (0,1), de tal forma que os modelos sejam próprios. Logo, uma escolha de distribuição a priori natural para esses parâmetros é

$$p(\rho) \sim U(0, 1).$$

Lu et al. (2007) sugere o uso de distribuições a priori normais para γ_0 e γ_1,

$$p(\gamma_0) \sim N(\mu_0, \sigma_0^2),$$

$$p(\gamma_1) \sim N(\mu_1, \sigma_1^2).$$

O parâmetro ϵ, artifício utilizado para tornar o determinante da matriz de variâncias do modelo de Lu positiva definida, foi fixado em 0,5. Lu et al. (2007) apontam que valores de ϵ próximos de zero aumentam o vício na estimação dos riscos. Por outro lado, valores próximos de um aumentam o vício na estimação de σ^2. Logo, a escolha supracitada fixada para ϵ representa um ponto intermediário encontrado para balancear o problema enfrentado.

Assumindo o modelo especificado em (3.1) e (3.2) com $\lambda_i = E_i R_i$, $i = 1, ..., n$, a função de verossimilhança é dada por

$$\begin{aligned} L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!} = \prod_{i=1}^{n} \frac{(E_i R_i)^{y_i} \exp(-E_i R_i)}{y_i!} \\ &= \prod_{i=1}^{n} \frac{[E_i \exp(\alpha + \boldsymbol{x}_i^t \boldsymbol{\beta} + \phi_i)]^{y_i} \exp\left(E_i \exp(\alpha + \boldsymbol{x}_i^t \boldsymbol{\beta} + \phi_i)\right)}{y_i!}. \end{aligned} \quad (3.21)$$

Logo, as distribuições a posteriori conjuntas para o vetor de parâmetros dos modelos, assumindo independência entre as distribuições a priori, é:

1. Para o modelo intrínseco

$$\begin{aligned} p(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma|\boldsymbol{y}) &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma) \\ &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha) p(\boldsymbol{\beta}) p(\boldsymbol{\phi}) p(\sigma). \end{aligned} \tag{3.22}$$

2. Para o modelo de convolução

$$\begin{aligned} p(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma_\psi, \sigma_\theta|\boldsymbol{y}) &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma_\psi, \sigma_\theta) \\ &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha) p(\boldsymbol{\beta}) p(\boldsymbol{\phi}) p(\sigma_\psi) p(\sigma_\theta). \end{aligned} \tag{3.23}$$

3. Para os modelos de Cressie e de Leroux

$$\begin{aligned} p(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma, \rho|\boldsymbol{y}) &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma, \rho) \\ &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha) p(\boldsymbol{\beta}) p(\boldsymbol{\phi}) p(\sigma) p(\rho). \end{aligned} \tag{3.24}$$

4. Para o modelo de Lu

$$\begin{aligned} p(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma, w_{ij}, \boldsymbol{\gamma}|\boldsymbol{y}) &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma, w_{ij}, \boldsymbol{\gamma}) \\ &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha) p(\boldsymbol{\beta}) p(\boldsymbol{\phi}) p(\sigma) p(w_{ij}) p(\boldsymbol{\gamma}). \end{aligned} \tag{3.25}$$

De (3.22), (3.23), (3.24) e (3.25) pode-se observar que as expressões para as distribuições a posteriori são complexas e não apresentam formas fechadas. Dessa forma, para a obtenção das distribuições marginais a posteriori se faz necessário o uso de métodos numéricos e o amostrador de Gibbs será utilizado. Embora não sejam necessárias as especificações das distribuições condicionais completas devido à utilização do software OpenBUGS, as mesmas são listadas a seguir:

1. para $\boldsymbol{\alpha}$

$$p(\alpha|\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{y}) \propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\alpha);$$

2. para $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}|\alpha, \boldsymbol{\phi}, \boldsymbol{y}) \propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\beta);$$

3. para $\boldsymbol{\phi}$

   a) modelo intrínseco

$$\begin{aligned} p(\phi_i|\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}_{-i}, \sigma, \boldsymbol{y}) &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) p(\phi_i|\boldsymbol{\phi}_{-i}) \\ &\propto L(\alpha, \boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) \exp\left[-\frac{n_i}{\sigma^2} \left(\phi_i - \frac{\sum_{j \sim i} \phi_j}{n_i} \right)^2 \right]; \end{aligned}$$

b) modelo de convolução

$$\begin{aligned} p(\psi_i|\alpha,\boldsymbol{\beta},\boldsymbol{\psi}_{-i},\boldsymbol{y}) &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\psi};\boldsymbol{y})p(\psi_i|\boldsymbol{\psi}_{-i}) \\ &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\psi};\boldsymbol{y})\exp\left[-\frac{n_i}{\sigma_\psi^2}\left(\psi_i-\frac{\sum_{j\sim i}\psi_j}{n_i}\right)^2\right]; \end{aligned}$$

$$\begin{aligned} p(\boldsymbol{\theta}|\alpha,\boldsymbol{\beta},\boldsymbol{y}) &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y})p(\boldsymbol{\theta}) \\ &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y})\exp\left[-\frac{1}{\sigma_\theta^2}\sum_{i=1}^{n}\theta_i^2\right]; \end{aligned}$$

c) modelo de Cressie

$$\begin{aligned} p(\phi_i|\alpha,\boldsymbol{\beta},\boldsymbol{\phi}_{-i},\rho,\sigma,\boldsymbol{y}) &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\phi};\boldsymbol{y})p(\phi_i|\boldsymbol{\phi}_{-i}) \\ &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\phi};\boldsymbol{y})\exp\left[-\frac{n_i}{\sigma^2}\left(\phi_i-\frac{\rho\sum_{j\sim i}\phi_j}{n_i}\right)^2\right]; \end{aligned}$$

d) modelo de Leroux

$$\begin{aligned} p(\phi_i|\alpha,\boldsymbol{\beta},\boldsymbol{\phi}_{-i},\rho,\sigma,\boldsymbol{y}) &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\phi};\boldsymbol{y})p(\phi_i|\boldsymbol{\phi}_{-i}) \\ &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\phi};\boldsymbol{y})\exp\left[-\frac{n_i\rho+1-\rho}{\sigma^2}\left(\phi_i-\frac{\rho\sum_{j\sim i}\phi_j}{n_i\rho+1-\rho}\right)^2\right]; \end{aligned}$$

e) modelo de Lu

$$\begin{aligned} p(\phi_i|\alpha,\boldsymbol{\beta},\boldsymbol{\gamma},w_{ij},\boldsymbol{\phi}_{-i},\sigma,\boldsymbol{y}) &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\phi};\boldsymbol{y})p(\phi_i|\boldsymbol{\phi}_{-i}) \\ &\propto L(\alpha,\boldsymbol{\beta},\boldsymbol{\phi};\boldsymbol{y})\exp\left[-\frac{n_i}{\sigma^2}\left(\phi_i-\frac{\sum_{j\sim i}w_{ij}\phi_j}{n_i}\right)^2\right]; \end{aligned}$$

4. para σ

$$p(\sigma|\boldsymbol{\phi}) \propto \prod_{i=1}^{n} p(\phi_i|\boldsymbol{\phi}_{-i})p(\sigma);$$

5. para ρ

$$p(\rho|\boldsymbol{\phi}) \propto \prod_{i=1}^{n} p(\phi_i|\boldsymbol{\phi}_{-i})p(\rho);$$

6. para $\boldsymbol{\gamma}$

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{\phi}) &\propto \prod_{i=1}^{n} p(\phi_i|\boldsymbol{\phi}_{-i})p(\boldsymbol{\gamma}) \\ &= \prod_{i=1}^{n}\left[\frac{\exp(\gamma_0+\gamma_1 z_{ij})}{1+\exp(\gamma_0+\gamma_1 z_{ij})}\right]^{w_{ij}}\left[\frac{1}{1+\exp(\gamma_0+\gamma_1 z_{ij})}\right]^{1-w_{ij}}\exp\left(-\frac{\gamma_0^2}{2\sigma_0^2}-\frac{\gamma_1^2}{2\sigma_1^2}\right); \end{aligned}$$

7. para w_{ij}

$$\begin{aligned} p(w_{ij}|\boldsymbol{\phi}) &\propto \prod_{i=1}^{n} p(\phi_i|\boldsymbol{\phi}_{-i})p(w_{ij}) \\ &\propto \prod_{i=1}^{n} p(\phi_i|\boldsymbol{\phi}_{-i})p_{ij}^{w_{ij}}(1-p_{ij})^{(1-w_{ij})} \end{aligned} \tag{3.26}$$

A estimação foi feita utilizando o software OpenBUGS (LUNN et al., 2009). O BUGS é um programa direcionado para a Inferência Bayesiana utilizando o amostrador de Gibbs. Permite, ao usuário, a especificação do modelo ao estabelecer relações entre as variáveis. Além disso, é dotado da capacidade de determinar o melhor esquema MCMC baseado no amostrador de Gibbs para a simulação e análise do modelo especificado. A especificação de modelos pertence à classe conhecida como Grafos acíclicos dirigidos (*Directed Acyclic Graphs* — em inglês), que permite evitar a análise de estruturas complexas para uma sequência de cálculos relativamente simples.

A interface utilizada dentro do OpenBUGS foi o GeoBUGS, sistema especializado para modelagem de dados espaciais. Algumas distribuições CAR já se encontram disponíveis, como o modelo intrínseco e o de convolução. Entretanto, as distribuições de Cressie, Leroux e Lu foram implementados nessa linguagem para tornar possível as aplicações e simulação conduzidas.

4 Aplicações a Dados Epidemiológicos

Neste capítulo, são exploradas duas aplicações utilizando os modelos para mapeamento de doenças, descritos no Capítulo 3, para dados disponíveis no portal DATASUS (http://www.datasus.gov.br). As variáveis de interesse referem-se a dados epidemiológicos para o Estado de São Paulo, no período compreendido de 2008 a 2012, para mortalidade por: 1) Doença de Crohn e Colite Ulcerativa; 2) Câncer de traqueia, brônquios e pulmões. A base cartográfica considerada neste trabalho consiste na divisão administrativa do Estado de São Paulo por microrregiões, segundo o Instituto Brasileiro de Geografia e Estatística (IBGE). Essa escolha reside no fato de que muitos municípios do estado têm uma população menor que 10.000 habitantes, além de eventualmente não apresentarem caso de mortalidade por câncer num determinado período. Dessa forma, os 645 municípios foram agrupados em 63 microrregiões. Optou-se pelo óbito por residência, isto é, considerou-se o local em que o indivíduo residia ao invés do local onde foi registrado o seu óbito. O interesse na escolha da mortalidade por Doença de Crohn e Colite Ulcerativa se deve à carência de estudos nesse contexto. Por outro lado, o Câncer de traqueia, brônquios e pulmões é o tipo de câncer que mais mata no mundo, o que desperta a atenção para a melhor compreensão da distribuição espacial da doença.

Em relação ao modelo de Lu, adotou-se a proposta feita em (3.20) para estimar as probabilidades de duas regiões serem vizinhas, com uma só variável explicativa. Como covariável z_{ij}, escolheu-se a distância entre os centroides das microrregiões i e j. Dessa forma, tal variável é uma medida de dissimilaridade. A matriz de distâncias foi obtida a partir de uma base georreferenciada das microrregiões do Estado de São Paulo fornecida pelo IBGE. As distâncias foram calculadas a partir de ferramentas disponíveis nos pacotes ArcView e ArcGis. Os valores da matriz foram normalizados para o intervalo (0,1).

As escolhas propostas para as hiperprioris de γ_0 e γ_1, baseadas em Lu et al. (2007), foram

$$p(\gamma_0) \sim N(1, 1),$$
$$p(\gamma_1) \sim N(-5, 1).$$

Embora distribuições a priori não informativas sejam frequentemente utilizadas em mapeamento de doenças, a utilização de distribuições a priori informativas foram um recurso necessário para induzir o efeito desejado para a covariável em estudo. O valor negativo na média de γ_1 é usado justamente pelo fato da distância entre os centroides ser uma medida de dissimilaridade.

4.1 Doença de Crohn e Colite Ulcerativa

A Doença de Crohn e a Colite Ulcerativa são doenças inflamatórias intestinais distintas, mas que apresentam certas semelhanças clínicas. Embora não haja consenso científico, ambas possuem características que se assemelham a uma doença auto-imune. Em uma doença auto-imune, o sistema imunológico passa a produzir anticorpos que não conseguem diferenciar antígenos (agentes invasores como bactérias e vírus) de tecidos saudáveis. Isto é, o corpo ataca a si mesmo.

A Doença de Crohn afeta predominantemente a parte inferior do intestino delgado (íleo) e intestino grosso (cólon), mas pode afetar qualquer parte do trato gastrointestinal, desde a boca até o ânus. Já a colite ulcerativa ataca somente o cólon e o reto, apresentando inflamação e ulceração da camada mais superficial do cólon.

As causas para a incidência dessas doenças não são claras. Entre os especialistas, acredita-se que fatores ambientais e genéticos exerçam influência. Sabe-se também que as doenças inflamatórias intestinais são mais predominantes em países do hemisfério norte e afetam igualmente homens e mulheres. Embora sejam doenças não infecciosas, uma dependência espacial entre as regiões pode ocorrer justamente devido a presença de fatores ambientais que, por si só, são correlacionados entre si. Como uma análise exploratória, na Figura 1 estão dispostas as razões de mortalidade padronizadas referentes à Doença de Crohn e Colite Ulcerativa, calculadas para cada microrregião. A partir desse mapa, é possível obter um panorama geral de como o risco se comporta ao longo das microrregiões. Entretanto, o mapa, por si só, não fornece informações relevantes, como as relações de vizinhança entre as regiões ou a mensuração da estrutura espacial, motivo pelo qual os modelos CAR são necessários.

4.1.1 Resultados

Aplicando os modelos definidos em (3.1) e (3.2), a estimação foi realizada via MCMC através do software OpenBUGS, com 5000 iterações iniciais descartadas para aquecimento da cadeia e um total de 20000 iterações subsequentes. O salto (*thin*) utilizado foi igual a 30. Em um primeiro momento, um estudo para verificar a indicação de convergência dos parâmetros foi conduzido. O pacote CODA (*Convergence Diagnosis and Output Analysis*), que corresponde a um conjunto de funções implementadas na linguagem de programação estatística R (R Core Team, 2013), foi utilizado para obtenção de diagnósticos de convergência. Além disso, o gráfico com as trajetórias das cadeias, bem como os gráficos de autocorrelações foram obtidos (Figuras 2, 3, 4, 5 e 6). Aqui, omitiu-se os gráficos referentes às cadeias dos riscos de cada microrregião devido ao elevado número de figuras que tal tarefa demandaria. Entretanto, tais indicações de convergência também foram verificadas.

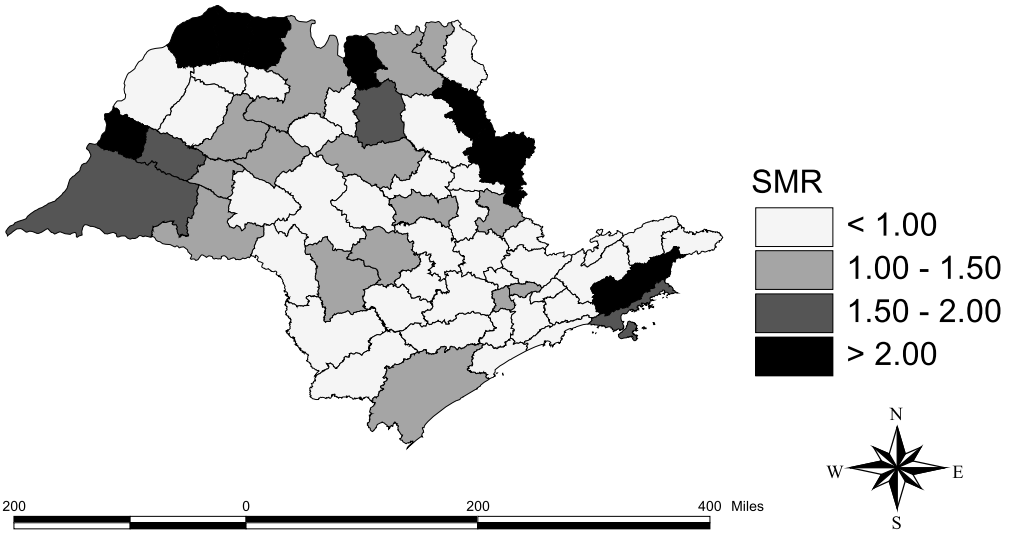


Figura 1 – Razões de mortalidade padronizadas de Doença de Crohn e Colite Ulcerativa calculadas para cada microrregião do Estado de São Paulo no período de 2008 a 2012

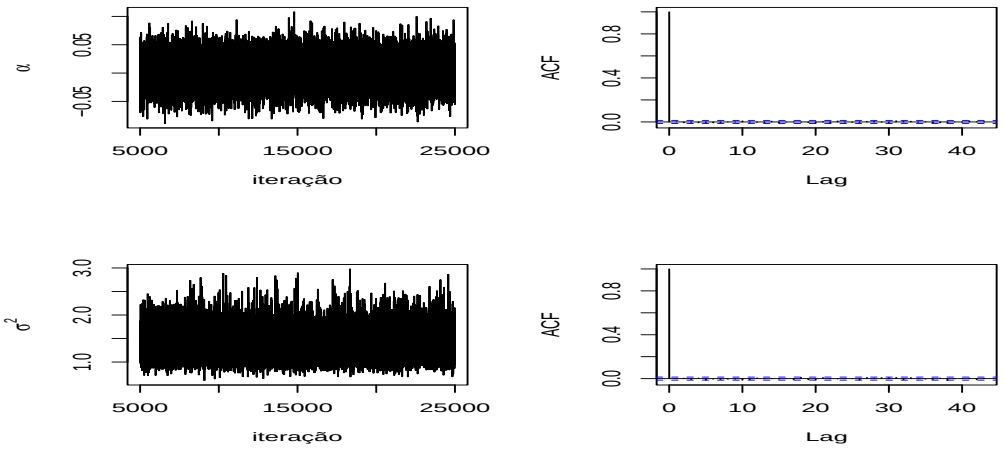


Figura 2 – Trajetória das cadeias e gráficos de autocorrelação: modelo intrínseco

Realizando uma inspeção visual, em regra, a indicação de convergência foi aparentemente assegurada para os parâmetros. A exceção mais evidente foi o intercepto do modelo de Leroux, que apresentou uma alta autocorrelação da cadeia mesmo em *lags* mais elevados. Também, o intercepto do modelo de Cressie, bem como os parâmetros de variância do modelo de convolução (σ^2_θ e σ^2_ψ) merecem certa atenção. Para verificar tais convergências de maneira mais formal, foram aplicados dois testes de diagnóstico: Geweke e Heidelberger e Welch.

O diagnóstico de Geweke et al. (1991), baseado em métodos de séries temporais, testa se diferentes partições da cadeia têm médias iguais, a partir de um teste Z. Por *default*, utilizou-se as frações 0,1 e 0,5 da cadeia. De acordo com o teste, todos as estatísticas

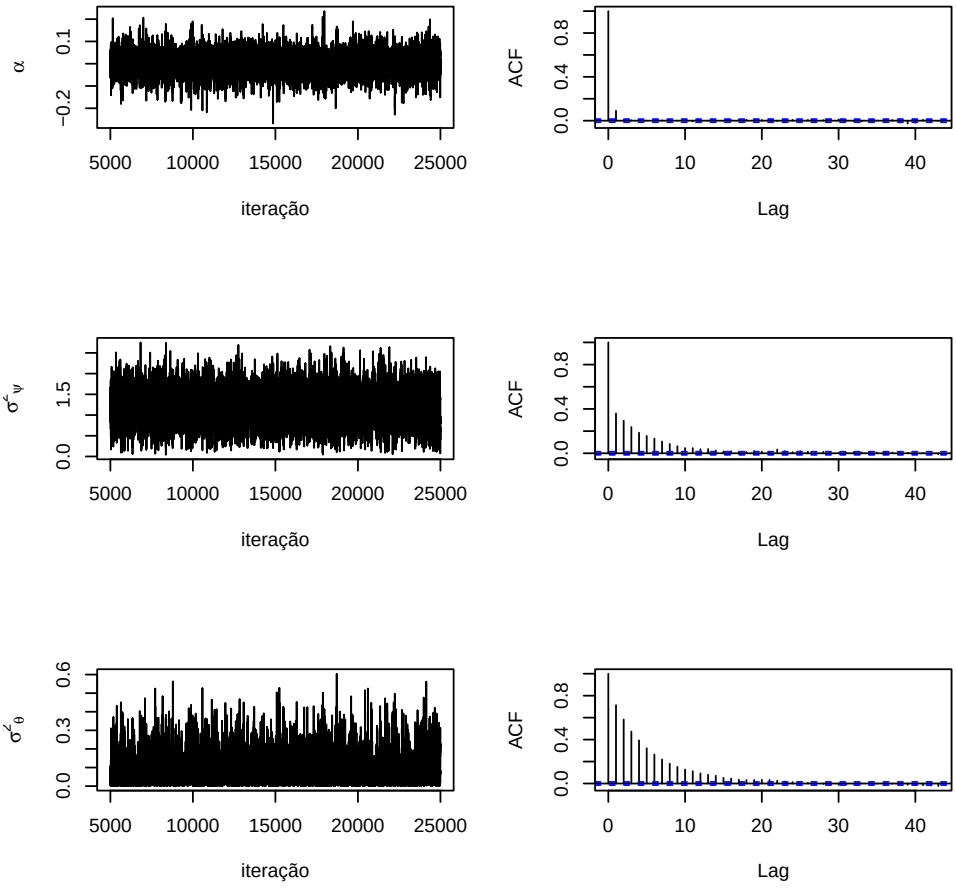


Figura 3 – Trajetória das cadeias e gráficos de autocorrelação: modelo de convolução

calculadas situaram-se dentro do intervalo de (-1,96;1,96). Com um nível de confiança de 95%, pode-se dizer que a indicação de convergência não é rejeitada para todos os parâmetros, dos cinco modelos.

Já Heidelberger e Welch (1983) desenvolveram um método para detectar um estado inicial transiente em sequências simuladas de eventos discretos. O diagnóstico proposto é baseado na estatística de Cramer-Von-Mises para testar a hipótese de que os valores simulados formam um processo estacionário. Se a hipótese de estacionariedade não for rejeitada para a cadeia de interesse, o teste *half-width* é aplicado, consistindo em um intervalo de 95% de confiança para a média. Metade do tamanho desse intervalo é comparado com a média estimada. Se a razão entre a metade do tamanho e a média estimada for inferior a um certo valor alvo, o parâmetro em teste não é rejeitado. Uma rejeição nesse teste pode ser um indício que o tamanho da cadeia não é grande o suficiente para estimar a média com precisão. Aplicando o teste de Heidelberger e Welch aos parâmetros estimados da cadeia, todos tiveram indicação de estacionariedade na parte inicial do teste. Entretanto, no teste *half-width*, adotando um valor alvo padrão de 0,1, os

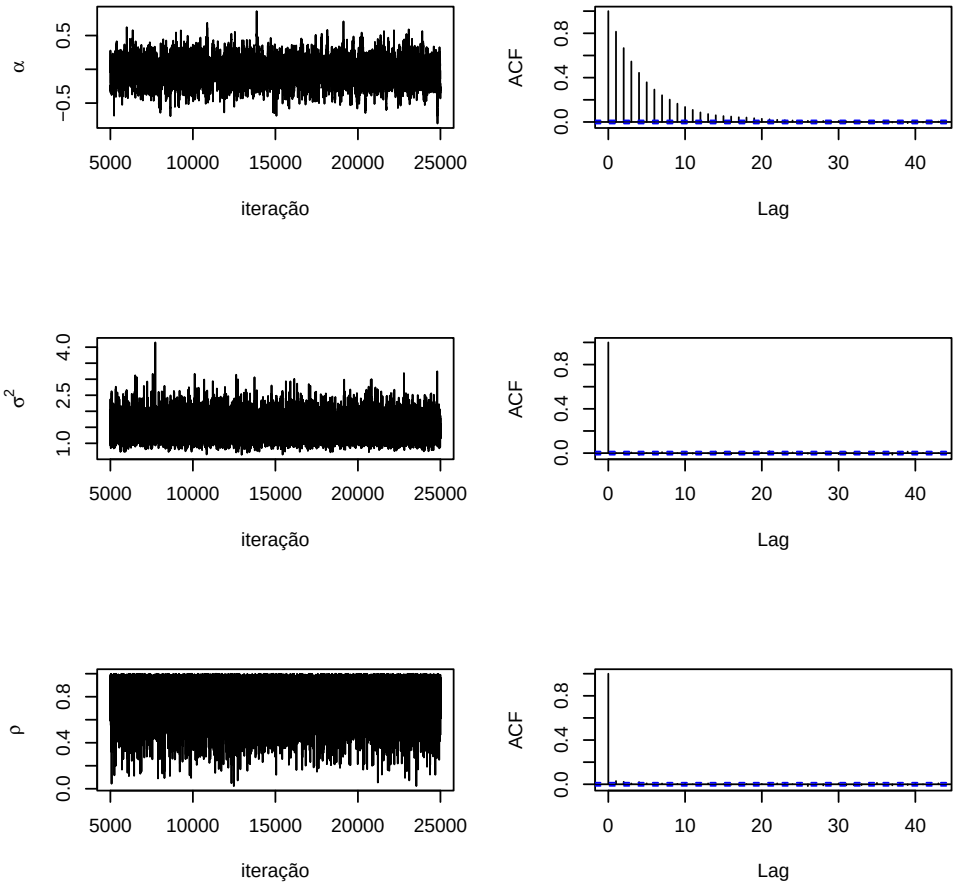


Figura 4 – Trajetória das cadeias e gráficos de autocorrelação: modelo de Cressie

interceptos dos modelos de Cressie e Leroux são rejeitados. Uma nova estimação foi feita via MCMC para esses modelos, dessa vez com o dobro do tamanho da cadeia (40000). Ainda assim, o teste persistiu em rejeitar esses dois parâmetros. Logo, apesar de o teste de Geweke ter assegurado uma possível indicação de convergência, o mesmo não se pode dizer para esse teste.

A Tabela 1 fornece as estimativas (médias a posteriori) de cada um dos parâmetros dos modelos. Para diferenciação entre os parâmetros dos modelos, índices foram adotados: In refere-se ao modelo intrínseco, Cv ao modelo de convolução, Cr ao modelos de Cressie, Le ao modelo de Leroux e Lu ao modelo de Lu.

É interessante notar que, apesar de α_{Cr} e α_{Le} terem apresentado certos problemas de convergência, suas estimativas situam-se próximas aos valores de α_{In}, α_{Cv} e α_{Lu}. Além disso, os interceptos de todos os modelos podem ser considerados não significativos, já que seus respectivos intervalos de credibilidade contém o valor zero. Entre os parâmetros de variância (excetuando-se σ^2_θ, que representa a variabilidade do efeito não estruturado espacialmente), todos os valores podem ser considerados relativamente próximos, já que seus

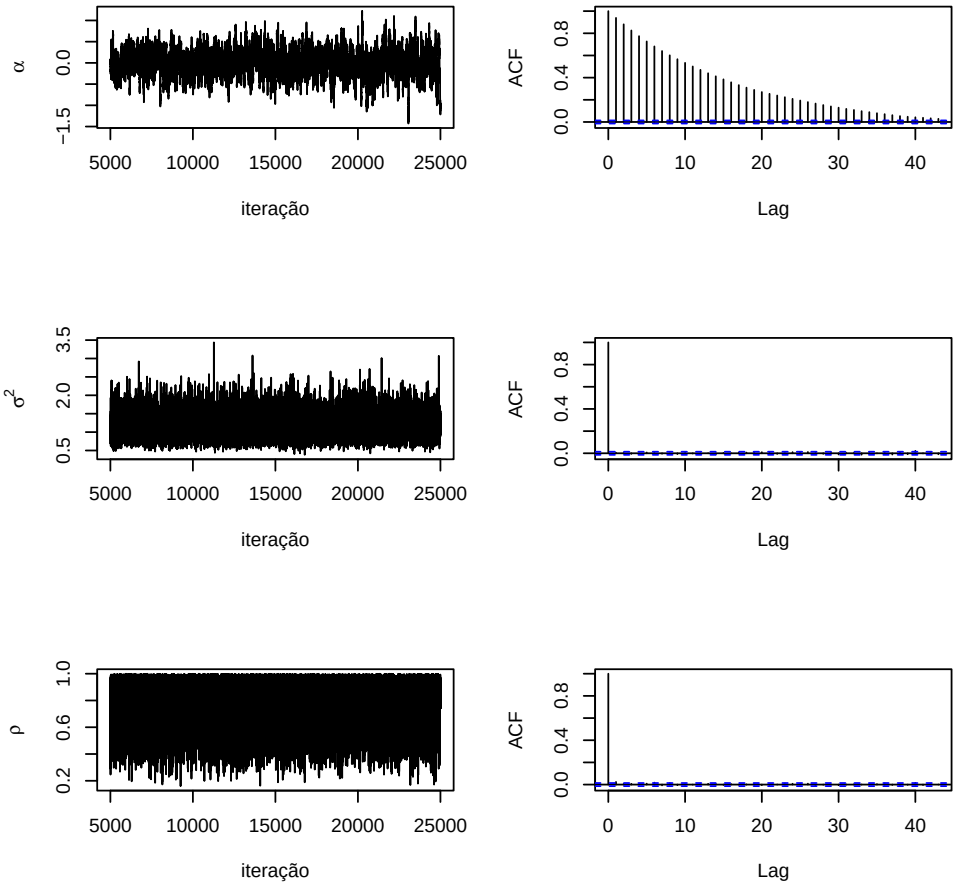


Figura 5 – Trajetória das cadeias e gráficos de autocorrelação: modelo de Leroux

intervalos de credibilidade de 95% se sobrepõem. Já os parâmetros de dependência espacial — principal atrativo dos modelos CAR próprios (Cressie e Leroux) — foram estimados em aproximadamente 0,77, o que caracteriza uma dependência espacial relativamente alta, embora não exista um critério definitivo para classificar a força da dependência espacial baseado em sua estimativa.

Para o modelo de Lu, as estimativas para γ_0 e γ_1 foram 0,9191 e -4,9840, respectivamente. Tais valores levam a um intervalo de credibilidade médio para p_{ij} de aproximadamente (0,36; 0,86). Dessa forma, o teor informativo das distribuições a priori de γ_0 e γ_1 , fornecidas na Seção 3.3, é um preço a se pagar para a obtenção de informações adicionais sobre a dinâmica existente entre regiões vizinhas, a serem estudadas em uma possível análise de fronteira.

A Figura 7 apresenta os riscos de óbito por Doença de Crohn e Colite Ulcerativa estimados pelos cinco modelos. De acordo com os mapas produzidos, nota-se que os riscos estimados para cada microrregião são próximos, para todos os modelos, já que a maioria das microrregiões foi classificada dentro do mesmo intervalo.

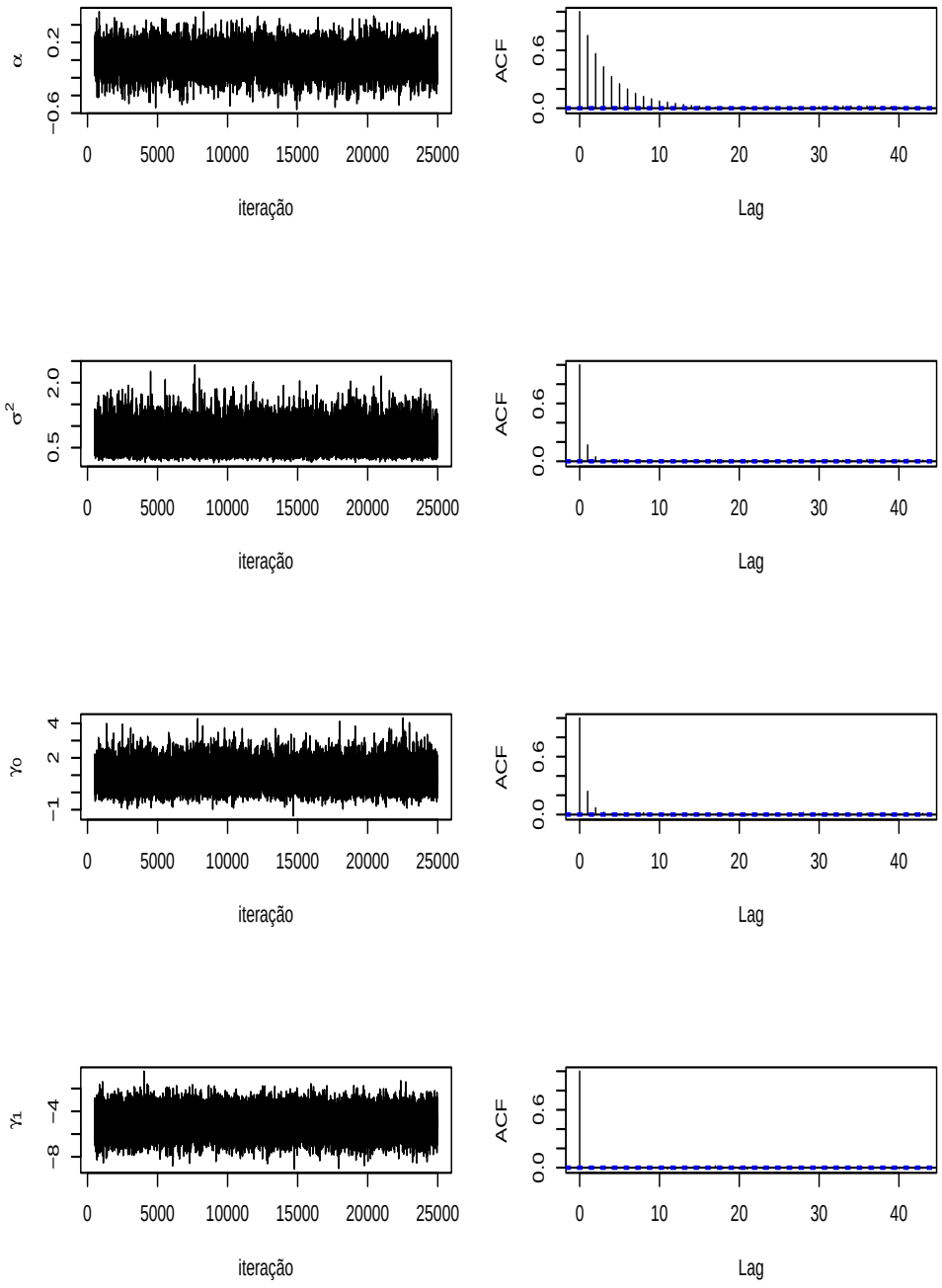


Figura 6 – Trajetória das cadeias e gráficos de autocorrelação: modelo de Lu

Para a seleção do modelo mais adequado para esse conjunto de dados, foi empregado em um primeiro momento o DIC (*Deviance Information Criterion*). O DIC é uma medida proposta por Spiegelhalter et al. (2002) bastante apropriada para modelos bayesianos, pois pode ser calculado diretamente através das amostras simuladas via MCMC. Sua

Tabela 1 – Estimativas dos parâmetros, desvios padrões e intervalos de credibilidade de 95% referentes aos dados de óbito pela Doença de Crohn e Colite Ulcerativa

Parâmetro	Média	Desvio padrão	I.Cr.(95%)
α_{In}	0,0044	0,0254	(-0,0460;0,0054)
α_{Cv}	0,0026	0,0399	(-0,0808;0,0806)
α_{Cr}	-0,0344	0,1654	(-0,3637;0,2998)
α_{Le}	0,0058	0,2697	(-0,5305;0,5157)
α_{Lu}	0,0029	0,1408	(-0,2786;0,2731)
σ^2_{In}	1,3559	0,2835	(0,9014;2,0060)
σ^2_{θ}	0,0725	0,0748	(0,0012;0,2807)
σ^2_{ψ}	1,1222	0,3649	(0,3843;1,8820)
σ^2_{Cr}	1,4440	0,3090	(0,9516;2,1640)
σ^2_{Le}	1,1734	0,3017	(0,6739;1,8641)
σ^2_{Lu}	0,6579	0,2578	(0,2926;1,2740)
ρ_{Cr}	0,7797	0,1580	(0,4070;0,9898)
ρ_{Le}	0,7713	0,1674	(0,3874;0,9911)
γ_0	0,9191	0,6165	(-0,1498;2,2710)
γ_1	-4,9840	0,9955	(-6,9395;-3,0395)

formulação é dada por

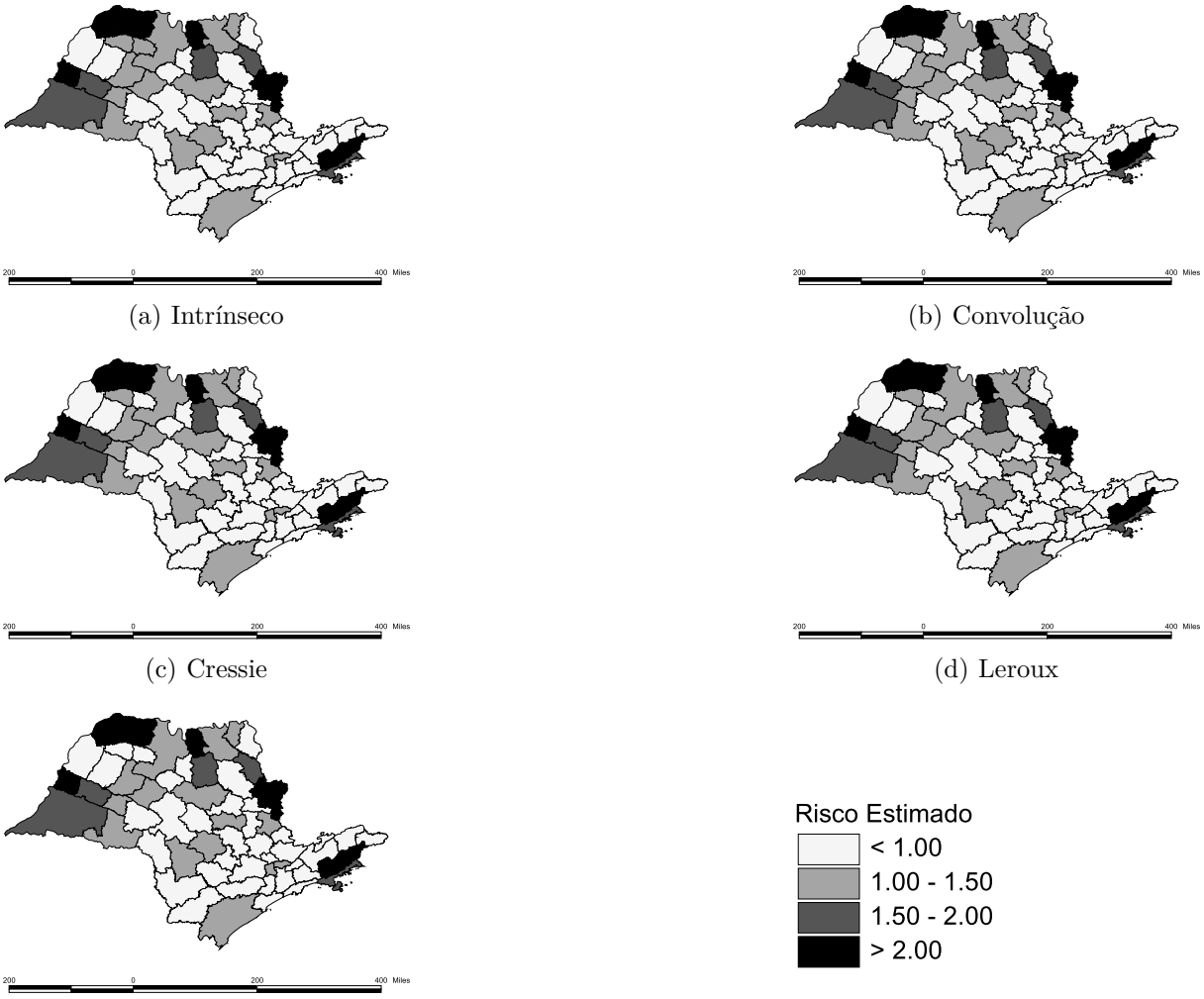
$$\begin{aligned} DIC &= p_D + \hat{D}, \\ p_D &= \hat{D} - D(\hat{\boldsymbol{\theta}}), \\ D(\hat{\theta}) &= -2\log\left(p\left(y|\hat{\boldsymbol{\theta}}\right)\right), \end{aligned}$$

em que p_D representa o número efetivo de parâmetros do modelo, $D(\hat{\boldsymbol{\theta}})$ é a estimativa pontual do desvio (*deviance*), $p\left(y|\hat{\boldsymbol{\theta}}\right)$ é a função de verossimilhança, $\boldsymbol{\theta}$ é o vetor de parâmetros do modelo e $\hat{D}$ é a média a posteriori do desvio. O modelo mais adequado é aquele que apresenta o menor DIC.

Uma crítica apontada ao DIC é que esse pode não ser adequado para casos em que as distribuições a posteriori são assimétricas, já que sua formulação envolve o cálculo de médias a posteriori. Logo, utilizá-lo individualmente pode levar a escolhas inapropriadas. Uma segunda medida para avaliar o desempenho dos modelos foi a soma de quadrados dos resíduos (SQR). Os valores calculados, tanto para o DIC quanto para a SQR estão dispostos na Tabela 2.

Tabela 2 – DIC e Soma de quadrados dos resíduos fornecidos pelos cinco modelos CAR para os dados de mortalidade pela Doença de Crohn e Colite Ulcerativa

Modelo	Intrínseco	Convolução	Cressie	Leroux	Lu
DIC	409,1	336,1	464	464,6	463,7
SQR	180,3	155,7	159,4	164,6	216,2



(a) Intrínseco (b) Convolução

(c) Cressie (d) Leroux

(e) Lu (f) Legenda

Figura 7 – Riscos de óbito de Doença de Crohn e Colite Ulcerativa estimados para o modelo intrínseco, de convolução, de Cressie, de Leroux e de Lu para cada microrregião do Estado de São Paulo

O menor DIC e por consequência o modelo apontado como mais adequado é o de convolução. Contudo, tal diferença em relação aos demais não parece representar adequadamente uma superioridade definitiva desse modelo. O modelo com a menor soma de quadrado de resíduos também é o de convolução, mas não há uma diferença acentuada em relação aos modelos de Cressie e Leroux. Isso pode ser analisado visualmente através da Figura 8, que traz os resíduos calculados para cada um dos modelos, em cada microrregião do Estado de São Paulo. Nota-se uma clara proximidade entre os valores, evidenciando que os modelos, na verdade, produziram valores estimados similares.

De qualquer modo, tomando o modelo de convolução como referência, os maiores riscos de mortalidade por Doença de Crohn e Colite Ulcerativa foram observados em Dracena (15,34), Fernandópolis (4,37), Votuporanga (3,41), Barretos (2,87) e Jales (2,84). Já os menores foram encontrados em Amparo (0,29), Araçatuba (0,41), Santos (0,45),

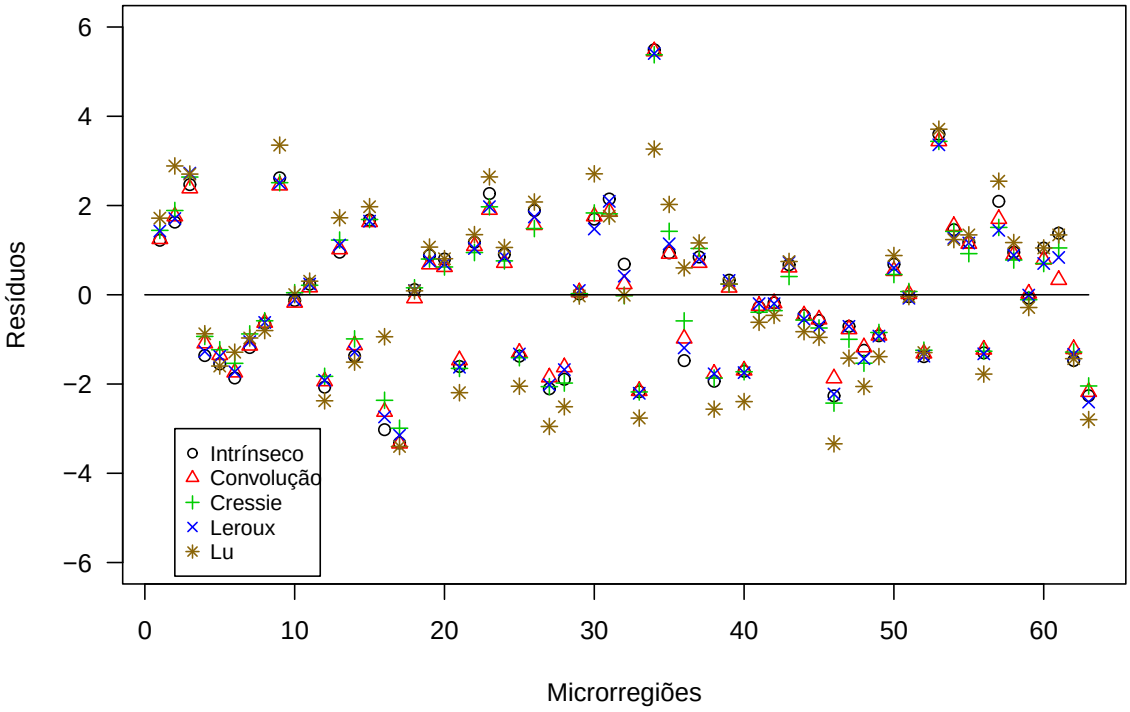


Figura 8 – Resíduos dos cinco modelos CAR de cada microrregião do Estado de São Paulo de acordo com os dados de mortalidade de Doença de Crohn e Colite Ulcerativa

Bragança (0,50) e Limeira (0,51). O elevado risco de Dracena pode ser considerado um valor discrepante, já que é demasiado superior aos demais. Um estudo específico dessa região é recomendado para uma possível causa que explique tamanha diferença.

A Figura 9 sintetiza a estimação sobre os riscos, classificando cada microrregião em três categorias. Áreas em branco são aquelas cujos intervalos de credibilidade de 95% contém valores inferiores a um. Regiões em cinza claro possuem intervalos de credibilidade que contém o valor um. Por fim, áreas em cinza escuro podem ser consideradas regiões de maior risco, pois possuem intervalos de credibilidade com valores superiores a um. Pode-se notar alguns grupos de regiões de alto risco no oeste paulista, no norte do estado e até mesmo em uma faixa litorânea.

Com o auxílio do programa GeoDa (ANSELIN; SYABRI; KHO, 2006), adotou-se o Índice de Moran (um coeficiente descritivo que mensura a autocorrelação espacial) para os efeitos aleatórios estimados pelo modelo de convolução para os dados de Doença de Crohn e Colite Ulcerativa e obteve-se o valor de 0,2460. Aplicando-se um teste de permutação (999 permutações), um valor p inferior a 0,001 foi obtido, o que caracteriza um valor significativo desse índice e a captura da estrutura espacial pelos efeitos aleatórios. A Figura 10 ilustra os agrupamentos formados de acordo com o Índice de Moran Local dos mesmos

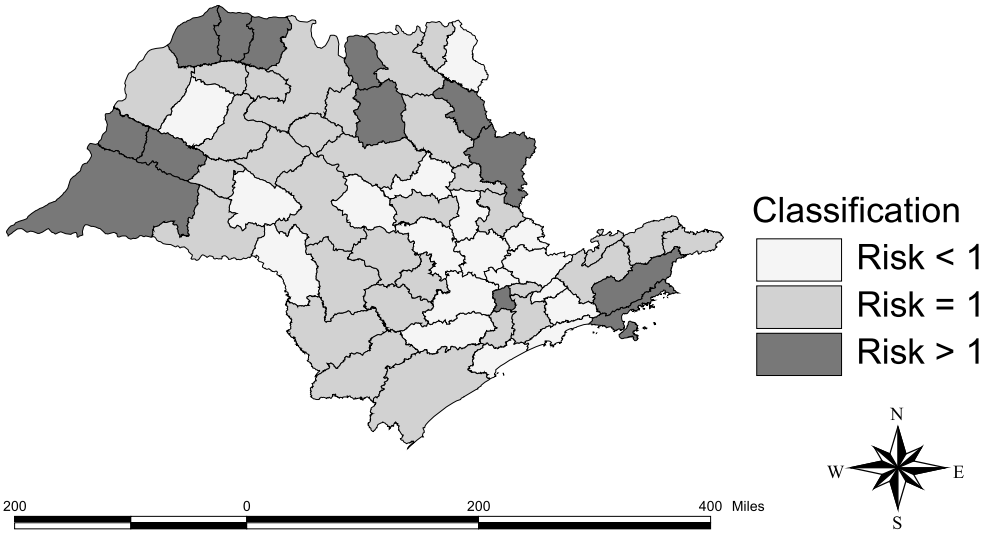


Figura 9 – Microrregiões classificadas em três grupos baseados em seus intervalos de credibilidade de 95% para os dados de óbito pela Doença de Crohn e Colite Ulcerativa: baixo risco (branco), risco dentro do esperado (cinza claro), alto risco (cinza escuro)

efeitos aleatórios. Percebe-se que a presença de autocorrelação é significativa em poucas microrregiões.

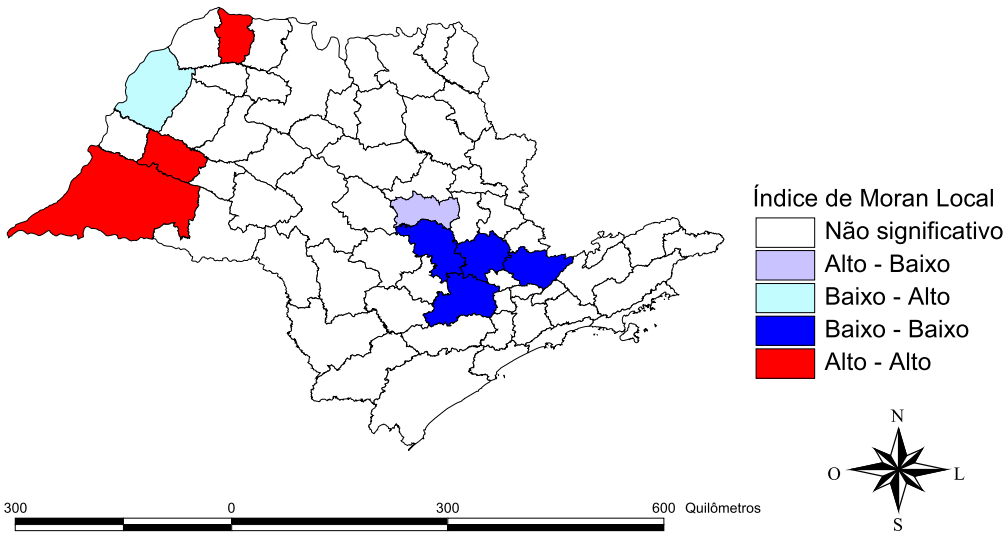


Figura 10 – Agrupamentos de acordo com o Índice de Moran Local dos efeitos aleatórios estimados pelo modelo de convolução para os dados de mortalidade por Doença de Crohn e Colite Ulcerativa

4.2 Câncer de traqueia, brônquios e pulmões

A traqueia, os brônquios e os pulmões são componentes do sistema respiratório, motivos que os fazem serem enquadrados em um mesmo grupo. Esse câncer é de difícil

diagnóstico e costuma ser descoberto em estágios avançados, o que faz o índice de mortalidade chegar a 86%. Sabe-se também que possui estreita ligação com o hábito de fumar. No início do século XX, quando o consumo do cigarro ainda não tinha se disseminado, era uma enfermidade rara. Hoje, de acordo com a Organização Mundial da Saúde (OMS), esse é o tipo de câncer que mais mata no mundo.

Como uma análise exploratória, na Figura 11 estão dispostas as razões de mortalidade padronizadas referentes ao Câncer de traqueia, brônquios e pulmões, calculados para cada microrregião. Com esse mapa, pode-se ter uma ideia inicial do comportamento do risco ao longo das microrregiões.

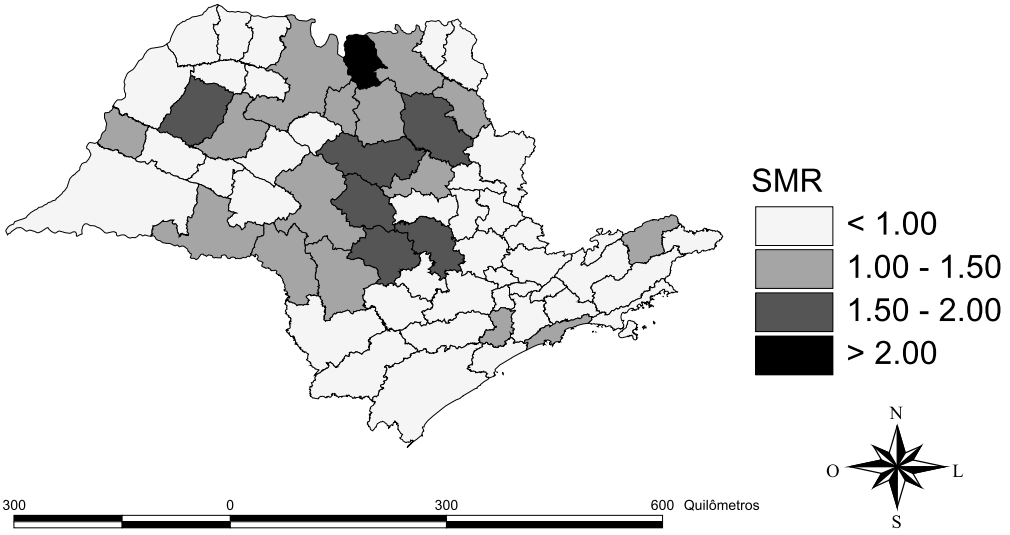


Figura 11 – Razões de mortalidade padronizadas para o Câncer de traqueia, brônquios e pulmões calculadas para cada microrregião do Estado de São Paulo no período de 2008 a 2012

4.2.1 Resultados

Assim como na aplicação apresentada na Subseção 4.1, a estimação foi feita via MCMC através do software OpenBUGS, com 5000 iterações iniciais descartadas para aquecimento da cadeia e um total de 20000 iterações subsequentes. O salto (*thin*) utilizado foi igual a 30. Novamente, a indicação de convergência para os parâmetros foi verificada através de inspeções visuais (trajetórias dos parâmetros e autocorrelações das cadeias) e diagnósticos de convergência (Geweke e Heidelberger e Welch). Os resultados detalhados para os diagnósticos serão omitidos, já que as conclusões obtidas são análogas às descritas na Subseção 4.1.1, isto é, apenas os interceptos dos modelos de Cressie e Leroux apresentaram certos problemas no teste de *half-width* (diagnóstico de Heidelberger e Welch).

A Tabela 3 fornece as estimativas (médias a posteriori) de cada um dos parâmetros dos modelos. Novamente, apesar de α_{Cr} e α_{Le} terem apresentado certos problemas de

convergência, suas estimativas situam-se próximas aos valores de α_{In} e α_{Cv}. Entre os parâmetros de variância os menores valores são observados no modelo de convolução e de Leroux, indicando uma maior precisão na estimação dos efeitos aleatórios desses modelos. Porém, excetuando-se σ^2_θ, todos os parâmetros de variância podem ser considerados relativamente próximos, já que seus intervalos de credibilidade de 95% se sobrepõem. Já os parâmetros de dependência espacial foram estimados em aproximadamente 0,84 e 0,66 para os modelos de Cressie e Leroux, respectivamente, o que caracteriza uma dependência espacial moderada/alta.

Tabela 3 – Estimativas dos parâmetros, desvios padrões e intervalos de credibilidade de 95% referentes aos dados de óbito de Câncer de traqueia, brônquios e pulmões

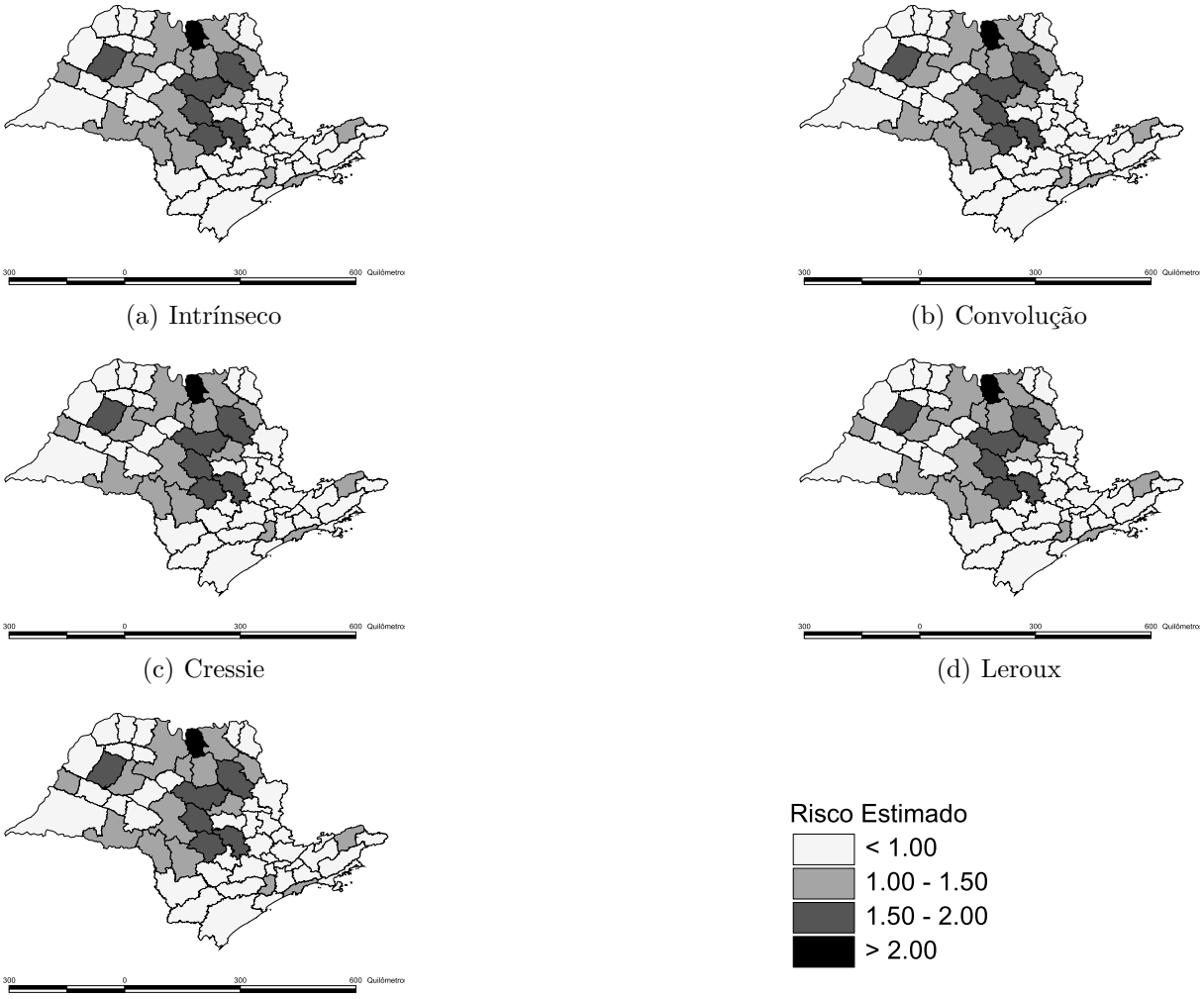
Parâmetro	Média	Desvio padrão	I.Cr.(95%)
α_{In}	-0,0798	0,0135	(-0,1067;-0,0535)
α_{Cv}	-0,0826	0,0308	(-0,1486;-0,0226)
α_{Cr}	-0,0622	0,1184	(-0,3026;0,1715)
α_{Le}	-0,0590	0,1430	(-0,3389;0,2429)
α_{Lu}	-0,0760	0,0798	(-0,2321;0,0839)
σ^2_{In}	0,5334	0,1087	(0,3566;0,7810)
σ^2_θ	0,0500	0,0369	(0,0019;0,1367)
σ^2_ψ	0,3309	0,1583	(0,0540;0,6520)
σ^2_{Cr}	0,5479	0,1163	(0,3636;0,8179)
σ^2_{Le}	0,3921	0,1181	(0,2007;0,6568)
σ^2_{Lu}	0,2207	0,1114	(0,0780;0,5009)
ρ_{Cr}	0,8419	0,1186	(0,5605;0,9929)
ρ_{Le}	0,6572	0,2065	(0,2512;0,983)
γ_0	0,6019	0,7105	(-0,5799;2,2040)
γ_1	-5,0241	0,9984	(-6,9961;-3,0670)

Para o modelo de Lu, as estimativas para γ_0 e γ_1 foram 0,6019 e -5,0241, respectivamente. Tais estes valores levam a um intervalo de credibilidade médio para p_{ij} de aproximadamente (0,26; 0,86).

A Figura 12 apresenta os riscos de óbito por Câncer de traqueia, brônquios e pulmões estimados pelos cinco modelos. De acordo com os mapas produzidos, nota-se que os riscos estimados para cada microrregião são próximos, para todos os modelos, já que a maioria das microrregiões foi classificada dentro do mesmo intervalo.

Para a seleção do modelo mais adequado para esse conjunto de dados, foi empregado novamente o DIC e a SQR. Os valores calculados de cada modelo estão dispostos na Tabela 4.

O menor DIC e portanto o modelo apontado como mais adequado é novamente o de convolução, com uma diferença acentuada em relação aos demais. O mesmo modelo também possui a menor soma de quadrado de resíduos, sugerido portanto como o mais



(a) Intrínseco (b) Convolução

(c) Cressie (d) Leroux

(e) Lu (f) Legenda

Figura 12 – Riscos de óbito de Câncer de traqueia, brônquios e pulmões estimados para o modelo intrínseco, de convolução, de Cressie, de Leroux e de Lu para cada microrregião do Estado de São Paulo

Tabela 4 – DIC e Soma de quadrados dos resíduos (SQR) fornecidos pelos cinco modelos CAR para os dados de mortalidade de Câncer de traqueia, brônquios e pulmões

Modelo	Intrínseco	Convolução	Cressie	Leroux	Lu
DIC	507,4	166,3	549,6	549,4	549,4
SQR	562,6	450,1	543,1	548,7	707,7

adequado para representar os dados de mortalidade por Câncer de traqueia, brônquios e pulmões. A Figura 8 traz os resíduos calculados para cada um dos modelos, em cada microrregião do Estado de São Paulo.

Novamente percebe-se a similaridade dos modelos devido a proximidade dos valores dos resíduos para cada microrregião. Entretanto, dessa vez percebe-se uma pequena vantagem aparente do modelo de convolução, já que para algumas microrregiões o resíduo

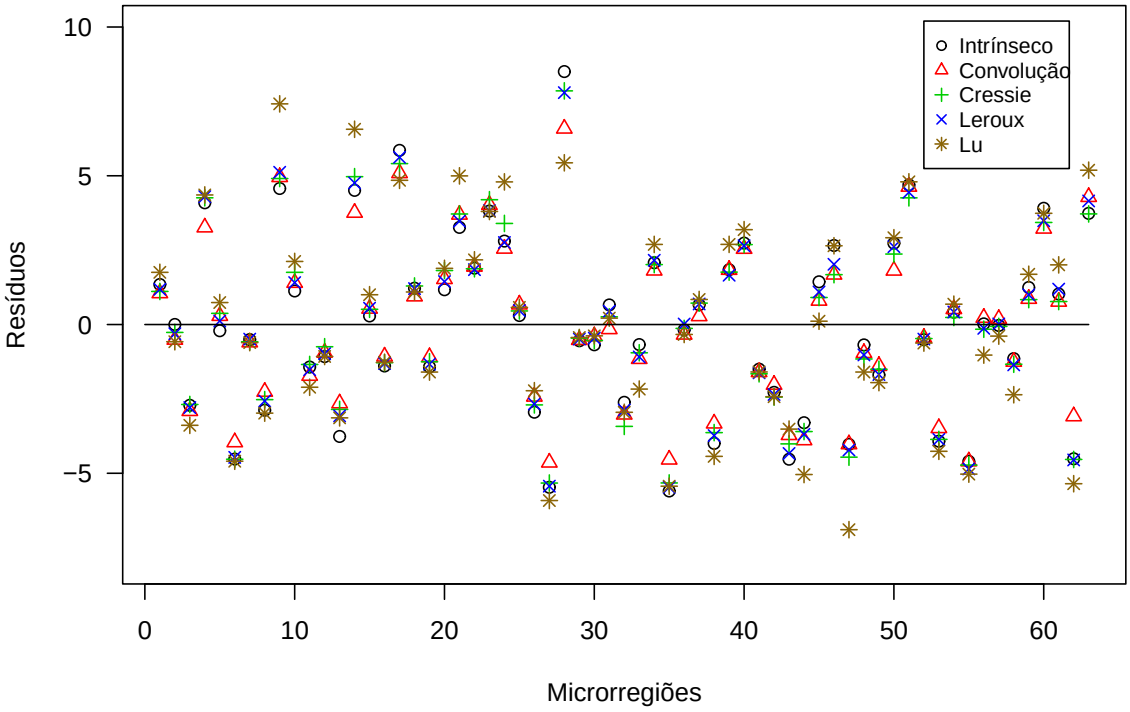


Figura 13 – Resíduos para os cinco modelos CAR de cada microrregião do Estado de São Paulo de acordo com os dados de mortalidade pela Câncer de traqueia, brônquios e pulmões

desse modelo é o que mais se aproxima da linha horizontal (zero). Tomando o modelo de convolução como referência, os maiores riscos de mortalidade por Câncer de traqueia, brônquios e pulmões foram observados em Barretos (2,37), Jaú (1,71), Araçatuba (1,64), Ribeirão Preto (1,64) e Piracicaba (1,63). Já os menores foram encontrados em Registro (0,40), Capão Bonito (0,41), Jundiaí (0,46), Limeira (0,57) e Amparo (0,59). Suspeita-se que o elevado valor presente em Barretos seja explicado por um fator não necessariamente relacionado ao risco de câncer. Essa região possui o maior hospital para tratamento de câncer da América Latina, sendo uma referência em seu tratamento. Isso provavelmente motiva indivíduos acometidos com a doença a se mudaram para a região para ter mais facilidade no acesso ao tratamento. Com o interesse de melhor investigar esse fenômeno, novos modelos foram propostos introduzindo uma covariável que indica a presença de centros especializados no tratamento de câncer no Estado de São Paulo. Das 63 microrregiões do estado, 31 possuem tais centros. Entretanto, a covariável não mostrou-se significativa em nenhum dos cinco modelos, ao passo que os intervalos de credibilidade de 95% para o coeficiente associado contemplaram o valor zero.

A Figura 14 sintetiza a estimação sobre os riscos, classificando cada microrregião em

três categorias. Áreas em branco são aquelas cujos valores nos intervalos de credibilidade de 95% são inferiores a um. Regiões em cinza claro possuem intervalos de credibilidade que contém o valor um. Por fim, áreas em cinza escuro podem ser consideradas regiões de maior risco, pois possuem intervalos de credibilidade cujos valores são superiores a um.

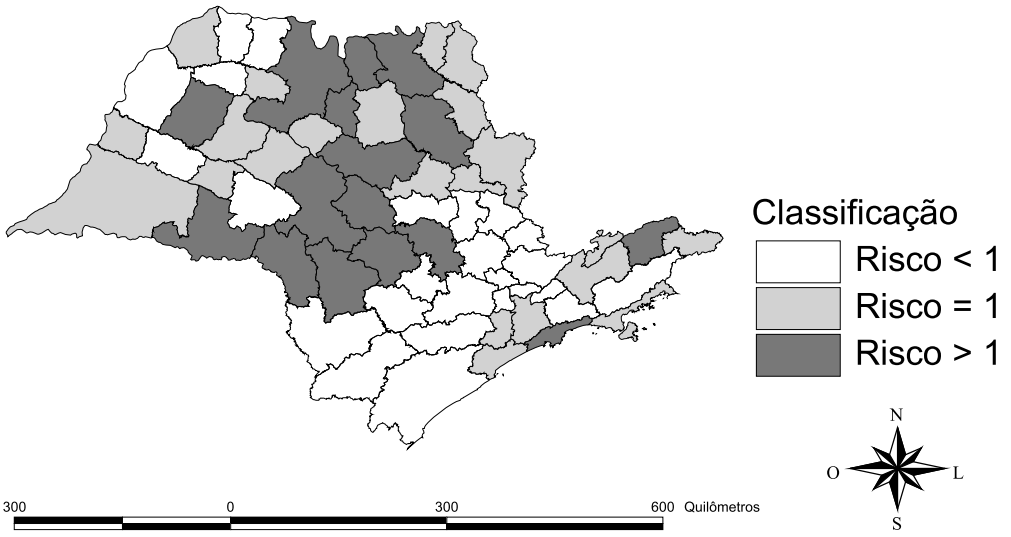


Figura 14 – Microrregiões classificadas em três grupos baseados em seus intervalos de credibilidade de 95% para os dados de óbito de Câncer de traqueia, brônquios e pulmões: baixo risco (branco), risco dentro do esperado (cinza claro), alto risco (cinza escuro)

Pode-se notar uma grande concentração de regiões de alto risco no centro e norte do Estado de São Paulo. Áreas de baixo risco estão dispostas principalmente na porção sul e leste.

Adotando-se o Índice de Moran para os efeitos aleatórios estimados pelo modelo de convolução para os dados de óbito por Câncer de traqueia, brônquios e pulmão, obteve-se o valor de 0,3178. Aplicando-se um teste de permutação (999 permutações), um valor p inferior a 0,001 foi obtido, o que caracteriza um valor significativo desse índice e a captura da estrutura espacial pelos efeitos aleatórios. A Figura 15 ilustra os agrupamentos formados de acordo com o Índice de Moran Local dos mesmos efeitos aleatórios. Percebe-se que a presença de autocorrelação é significativa em três grupos bem definidos.

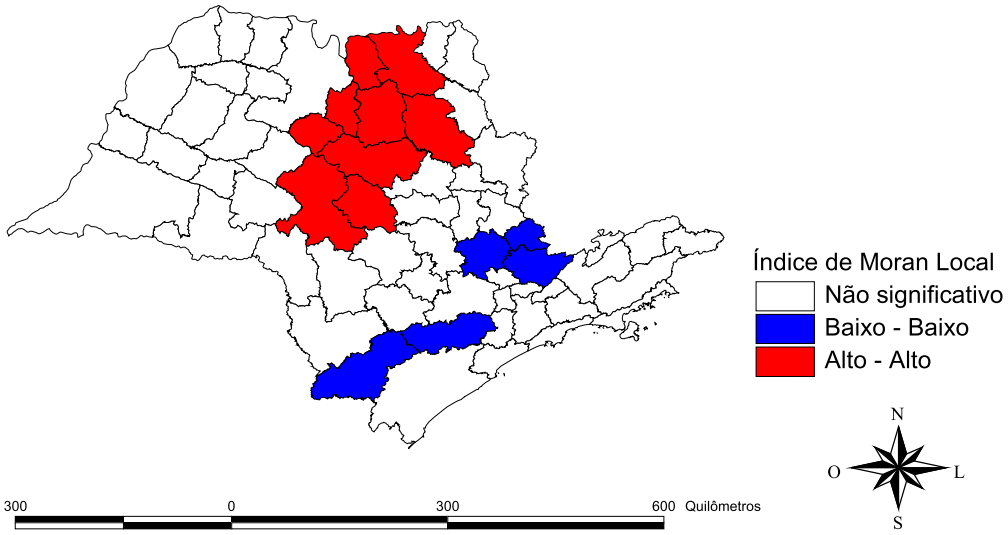


Figura 15 – Agrupamentos de acordo com o Índice de Moran Local dos efeitos aleatórios estimados pelo modelo de convolução para os dados de mortalidade por Câncer de traqueia, brônquios e pulmões

5 Estudo de Simulação

Com o objetivo de avaliar o desempenho dos modelos CAR descritos no Capítulo 3, um estudo de simulação baseado no proposto por Lee (2011) foi conduzido. Os dados foram gerados utilizando como referência o mapa das 63 microrregiões do Estado de São Paulo. Adotou-se os valores esperados referentes ao número de óbitos por câncer de traqueia, brônquios e pulmões, no Estado de São Paulo, no período de 2008 a 2012. O termo intercepto foi arbitrariamente fixado em $\alpha = 0,1$. Três cenários distintos foram traçados, com o intuito de averiguar como os modelos se comportam na presença de diferentes níveis de dependência espacial, incluindo-se também o caso de independência. Para gerar dados com dependência espacial, foi empregada a função de covariância Matérn (MINASNY; MCBRATNEY, 2005). Essa função pode ser definida como

$$C(d) = \tau^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu}\frac{d}{\eta}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{d}{\eta}\right),$$

em que d é a distância entre duas regiões, η é a amplitude, ν é um parâmetro de alisamento e τ^2 é a variância marginal, com d, η, ν e $\tau > 0$. K_{ν} é a função de Bessel modificada do segundo tipo de ordem ν, que pode ser escrita como

$$K_q(z) = \left(\frac{\pi}{2}\right) \frac{I_{-q}(z) - I_q(z)}{sen(\pi q)},$$

em que

$$I_q(z) = \left(\frac{z}{2}\right)^q \sum_{p=0}^{\infty} \frac{\left(\frac{z^2}{4}\right)^p}{p!\Gamma(q+p+1)}.$$

A partir de uma matriz de distâncias — baseada nos centroides de cada microrregião —, a função Matérn é empregada com o objetivo de se obter uma matriz de correlação a ser utilizada para gerar os efeitos aleatórios de uma distribuição normal multivariada. Com os valores gerados para os efeitos aleatórios, é possível calcular uma taxa média para cada microrregião, usadas como parâmetros da distribuição de Poisson para gerar um vetor de contagens que emulam o número de óbitos da doença.

Finalmente, os cenários criados para o estudo de simulação foram os seguintes:

- Cenário 1: Independência. Os efeitos aleatórios $\boldsymbol{\phi}$ foram gerados de uma distribuição normal multivariada com vetor de médias $\mathbf{0}$ e matriz de correlação dada pela classe Matérn, com parâmetros $\tau^2 = 1$, $\nu = 2,5$ e amplitude desprezível ($\eta = 0,0001$), o que, com essas especificações, fornece uma matriz identidade.

- Cenário 2: Dependência espacial moderada. Os efeitos aleatórios $\boldsymbol{\phi}$ foram gerados de uma distribuição normal multivariada com vetor de médias $\mathbf{0}$ e matriz de correlação

dada pela classe Matérn, com parâmetros $\tau^2 = 1$, $\nu = 2,5$ e $\eta = 0,15$, o que, com essas especificações, fornecem uma correlação média entre (todas) as áreas de 0,2.

- Cenário 3: Dependência espacial forte. Os efeitos aleatórios $\boldsymbol{\phi}$ foram gerados de uma distribuição normal multivariada com vetor de médias $\mathbf{0}$ e matriz de correlação dada pela classe Matérn, com parâmetros $\tau^2 = 1$, $\nu = 2,5$ e $\eta = 0,6$, o que, com essas especificações, fornecem uma correlação média entre (todas) as áreas de 0,4.

Ressalta-se que as correlações médias definidas nos cenários da simulação diferem do parâmetro de dependência ρ dos modelos de Cressie e de Leroux. As primeiras referem-se às correlações médias entre todas as regiões, inclusive àquelas não adjacentes. Já a segunda representa um parâmetro de dependência restrito apenas às áreas vizinhas. Portanto, é natural esperar que as correlações médias utilizadas para a geração de dados na simulação sejam acentuadamente menores que o valor estimado para ρ.

Para cada cenário, foram gerados um total de 500 conjuntos de dados simulados. Para cada um desses conjuntos de dados, uma estimação via MCMC foi realizada com 2000 iterações descartadas para aquecimento da cadeia e um total de 8000 iterações subsequentes, com um salto (*thin*) igual a 15. A Tabela 5 fornece uma síntese dos resultados obtidos pela simulação, através de estatísticas como o valor médio dos parâmetros α, σ^2 e ρ, além do erro quadrático médio (EQM) e vício para os riscos estimados. Os valores do EQM e do vício são calculados como

$$\text{EQM}(R_k) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(\hat{R_{ij}} - R_{ij})^2}{m}$$

$$\text{Vício}(R_k) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(\hat{R_{ij}} - R_{ij})}{m}$$

em que $n = 63$ é o número de microrregiões e $m = 500$ é o total de iterações realizadas na simulação.

Em regra, todos os modelos estimaram eficientemente o parâmetro α, já que, na média, as estimativas situaram-se próximas ao seu verdadeiro valor ($\alpha = 0,1$). A exceção ficou com o modelo de Cressie para o cenário 3 (média de $\alpha = 0,557$). Em relação aos parâmetros de variância, os modelos com mais precisão foram o de convolução, Leroux e Lu. Ressalta-se que o modelo de convolução possui dois parâmetros de variância, um estruturado e outro não estruturado espacialmente. É interessante notar que no cenário de independência, o parâmetro de variância referente ao efeito aleatório não estruturado espacialmente é superior (0,868 contra 0,714). Por outro lado, para o cenário 3, com a presença de forte dependência espacial, essa característica é invertida, sendo agora o parâmetro referente ao efeito aleatório espacial fortemente dominante (0,948 contra 0,026). Em relação ao parâmetro ρ, nota-se que ele consegue refletir diferentes níveis de correlação espacial. Entretanto, é possível perceber a existência de uma certa superestimação para

Tabela 5 – Resultados obtidos pelo estudo de simulação considerando três diferentes cenários

Métricas	Cenário	Intrínseco	Convolução	Cressie	Leroux	Lu
	1	0,0963	0,0930	0,0993	0,0957	0,1024
	2	0,0977	0,0873	0,1069	0,0771	0,1064
Valor médio de α	3	0,0868	0,0850	0,5570	0,0952	0,1114
	1	5,3090	0,7141	5,3265	1,7889	0,9103
	2	4,2277	1,3526	4,4826	2,4734	1,0258
Valor médio de σ^2	3	0,9755	0,9478	0,5815	0,5448	0,4618
	1	-	0,8680	-	-	-
	2	-	0,6253	-	-	-
Valor médio de σ_θ^2	3	-	0,0261	-	-	-
1	-	-	0,2775	0,1644	-	
		2	-	-	0,6183	0,4550
Valor médio de ρ						
	3	-	-	0,9902	0,9010	-
	1	2,6065	2,6070	2,5966	2,5937	2,5985
	2	2,6068	2,6043	2,6065	2,6080	2,6114
EQM(R_k)	3	2,3302	2,3465	2,3021	2,3054	2,3477
	1	-0,0571	-0,0489	-0,0556	-0,0606	-0,0661
	2	0,0671	0,0498	0,0595	0,0538	0,0434
Vício(R_k)	3	0,0806	0,0725	0,0588	0,0510	0,0243

o mesmo, já que no cenário de independência suas médias foram 0,277 e 0,164 para os modelos de Cressie e Leroux, respectivamente, quando na verdade essas deveriam situar-se mais próximas de zero.

A Figura 16 ilustra os histogramas para os valores estimados na simulação para o parâmetro ρ. No cenário 1 é possível perceber uma clara assimetria positiva, indicando que a média talvez não seja o estimador mais adequado para representar os valores estimados. Utilizando a moda, um estimador mais robusto na presença de assimetria, as estimativas do parâmetro ρ para os modelos de Cressie e Leroux são menores (0,1686 e 0,0732, respectivamente), o que tornam os resultados mais condizentes com o verdadeiro valor do parâmetro. O método utilizado para o cálculo da moda foi o de Robertson e Cryer (1974).

Em relação ao desempenho dos modelos estimando os riscos de mortalidade, adotou-se o erro quadrático médio e o vício como métricas. A característica observada na inferência com dados reais apresentada na Subseção 4.1.1 se repete e todos os modelos produzem riscos bastante similares, em todos os cenários. Tais resultados são de certa forma interessantes. Por exemplo, poderia se esperar que alguns cenários favoreceriam alguns modelos em detrimento de outros, como no caso do modelo intrínseco, em que o parâmetro ρ é fixo ($\rho = 1$). Contudo, o desempenho de todos os modelos, no que se refere à estimação dos riscos, não diferiu de forma acentuada. Logo, pode-se concluir que, apesar de suas

particularidades, vantagens e desvantagens, qualquer que seja o modelo adotado, o risco de mortalidade será satisfatoriamente estimado.

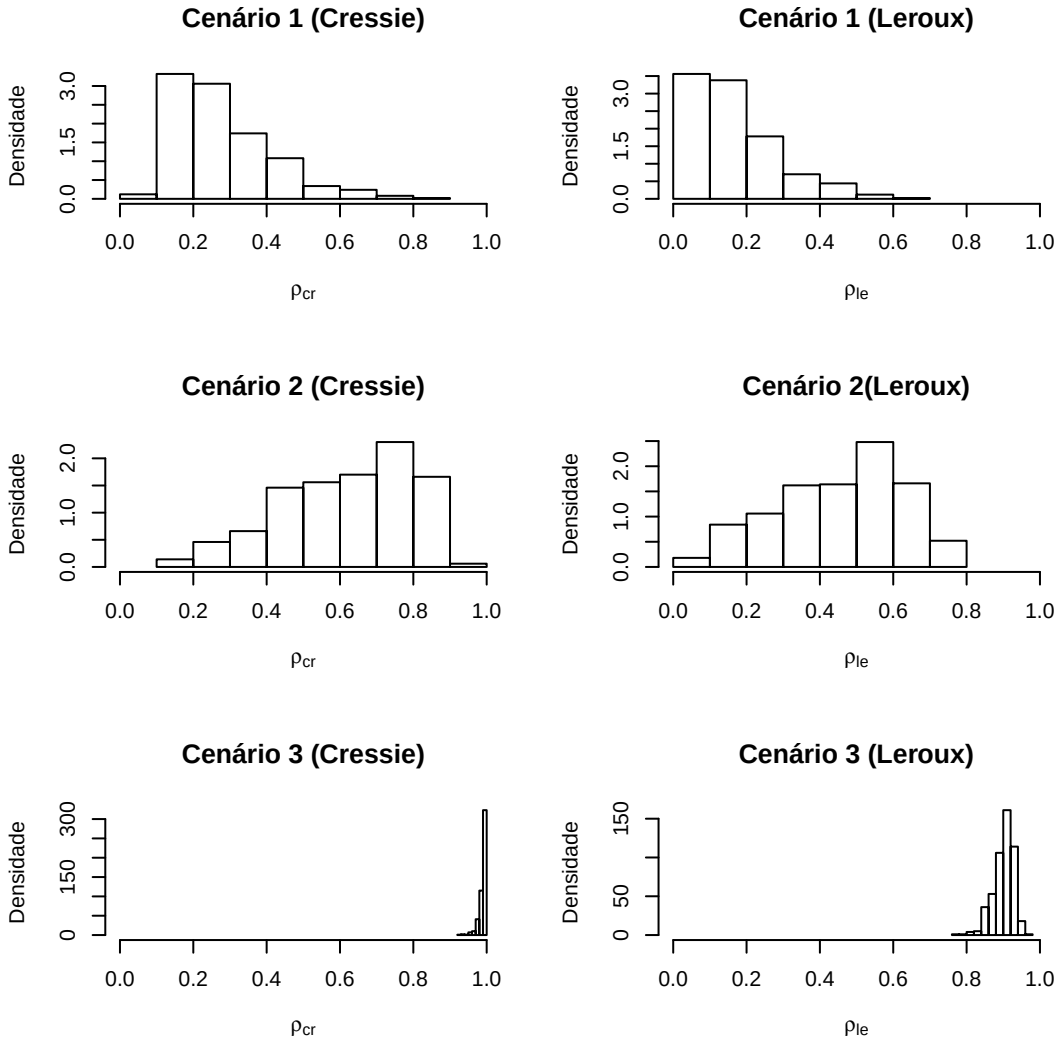


Figura 16 – Histogramas dos valores estimados de ρ_{Cr} e ρ_{Le} em cada iteração do estudo de simulação

6 Considerações Finais

Neste trabalho, estudou-se diferentes modelos autorregressivos condicionais com duas aplicações em mapeamento de doenças. Para ambos os conjuntos de dados (Doença de Crohn e Colite Ulcerativa e Câncer de traqueia, brônquios e pulmões) observou-se a existência de uma aparente estrutura espacial, evidenciada pelos valores estimados do parâmetro de dependência espacial nos modelos de Cressie e Leroux, além da variabilidade relativa entre os parâmetros de variância do modelo de convolução. As áreas de alto risco foram identificadas.

Todos os modelos produziram riscos estimados próximos. Essa característica foi reforçada por um estudo de simulação no qual se criou três cenários, com diferentes níveis de dependência espacial, através do auxílio de uma função de covariância da classe Matérn. Isso minimiza a necessidade de escolher um modelo definitivo que acomode os efeitos aleatórios, pois, no que concerne à estimação dos riscos de óbito, todos produziram resultados similares. As vantagens e ganhos trazidos pela escolha de cada modelo foram levantadas. Cabe então ao pesquisador a definição da característica mais relevante que atenda aos seus anseios.

Algumas janelas de interesse permanecem abertas. Uma delas é considerar medidas alternativas ao centroide para representar a informação da região como covariável no modelo de Lu. Uma alternativa mais plausível é, ao invés de simplesmente atribuir ao centro da área toda informação, considerar um centroide direcionado para áreas de maior densidade populacional. Outras covariáveis também poderiam ser adotadas, tanto no modelo logístico de Lu, quanto na formulação geral apresentada em (3.1). Finalmente, uma análise de fronteira pode ser conduzida para os dados aqui apresentados, de forma a melhor compreender a natureza das relações de vizinhança entre as microrregiões do Estado de São Paulo.

Adicionalmente, pesquisas mais aprofundadas envolvendo os modelos CAR podem ser conduzidas. Estudos de sensibilidade para γ_0 e γ_1 são de interesse para buscar alternativas às distribuições a priori com o teor informativo. Finalmente, estudos de desempenho que envolvam generalizações dos modelos CAR aqui descritos para o caso multivariado poderiam ser realizados para verificar o comportamento dos modelos para o caso de interações entre diferentes tipos de doenças. O mesmo se aplica para a generalização temporal.

Referências

ACHCAR, J. A. et al. Use of poisson spatiotemporal regression models for the brazilian amazon forest: malaria count data. *Revista da Sociedade Brasileira de Medicina Tropical*, SciELO Brasil, v. 44, n. 6, p. 749–754, 2011. Citado na página 11.

ANSELIN, L.; SYABRI, I.; KHO, Y. Geoda: an introduction to spatial data analysis. *Geographical analysis*, Wiley Online Library, v. 38, n. 1, p. 5–22, 2006. Citado na página 35.

ASSUNÇÃO, R.; KRAINSKI, E. Neighborhood dependence in bayesian spatial models. *Biometrical Journal*, Wiley Online Library, v. 51, n. 5, p. 851–869, 2009. Citado na página 11.

BANERJEE, S.; GELFAND, A. E.; CARLIN, B. P. *Hierarchical modeling and analysis for spatial data*. [S.l.]: Crc Press, 2003. Citado na página 20.

BESAG, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 192–236, 1974. Citado 2 vezes nas páginas 10 e 15.

BESAG, J.; KOOPERBERG, C. On conditional and intrinsic autoregressions. *Biometrika*, Biometrika Trust, v. 82, n. 4, p. 733–746, 1995. Citado na página 17.

BESAG, J.; YORK, J.; MOLLIÉ, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, Springer, v. 43, n. 1, p. 1–20, 1991. Citado 3 vezes nas páginas 10, 16 e 17.

BEST, N.; RICHARDSON, S.; THOMSON, A. A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, SAGE Publications, v. 14, n. 1, p. 35–59, 2005. Citado na página 17.

BRESLOW, N. E. Extra-poisson variation in log-linear models. *Applied Statistics*, JSTOR, p. 38–44, 1984. Citado na página 18.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, Taylor & Francis, v. 88, n. 421, p. 9–25, 1993. Citado na página 10.

CHRISTAKOS, G. On the problem of permissible covariance and variogram models. *Water Resources Research*, Wiley Online Library, v. 20, n. 2, p. 251–265, 1984. Citado na página 12.

CLAYTON, D.; KALDOR, J. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, JSTOR, p. 671–681, 1987. Citado 2 vezes nas páginas 10 e 13.

CRESSIE, N. A. *Statistics for Spatial Data, revised edition*. [S.l.]: Wiley, New York, 1993. Citado 5 vezes nas páginas 10, 13, 15, 16 e 18.

EBERLY, L. E.; CARLIN, B. P. et al. Identifiability and convergence issues for markov chain monte carlo fitting of spatial models. *Statistics in Medicine*, v. 19, n. 1718, p. 2279–2294, 2000. Citado na página 18.

GELFAND, A. E.; SMITH, A. F. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, Taylor & Francis Group, v. 85, n. 410, p. 398–409, 1990. Citado na página 10.

GELFAND, P. J. D. A. E.; FUENTES, P. D. M. Handbook of spatial statistics. *Chapman and hall/CRC handbooks of modern statistical methods.*, Boca Raton, FL. CRC Press, 2010. Citado na página 13.

GELMAN, A. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, International Society for Bayesian Analysis, v. 1, n. 3, p. 515–534, 2006. Citado na página 21.

GEWEKE, J. et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. [S.l.]: Federal Reserve Bank of Minneapolis, Research Department, 1991. Citado na página 28.

HAINING, R. *Spatial data analysis in the social and environmental sciences*. [S.l.]: Cambridge University Press, 1993. Citado 2 vezes nas páginas 10 e 12.

HAMMERSLEY, J.; CLIFFORD, P. Markov fields on finite graphs and lattices. *Unpublished manuscript*, 1971. Citado na página 15.

HEIDELBERGER, P.; WELCH, P. D. Simulation run length control in the presence of an initial transient. *Operations Research*, INFORMS, v. 31, n. 6, p. 1109–1144, 1983. Citado na página 29.

LEE, D. A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, Elsevier, v. 2, n. 2, p. 79–89, 2011. Citado 5 vezes nas páginas 8, 11, 18, 22 e 42.

LEE, D.; FERGUSON, C.; MITCHELL, R. Air pollution and health in scotland: a multicity study. *Biostatistics*, Biometrika Trust, v. 10, n. 3, p. 409–423, 2009. Citado na página 11.

LEROUX, B. G.; LEI, X.; BRESLOW, N. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. [S.l.]: Springer, 2000. p. 179–191. Citado 3 vezes nas páginas 10, 17 e 19.

LU, H.; CARLIN, B. P. Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, Wiley Online Library, v. 37, n. 3, p. 265–285, 2005. Citado na página 10.

LU, H. et al. Bayesian areal wombling via adjacency modeling. *Environmental and ecological statistics*, Springer, v. 14, n. 4, p. 433–452, 2007. Citado 5 vezes nas páginas 10, 20, 21, 22 e 26.

LUNN, D. et al. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, Wiley Online Library, v. 28, n. 25, p. 3049–3067, 2009. Citado na página 24.

MACNAB, Y. C. et al. An innovative application of bayesian disease mapping methods to patient safety research: A canadian adverse medical event study. *Statistics in medicine*, Wiley Online Library, v. 25, n. 23, p. 3960–3980, 2006. Citado na página 20.

MARTIN, R. The use of time-series models and methods in the analysis of agricultural field trials. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 19, n. 1, p. 55–81, 1990. Citado na página 11.

MINASNY, B.; MCBRATNEY, A. B. The matérn function as a general model for soil variograms. *Geoderma*, Elsevier, v. 128, n. 3, p. 192–207, 2005. Citado na página 42.

MOLINA, R.; KATSAGGELOS, A. K.; MATEOS, J. Bayesian and regularization methods for hyperparameter estimation in image restoration. *Image Processing, IEEE Transactions on*, IEEE, v. 8, n. 2, p. 231–246, 1999. Citado na página 11.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>. Citado na página 27.

RIPLEY, B. Spatial statistics. 1981. *Willey, New York*, 1981. Citado na página 12.

ROBERTSON, T.; CRYER, J. D. An iterative procedure for estimating the mode. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 69, n. 348, p. 1012–1016, 1974. Citado na página 44.

RODRIGUES, E. C. *Inferindo a Estrutura de Vizinhança em Modelos Bayesianos Espaciais*. Dissertação (Mestrado) — Pos-Graduação em Estatistica, UFMG, 2011. Citado na página 20.

RODRIGUES, E. C.; ASSUNÇÃO, R. Bayesian spatial models with a mixture neighborhood structure. *Journal of Multivariate Analysis*, Elsevier, v. 109, p. 88–102, 2012. Citado 2 vezes nas páginas 11 e 18.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Citado na página 32.

STERN, H. S.; CRESSIE, N. Posterior predictive model checks for disease mapping models. *Statistics in medicine*, Wiley Online Library, v. 19, n. 17-18, p. 2377–2397, 2000. Citado 4 vezes nas páginas 8, 10, 18 e 19.

THOMAS, A. et al. Geobugs user manual. 2004. Disponível em: <http://www-.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml>. Citado 2 vezes nas páginas 10 e 17.

WALL, M. M. A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference*, Elsevier, v. 121, n. 2, p. 311–324, 2004. Citado na página 12.

WIMBERLY, M. C. et al. Assessing fuel treatment effectiveness using satellite imagery and spatial statistics. *Ecological Applications*, Eco Soc America, v. 19, n. 6, p. 1377–1384, 2009. Citado na página 11.

WOMBLE, W. H. Differential systematics. *Science*, American Association for the Advancement of Science, v. 114, n. 2961, p. 315–322, 1951. Citado na página 21.

Apêndices

APÊNDICE A – Códigos R-OpenBUGS Utilizados nas Aplicações

Abaixo estão listados duas funções que implementam os modelos CAR na linguagem de programação *R*. A ligação com o OpenBUGS foi feita com o auxílio do pacote R2OpenBUGS.

```
 intrinseco <- function()
{
        for (i in 1 : n)
        {
                Y[i] ~ dpois(mu[i])
                log(mu[i]) <- alpha + log(E[i]) + phi[i]
                RR[i] <- exp(alpha + phi[i])
        }
        phi[1:n] ~ car.normal(adj[], weights[], num[], tau)
        for(k in 1:sumNumNeigh)
                weights[k] <- 1
#Prioris
        alpha ~ dflat()
        sigma ~ dunif(0, 10)
        sigma2 <- pow(sigma,2)
        tau <- 1/sigma2
}

convolucao <- function()
{
        for (i in 1 : n)
        {
                Y[i] ~ dpois(mu[i])
                log(mu[i]) <- alpha + log(E[i]) + phi[i] + psi[i]
                RR[i] <- exp(alpha + phi[i] + psi[i])
        psi[i] ~ dnorm(0, tau.psi)
        }
        phi[1:n] ~ car.normal(adj[], weights[], num[], tau.phi)
        for(k in 1:sumNumNeigh)
```

```
            weights[k] <- 1
#Prioris
    alpha ~ dflat()
    sigma.phi ~ dunif(0, 10)
    sigma.phi2 <- pow(sigma.phi,2)
    tau.phi <- 1/sigma.phi2
    sigma.psi ~ dunif(0, 10)
    sigma.psi2 <- pow(sigma.psi,2)
    tau.psi <- 1/sigma.psi2
}

cressie <- function()
{
    for (i in 1 : n)
    {
        Y[i] ~ dpois(mu[i])
        log(mu[i]) <- alpha + log(E[i]) + phi[i]
        RR[i] <- exp(alpha + phi[i])
    }
    for (a in 1:sumNumNeigh)
    {
        phiaux[a] <- phi[adj[a]]
    }
    for (j in 1:n)
    {
        phi[j] ~ dnorm(phibar[j],phivar[j])
        phibar[j] <- rho*sum(phiaux[off1[j]:off2[j]])/(num[j])
        phivar[j] <- tau*num[j]
    }
    #Prioris
    alpha ~ dflat()
    sigma ~ dunif(0, 10)
    sigma2 <- pow(sigma,2)
    tau <- 1/sigma2
    rho ~ dunif(0,1)
}

leroux <- function()
{
```

```
        for (i in 1 : n)
        {
                Y[i] ~ dpois(mu[i])
                log(mu[i]) <- alpha + log(E[i]) + phi[i]
                RR[i] <- exp(alpha + phi[i])
        }
        for (a in 1:sumNumNeigh)
        {
                phiaux[a] <- phi[adj[a]]
        }
        for (j in 1:n)
        {
                phi[j] ~ dnorm(phibar[j],phivar[j])
                phibar[j] <- rho*sum(phiaux[off1[j]:off2[j]])/ (1-rho+rho*num[j])
                phivar[j] <- tau*(1-rho+rho*num[j])
        }
        #Prioris
        alpha ~ dflat()
        sigma ~ dunif(0, 10)
        sigma2 <- pow(sigma,2)
        tau <- 1/sigma2
        rho ~ dunif(0,1)
}

lu <- function()
{
        for (i in 1 : n)
        {
                Y[i] ~ dpois(mu[i])
                log(mu[i]) <- alpha + log(E[i]) + phi[i]
                RR[i] <- exp(alpha + phi[i])
        }
        for (i in 1 : n)
        {
                for (j in i : n)
                {
                        logit(p[i,j]) <- g0 + g1*D[i,j]
                        w2[i,j] ~ dbern(p[i,j])
                        w[i,j] <- w1[i,j]*w2[i,j]
```

```
            }
        }
        for (i in 2 : n)
        {
            for (j in 1 : (i-1))
            {
                w[i,j] <- w[j,i]
            }
        }
        for (j in 1:n)
        {
            b[j] ~ dnorm(bbar[j],bvar[j])
            bbar[j] <- inprod(w[j,],b[1:n])/(sum(w[j,])+0.5)
            bvar[j] <- tau*(sum(w[j,])+0.5)
        }
#Prioris
        alpha ~ dflat()
        sigma ~ dunif(0, 10)
        sigma2 <- pow(sigma,2)
        tau <- 1/sigma2
        g0 ~ dnorm(1,1)
        g1 ~ dnorm(-5,1)
}
```

APÊNDICE B – Códigos R-OpenBUGS Utilizados na Simulação

Abaixo está listado o código utilizado no estudo de simulação. Novamente, a ligação com o OpenBUGS foi feita com o auxílio do pacote R2OpenBUGS.

```
 #Definindo sementes
#Cenário 1
set.seed(2983472)
u1 <- round(runif(iter,1,100000),0)
#Cenário 2
set.seed(2124456)
u2 <- round(runif(iter,1,100000),0)
#Cenário 3
set.seed(987452)
u3 <- round(runif(iter,1,100000),0)

#Simulação cenário 1
for(i in 1:iter)
{
        set.seed(u1[i])
        b <- mvrnorm(n=1, rep(0,n), Rho1)
        mu <- E*exp(b0+b)
        Y <- rpois(n,mu)
        R1[i,] <- exp(b0+b)

        data <- list(n=n,
        Y=Y,
        E=E,
        off1 = off1,
        off2 = off2,
        num = num,
        adj = adj,
        sumNumNeigh = sumNumNeigh)

        model1 <- bugs(data, inits1, model.file = intrinseco,parameters = c("sigma2",
```

```
"alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
coda1 <- read.bugs(model1)
for(j in 1:n)
        R_est11[i,j] <- mean(coda1[,j][[1]])
est11[i,1] <- mean(coda1[,64][[1]])
est11[i,2] <- sd(coda1[,64][[1]])
est11[i,3] <- mean(coda1[,66][[1]])
est11[i,4] <- sd(coda1[,66][[1]])

model2 <- bugs(data, inits2, model.file = convolucao,parameters = c("sigmab2",
"sigmah2","alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
coda2 <- read.bugs(model2)
for(j in 1:n)
        R_est21[i,j] <- mean(coda2[,j][[1]])
est21[i,1] <- mean(coda2[,64][[1]])
est21[i,2] <- sd(coda2[,64][[1]])
est21[i,3] <- mean(coda2[,66][[1]])
est21[i,4] <- sd(coda2[,66][[1]])
est21[i,5] <- mean(coda2[,67][[1]])
est21[i,6] <- sd(coda2[,67][[1]])

model3 <- bugs(data, inits1, model.file = cressie,parameters = c("rho", "sigma2",
"alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
coda3 <- read.bugs(model3)
for(j in 1:n)
        R_est31[i,j] <- mean(coda3[,j][[1]])
est31[i,1] <- mean(coda3[,64][[1]])
est31[i,2] <- sd(coda3[,64][[1]])
est31[i,3] <- mean(coda3[,66][[1]])
est31[i,4] <- sd(coda3[,66][[1]])
est31[i,5] <- mean(coda3[,67][[1]])
est31[i,6] <- sd(coda3[,67][[1]])

model4 <- bugs(data, inits1, model.file = leroux,parameters = c("rho", "sigma2",
"alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
```

```
        coda4 <- read.bugs(model4)
        for(j in 1:n)
                R_est41[i,j] <- mean(coda4[,j][[1]])
        est41[i,1] <- mean(coda4[,64][[1]])
        est41[i,2] <- sd(coda4[,64][[1]])
        est41[i,3] <- mean(coda4[,66][[1]])
        est41[i,4] <- sd(coda4[,66][[1]])
        est41[i,5] <- mean(coda4[,67][[1]])
        est41[i,6] <- sd(coda4[,67][[1]])

        model5 <- bugs(data, initslu, model.file = lu,parameters = c("alpha",
        "sigma2","RR","g0","g1"),debug=F,DIC=T,n.burnin=burnin,n.thin=thin,
        n.chains = 1,n.iter = cadeia,codaPkg = T)
        coda5 <- read.bugs(model5)
        for(j in 1:n)
                R_est51[i,j] <- mean(coda5[,j][[1]])
        est51[i,1] <- mean(coda5[,64][[1]])
        est51[i,2] <- sd(coda5[,64][[1]])
        est51[i,3] <- mean(coda5[,66][[1]])
        est51[i,4] <- sd(coda5[,66][[1]])
        est51[i,5] <- mean(coda5[,67][[1]])
        est51[i,6] <- sd(coda5[,67][[1]])
        est51[i,7] <- mean(coda5[,68][[1]])
        est51[i,8] <- sd(coda5[,68][[1]])
}

#Simulação cenário 2
tempo2 <- proc.time()
for(i in 1:iter)
{
        set.seed(u2[i])
        b <- mvrnorm(n=1, rep(0,n), Rho2)
        mu <- E*exp(b0+b)
        Y <- rpois(n,mu)
        R2[i,] <- exp(b0+b)

        data <- list(n=n,
        Y=Y,
        E=E,
```

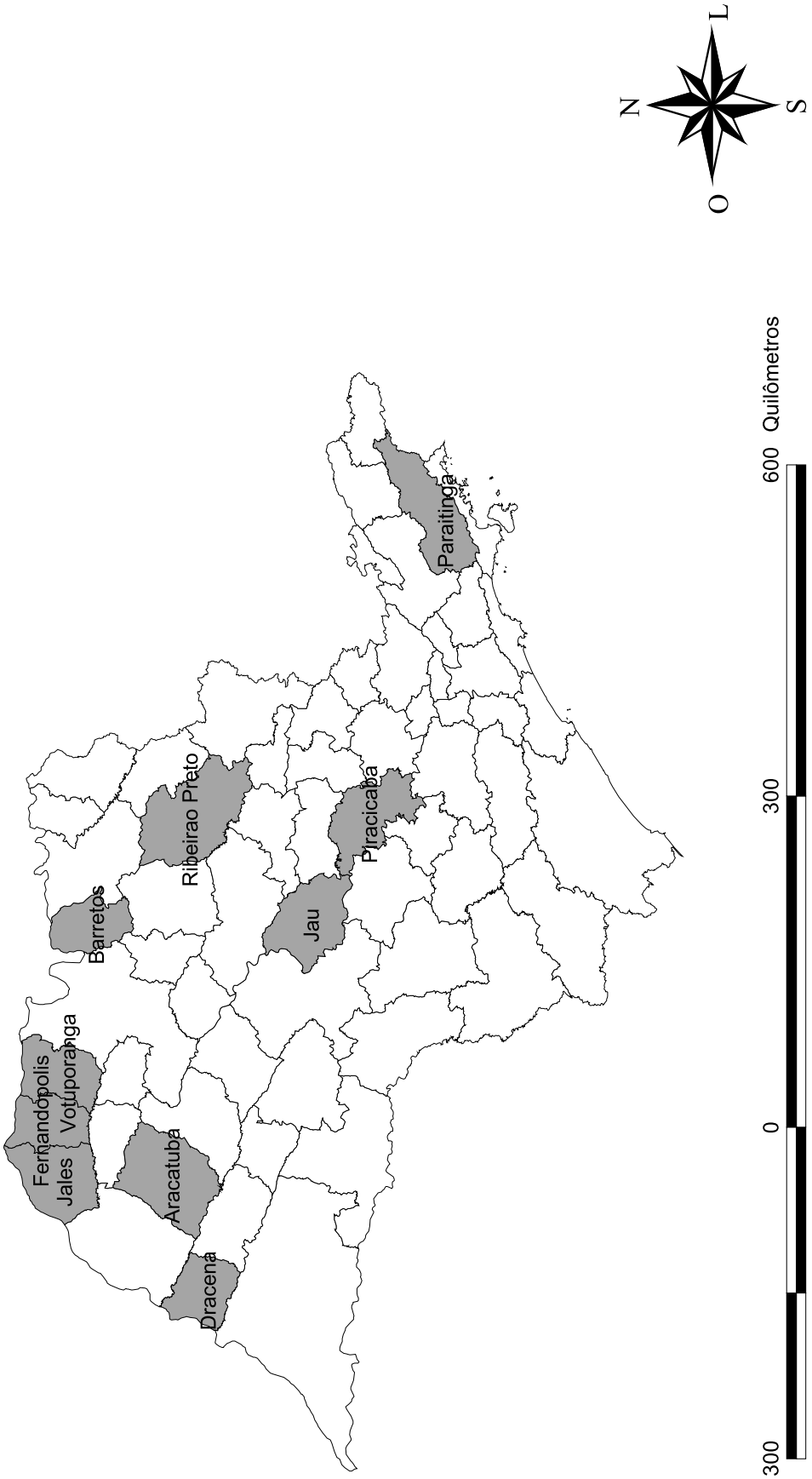
```
off1 = off1,
off2 = off2,
num = num,
adj = adj,
sumNumNeigh = sumNumNeigh)

model1 <- bugs(data, inits1, model.file = intrinseco,parameters = c("sigma2",
"alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
coda1 <- read.bugs(model1)
for(j in 1:n)
        R_est12[i,j] <- mean(coda1[,j][[1]])
est12[i,1] <- mean(coda1[,64][[1]])
est12[i,2] <- sd(coda1[,64][[1]])
est12[i,3] <- mean(coda1[,66][[1]])
est12[i,4] <- sd(coda1[,66][[1]])

model2 <- bugs(data, inits2, model.file = convolucao,parameters = c("sigmab2",
"sigmah2","alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
coda2 <- read.bugs(model2)
for(j in 1:n)
        R_est22[i,j] <- mean(coda2[,j][[1]])
est22[i,1] <- mean(coda2[,64][[1]])
est22[i,2] <- sd(coda2[,64][[1]])
est22[i,3] <- mean(coda2[,66][[1]])
est22[i,4] <- sd(coda2[,66][[1]])
est22[i,5] <- mean(coda2[,67][[1]])
est22[i,6] <- sd(coda2[,67][[1]])

model3 <- bugs(data, inits1, model.file = cressie,parameters = c("rho", "sigma2",
"alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
coda3 <- read.bugs(model3)
for(j in 1:n)
        R_est32[i,j] <- mean(coda3[,j][[1]])
est32[i,1] <- mean(coda3[,64][[1]])
est32[i,2] <- sd(coda3[,64][[1]])
est32[i,3] <- mean(coda3[,66][[1]])
```

```
        est32[i,4] <- sd(coda3[,66][[1]])
        est32[i,5] <- mean(coda3[,67][[1]])
        est32[i,6] <- sd(coda3[,67][[1]])

        model4 <- bugs(data, inits1, model.file = leroux,parameters = c("rho", "sigma2",
        "alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
        n.iter = cadeia,codaPkg = T)
        coda4 <- read.bugs(model4)
        for(j in 1:n)
                R_est42[i,j] <- mean(coda4[,j][[1]])
        est42[i,1] <- mean(coda4[,64][[1]])
        est42[i,2] <- sd(coda4[,64][[1]])
        est42[i,3] <- mean(coda4[,66][[1]])
        est42[i,4] <- sd(coda4[,66][[1]])
        est42[i,5] <- mean(coda4[,67][[1]])
        est42[i,6] <- sd(coda4[,67][[1]])

        model5 <- bugs(data, initslu, model.file = lu,parameters = c("alpha",
        "sigma2","RR","g0","g1"),debug=F,DIC=T,n.burnin=burnin,n.thin=thin,
        n.chains = 1,n.iter = cadeia,codaPkg = T)
        coda5 <- read.bugs(model5)
        for(j in 1:n)
                R_est52[i,j] <- mean(coda5[,j][[1]])
        est52[i,1] <- mean(coda5[,64][[1]])
        est52[i,2] <- sd(coda5[,64][[1]])
        est52[i,3] <- mean(coda5[,66][[1]])
        est52[i,4] <- sd(coda5[,66][[1]])
        est52[i,5] <- mean(coda5[,67][[1]])
        est52[i,6] <- sd(coda5[,67][[1]])
        est52[i,7] <- mean(coda5[,68][[1]])
        est52[i,8] <- sd(coda5[,68][[1]])
}

#Simulação cenário 3
tempo3 <- proc.time()
for(i in 1:iter)
{
        set.seed(u3[i])
        b <- mvrnorm(n=1, rep(0,n), Rho3)
```

```
mu <- E*exp(b0+b)
Y <- rpois(n,mu)
R3[i,] <- exp(b0+b)

data <- list(n=n,
Y=Y,
E=E,
off1 = off1,
off2 = off2,
num = num,
adj = adj,
sumNumNeigh = sumNumNeigh)

model1 <- bugs(data, inits1, model.file = intrinseco,parameters = c("sigma2",
"alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
coda1 <- read.bugs(model1)
for(j in 1:n)
        R_est13[i,j] <- mean(coda1[,j][[1]])
est13[i,1] <- mean(coda1[,64][[1]])
est13[i,2] <- sd(coda1[,64][[1]])
est13[i,3] <- mean(coda1[,66][[1]])
est13[i,4] <- sd(coda1[,66][[1]])

model2 <- bugs(data, inits2, model.file = convolucao,parameters = c("sigmab2",
"sigmah2","alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
n.iter = cadeia,codaPkg = T)
coda2 <- read.bugs(model2)
for(j in 1:n)
        R_est23[i,j] <- mean(coda2[,j][[1]])
est23[i,1] <- mean(coda2[,64][[1]])
est23[i,2] <- sd(coda2[,64][[1]])
est23[i,3] <- mean(coda2[,66][[1]])
est23[i,4] <- sd(coda2[,66][[1]])
est23[i,5] <- mean(coda2[,67][[1]])
est23[i,6] <- sd(coda2[,67][[1]])

model3 <- bugs(data, inits1, model.file = cressie,parameters = c("rho", "sigma2",
"alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
```

```
        n.iter = cadeia,codaPkg = T)
        coda3 <- read.bugs(model3)
        for(j in 1:n)
                R_est33[i,j] <- mean(coda3[,j][[1]])
        est33[i,1] <- mean(coda3[,64][[1]])
        est33[i,2] <- sd(coda3[,64][[1]])
        est33[i,3] <- mean(coda3[,66][[1]])
        est33[i,4] <- sd(coda3[,66][[1]])
        est33[i,5] <- mean(coda3[,67][[1]])
        est33[i,6] <- sd(coda3[,67][[1]])

        model4 <- bugs(data, inits1, model.file = leroux,parameters = c("rho", "sigma2",
        "alpha","RR"),debug=FALSE ,n.burnin=burnin,n.thin=thin,n.chains = 1,
        n.iter = cadeia,codaPkg = T)
        coda4 <- read.bugs(model4)
        for(j in 1:n)
                R_est43[i,j] <- mean(coda4[,j][[1]])
        est43[i,1] <- mean(coda4[,64][[1]])
        est43[i,2] <- sd(coda4[,64][[1]])
        est43[i,3] <- mean(coda4[,66][[1]])
        est43[i,4] <- sd(coda4[,66][[1]])
        est43[i,5] <- mean(coda4[,67][[1]])
        est43[i,6] <- sd(coda4[,67][[1]])

        model5 <- bugs(data, initslu, model.file = lu,parameters = c("alpha",
        "sigma2","RR","g0","g1"),debug=F,DIC=T,n.burnin=burnin,n.thin=thin,
        n.chains = 1,n.iter = cadeia,codaPkg = T)
        coda5 <- read.bugs(model5)
        for(j in 1:n)
                R_est53[i,j] <- mean(coda5[,j][[1]])
        est53[i,1] <- mean(coda5[,64][[1]])
        est53[i,2] <- sd(coda5[,64][[1]])
        est53[i,3] <- mean(coda5[,66][[1]])
        est53[i,4] <- sd(coda5[,66][[1]])
        est53[i,5] <- mean(coda5[,67][[1]])
        est53[i,6] <- sd(coda5[,67][[1]])
        est53[i,7] <- mean(coda5[,68][[1]])
        est53[i,8] <- sd(coda5[,68][[1]])
}
```

APÊNDICE C – Microrregiões de Maiores Riscos nas Aplicações


L
N
S
O
Paraitinga
Piracicaba
Ribeirao Preto
Barretos
Jau
Votuporanga
Fernandópolis
Jales
Aracatuba
Dracena
300
0
300
600 Quilômetros