Universidade Estadual Paulista - UNESP

Faculdade de Ciências e Tecnologia

Suelen Umbelino da Silva

Modelo dinâmico bayesiano multivariado para análise espaço-temporal de dados de área

Presidente Prudente

2014

Universidade Estadual Paulista - UNESP

Faculdade de Ciências e Tecnologia

Suelen Umbelino da Silva

Modelo dinâmico bayesiano multivariado para análise espaço-temporal de dados de área

Relatório final para obtenção do título de Mestre em Matemática Aplicada e Computacional pela Universidade Estadual Paulista, sob orientação da Profa. Dra. Aparecida Doniseti Pires de Souza

Programa de Pós Graduação em Matemática Aplicada e Computacional

Orientadora: Profa. Dra. Aparecida Doniseti Pires de Souza

Presidente Prudente
2014

Silva, Suelen Umbelino da.
S583m Modelo dinâmico bayesiano multivariado para análise espaço-temporal de dados de área / Suelen Umbelino da Silva. - Presidente Prudente : [s.n], 2014
x, 111 f. : il.

Orientador: Aparecida Doniseti Pires de Souza
Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Ciências e Tecnologia
Inclui bibliografia

1. Modelo Hierárquico Bayesiano Dinâmico. 2. Razão de Mortalidade Padronizada. 3. MVCAR. I. Souza, Aparecida Doniseti Pires de. II Universidade Estadual Paulista. Faculdade de Ciências e Tecnologia. III. Título.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Presidente Prudente

BANCA EXAMINADORA

PROFA. DRA. APARECIDA DONISETI P. DE SOUSA
ORIENTADORA

PROF. DR. LEONARDO SOARES BASTOS
FIOCRUZ

PROF. DR. RENATO MARTINS ASSUNÇÃO
UFMG

SUELEN UMBELINO DA SILVA

Presidente Prudente (SP), 29 de agosto de 2014.

Resultado: aprovada

Faculdade de Ciências e Tecnologia
Seção Técnica de Pós-Graduação
Rua Roberto Simonsen, 305 CEP 19060-900 Presidente Prudente SP
Tel 18 3229-5318 fax 18 3223-4519 posgrad@fct.unesp.br

Agradecimentos

Não poderia deixar de agradecer em primeiro lugar a Deus, aquele que “me confere poder” para que eu tenha “forças para enfrentar todas as coisas” (Filipenses 4:13), e que “me faz pisar no caminho em que devo andar” (Isaías 48:17).

Agradeço também à minha família, principalmente aos meus pais, que sempre me incentivaram o meu gosto pelos estudos, embora eles mesmos não tenham tido oportunidades.

Ao meu namorado, Jorge, pela paciência e conforto fornecido nos momentos difíceis.

À professora Aparecida Doniseti Pires de Souza, pela orientação.

À professora Vilma Mayumi Tachibana, por sempre ser prestativa e paciente em suas contribuições.

Aos membros da banca, pela disposição em dispor de seu tempo e energia pra contribuir com correções e enriquecimento deste trabalho.

Ao professor Sérgio Minoru Oikawa, pelas sugestões e contribuições no exame de qualificação.

À todos aqueles que de alguma forma me ajudaram direta ou indiretamente, tanto no decorrer do mestrado, para o meu crescimento profissional, quanto no decorrer da minha vida, para o meu crescimento pessoal.

“ Dê a um homem um peixe, e você o alimentará por um dia.
Ensine-o a pescar, e você o alimentará por toda a vida. ”

Provérbio Chinês.

Resumo

Modelagem de dados de área tem sido tema de diversas pesquisas em Estatística nas últimas décadas. Modelos espaço-temporais têm sido utilizados para lidar com esse tipo de dados de um modo natural, uma vez que muitas vezes envolvem processos que têm transições no tempo e no espaço. O avanço da tecnologia e, simultaneamente, de métodos estatísticos, têm permitido a elaboração de modelos cada vez mais estruturados para a descrição de fenômenos aleatórios complexos, cuja ideia é descrever, de forma realista, a estrutura de correlação presente nos dados, o que pode ser feito através do uso de modelos hierárquicos. Dada a importância atual da modelagem de fenômenos espaço-temporais, neste trabalho são estudadas propostas recentes apresentadas na literatura para dados espaciais de área, envolvendo modelos autorregressivos condicionais multivariados para capturar a estrutura espacial e modelos dinâmicos para capturar a estrutura temporal. Como aplicação da metodologia é estudada a distribuição espacial da mortalidade pelos cânceres de maior importância quantitativa, segundo as microrregiões administrativas do estado de São Paulo, considerando o período 1998 até 2010. Os resultados da aplicação de um modelo bayesiano hierárquico para os dados evidenciaram quais as regiões de maior risco de mortalidade no estado de São Paulo para cada um dos cânceres estudados, além de mostrar que existe forte correlação espacial entre algumas das doenças, o que constitui um resultado muito importante para os órgãos do sistema de saúde, que têm como função direcionar e alocar recursos para o tratamento e diagnóstico de tais doenças. Na aplicação de um modelo bayesiano hierárquico dinâmico, com passeio aleatório de ordem um assumido como distribuição a priori para os efeitos espaciais, tais efeitos não se mostraram significativos na aplicação do modelo aos dados em estudo. No entanto, a inclusão do domínio temporal proporcionou a produção de informação acerca das doenças ano a ano do período, levando a conclusões similares ao modelo sem efeito temporal, além da produção de estimativas mais suaves e de mais fácil interpretação para o risco relativo do que as obtidas através do modelo clássico.

Palavras-chave: Modelo Hierárquico Bayesiano Dinâmico, Razão de Mortalidade Padronizada, MVCAR.

Abstract

Data modeling area has been the subject of several studies in Statistics in recent decades. Spatio-temporal models have been used to deal with this kind of data in a natural way, since they often involve processes that have transitions in time and space. The advancement of technology and simultaneously statistical methods have allowed the development of increasingly structured models for the description of complex random phenomena, whose idea is to describe realistically, the structure of this correlation in the data, which can be done through the use of hierarchical models. Given the current importance of modeling spatio-temporal phenomena, the aim of this work is study recent paper that involve multivariate conditional autoregressive models to capture the spatial and dynamic structure models to capture the temporal structure. As an application of the methodology is the spatial distribution of mortality for cancers of greater quantitative importance studied, according to the administrative microregions of the state of São Paulo, considering the period 1998 to 2010. The results of the application of a Bayesian hierarchical model to the data showed that the regions of greatest risk of mortality in São Paulo for each one of the cancers studied, and show that there is a strong spatial correlation between some of the diseases, which is a very important result for the organs of the health system, whose function is to direct and allocate resources for the treatment and diagnosis of such diseases. In the application of a dynamic Bayesian hierarchical model with random walk of order as an assumed prior distribution for spatial effects, such effects were not significant in applying the model to the data in the study. However, the inclusion of the temporal domain provides the production of information about the disease every year in the period, leading to similar model without the time effect conclusions, as well as producing smoother estimates and easier to interpret than the relative risk those obtained through the classical model.

Keywords: Dynamic Bayesian Hierarchical Model, Standardized Mortality Ratio, MVCAR.

Lista de Figuras

Sumário

Capítulo 1

Introdução

Modelagem de dados de área tem sido tema de diversas pesquisas em Estatística nas últimas décadas. Modelos espaço-temporais têm sido utilizados para lidar com esse tipo de dados de um modo natural, uma vez que muitas vezes envolvem processos ambientais, epidemiológicos, ecológicos, entre outros que têm, em geral, transições no tempo e no espaço. O avanço da tecnologia e, simultaneamente, de métodos estatísticos, têm permitido a elaboração de modelos cada vez mais estruturados para a descrição de fenômenos aleatórios complexos. A ideia é descrever, de forma realista, a estrutura de correlação presente nos dados, o que pode ser feito através do uso de modelos hierárquicos bayesianos.

Dada a importância atual da modelagem de fenômenos espaço-temporais, este trabalho tem por objetivo o estudo de modelos propostos na literatura para analisar dados espaciais de área, os modelo hierárquicos bayesianos, que envolvem o uso de modelos autorregressivos condicionais multivariados para capturar a estrutura espacial dos dados, e modelos dinâmicos lineares generalizados para capturar a estrutura temporal. Devido à complexidade dos modelos em estudo e o uso da abordagem Bayesiana como procedimento de inferência, métodos de Monte Carlo via Cadeias de Markov (MCMC) são utilizados na estimação dos parâmetros de interesse. Para a implementação destes modelos foi utilizado o Software OpenBUGS (Bayesian Analysis Using Gibbs Sampler) (Lunn *et al.* (2009)), e o seu módulo GeoBUGS, que permite o mapeamento das amostras a posteriori de parâmetros de interesse.

Para a aplicação dessa metodologia considera-se a distribuição espacial da mortalidade pelos cânceres de maior importância quantitativa, segundo as microrregiões administrativas do estado de São Paulo, englobando o período compreendido de 1998 até 2010. O objetivo na aplicação é verificar a existência de padrões na distribuição espaço-temporal dos óbitos, a presença de correlação entre os diferentes tipos da doença e, por consequência, determinar regiões de maior risco.

A divisão deste trabalho encontra-se de modo que, no segundo Capítulo são considerados modelos apropriados para o mapeamento de doenças, o que inclui o Modelo Clássico de Riscos Relativos, o Modelo Hierárquico Bayesiano, e o Modelo Hierárquico Bayesiano Dinâmico, quanto às suas características e aplicabilidade a dados de área. No terceiro Capítulo apresenta-se uma introdução sobre Campos Aleatórios Markovianos Gaussianos (CAMG), uma vez que estes estão

diretamente ligados à construção da estrutura dos modelos Condicionais Autorregressivos (CAR), sendo estes apresentados na sequência, na qual também é considerada sua versão multivariada. No Capítulo 4, relata-se alguns dos algoritmos da classe MCMC, implementados no Software OpenBUGS, utilizados na obtenção de amostras da distribuição a posteriori de parâmetros de interesse, bem como diagnósticos para a software OpenBUGS. No Capítulo 5 estão dispostos os resultados de uma análise exploratória útil para a compreensão dos resultados do modelo. No Capítulo 6 constam os resultados da aplicação do Modelo Hierárquico Bayesiano para os dados de câncer agrupados, seguidos da aplicação do Modelo Hierárquico Bayesiano Dinâmico, ajustado aos dados ano a ano. As áreas de maior risco para as doenças são apresentadas, e a correlação entre elas discutida.

Finalmente, no Capítulo 7 são expostos os resultados e as conclusões obtidas neste estudo, em conexão com perspectivas futuras ligadas à este trabalho.

Capítulo 2

Modelos para o mapeamento de doenças

O termo *mapeamento de doenças* é utilizado para denominar uma área da epidemiologia que tem por objetivo estudar o padrão espacial do risco de uma doença em determinada região geográfica, de modo que as áreas de alto risco possam ser identificadas. A maior parte dos mapas são temáticos ou coropléticos, nos quais um conjunto de áreas são sombreadas de acordo com seus valores na variável de interesse. A variável a ser modelada geralmente é a taxa de mortalidade para a doença, ou o risco associado à mesma. Mapas de taxas de incidência constituem a principal ferramenta na análise da dispersão do risco de uma doença, pois além de permitir a visualização da distribuição espacial do fenômeno, são importantes instrumentos em apontar fatores etiológicos desconhecidos e potenciais fontes de contaminação, resultantes da presença de evidentes áreas de risco elevado.

Apesar do atrativo, Stern e Cressie (2000) discutem que os mapas de taxas de mortalidade não são confiáveis, devido à variância não constante associada à heterogeneidade do tamanho da população, que em algumas áreas podem ser muito pequenas. Sendo assim, uma análise mais fidedigna leva em conta não só o tamanho da população, como também a sua estrutura, e considera o mapeamento do risco a partir de medidas mais condizentes com a realidade do fenômeno, o que pode ser feito através de uma modelagem adequada para os riscos relativos de cada região.

Na sequência são consideradas duas abordagens para a estimativa do risco relativo a ser mapeado: o modelo clássico de riscos relativos e o modelo hierárquico bayesiano. Por fim, o modelo hierárquico bayesiano dinâmico é definido para incluir o domínio temporal na análise.

2.1 Modelo Clássico de Riscos Relativos

Na abordagem clássica, um procedimento comumente utilizado para lidar espacialmente com taxas de mortalidade de doenças consiste em mapear as Razões de Mortalidade Padronizadas (Standardized Mortality Ratio - SMR, em inglês). A SMR é a razão entre o número observado de mortes na população em estudo e o número de mortes que seria de se esperar nesta baseado na distribuição de idade e sexo da respectiva população ou de uma população padrão com a qual se deseja comparar a região em estudo. Se a proporção de mortes observadas em relação às esperadas

é maior do que 1 em determinada área, então é dito haver mortes em excesso em sua população.

Essas razões também são as estimativas de máxima verossimilhança do risco relativo de mortalidade da doença nas áreas, considerando que a contagem dos óbitos segue o modelo de Poisson. Mais formalmente, seja o número de óbitos devido à uma determinada doença em n áreas de uma região geográfica denotado por $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, sendo Y_i o número de casos na área i, para $i = 1, ..., n$. Isto é,

$$Y_i \sim Poisson(\lambda_i), \quad i = 1, ..., n,$$

com $\lambda_i = E_i R_i$, sendo E_i o elemento do vetor $\mathbf{E} = (E_1, E_2, ..., E_n)$ correspondente à área i, assumido conhecido, e que representa o número esperado de casos de acordo com as características da área em questão e R_i é o risco relativo de óbito na área i a ser modelado.

A função de verossimilhança de λ_i dado y_i é dada por

$$L(\lambda_i; y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \quad i = 1, ..., n.$$

Aplicando o logaritmo nessa função, obtém-se

$$l(\lambda_i, y_i) = \log(L(\lambda_i, y_i)) = -\lambda_i + y_i \log(\lambda_i) - \log(y_i!).$$

Como

$$\frac{\partial l(\lambda_i, y_i)}{\partial \lambda_i} = -1 + \frac{y_i}{\hat{\lambda i}},$$

verifica-se facilmente que o estimador de máxima verossimilhança de λ_i é dado por

$$\hat{\lambda_i} = y_i$$

Isto é,

$$E_i \hat{R}_i = y_i \Rightarrow \hat{R}_i = \frac{y_i}{E_i} = SMR_i, \quad i = 1, ..., n.$$

Portanto, a SMR é o EMV para o risco. Levando-se em conta que se trata de uma razão padronizada, é necessário esclarecer que a padronização ocorre durante o cálculo do vetor esperado **E**. Isso permite que a estrutura demográfica das áreas sejam levadas em consideração, e, portanto, le-

vem a uma interpretação mais realística dos riscos relativos, o que é especialmente importante nas áreas cujas populações incluem uma grande parcela de idosos ou de indivíduos do sexo masculino, uma vez que, conhecidamente, tais fatores aumentam o risco para determinadas doenças.

Para a padronização indireta, considere i o índice da área e j o índice da classe de idade-sexo. Por exemplo, $j = 1$ indica os óbitos femininos na faixa etária de menores de 1 ano de idade, $j = 2$ indica os óbitos femininos na faixa etária de 1 a 4 anos de idade, e assim por diante. Atente para uma determinada classe j, por exemplo, $j = 5$, que significa mulheres de 15 a 19 anos de idade. Seja y_{ij} o número de óbitos que ocorreram entre pessoas da classe j na área i, e N_{ij} a respectiva população na classe j e área i. O risco global em todo o mapa referente apenas à classe de idade-sexo j é dado por

$$r_j = \frac{\sum_i y_{ij}}{\sum_i N_{ij}}$$

Então, $E_{ij} = r_j N_{ij}$ é o número esperado de óbitos na classe j e área i se o risco na classe j fosse constante no espaço. Consequentemente, o número total esperado de óbitos na área i para o risco de cada classe de idade-sexo constante no espaço é dado por

$$E_i = \sum_j E_{ij}.$$

Com isso, a SMR é calculada como a razão entre o número observado de óbitos e o número esperado, se o risco fosse constante no espaço. Isto é,

$$SMR_i = \hat{R}_i = \frac{y_i}{E_i}$$

Assim, na hipótese de que o risco seja constante no espaço em cada classe de idade-sexo, temos que $Y_i \sim Poisson(E_i R_i)$, sendo E_i calculado de acordo com a explicação anterior.

Uma das críticas associada ao uso da SMR como estimador do risco relativo é a flutuação aleatória associada a áreas com pequenas populações. Observe que este estimador possui variância inversamente proporcional ao número esperado de eventos E_i. Assim, quando este número for pequeno, o que geralmente acontece para regiões pequenas, a variabilidade do estimador pode ser muito grande. Daí o motivo de se evitar o uso de unidades territoriais pequenas, como divisão por bairros ou municípios (o que abrange grande parte dos dados epidemiológicos e demográficos disponíveis). Neste caso, as populações de tais áreas são pequenas, gerando valores pequenos para E_i. Como consequência, os valores extremos de $\hat{R}_i$ tendem a ocorrer nestas áreas. O erro de interpretação a que isso induz é que aquilo que mais chamará a atenção no mapa, que são os seus valores extremos, será o menos precisamente estimado. Assim, as maiores oscilações do risco relativo, em geral, não estarão associadas ao verdadeiro risco da doença subjacente à população,

mas sim à mera flutuação aleatória. Além disso, esse tipo de estimativa não leva em consideração a possível dependência espacial entre as áreas, presente em muitas situações.

Uma alternativa para lidar com esse problema é a abordagem Bayesiana, através dos modelos hierárquicos bayesianos. A proposta, feita inicialmente por Clayton e Kaldor (1987), é modelar o logaritmo do risco relativo por meio de uma regressão linear. A ideia dos autores era impor uma estrutura de relação espacial plausível entre as áreas, por meio da modelagem conjunta dos riscos como um processo espacial. Em outras palavras, usar a informação das áreas vizinhas para estimar o risco relativo de uma unidade territorial. Isso também pode ser visto no trabalho de Assunção e Castro (2004), que estimaram o risco para os seis tipos de câncer mais comuns em homens e mulheres, em 18 cidades brasileiras do estado de São Paulo no ano de 1991, por meio da SIR (*Standardized Incidence Rate*) - estimativa similar à SMR, porém, referente à incidência da doença, e não à mortalidade - e alternativamente, através de um modelo bayesiano multivariado. Seus resultados mostraram que as estimativas obtidas pelas taxas de incidência padronizadas indiretas usuais tinham intervalos de confiança muito grandes para muitos tipos de câncer e cidades, devido ao pequeno número de casos esperados. O uso do método bayesiano levou a estimativas mais precisas.

Justificada a importância do método bayesiano em relação ao clássico, considere a seguir uma possível estrutura para um modelo nesse contexto.

2.2 Modelo Hierárquico Bayesiano para dados de área

Como já mencionado, o modelo clássico de riscos relativos, que usa a SMR como estimador do risco, assume uma densidade Poisson com risco de mortalidade constante sobre as áreas e independentes entre delas. Na prática, porém, Congdon (2007) alerta que os riscos variam tanto dentro como entre as áreas, de modo que as contagens nas áreas tem mais variabilidade do que a densidade que a Poisson estipula, o que é conhecido como variabilidade extra-Poisson.

Uma alternativa para contornar tal problema seria modelar as contagens de acordo com uma distribuição Binomial negativa, visto que ela apresenta um parâmetro adicional, chamado parâmetro de heterogeneidade ou superdispersão (Hilbe (2011)). Além disso, a distribuição Binomial negativa generaliza a distribuição de Poisson quando esse parâmetro tende a zero.

Outra possibilidade para modelar a variação extra é incluir efeitos aleatórios no modelo para o risco relativo de doença ou mortalidade. Tais efeitos podem ou não ser estruturados espacialmente, sendo que os últimos têm sido denotados como “excesso de heterogeneidade”, segundo Best *et al.* (1999). Os autores ainda afirmam que, por outro lado, também pode ocorrer sobredispersão devido a efeitos espacialmente correlacionados, uma vez que tais efeitos espaciais frequentemente procuram capturar fatores de risco não observados, os quais variam suavemente no espaço.

Mesmo assim, o contínuo uso de tal abordagem em suas mais diversas variações por inúmeros pesquisadores em todo o mundo revela que ainda constitui-se um bom método de análise, além de evidentemente ser mais realística que a abordagem clássica considerada anteriormente,

devido aos motivos já mencionados. Dentre trabalhos interessantes neste contexto encontra-se o de Waller *et al.* (1997), que estende os modelos hierárquicos espaciais para explicar os efeitos temporais e interações espaço-temporais, e ilustra a abordagem usando um conjunto de dados de taxas de câncer de pulmão em Ohio, EUA. No mesmo contexto está o artigo de Xia *et al.* (1997), que relaciona a incidência de determinada doença com variáveis sócio-demográficas. Song *et al.* (2006) consideram um modelo bayesiano espacial hierárquico para estimar taxas de acidentes no Texas, EUA. Isso apenas para citar alguns trabalhos de destaque nessa área de pesquisa tão ampla e crescente.

Na modelagem Bayesiana os parâmetros de um modelo seguem distribuições. Tais distribuições controlam sua forma e são especificadas pelo pesquisador baseado, geralmente, nas suas crenças a priori sobre seu comportamento. A ideia de que os valores dos parâmetros ocorrem a partir de distribuições (a priori) leva naturalmente ao uso de modelos nos quais os parâmetros surgem dentro de hierarquias - os Modelos Hierárquicos Bayesianos. O princípio nesses modelos é dividir a especificação da distribuição a priori em estágios. Além de facilitar a especificação, essa abordagem é natural em determinadas situações experimentais.

Seja φ o parâmetro de interesse, e ϑ os valores dos *hiperparâmetros* - denominação que se dá aos parâmetros pertencentes à distribuição a priori especificada para φ. Como a distribuição a priori de φ depende dos valores de ϑ, é possível especificar $p(\varphi|\vartheta)$ em vez de $p(\varphi)$. Além disso, ao invés de fixar valores para os hiperparâmetros, é possível especificar uma distribuição a priori $p(\vartheta)$ para eles, completando a especificação do segundo estágio da hierarquia. E então, a distribuição a priori marginal de φ pode ser obtida por integração

$$p(\varphi) = \int p(\varphi, \vartheta) d\vartheta = \int p(\varphi|\vartheta) p(\vartheta) d\vartheta.$$

Voltando ao contexto dos dados de interesse, no mapeamento de doenças o modelo mais comumente utilizado para dados de contagem em pequenas áreas é o modelo de Poisson. Segundo Lawson (2008), este modelo é apropriado quando existe uma contagem relativamente baixa da doença e a população é relativamente grande nas áreas. A contagem da doença Y_i nas $i = 1, ..., n$ áreas é assumida como tendo uma média λ_i e sendo independentemente distribuída como

$$Y_i|R_i \sim Poisson\left(\lambda_i\right),$$

em que $\lambda_i = E_i R_i$.

Assim como anteriormente, a média é considerada consistindo em dois componentes: *i)* um componente representando o efeito da população (valores esperados), e *ii)* um componente representando o excesso de risco na área (risco relativo). O cálculo dos valores esperados se dá como explicado na seção anterior. Assim, os dados são independentemente distribuídos com esperança

$$E(Y_i|R_i) = \lambda_i = E_i R_i$$

em que E_i é o valor esperado para a i-ésima área, e R_i o respectivo risco relativo. Como o interesse é desenvolver um modelo bayesiano hierárquico, Y_i é considerado independente dado o conhecimento de R_i.

A abordagem mais comum para a modelagem do risco relativo é assumir função de ligação logarítmica para o preditor linear, isto é,

$$\log(R_i) = \eta_i.$$

Diferentes especificações para η_i podem ser adotadas. Definir efeitos aleatórios com distribuições a priori Gama ou Beta para o risco relativo pode ser útil, mas têm uma série de inconvenientes. Primeiro, a distribuição Gama não permite que se obtenha facilmente adaptações para a inclusão de covariáveis no modelo, e, segundo, não há generalização simples e adaptável de tal distribuição para parâmetros espacialmente correlacionados. Best *et al.* (2005) fornecem um exemplo do uso de modelos Gama correlacionados, mas esses modelos mostraram ter um desempenho ruim num estudo de simulação. Além disso, de acordo com Lawson (2008), as vantagens de incorporar uma especificação Gaussiana são muitas. A principal é que um efeito aleatório com distribuição Gaussiana se comporta de maneira similar a um com distribuição Gama, mas o modelo gaussiano pode incluir uma estrutura de correlação. Assim, para o caso em que suspeita-se que os efeitos aleatórios são correlacionados espacialmente (o que é bastante razoável), o mais comum é especificar uma forma Gaussiana para qualquer variação extra presente. Uma alternativa é considerar componentes aditivos descrevendo diferentes aspectos da variação que se presume haver nos dados, e atribuir a um desses componentes distribuição a priori Gaussiana. Enfim, existem muitas maneiras de se incorporar tal heterogeneidade no modelo, e uma dessas é apresentada a seguir.

Besag *et al.* (1991) primeiramente sugeriram a seguinte forma para η_i

$$\eta_i = \alpha + \boldsymbol{x}_i^{'}\boldsymbol{\beta} + \phi_i$$

em que α é um termo comum a todas as áreas, $\boldsymbol{x}_i^{'} = (x_{i1}, ..., x_{ip})$ é um conjunto de p covariáveis associadas à y_i, e ϕ_i é o efeito aleatório estruturalmente espacial, que tem por finalidade capturar a dependência espacial da região e a variabilidade devido à ausência de algum fator de risco no modelo.

A inclusão do termo de intercepto α no modelo exige que seja atribuída a este distribuição a priori Uniforme na reta real, conhecida como distribuição *flat* (Thomas *et al.* (2004)). Para o vetor de parâmetros de regressão $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ atribuí-se, em geral, distribuição a priori Normal Multivariada com baixa precisão e vetor de médias $\mathbf{0}$. Para modelar os efeitos aleatórios ϕ_i, geralmente

usa-se como distribuição a priori a classe de modelos condicionais autorregressivos (Conditional Autoregressive - CAR, em inglês). Em praticamente todas as vezes, o seu uso como priori requer o uso de métodos numéricos, incluindo técnicas de Monte Carlo via Cadeia de Markov (ou do inglês Monte Carlo via Markov Chain - MCMC) para obter amostras da distribuição a posteriori. Apenas em poucos casos particulares é possível encontrar distribuições a posteriori conhecidas (por exemplo, quando $\mathbf{Y}$ é gaussiano). Atualmente, vários modelos considerando o modelo CAR como distribuição a priori foram implementados e estão disponíveis em programas computacionais populares, tais como OpenBUGS (e seu módulo GeoBUGS) [Lunn *et al.* (2000)], e BayesX, [Lang e Brezger (2000)].

O modelo CAR (Besag (1974)), definido para modelar os efeitos aleatórios estruturados, é dado por

$$\phi_i|\phi_{-i} \sim N\left(\mu_i + \rho\sum_{j\sim i} c_{ij}(\phi_j - \mu_j), \sigma^2 m_{ii}\right), \tag{2.1}$$

em que $\phi_{-i} = (\phi_1, ...\phi_{i-1}, \phi_{i+1}, ..., \phi_n)$ é o vetor de efeitos aleatórios sem o elemento ϕ_i, e $\sigma^2 > 0$ é um termo de variância. $\boldsymbol{C} = (c_{ij})$, é uma matriz de associação espacial com zeros na diagonal (isto é, $c_{ii} = 0$), $\boldsymbol{M} = (m_{ii})$ é uma matriz diagonal conhecida; e ρ é um parâmetro que mede a força da dependência espacial de ϕ_i nos seus vizinhos. C_{ij} e M_{ij} são definidas de modo que a matriz $(\boldsymbol{I} - \rho\boldsymbol{C})^{-1}\boldsymbol{M}$ seja simétrica e positiva-definida, tornando o modelo válido. Observe que $(\boldsymbol{I} - \rho\boldsymbol{C})^{-1}M$ é simétrica somente se $c_{ij}m_{jj} = c_{ji}m_{ii}$, $i, j = 1, ..., n$. E para que esta matriz seja positiva-definida, o parâmetro ρ deve pertencer ao intervalo (ρ_{min}, ρ_{max}), em que $1/\rho_{min}$ e $1/\rho_{max}$ são o menor e maior autovalores da matriz $\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{C}\boldsymbol{M}^{\frac{1}{2}}$. Mais detalhes sobre a especificação deste modelo são apresentados no próximo Capítulo.

Resumindo, o modelo hierárquico bayesiano definido é tal que

$$\begin{aligned}
Y_i|\lambda_i &\sim Poisson(\lambda_i) \\
\lambda_i &= E_i R_i \\
log(R_i) &= \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + \phi_i \\
\alpha &\sim U(-\infty, +\infty) \\
\boldsymbol{\beta} &\sim N(0, \sigma_\beta^2) \\
\phi_i &\sim CAR(\sigma^2).
\end{aligned} \tag{2.2}$$

Dada a importância da modelagem dos efeitos aleatórios estruturados, no Capítulo 3 são apresentados mais detalhadamente o modelo Condicional Autorregressivo Intrínseco (ICAR), e o modelo de convolução, que assume priori ICAR para um de seus efeitos aleatórios. Também é abordado o CAR multivariado, que acomoda a especificação do CAR para o caso multivariado, e é utilizado nos dados de aplicação deste trabalho mais à frente, no Capítulo 5.

2.3 Modelo Hierárquico Bayesiano Dinâmico para dados de área

Suponha que, além de analisar a ocorrência dos riscos relativos no espaço, também exista interesse em compreender a sua dinâmica ao longo do tempo. Talvez a ideia mais intuitiva seja modelar os riscos como uma série temporal, cujo comportamento futuro é analisado com base em um conjunto de informações já existentes. Um dos principais objetivos de uma análise de séries temporais é o entendimento de seu mecanismo gerador, e também a predição para tempos futuros. O conhecimento sobre o mecanismo de geração da série possibilita uma melhor descrição destas, enquanto que a previsão contribui para a tomada de decisões.

Uma possibilidade de modelagem para uma série temporal, seguindo o enfoque bayesiano, são os modelos dinâmicos lineares generalizados (MDLG's), propostos por West *et al.* (1985) como uma generalização dos modelos dinâmicos lineares (MDL's), introduzidos por Harrison e Stevens (1976). Ambos estão bem documentados em West e Harrison (1997).

Os MDL's, também conhecidos como modelos de espaço de estados, tem por objetivo analisar uma variável latente com base em uma variável observável que segue distribuição Normal. No caso dos MDLG's a ideia é a mesma, com a vantagem de que a variável resposta não precisa ser normalmente distribuída, mas apenas pertencer à família exponencial de distribuições. Um dos modelos dinâmicos lineares generalizados mais simples é o passeio aleatório de primeira ordem. O princípio básico que rege tal modelo é a flutuação aleatória dos valores da série temporal em torno de um ponto médio, sendo este também sujeito a variações ao longo do tempo. Basicamente, atribui-se às observações $\{y_t|\lambda_t\}$ uma distribuição pertencente à família exponencial. Então, para $t = 1,2,...,T$, λ_t segue um passeio aleatório tal que

$$\begin{aligned} \lambda_t &= \lambda_{t-1} + \omega_t, \\ \text{com} \quad \omega_t &\sim N(0, \sigma_\omega). \end{aligned} \tag{2.3}$$

Apesar de ser relativamente simples, esse modelo incorpora o conceito de evolução temporal de forma que a média possa variar ao longo do tempo, o que o torna bastante atrativo.

Agora, suponha que a variável Y seja observada no tempo e no espaço, e que para $i = 1,2,...,n$ e $t = 1,2,...,T$

$$Y_{it}|\lambda_{it} \sim Poisson(\lambda_{it}). \tag{2.4}$$

Sabe-se que a distribuição de Poisson pertence à família exponencial, portanto, a modelagem proposta para a média em (2.3) é válida. Neste caso, porém, o modelo é hierárquico, sendo que

$$\lambda_{it} = E_{it} R_{it}, \tag{2.5}$$

com E_{it} representando o valor esperado para a área i no tempo t, e R_{it} o respectivo risco relativo.

Aplicando o log nessa expressão, que é função de ligação natural para o modelo de Poisson, obtém-se

$$\log(\lambda_{it}) = \log(E_{it}) + \log(R_{it}). \tag{2.6}$$

Como os valores para E_{it} são conhecidos, a modelagem da média acerca do $\log(\lambda_{it})$ se resume à especificação do $\log(R_{it})$. Utilizando a hierarquia do modelo proposto em (2.2), e com a adição de um parâmetro para capturar a dinâmica temporal da média, os riscos relativos definidos são tais que

$$\log(R_{it}) = \alpha + x_{it}^{'}\boldsymbol{\beta} + \theta_t + \phi_i, \tag{2.7}$$

no qual tanto os parâmetros α como β também poderiam variar no tempo. No entanto, a inclusão do domínio temporal em tais parâmetros tem se mostrado pouco vantajosa no sentido de custo computacional e de não apresentar significância no modelo. Além disso, nem sempre faz sentido que estes variem no tempo.

A distribuição especificada para θ_t, de modo similar à (2.3), possui estrutura dinâmica dada por

$$\begin{aligned} \theta_t &= \theta_{t-1} + \omega_t, \\ \omega_t &\sim N(0, \sigma_\omega). \end{aligned} \tag{2.8}$$

Em outras palavras, o nível da série é modelado como um passeio aleatório, no qual o valor inicial é

$$\theta_0 \sim N(0, \sigma_\omega). \tag{2.9}$$

Usando termos dos modelos dinâmicos lineares, a equação 2.7 é conhecida como *equação de observação*, a equação 2.8 como *equação do sistema*, e θ_t como o *estado*. Para completar a modelagem do $\log(R_{it})$, especifica-se uma distribuição a priori para os efeitos aleatórios estruturados espacialmente ϕ_i. Como já mencionado, essa distribuição pertence à classe dos Modelos Condicionais Autorregressivos, apresentada com mais detalhes no próximo Capítulo. Desta forma está

especificado um modelo espaço-temporal para análise de dados de área, cujos resultados de uma aplicação em dados reais pode ser vista mais a frente, no Capítulo 6.

No próximo Capítulo, apresenta-se uma introdução sobre Campos Aleatórios Markovianos Gaussianos, utilizados no desenvolvimento do CAR, bem como aspectos da formulação condicional para modelos Gaussianos espaciais, de modo a tornar válida a sua estrutura. Em seguida, considera-se algumas formas de especificação para um modelo Condicional Autorregressivo.

Capítulo 3

Campos Aleatórios Markovianos Gaussianos (CAMG) e os Modelos CAR

Seja $D \subset R^2$ a região geográfica em estudo e $s_1, s_2, ..., s_n \in D$ as n áreas amostrais sobre as quais é observada a variável aleatória $y(s)$. É possível escrever

$$y(s_i) = \mu(s_i) + \phi(s_i), \quad i = 1, 2, ..., n, \tag{3.1}$$

com $\boldsymbol{\mu} = (\mu(s_1), \mu(s_2), ..., \mu(s_n))$ representando as médias gerais que podem ou não depender dos locais de observação s_i, para $i = 1, ..., n$, e $\boldsymbol{\phi} = (\phi(s_1), \phi(s_2), ..., \phi(s_n))$ os erros aleatórios, isto é, o componente estocástico do modelo.

Se $\boldsymbol{\mu}$ for modelado como num modelo de regressão linear (simples ou múltipla), explicada por uma ou mais covariáveis, então assume-se que os erros são independentes, caso em que não existe autocorrelação espacial. Por outro lado, se o componente estocástico apresenta uma estrutura espacial (que é o caso de interesse neste estudo), então não é possível assumir independência entre tais erros, e torna-se necessário definir no modelo uma estrutura que acomode essa dependência espacial. Neste caso, uma alternativa é utilizar os Campos Aleatórios Markovianos (CAM) (Mollié (1996)) para definir a distribuição a priori para os efeitos, assunto que é tratado a seguir. Na sequência são apresentados alguns aspectos da modelagem condicional e sua relação com os modelos CAR.

3.1 Campos Aleatórios Markovianos Gaussianos

A principal solução para o problema de se determinar a existência e especificação da distribuição conjunta associada com as distribuições condicionais foi encontrada na década de 70 por Hammersley e Clifford. Eles descobriram uma ligação fundamental entre o problema teórico da especificação de uma distribuição via suas condicionais e os campos aleatórios de Markov, embora eles mesmos não tenham publicado a prova de seu teorema, que só veio a ser conhecida no meio

estatístico através de Besag (1974). Para compreender a demonstração, no entanto, é necessário conhecimento sobre conceitos da teoria dos grafos.

Os campos aleatórios de Markov são uma generalização das cadeias de Markov, substituindo o espaço-índice do tempo por um espaço mais genérico, como o espaço geográfico. Trata-se de um conjunto de variáveis aleatórias contendo a propriedade de Markov, que está associada com distribuições condicionais. Assim, em Estatística Espacial, essa propriedade implica que a distribuição condicional de uma área (em qualquer estado) dado todo o restante do mapa depende apenas de suas áreas vizinhas, e não dos valores de áreas mais distantes.

A distribuição mais utilizada sobre os CAMs para $\boldsymbol{\phi}$ é a Normal, levando aos Campos Aleatórios Markovianos Gaussianos (CAMG), que tem a seguinte notação:

$$\boldsymbol{\phi} \sim CAMG(\boldsymbol{\mu}, \boldsymbol{P}), \tag{3.2}$$

em que $\boldsymbol{\mu}$ representa um vetor de médias $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ e $\boldsymbol{P}$ uma matriz de precisões, tal que

$$p(\boldsymbol{\phi}) \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\mu})' \boldsymbol{P} (\boldsymbol{\phi} - \boldsymbol{\mu}). \right\} \tag{3.3}$$

Os modelos propostos por Besag *et al.* (1991) são um caso particular deste modelo, obtidos fazendo $\boldsymbol{P} = \boldsymbol{M}$, com

$$M_{ij} = \begin{cases} m_i, & i = j \\ -w_{ij}, & i \sim j \\ 0, & \text{caso contrário} \end{cases}$$

em que $i \sim j$ denota que as áreas s_i e s_j são vizinhas, para $i, j = 1, ..., n$; m_i é o número de vizinhos da região i e $w_{ij} > 0$ é uma medida de similaridade entre s_i e s_j. Uma das escolhas mais comuns para w_{ij} é baseado em fronteiras, fazendo $w_{ij} = 1$ se s_i faz fronteira com s_j e $w_{ij} = 0$ caso contrário. Outro critério bastante utilizado para w_{ij} é o inverso da distância entre os centroides das áreas s_i e s_j.

Observe que, como a matriz $\boldsymbol{M}$ especificada é singular, sua inversa, a matriz de covariâncias, não existe. Em resultado disso, a distribuição conjunta de $\boldsymbol{\phi}$ não é própria nestes modelos. Nos tópicos seguintes considera-se como lidar com este problema.

Antes de introduzir os modelos condicionais autorregressivos de Besag *et al.* (1991), porém, são abordados alguns aspectos importantes da modelagem condicional.

3.1.1 Modelos Gaussianos Espaciais especificados condicionalmente

Como anteriormente, assuma que $\boldsymbol{\phi}(s) : s \in D$, com $\phi(s_i)$; $i = 1,...,n$, representa o vetor de efeitos espaciais definido sobre a região geográfica em estudo. Utilizando a notação de Cressie (1993), seja $\boldsymbol{\varepsilon} \sim NM(\mathbf{0}, \boldsymbol{\Lambda})$ uma distribuição conjunta (n-dimensional) com média $\mathbf{0}$ e matriz de covariâncias diagonal $\boldsymbol{\Lambda}$ (por exemplo, $\boldsymbol{\Lambda} = \sigma^2 I$), sendo que os elementos de $\boldsymbol{\varepsilon}$ também são indexados de acordo com suas localizações $\{s_i : i = 1,...,n\}$.

Seja $\boldsymbol{B} = (b_{ij})$ a matriz que acomoda a dependência espacial. É possível afirmar, mesmo através de um pensamento intuitivo, que, se existe a crença de que $\phi(s_1)$ é correlacionado espacialmente com $\phi(s_2)$, por exemplo, então o elemento $b_{12} > 0$, mas, se esses efeitos são pensados como sendo independentes no espaço, então $b_{12} = 0$. Além disso, assume-se que $b_{ii} = 0$ para $i = 1,...,n$, e que $(\boldsymbol{I} - \boldsymbol{B})^{-1}$ existe (Ripley (2005)). Não é um requisito para o modelo que $b_{ij} = b_{ji}$.

Então, uma maneira de definir $\boldsymbol{\phi} = (\phi(s_1),...,\phi(s_n))$, através de uma especificação simultânea, seria utilizar o fato de que

$$(\boldsymbol{I} - \boldsymbol{B})(\boldsymbol{\phi} - \boldsymbol{\mu}) = \boldsymbol{\varepsilon}. \tag{3.4}$$

É evidente que $E(\boldsymbol{\phi}) = \boldsymbol{\mu}$ e $var(\boldsymbol{\phi}) = E[(\boldsymbol{\phi} - \boldsymbol{\mu})(\boldsymbol{\phi} - \boldsymbol{\mu})'] = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B}')^{-1}$. Observe que, como $(\boldsymbol{\phi} - \boldsymbol{\mu})$ é uma combinação linear de $\boldsymbol{\varepsilon}$, que é Normal multivariado, então

$$\boldsymbol{\phi} \sim NM(\boldsymbol{\mu}, (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B}')^{-1}). \tag{3.5}$$

Com isso, a equação (3.4) pode ser escrita equivalentemente como

$$\phi(s_i) = \mu_i + \sum_{j=1}^{n} b_{ij}(\phi(s_j) - \mu_j) + \varepsilon_i, \quad i = 1,...,n. \tag{3.6}$$

A respectiva verossimilhança é, então,

$$(2\pi)^{-\frac{n}{2}}|\boldsymbol{\Lambda}|^{-\frac{1}{2}}|\boldsymbol{I} - \boldsymbol{B}| \exp\left\{-\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\mu})'(\boldsymbol{I} - \boldsymbol{B}')\boldsymbol{\Lambda}^{-1}(\boldsymbol{I} - \boldsymbol{B})(\boldsymbol{\phi} - \boldsymbol{\mu})\right\}. \tag{3.7}$$

Na abordagem clássica, essa função é maximizada a fim de estimar os parâmetros $\boldsymbol{\mu}$, $\mathbf{B}$ e $\boldsymbol{\Lambda}$. O interesse aqui, no entanto, é obter a distribuição condicional dos efeitos aleatórios espacialmente estruturados, ao passo que no decorrer de todo o trabalho utiliza-se métodos bayesianos para a obtenção das estimativas de interesse.

Apesar de existir a possibilidade de modelar $\boldsymbol{\phi}$ através da especificação simultânea, como já mencionado na seção anterior, foi descoberto que realizações de uma variável aleatória espacial com a propriedade de Markov são mais satisfatoriamente modeladas via abordagem condicional.

Assim, para dados Gaussianos, Cressie (1993) mostrou que a distribuição condicional pode ser escrita como

$$f(\phi(s_i)|\phi(s_j) : j \sim i, j \neq i) = (2\pi\tau_i^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\tau_i^2}\left\{\phi(s_i) - \overline{\phi_i}(\phi(s_j))\right\}^2\right], \quad i = 1, ..., n, \quad (3.8)$$

em que f denota a densidade condicional de $\phi(s_i)$ dado $\left\{\phi(s_j) : j \sim i, j \neq i, j = 1, ..., n\right\}$ e $\overline{\phi_i}$ e τ_i^2 são sua média e variância condicionais, respectivamente. Sob uma condição de regularidade de dependência somente "aos pares" entre as áreas, é possível escrever

$$\overline{\phi_i}(\phi(s_j) : j \sim i, j \neq i) = \mu_i + \sum_{j=1}^{n} c_{ij}(\phi(s_j) - \mu_j), \qquad i = 1, ..., n, \quad (3.9)$$

em que $c_{ij}\tau_j^2 = c_{ji}\tau_i^2$, $c_{ii} = 0$ e $c_{ik} = 0$ se não houver dependência entre os efeitos das áreas i e k.

Será mostrado que, a partir dessa formulação, a distribuição conjunta dos efeitos aleatórios estruturados é tal que

$$\boldsymbol{\phi} \sim NM(\boldsymbol{\mu}, (\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{M}), \quad (3.10)$$

em que $(\boldsymbol{I} - \boldsymbol{C})$ é invertível e $(\boldsymbol{I} - \boldsymbol{C})^{-1}M$ é simétrica e positiva definida. Aqui, $\boldsymbol{\phi} = (\phi(s_1), ..., \phi(s_n))'$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$, $\boldsymbol{C} = (c_{ij})$ é uma matriz $n \times n$ na qual o (i, j)-ésimo elemento é c_{ij}, e $\boldsymbol{M} = diag(\tau_1^2, ..., \tau_n^2)$ é uma matriz diagonal também $n \times n$.

Apenas para constar, a verossimilhança passa a ser

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

A matriz de variâncias em (3.10) não é a mesma de (3.5). É claro que, quando

$$(\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B}')^{-1}$$

elas produzem o mesmo modelo, mas é evidente que $\left\{b_{ij}\right\}$ e $\left\{c_{ij}\right\}$ não podem ter a mesma interpretação.

Para mostrar que a distribuição conjunta dos efeitos aleatórios segue a distribuição em (3.10) e satisfaz as condições mencionadas, antes é enunciado um teorema que deve ser satisfeito por qualquer especificação condicional - o Teorema da Fatoração.

Teorema da Fatoração (Besag (1974)) *Suponha que as variáveis* $\{Z(s_i) : i = 1,...,n\}$ *tem função densidade de probabilidade conjunta Pr(.) cujo suporte* ζ *satisfaz a condição de positividade, isto é, de que* $P(z_i) > 0 \Rightarrow P(z_i,...,z_n) > 0 \; \forall$ *i. Então,*

$$\frac{Pr(z)}{Pr(y)} = \prod_{i=1}^{n} \frac{Pr(z(s_i)|z(s_1),...,z(s_{i-1}),y(s_{i+1}),...,y(s_n))}{Pr(y(s_i)|z(s_1),...,z(s_{i-1}),y(s_{i+1}),...,y(s_n))}, \; z,y \in \zeta, \tag{3.11}$$

em que $y = (y(s_1),...,y(s_n))'$, $z = (z(s_1),...,z(s_n))'$ *são possíveis realizações de Z.*

Prova

Para $y(s_n) \in \zeta_n$,

$$\begin{aligned} Pr(z) &= Pr\left(z(s_n)|\left\{z(s_j) : j \neq n\right\}\right) Pr\left(\left\{z(s_j) : j \neq n\right\}\right) \\ &= \frac{Pr\left(z(s_n)|\left\{z(s_j) : j \neq n\right\}\right) Pr\left(\left\{z(s_j) : j \neq n\right\}, y(s_n)\right)}{Pr\left(y(s_n)|\left\{z(s_j) : j \neq n\right\}\right)}, \end{aligned}$$

Sob a condição de positividade, o denominador desta expressão é positivo. Agora,

$$\begin{aligned} Pr\left(\left\{z(s_j) : j \neq n\right\}, y(s_n)\right) &= Pr\left(z(s_{n-1}|\left\{z(s_i) : i \neq n-1, n\right\}, y(s_n)\right) Pr\left(\left\{z(s_i) : i \neq n-1, n\right\}, y(s_n)\right) \\ &= \frac{Pr\left(z(s_{n-1})|z(s_1),...,z(s_{n-2}),y(s_n)\right) Pr\left(z(s_1),...,z(s_{n-2}),y(s_{n-1}),y(s_n)\right)}{Pr\left(y(s_{n-1})|z(s_1),...,z(s_{n-2}),y(s_n)\right)}, \end{aligned}$$

para algum $y(s_{n-1}) \in \zeta_{n-1}$. Novamente, a condição de positividade é usada para garantir que a última expressão esteja bem definida. Prosseguindo desta maneira, o teorema está provado. ■

Proposição A especificação condicional em (3.8) e (3.9) implicam que

$$\boldsymbol{Z} \sim NM\left(\boldsymbol{\mu}, (\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{M}\right),$$

sendo $(\boldsymbol{I} - \boldsymbol{C})$ invertível e $(\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{M}$ simétrica e positiva-definida.

Prova

Usando o teorema da fatoração para densidades e fazendo $\boldsymbol{y} = \boldsymbol{\mu}$ em (3.11), obtém-se

$$\begin{aligned} log(f(\boldsymbol{z})/f(\boldsymbol{\mu})) &= -\frac{1}{2\tau_i^2}\sum_{i=1}^{n}\left\{z(s_i)-\mu(s_i)-\sum_{j=1}^{i-1}c_{ij}\left(z(s_j)-\mu(s_j)\right)\right\}^2 \\ &+ \frac{1}{2\tau_i^2}\sum_{i=1}^{n}\left\{\sum_{j=1}^{i-1}c_{ij}\left(z(s_j)-\mu(s_j)\right)\right\}^2 \\ &= -\frac{1}{2\tau_i^2}\sum_{i=1}^{n}\left(z(s_i)-\mu(s_i)\right)^2 \\ &+ \frac{1}{\tau_i^2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}c_{ij}\left(z(s_i)-\mu(s_i)\right)\left(z(s_j)-\mu(s_j)\right) \\ &= -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})'\boldsymbol{M}^{-1}(\boldsymbol{I}-\boldsymbol{C})(\boldsymbol{z}-\boldsymbol{\mu}). \end{aligned}$$

O lado direito da equação é o expoente de uma distribuição Gaussiana n-dimensional com média $\boldsymbol{\mu}$ e matriz de variâncias $(\boldsymbol{I}-\boldsymbol{C})^{-1}\boldsymbol{M}$. ■

O teorema da fatoração mostra o quão severas as condições de consistência para probabilidades condicionais podem ser. Uma vez que existem $n!$ maneiras de se ordenar as áreas, existem $n!$ fatorações para $Pr(\boldsymbol{z})/Pr(\boldsymbol{y})$, que devem ser todas iguais.

Toda essa formulação serve de auxílio para a definição das distribuições dos modelos condicionais autorregressivos na próxima Seção.

3.2 Modelos Condicionais Autorregressivos (CAR)

O conceito de modelo condicional autorregressivo para dados de área foi introduzido por Besag (1974), que mostrou que a abordagem de probabilidade condicional para a especificação e análise da interação espacial é mais atraente do que a abordagem de probabilidade conjunta alternativa. A base de seu desenvolvimento vem do modelo de rede de Künsch (1987), que usa a definição de distribuição espacial em termos de diferenças e permite o uso de uma distribuição conjunta Normal singular. Veio a ser explorada mais profundamente pela primeira vez por Besag e Kooperberg (1995), com a abordagem do modelo Condicional Autorregressivo Intrínseco (ICAR, Intrinsic Conditional Autoregressive) para análise Bayesiana de imagens, mostrando as restrições necessárias para se obter distribuição a posteriori própria. Este modelo é apresentado no próximo tópico. Em seguida, é considerado o Modelo de Convolução, que utiliza como distribuição a priori para um de seus efeitos o modelo CAR.

3.2.1 Modelo CAR intrínseco (ICAR)

Utilizando notação semelhante à de Stern e Cressie (2000), num modelo CAR, o vetor $\boldsymbol{\phi}$ dos efeitos aleatório espaciais ϕ_i, $i = 1, ..., n$ segue distribuição Normal multivariada

$$\boldsymbol{\phi} \sim NM\left(\boldsymbol{\mu}, (\boldsymbol{I} - \rho\boldsymbol{C})^{-1}\boldsymbol{M}\right), \tag{3.12}$$

em que $\boldsymbol{C} = \left(c_{ij}\right)$ é uma matriz de associação espacial com zeros na diagonal; ρ é o parâmetro que mede a força da dependência espacial de $\boldsymbol{\phi}$ nos seus vizinhos; $\boldsymbol{I}$ é a matriz Identidade $n \times n$; $\boldsymbol{M}$ é uma matriz diagonal conhecida, escolhida de modo que a matriz de covariâncias $\boldsymbol{\Sigma} = (\boldsymbol{I} - \rho\boldsymbol{C})^{-1}\boldsymbol{M}$ seja positiva-definida.

Sendo $\boldsymbol{M} = (m_{ii})$ e por inspeção na matriz $\boldsymbol{\Sigma}^{-1}$, conclui-se que $\boldsymbol{\Sigma}$ é simétrica quando $m_{jj}C_{ij} = m_{ii}C_{ji}$. Observe também que a matriz de covariâncias pode ser expressa como

$$\boldsymbol{\Sigma} = \boldsymbol{M}^{\frac{1}{2}}\left(\boldsymbol{I} - \rho\boldsymbol{M}^{\frac{-1}{2}}\boldsymbol{C}\boldsymbol{M}^{\frac{1}{2}}\right)^{-1}\boldsymbol{M}^{\frac{1}{2}}. \tag{3.13}$$

Então, ela será definida-positiva quando $\rho \in (\rho_{min}, \rho_{max})$, sendo que $1/\rho_{min}$ e $1/\rho_{max}$ são o menor e maior autovalores de $\boldsymbol{M}^{\frac{-1}{2}}\boldsymbol{C}\boldsymbol{M}^{\frac{1}{2}}$, respectivamente.

É importante ressaltar que a inclusão de ρ no modelo - o parâmetro que mede a força da dependência de $\boldsymbol{\phi}$ nos seus vizinhos - não faz com que se perca a generalidade do resultado em (3.10) e, portanto, a distribuição conjunta de $\boldsymbol{\phi}$ e as distribuições condicionais dos efeitos estão garantidas por este.

Assim, as condicionais completas para o modelo CAR podem ser expressas como

$$\phi_i|\phi_{-i} \sim N\left(\mu_i + \rho\sum_{j\sim i} c_{ij}(\phi_j - \mu_j), \sigma^2 m_{ii}\right), \tag{3.14}$$

em que $j \sim i$ indica que j pertence à vizinhança de i (j faz fronteira com i, se esse for o critério adotado).

A escolha de $\rho = 0$ implica em independência espacial dos efeitos aleatórios, ao passo que ao se escolher $\rho = 1$ admite-se máxima autocorrelação espacial. Esta última opção leva ao modelo CAR intrínseco (ICAR, do inglês Intrinsic Conditional Autoregressive).

Assim, a distribuição a priori ICAR dada por Besag *et al.* (1991) para $\boldsymbol{\phi}$ é

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

em que n_i é o número de vizinhos da área i.

Note que esse modelo é uma variação de (3.14). Neste caso, considera-se que $\mu_i = 0$; $m_{ii} = \frac{1}{n_i}$; $c_{ij} = \frac{1}{n_i}$ se as áreas i e j forem adjacentes e 0 se não forem; e por fim, $\rho = 1$. O fato de $\rho = 1$ é o que leva ao termo "intrínseco" utilizado para se referir ao modelo, pois ele faz com que exista correlação espacial máxima entre os efeitos. Dessa forma, uma crítica associada a este modelo é que ele é adequado apenas quando existe forte autocorrelação espacial. Observe que, nessa formulação, a esperança condicional de ϕ_i é igual a média dos efeitos aleatórios das áreas vizinhas de i, enquanto a variância condicional é inversamente proporcional ao número de vizinhos n_i. O parâmetro de variância σ^2 controla a variação entre os efeitos aleatórios.

Essa é uma distribuição imprópria, com uma média geral indefinida para ϕ_i, uma vez que é possível adicionar uma constante para cada ϕ_i sem alterar a distribuição. Segundo Congdon (2007), isso pode resultar em problemas na convergência e de identificabilidade na estimação Bayesiana baseada em amostras repetidas. Eberly *et al.* (2000) afirmam que uma maneira de se obter distribuição própria é impor ao modelo a restrição de que $\sum_i \phi_i = 0$. Os autores trabalham, ainda, com a relação entre identificabilidade e as taxas de convergência do MCMC, de modo a fornecer orientação sobre a seleção de priori e melhoria no algoritmo. Adicionalmente, Besag e Kooperberg (1995) demonstraram que impondo que a soma dos efeitos aleatórios seja igual a zero, e especificando um intercepto com locação invariante, e priori Uniforme $(-\infty,+\infty)$, que é equivalente a uma nova parametrização do modelo incluindo um intercepto, garante-se a identificabilidade do modelo. No OpenBUGS, as distribuições *"car.normal"* e *"mv.car"* utilizadas para definir o modelo CAR intrínseco univariado e multivariado, respectivamente, são parametrizadas para incluir a restrição de soma a zero sobre os efeitos aleatórios. Isso significa que o usuário deve incluir um termo de intercepto separado no modelo, ao qual deve-se atribuir uma distribuição a priori Uniforme imprópria, usando a distribuição *"dflat()"* definida no programa.

Com respeito à especificação da estrutura da matriz de vizinhanças $\boldsymbol{C}$, apesar de o mais comum ser atribuir pesos normalizados, existem diversas maneiras de se construir tal estrutura, como por exemplo, criando elaborações de pesos como funções do comprimento das fronteiras. Muitos autores criticam a especificação da matriz de adjacências utilizando apenas $0's$ e $1's$ como não sendo consistente no caso em que o número de vizinhos varia (que é o caso da maioria das grades irregulares).

3.2.2 Modelo de Convolução

O modelo de convolução, também proposto por Besag *et al.* (1991), é bastante atrativo do ponto de vista prático. Basicamente, consiste num modelo CAR com dois efeitos aleatórios, um com estrutura espacial e priori ICAR, e outro para capturar a variabilidade dos dados que não tem relação com sua distribuição espacial. O modelo é dado por

$$\phi_i = \theta_i + \psi_i,$$

$$\theta_i|\sigma_\theta^2 \sim N(0,\sigma_\theta^2), \tag{3.16}$$

com

$$\boldsymbol{\psi} = (\psi_1,...,\psi_n)|\boldsymbol{W},\sigma_\psi^2 \sim ICAR(\boldsymbol{W},\sigma_\psi^2).$$

O termo $\boldsymbol{\psi}$ tem priori ICAR descrita em (3.15), na seção anterior. O segundo conjunto de efeitos aleatórios $\boldsymbol{\theta} = (\theta_1,...,\theta_n)$ é independente entre as áreas, e diferentes intensidades de correlação podem ser representadas por variar os tamanhos relativos das duas componentes $(\boldsymbol{\theta},\boldsymbol{\psi})$. A convolução entre as funções densidades de probabilidade para $\boldsymbol{\theta}$ e $\boldsymbol{\psi}$ resulta na densidade dos efeitos aleatórios. Do ponto de vista prático, é bastante atrativo usar dois efeitos aleatórios, sendo que, ao passo que um deles capta a estrutura de autocorrelação da região, o outro permite deter variabilidade oriunda de outras fontes de variação. Por outro lado, de acordo com Eberly *et al.* (2000), o uso de tais efeitos resulta num problema de identificabilidade, pois somente a soma dos erros é identificada pelos dados. Os autores mencionam que esse problema pode ser corrigido por incluir um termo de intercepto no preditor linear, ou por impor que a soma dos efeitos ϕ_i seja igual a zero. Xie e Carlin (2006) exploram medidas de resolver esse tipo de problema de identificabilidade com base em diferenças na precisão e na medida de divergência de Kullback-Leibler.

De acordo com Rodrigues e Assunção (2012), o termo "convolução" está ligado a este modelo porque a densidade conjunta do vetor de efeitos aleatórios $\boldsymbol{\phi}$ é obtida como uma convolução das densidades conjuntas dos vetores dos efeitos $\boldsymbol{\theta}$ e $\boldsymbol{\psi}$. Lembrando que, por definição, em Estatística, "convolução" é a distribuição de probabilidade da função soma de duas variáveis aleatórias.

3.2.3 Modelo CAR Multivariado (MVCAR)

O modelo CAR multivariado é uma extensão multivariada do modelo CAR já apresentado. É uma ótima ferramenta no estudo de morbidades, uma vez que permite modelar várias doenças simultaneamente, além de possibilitar a obtenção dos coeficientes de correlação entre estas, ajudando a elucidar possíveis relações entre óbitos decorrentes de duas ou mais doenças. Este resultado é importante aos sistemas de assistência em saúde, que podem melhor direcionar recursos para a prevenção e tratamento de tais doenças, bem como identificar maneiras de combater fatores de riscos relacionados ao ambiente.

Este modelo tem sido usado amplamente, e dentre os inúmeros trabalhos da literatura que o utilizam, podem ser citados Kramer e Williamson (2013), que utilizou o MVCAR (Multivariate Conditional Autoregressive, em inglês) num modelo espacial bayesiano multivariado cujo interesse era compreender a ocorrência e a relação entre a ocorrência de partos prematuros e doenças cardiovasculares em mulheres na Geórgia. Carlin e Banerjee (2003) utilizaram o CAR multivariado para modelar dados multivariados da área de análise de sobrevivência. Song *et al.* (2006) utilizaram MVCAR para explicar o efeito espacial na modelagem multivariada de taxas de acidentes no Texas, EUA. Assunção e Krainski (2009) fizeram uso do modelo na análise de dados de câncer, entre outros.

Assim como nos modelos CAR univariados, a sua versão multivariada, que pode ser utilizada para modelar efeitos aleatórios, é uma distribuição imprópria. No entanto, de acordo com Xie e Carlin (2006), isso na maioria das vezes não é visto como uma limitação para os bayesianos, uma vez que a distribuição a posteriori para $\boldsymbol{\phi}$ geralmente é própria.

O modelo MVCAR proposto por Gelfand e Vounatsou (2003) para K variáveis, utilizado como distribuição a priori para os efeitos espaciais no caso multivariado, para o modelo

$$\begin{gathered} Y_{ik}|R_{ik} \sim Poisson\left(E_{ik}R_{ik}\right), \\ R_{ik} = \exp\left(\alpha_k + x_{ik}^{'}\beta_k + \phi_{ik}\right), i = 1,...,n, \quad k = 1,...,K, \end{gathered}$$

especifica uma matriz $n \times K$ de efeitos aleatórios $\boldsymbol{\phi}$, definida com a restrição de que os efeitos espaciais, separados em efeitos não espaciais e espacialmente estruturados é especificada como

$$\boldsymbol{\phi} \sim N_{nK}(\mathbf{0}, \boldsymbol{H}_1),$$

em que $\boldsymbol{H_1} = [\boldsymbol{\Lambda} \otimes (\boldsymbol{D} - \rho \boldsymbol{W})]^{-1}$, com $\otimes$ denotando o produto de Kronecker, $\boldsymbol{D}$ uma matriz $n \times n$ diagonal cujos elementos são o número de vizinhos da i-ésima região, e $\boldsymbol{W} = \left(w_{ij}\right)$ é a matriz de adjacências, com $w_{ii} = 0$ e $w_{ij} = 1$ se as áreas i e j são adjacentes, (isto é, $i \sim j$), e 0 em outros casos. Aqui, $\boldsymbol{\Lambda}$ é uma matriz $K \times K$ positiva-definida de precisões não espaciais, definindo a relação entre as doenças, e ρ é um parâmetro comum de autocorrelação. Isso é denotado como o modelo $MVCAR(\rho, \boldsymbol{\Lambda})$. Segundo Lawson (2008), este modelo pode ser estendido para permitir a separação da autocorrelação para cada doença, fazendo

$$\boldsymbol{\phi} \sim N_{nK}(\mathbf{0}, \boldsymbol{H}_2),$$

em que $\boldsymbol{H}_2 = \left[Q(\boldsymbol{\Lambda} \otimes \boldsymbol{I}_{n \times n})\boldsymbol{Q}'\right]^{-1}$ e $\boldsymbol{Q} = diag(R_1,...,R_L)$, sendo $R_l = chol(\boldsymbol{D} - \rho_l \boldsymbol{W}), l = 1,...,L$, na qual $chol()$ denota a decomposição de Cholesky.

Suponha que deseja-se utilizar o MVCAR como distribuição a priori para os efeitos aleatórios de um modelo semelhante ao especificado em (2.2), mas sem considerar a presença de covariáveis, de modo que os riscos relativos sejam tais que

$$\log(R_{ik}) = \alpha_k + \phi_{ik} + \gamma_{ik}, \tag{3.17}$$

em que

$$\boldsymbol{\gamma} \sim NM(\mathbf{0}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\phi} \sim MVCAR(1, \boldsymbol{\Omega}).$$

O primeiro efeito, que não possui estrutura espacial, tem média igual a zero e matriz de covariâncias diagonal $\boldsymbol{\Sigma} = diag(\tau_1, ..., \tau_K)$. Para o segundo termo é assumido um modelo CAR intrínseco usando a distribuição MVCAR. Para Ω assume-se distribuição a priori $Wishart(\boldsymbol{R}, n)$ com matriz de parâmetros $\boldsymbol{R}$. A matriz de covariâncias é obtida, então, como Ω^{-1}. Outras suposições são feitas acerca dos parâmetros do modelo, as quais são distribuição a priori Wishart para as precisões dos efeitos não correlacionados ($\boldsymbol{\Sigma}^{-1}$), e Uniforme (*flat*) para os termos de intercepto α_k. No OpenBUGS a distribuição MVCAR está definida, permitindo ao usuário utilizá-la como priori para os efeitos aleatórios de um modelo multivariado similar ao apresentado. Neste trabalho, utiliza-se o MVCAR como distribuição a priori para os efeitos aleatórios de um modelo para os óbitos decorrentes de cinco tipos de câncer de maior importância quantitativa no estado de São Paulo, para o período de 1998 a 2010, cujos resultados constam no Capítulo 6.

Capítulo 4

Métodos Computacionais Intensivos

Como já mencionado no decorrer deste texto, na maioria dos casos, a distribuição a posteriori do vetor de parâmetros do modelo não pode ser obtida por meio de um método analítico devido à sua complexidade. Nessas situações, é necessário apelar para métodos numéricos que, graças à evolução computacional do último século, se tornaram facilmente disponíveis, implementáveis, e de obtenção de resultados em um período de tempo relativamente curto. O Método de Monte Carlo via Cadeias de Markov (MCMC), especificamente, o amostrador de Gibbs, está implementado no software OpenBUGS, utilizado neste trabalho.

Os métodos de MCMC são uma alternativa aos métodos não iterativos em problemas complexos (nos métodos não iterativos, os valores são gerados de forma independente e não há preocupação com a convergência do algoritmo, bastando que o tamanho da amostra seja suficientemente grande). A ideia é obter uma amostra da distribuição a posteriori e calcular estimativas amostrais de características de interesse desta distribuição. A diferença é que, neste caso, são usadas técnicas de simulação iterativa, baseadas em cadeias de Markov, implicando em que os valores gerados sejam dependentes, diferente do que acontece nos métodos não iterativos . Considere a seguir uma breve descrição sobre dois dos métodos de MCMC.

4.1 Algoritmo de Metropolis-Hastings

Seja $\Theta = (\theta_1, ..., \theta_d)$ o vetor (ou coleção) de parâmetros desconhecidos a estimar no modelo. A ideia básica é simular um passeio aleatório no espaço de Θ que converge para uma distribuição estacionária - a distribuição de interesse. Seguindo este princípio, os algoritmos de Metropolis-Hastings (Metropolis *et al.* (1953), Hastings (1970)) utilizam uma distribuição auxiliar para a geração de uma cadeia de pontos, que são aceitos ou rejeitados com uma determinada probabilidade. Isso garante que a cadeia convirja para uma distribuição de equilíbrio, que neste caso é a distribuição a posteriori de Θ, sobre a qual se tem interesse.

A partir de uma distribuição proposta $q\left(.|\Theta^j\right)$, suponha que a cadeia esteja no estado Θ^j, e que um valor Θ' é gerado a partir dela. Este novo valor é aceito com probabilidade

$$\alpha\left(\Theta^{j},\Theta'\right)=min\left\{1,\frac{\pi(\Theta')q(\Theta^{j}|\Theta')}{\pi(\Theta^{j})q(\Theta'|\Theta^{j})}\right\},$$

em que $\pi(\Theta)$ denota a densidade a posteriori de Θ.

Finalmente, o algoritmo de Metropolis-Hastings pode ser especificado pelos seguintes passos:

1. Inicialize o contador de iterações $t=0$ e especifique um valor inicial Θ^{0};

2. Gere um novo valor Θ' da distribuição $q(.|\Theta^{'})$;

3. Calcule a probabilidade de aceitação $\alpha\left(\Theta^{'},\Theta'\right)$ e gere $u\sim U(0,1)$;

4. Se $u\leq\alpha\left(\Theta^{'},\Theta'\right)$, então aceite o novo valor e faça $\Theta^{t+1}=\Theta'$, caso contrário, rejeite e faça $\Theta^{t+1}=\Theta^{'}$;

5. Incremente o contador de t para $t+1$ e volte ao passo 2.

O algoritmo deve ser executado até a convergência da cadeia. Feito isso, a cadeia resultante, isto é, os pontos gerados, podem ser considerados como uma amostra da distribuição a posteriori.

4.2 Amostrador de Gibbs

O amostrador de Gibbs, popularizado dentro de um contexto de reconstrução de imagens (Geman e Geman (1984)), é um caso especial do algoritmo de Metropolis-Hastings, mas com duas particularidades: todos os pontos gerados são aceitos, e são gerados a partir das distribuições condicionais completas $\pi(\theta_i|\Theta_{-i})$, onde $\Theta_{-i}=\{\theta_1,...,\theta_{i-1},\theta_{i+1},...,\theta_d\}$. Essa distribuição pode ser obtida a partir da distribuição conjunta, fazendo

$$\pi(\theta_i|\Theta_{-i})=\frac{\pi(\Theta)}{\int\pi(\Theta)\,d\theta_i}.$$

Sabe-se que, na maioria das situações, gerar uma amostra diretamente de $\pi(\Theta)$ pode ser difícil ou mesmo impossível. Felizmente, se as distribuições condicionais completas são conhecidas, então pode-se utilizar o amostrador de Gibbs, definido pelo seguinte esquema:

1. Faça $t=0$, $i=1$, e especifique um valor inicial Θ^{0};

2. Gere um novo valor θ_i^{t+1} da distribuição $\pi\left(\theta_i|\Theta^{'}_{-i}\right)$, que é a densidade condicional completa de θ_i;

3. Se $i<d$ faça $i=i+1$ e retorne ao passo 3;

4. Incremente o contador de t para $t+1$ e volte ao passo 2.

Pode-se mostrar que após a convergência, os valores resultantes formam uma amostra de $\pi(\Theta)$.

4.3 Diagnóstico de Convergência

Como visto, a partir da convergência, os pontos gerados da cadeia resultante passam a ser encarados como pontos gerados da distribuição a posteriori de interesse. Assim, são necessários alguns cuidados para verificar a convergência da cadeia. Primeiro, existe um período de aquecimento para a cadeia, denominado *burn-in*. Nesse período, a cadeia ainda não atingiu convergência, e, portanto, os pontos gerados até então devem ser descartados. Para saber quantas iterações são necessárias para o período de aquecimento, utilizam-se diagnósticos de convergência, tais como o de Raftery Lewis e Heidelberger Welch. Além disso, muitas vezes é possível verificar a indicação de convergência por meio de uma inspeção visual dos histogramas e densidades Kernel da estimativa da distribuição a posteriori do parâmetro de interesse, ainda assim é altamente recomendável que se utilize pelo menos um método de diagnóstico.

Outra questão importante é a verificação de uma possível autocorrelação na cadeia, nos pontos obtidos a partir do *burn-in*, pois cadeias autocorrelacionadas podem levar a subestimação da variância do parâmetro. Para corrigir esse problema, observamos o lag da autocorrelação e, a partir deste, selecionamos um ponto a cada k iterações para fazer parte da cadeia. Na próxima Seção, é considerado um pouco sobre o OpenBUGS, que utiliza os algoritmos de Metropolis-Hastings e o Amostrador de Gibbs para a obtenção de amostras da distribuição a posteriori do vetor de parâmetros de interesse, e permite monitoração da convergência através dos métodos citados, a partir do uso do pacote CODA.

4.4 OpenBUGS e CODA

BUGS (Bayesian Using Gibbs Sampler) é um pacote que permite a realização de inferência Bayesiana usando o amostrador de Gibbs. O usuário especifica um modelo estatístico de complexidade arbitrária, simplesmente expondo as relações entre as variáveis relacionadas. O software inclui um "sistema especialista" que determina um algoritmo adequado de MCMC para analisar o modelo especificado. Em seguida, o usuário pode controlar a execução do mecanismo e é livre para escolher entre uma vasta gama de tipos de saída. Considere um pouco sobre sua história e uma de suas ferramentas mais úteis - o pacote CODA.

4.4.1 História do projeto BUGS

O projeto BUGS foi desenvolvido a partir de um trabalho sobre inteligência artificial em 1980. A ideia para seu desenvolvimento surgiu a partir da compreensão de que os métodos de simulação poderiam ser usados para inferência, e do reconhecimento de que a programação orientada a objetos poderia ser explorada para generalizar o algoritmo de simulação. O programa BUGS iniciou-se em 1989 tendo como chefe programador Andrew Thomas, trabalhando com David Spiegelhalter para a Unidade de Bioestatística da MRC (Medical Research Council, em inglês, ou Conselho de Pesquisa Médica), em Cambridge. Coincidentemente, ao mesmo tempo, o relevante trabalho

de Gelfand e Smith (1990) estava sendo realizado em Nottingham, Reino Unido, mas de forma totalmente diferente e de um ponto de partida bem diferente.

Inicialmente, o BUGS só usou algoritmos especializados para a área do ambiente no qual se desenvolveu. Em 1996, no entanto, o projeto mudou-se para o Imperial College, Londres (liderado por Nicky Best, que já estava envolvida no projeto há alguns anos em Cambridge) e a capacidade do software passou a ser expandida. Em particular, Jon Wakefield e Dave Lunn aderiram ao projeto nessa fase, para trabalhar na implementação de modelos não-lineares, e o desenvolvimento de uma versão do software para Windows ganhou impulso. Nos anos seguintes, uma série de outros tipos de modelos desafiadores foram abordados, incluindo modelos espaciais, modelos dinâmicos (envolvendo equações diferenciais) e os modelos de dimensão variável (montados usando o algoritmo *reversible jump*).

Em 2004, Andrew Thomas mudou-se para Helsinki, Finlândia, para começar a trabalhar no OpenBUGS, enquanto Dave Lunn e Nicky Best permaneceram no Imperial College continuando na manutenção e desenvolvimento do WinBUGS. Com isso, os dois pacotes divergiram um pouco, cada um com suas próprias características avançadas não disponíveis no outro. No entanto, agora que o OpenBUGS progrediu de experimental para um pacote estável e confiável, todos os esforços de desenvolvimento estão concentrados sobre ele.

4.4.2 Diferenças entre WinBUGS e OpenBUGS

Ao longo do tempo foram aparecendo inúmeras pequenas diferenças entre o OpenBUGS e o WinBUGS à medida em que foram ampliadas as suas possibilidades de aplicações, como a inclusão de novas distribuições ou a correção no modo de leitura de outras já existentes. Uma diferença fundamental entre os software, porém, é a maneira em que o sistema seleciona o algoritmo de atualização a ser usado para a classe de distribuição condicional completa de cada nó. Enquanto o WinBUGS define um algoritmo para cada classe possível, o OpenBUGS permite ao usuário escolher entre as possibilidades disponíveis em cada caso, permitindo, assim, uma maior flexibilidade e extensibilidade em aplicações. O usuário pode selecionar o atualizador a ser utilizado para cada nó logo após a compilação.

Outro atrativo na diferença entre os programas é que o OpenBUGS pode ser executado de uma forma totalmente interativa a partir do R, através do pacote Brugs do R, permitindo maior manipulação e análise dos resultados obtidos pelo BUGS. Além disso, no módulo para análise de dados geográficos do programa, o GeoBUGS, a consistência do comprimento do vetor de pesos para o CAR com a dimensão dos dados agora é verificada, o que antes seria um problema no WinBUGS, que executaria o código do modelo com sucesso mesmo se o vetor de pesos fosse mais longo.

4.4.3 CODA

CODA (Convergence Diagnostic and Output Analysis) é um software direcionado para a análise de convergência das cadeias geradas via MCMC. É orientado por meio de funções na linguagem do S-Plus (mesma utilizada pelo R), e serve como um processador dos resultados de MCMC do BUGS. Nele estão implementados os principais diagnósticos de convergência, como os já citados Geweke e Gelman e Rubin, além dos de Raftery Lewis e Heidelberger Welch.

Também pode ser usado em conjunto com a saída do MCMC a partir do pacote CODA do R através do comando “read.openbugs”, que lê os resultados do MCMC no formato do CODA produzido pelo OpenBUGS. A partir disso, o usuário se depara com uma grande facilidade em utilizar os diagnósticos de convergência citados na Seção anterior, que já estão incluídos no pacote, bem como na produção de uma variedade de gráficos das amostras a posteriori de parâmetros de interesse do modelo, permitindo uma análise da trajetória da cadeia, e consequentemente, da convergência do algoritmo.

Capítulo 5

Análise Exploratória dos dados de aplicação: óbitos por câncer

Nas últimas décadas, o aumento do câncer se deu de tal modo que converteu-se em um evidente problema de saúde pública mundial. A Organização Mundial da Saúde (OMS) estima que, em 2030, haverá 27 milhões de casos incidentes de câncer no mundo, 17 milhões de mortes, e 75 milhões de pessoas vivas com a doença. No Brasil, o problema ganha relevância pelo seu perfil epidemiológico. Segundo o Instituto Nacional do Câncer (INCA), ao fim de 2012 foram registrados cerca 518.000 casos novos de câncer no país e mais de 50.000 óbitos. O Instituto ressalta ainda que, a prevenção e o controle do câncer precisam adquirir o mesmo foco e a mesma atenção que a área de serviços assistenciais, pois, quando o número de casos novos aumentar rapidamente, não haverá recursos suficientes para suprir as necessidades de diagnóstico, tratamento e acompanhamento. Dessa forma, as consequências poderão ser devastadoras nos aspectos social e econômico. O câncer pode se tornar um grande obstáculo para o desenvolvimento socioeconômico de países emergentes como o Brasil. Em face à dimensão do problema, profissionais de diversas áreas se empenham tanto em sugerir formas de tratamento e diagnóstico, como em fornecer informação de qualidade para subsidiar o conhecimento sobre a ocorrência da doença. No que tange a produzir informação, a Estatística pode valer-se de ferramentas poderosas para que as entidades de saúde pública possam estabelecer prioridades e alocar recursos de forma direcionada, modificando positivamente esse cenário na população brasileira. Como ponto de partida, o Ministério da Saúde fornece, através do Sistema de Informações sobre Mortalidade (SIM) (DATASUS (Visitada em junho/2013)), bases de dados geradas pelos Registros de Câncer de Base Populacional (RCBP). Esses são disponibilizados como dados de área e podem ser óbitos na página do Datasus (www.datasus.gov.br) segundo critérios definidos pelo usuário. Neste trabalho são considerados dados para as microrregiões do estado de São Paulo, para o período de 1998 a 2010, segundo local de residência, para os tipos de câncer de maior importância quantitativa. Neste Capítulo, a análise para os dados se dá com métodos exploratórios, buscando compreender as principais características da mortalidade decorrente dessas doenças no estado.

5.1 Análise Exploratória de dados

No meio Estatístico é um fato de comum aceitação que, o primeiro passo no conhecimento de um conjunto de dados, após a sua coleta, é a análise exploratória destes. Esta técnica, lançada em 1977 por John Wilder Tukey em seu livro Exploratory Data Analysis (sigla EDA, em inglês), emprega grande variedade de técnicas gráficas e quantitativas, visando obter informações sobre a estrutura dos dados, descobrir variáveis importantes em explicar sua variabilidade e tendências, detecção de comportamentos anômalos do fenômeno (*outliers*), testar se são válidas hipóteses assumidas, escolha de modelos ou determinação do número de variáveis a se utilizar.

Quando os dados são georreferenciados, porém, explorar os dados vai além de analisar gráficos e medidas quantitativas, mas inclui principalmente o mapeamento do fenômeno.

Quando o objeto de estudo é uma doença, tal mapeamento consiste em descrever a distribuição espacial, com o objetivo de avaliar a variação geográfica do processo, para identificar fatores de risco, levantar hipóteses sobre possíveis fatores etiológicos, e ainda, sugerir a escolha de modelos apropriados para uma compreensão mais profunda da variável de interesse.

Este mapeamento é informativo à medida que o mapa produzido estiver livre do "ruído" gerado pela flutuação aleatória de pequenas populações, ou de diferenças na estrutura demográfica da região.

A seguir estão dispostos os resultados de uma breve análise exploratória para cada uma das doenças em estudo, que incluem o mapeamento das SMRs para cada doença em cada ano do período em estudo, e uma análise de agrupamentos dessas razões para melhor compreender grupos de microrregiões com comportamento de óbitos semelhantes no decorrer da área em estudo. Num primeiro instante, os dados foram agrupados para todo o período, a fim de fornecer uma ideia geral sobre a mortalidade, mas o comportamento no decorrer do tempo também é explorado em busca de possíveis tendências. O cálculo das SMRs segue o procedimento descrito na Seção 2.1, bem como o cálculo dos valores esperados. Estes foram calculados através de um programa elaborado no R que coleta os dados diretamente do DATASUS, calcula os valores e os armazena em planilhas devidamente identificadas pelo usuário, diminuindo, assim, o risco de erro humano facilmente presente quando se trabalha apenas com planilhas. No Apêndice E encontra-se o mapa do estado de São Paulo, segundo as microrregiões definidas pelo IBGE, para consulta do leitor quanto aos resultados seguintes.

5.1.1 Câncer de traqueia, brônquios e pulmão

Sobre a doença: De acordo com o INCA (Instituto Nacional do Câncer), este é o mais comum de todos os tumores malignos, apresentando aumento de 2% por ano na sua incidência mundial. Em 90% dos casos diagnosticados, o câncer de pulmão está associado ao consumo de derivados de tabaco. No Brasil, foi responsável por 20.622 mortes em 2008, sendo o tipo que mais fez vítimas. Altamente letal, a sobrevida média cumulativa total em cinco anos varia entre 13 e 21% em países desenvolvidos e entre 7 e 10% nos países em desenvolvimento. No fim do século XX, o

câncer de pulmão se tornou uma das principais causas de morte evitáveis. Além disso, evidências na literatura mostram que pessoas que têm câncer de pulmão apresentam risco aumentado para o aparecimento de outros cânceres e que familiares primários de pessoas que tiveram câncer de pulmão apresentam risco levemente aumentado para o desenvolvimento dessa doença. Entretanto, ainda é difícil mesmo para a área médica estabelecer o quanto desse maior risco decorre de fatores hereditários e quanto é por conta do hábito de fumar tanto ativa como passivamente.

A Figura 5.1 apresenta os 5 grupos de microrregiões resultantes da análise de agrupamentos feita no *software* Minitab com base nas SMRs para os 13 anos do período. Para a utilização da técnica, optou-se por escolher como medida de distância a distância euclidiana, e como procedimento de aglomeração o método de Ward. Também chamado de "Mínima Variância", este método utiliza uma distância que leva em conta a diferença dos tamanhos dos conglomerados que estão sendo comparados e, com isso, produz grupos mais homogêneos e com aproximadamente o mesmo número de elementos.

É evidente que, a microrregião de Barretos, por conter um hospital de referência no tratamento de câncer, comportou-se como *outlier*. Isso se explica pelo fato de que, embora o cálculo das SMRs leve em conta a estrutura demográfica da microrregião, a quantidade de óbitos registrados no local não condiz apenas com a população residente, mas também acaba incluindo a população que migra para este local, e reside temporariamente em abrigos fornecidos pelo próprio hospital, mas que, ao falecer em meio ao tratamento, incrementa o registro de óbitos por residência da localidade.

No mapa da Figura 5.2 é possível visualizar a distribuição geográfica dos grupos formados.

Para compreender as particularidades de cada grupo, observe os boxplots da Figura 5.3.

Os grupos predominantes no mapa foram o grupo 1 (vermelho) e o grupo 2 (verde). O grupo 2 refere-se às microrregiões que apresentaram baixos valores para a SMR em todo o período, geralmente abaixo de 1, indicando que a mortalidade em tais regiões foi abaixo do que o esperado. No caso do grupo 1, os valores foram baixos no começo do período, mas sofreram crescimento de 2002 em diante, chegando a ter mortalidade até 3 vezes maior do que o esperado.

Os grupos 3 (azul) e 4 (amarelo) se concentraram no centro e centro-norte da região. Nota-se que o grupo 3, cujos valores para a SMR foram maiores que 1 em praticamente todo o período, encontra-se próximo ao grupo 5 (Barretos), que pode ser visualizado na Figura 4. O grupo 4, por outro lado, apresentou decréscimo dos valores no decorrer do período, além de uma aparente tendência de 4 em 4 anos.

Com respeito à microrregião de Barretos (Figura5.4), é válido chamar a atenção de que de 2008 em diante os valores se mostraram mais baixos, muito provavelmente devido à consolidação do sistema de informação.

Pode-se chamar a atenção para a importância de analisar as microrregiões segundo grupos específicos através dos boxplots da Figura 5.5. Observe que fica difícil detectar padrões ao olhar diretamente para o comportamento geral das SMRs. Nesses boxplots, as SMRs parecem apresentar variação suave no decorrer do período, enquanto que, através da análise de agrupamentos, pode ser visto que isso não é verdade para determinados grupos de microrregiões.

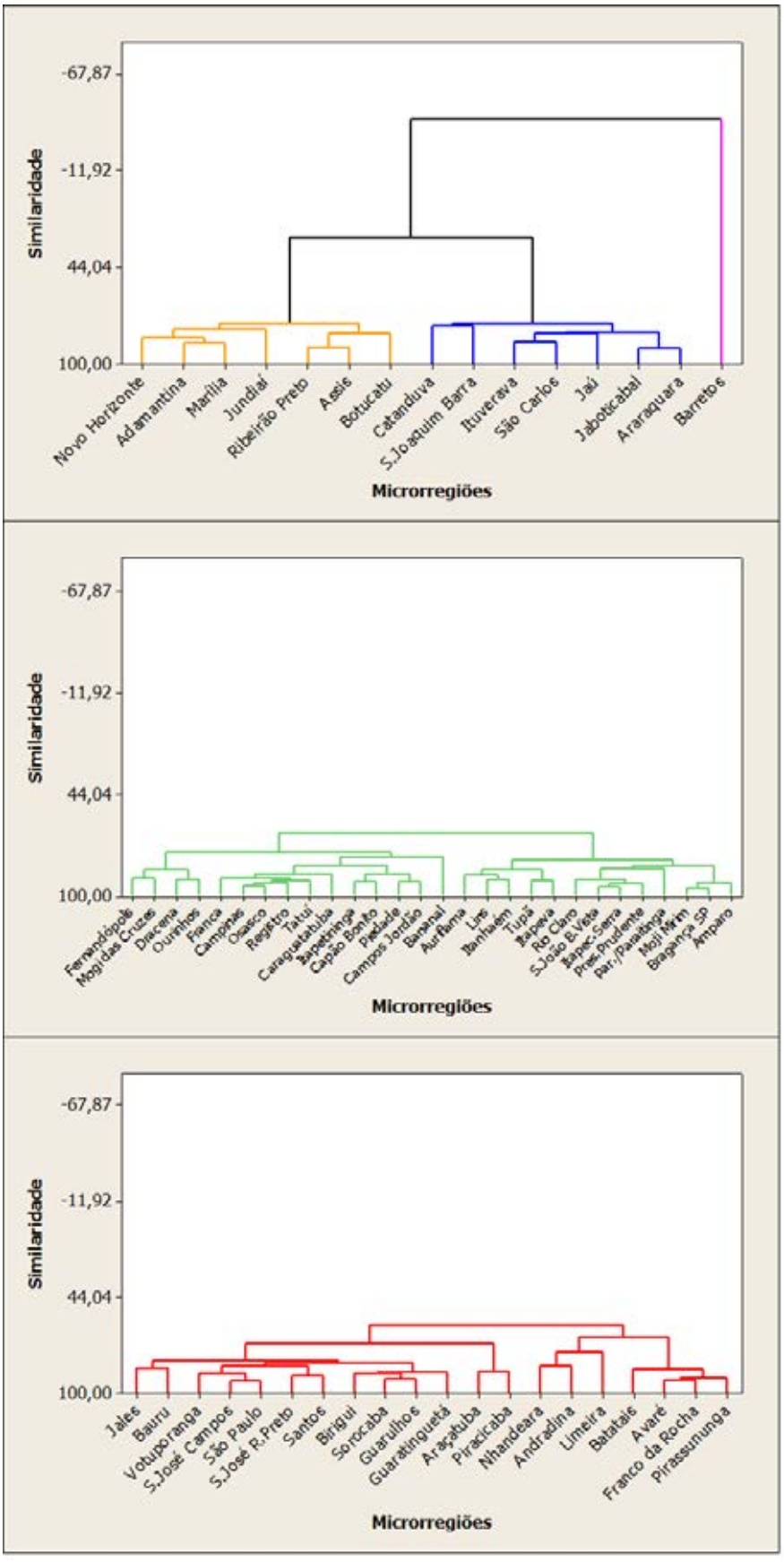

Figura 5.1: *Dendrogramas do agrupamento das SMRs referentes aos óbitos por câncer de traqueia, brônquios e pulmão de 1998 a 2010.*

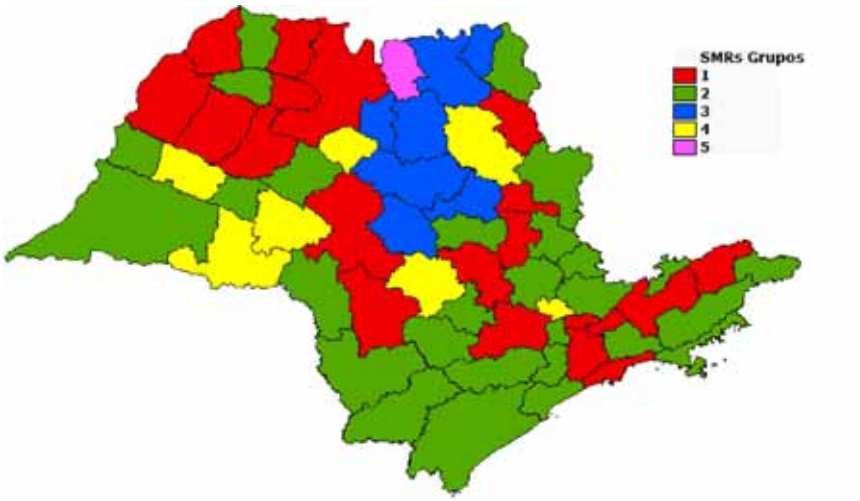

Figura 5.2: *Microrregiões do estado de São Paulo segundo os grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de traqueia, brônquios e pulmão de 1998 a 2010.*

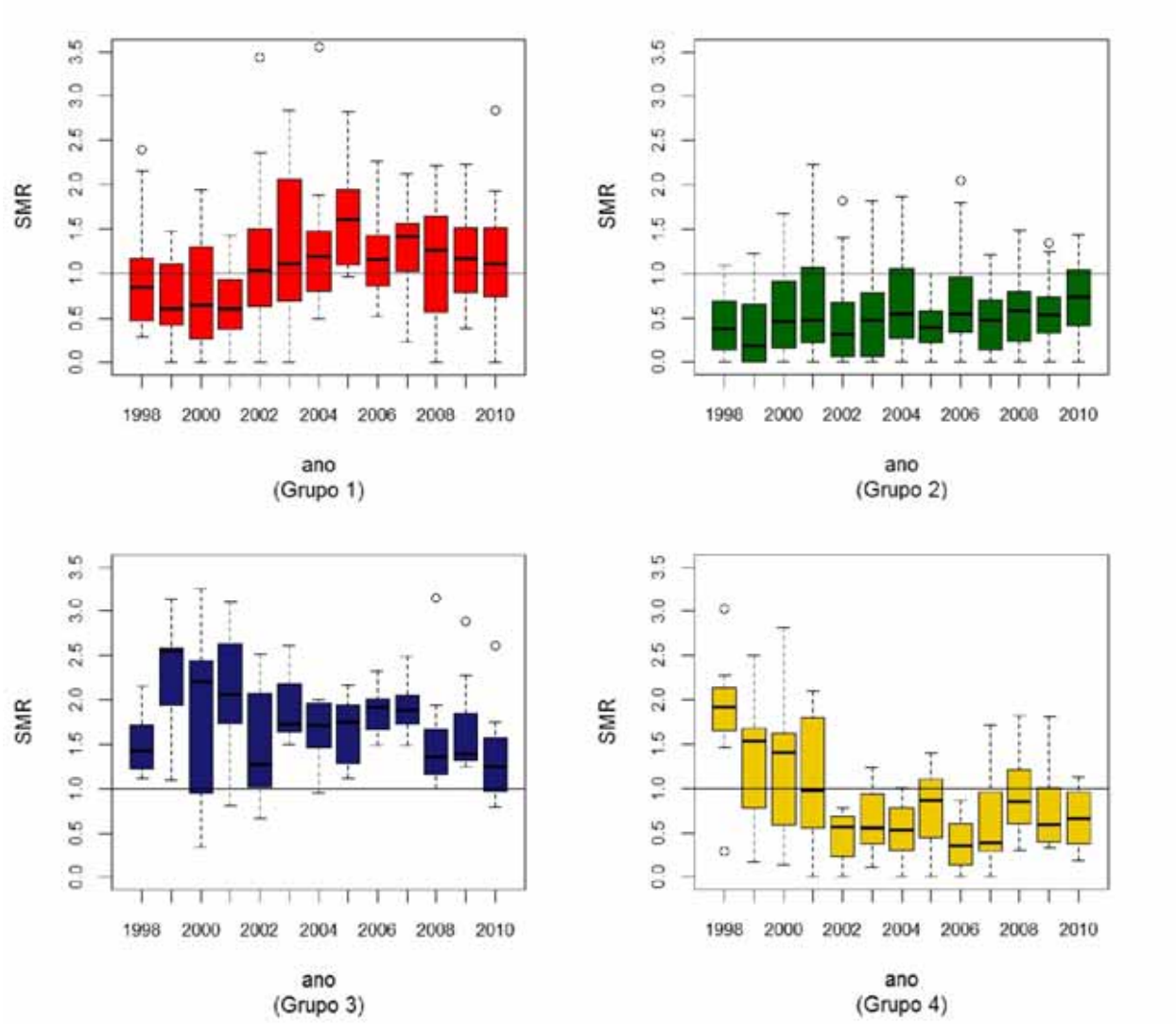

Figura 5.3: *Boxplots dos grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de traqueia, brônquios e pulmão de 1998 a 2010.*

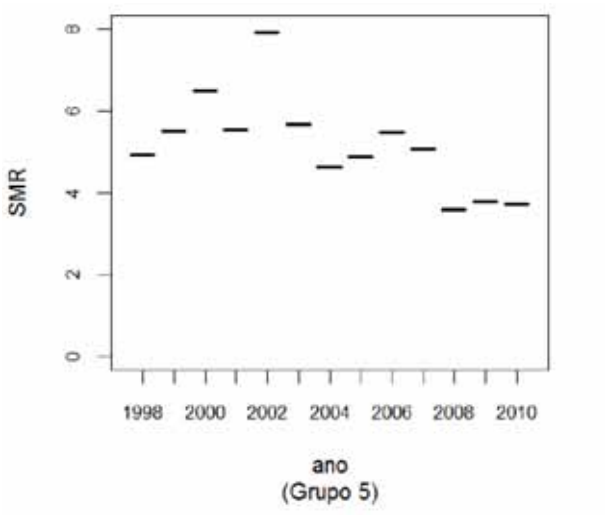

Figura 5.4: *Boxplots do grupo 5 (microrregião de Barretos) da análise de agrupamentos das SMRs referentes aos óbitos por câncer de traqueia, brônquios e pulmão de 1998 a 2010.*

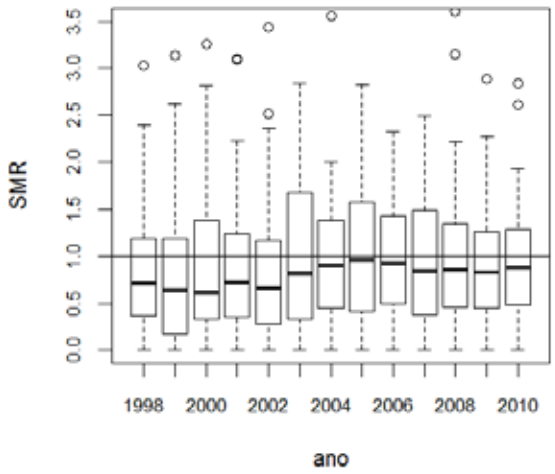

Figura 5.5: *Boxplots da distribuição das SMRs em todas as microrregiões do estado de São Paulo referentes aos óbitos por câncer de traqueia, brônquios e pulmão de 1998 a 2010.*

Na Figura 5.6 pode ser observado o comportamento das SMRs nas microrregiões em cada ano do período. Note que a análise de agrupamentos auxilia muito na compreensão de padrões na mortalidade ao longo do tempo, já que fica difícil observar tendências ou mudanças apenas por observação dos mapas para cada ano do período. Ainda assim, é possível notar que, na maior parte do período, os baixos valores para a SMR se distribuíram por todo o entorno do estado, enquanto as altas taxas, quase sempre estiveram apenas no centro e norte de São Paulo.

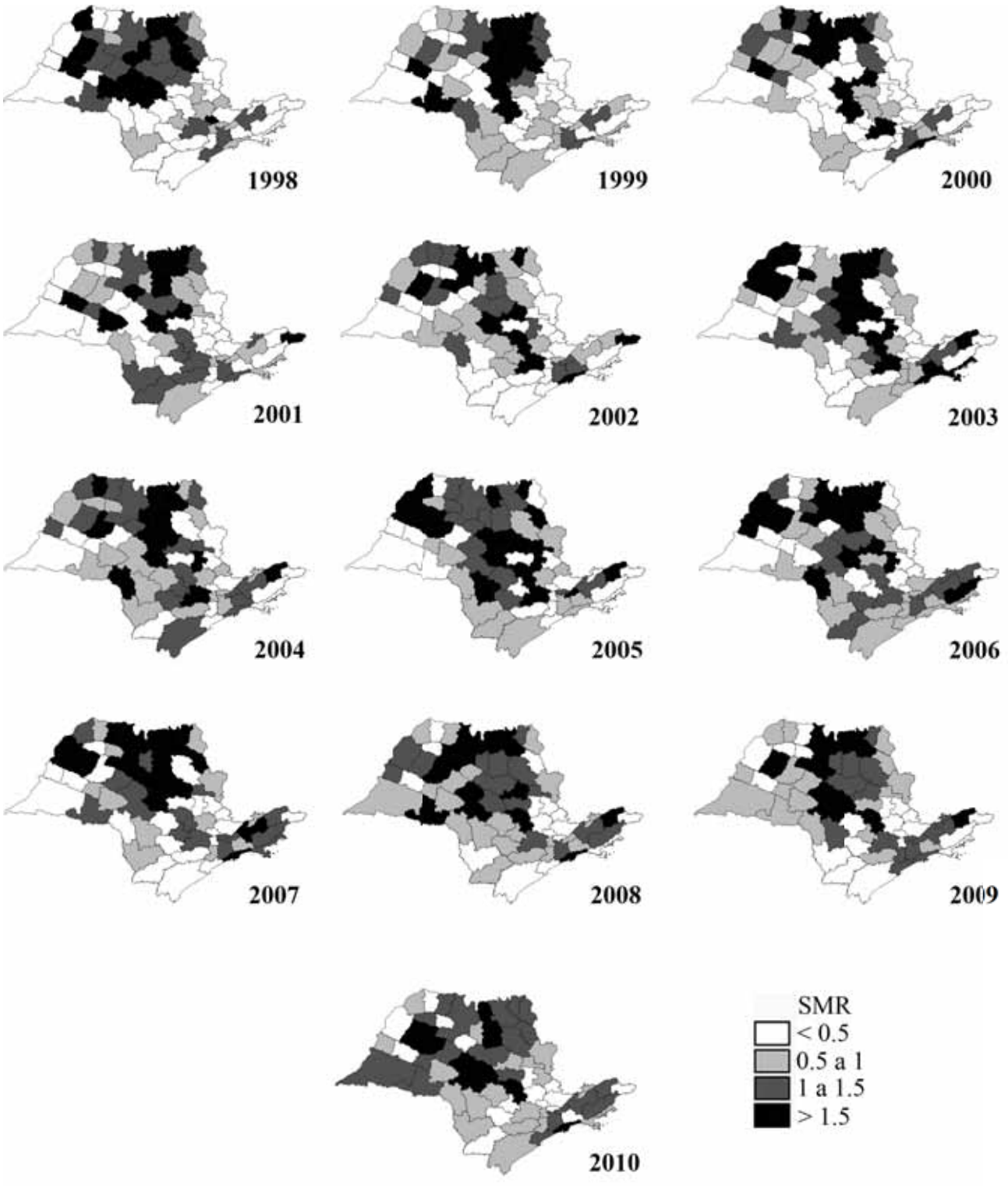

Figura 5.6: *Razões de Mortalidade Padronizadas (SMRs) para a mortalidade por câncer de traqueia, brônquios e pulmão nas microrregiões do estado de São Paulo, de 1998 a 2010.*

5.1.2 Câncer feminino de mama

Sobre a doença: É o câncer mais comum entre as mulheres, e o segundo tipo mais frequente no mundo, correspondendo a 22% dos casos novos a cada ano. Por ser diagnosticado em estágios avançados, pelo menos no Brasil, as taxas de mortalidade por câncer de mama continuam elevadas, e, segundo o INCA, na população mundial, a sobrevida média após cinco anos é de 61%.

Antes de passar para os resultados da análise de agrupamentos, atente para o gráfico da Figura 5.7. Note que a linha demarcando o valor 1 para a SMR passa por todas as caixas, sendo que a maioria delas tem sua maior parte abaixo desta, o que poderia levar erroneamente à conclusão de que a mortalidade por esse câncer foi constante na região de estudo, exceto pela presença de vários outliers, cuja interpretação só seria possível através da observação dos mapas com as SMRs para cada ano do período, relacionando referente outlier à sua microrregião no mapa. Tal maneira de analisar os dados seria tanto mais difícil quanto suscetível a maiores erros de interpretação, dificultando compreender o comportamento real do fenômeno na região em estudo.

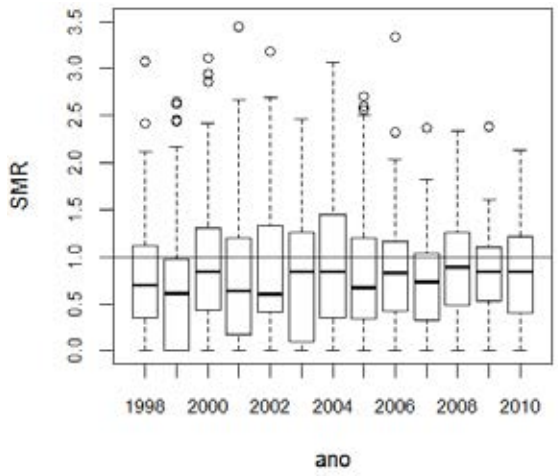

Figura 5.7: *Boxplots da distribuição das SMRs em todas as microrregiões do estado de São Paulo referentes aos óbitos por câncer feminino de mama, de 1998 a 2010.*

A análise de agrupamento das microrregiões para este tipo de câncer resultou na formação de grupos conforme o mapa da Figura 5.8. Na Figura 5.9 consta o respectivo dendrograma.

Analisando os boxplots das SMRs segundo os grupos ao longo do tempo para essa doença (Figura 5.10), é possível fazer algumas considerações. O grupo 1 (vermelho) corresponde às microrregiões cujo comportamento das SMRs oscilou no decorrer do período em estudo, mantendo-se, porém, abaixo de 1 em sua maior parte. O grupo 2 (verde) comportou-se de maneira muito similar ao mesmo na análise para o câncer de traqueia, brônquios e pulmão. Com as menores SMRs, geralmente abaixo de 1, tal grupo registrou menos óbitos do que o esperado para as microrregiões que o compõem. Ressalta-se, ainda, uma evidente semelhança na distribuição espacial deste quando comparado ao do câncer de traqueia, brônquios e pulmão.

O grupo 3 (azul), que distribuiu-se por todo o estado, apresentou os maiores valores para as

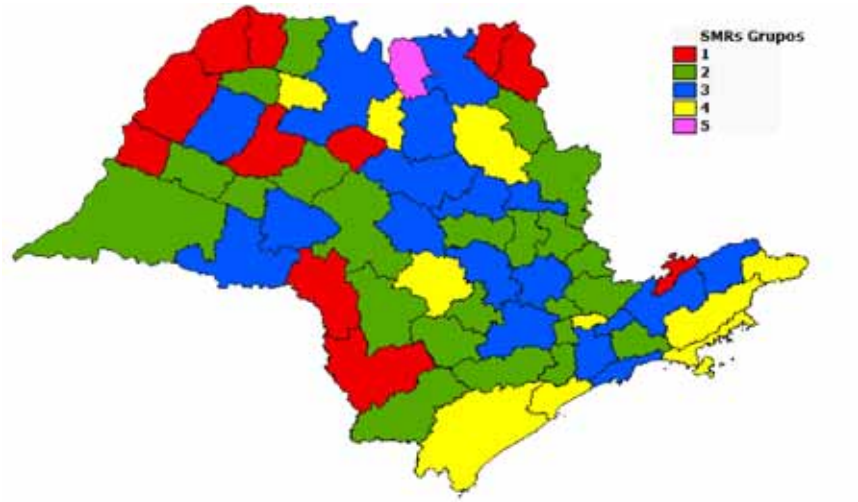

Figura 5.8: *Microrregiões do estado de São Paulo segundo os grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer feminino de mama, de 1998 a 2010.*

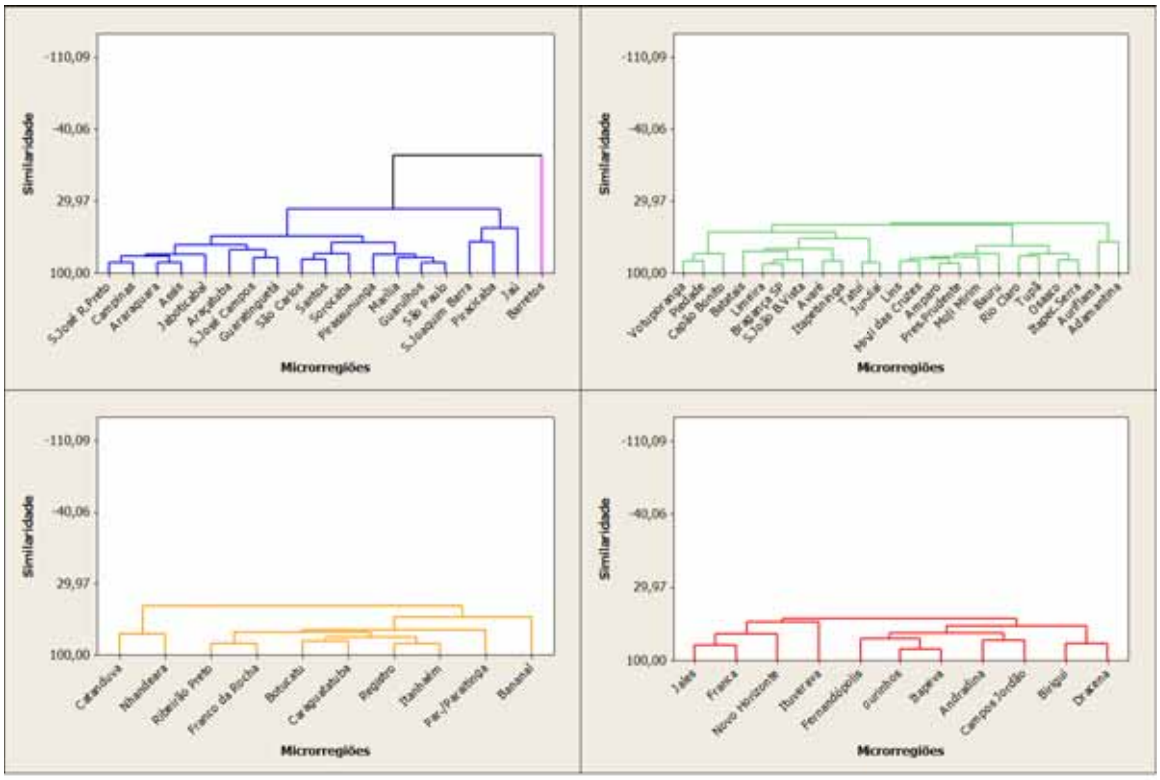

Figura 5.9: *Dendrogramas do agrupamento das SMRs referentes aos óbitos por câncer feminino de mama de 1998 a 2010.*

SMRs, e uma aparente tendência de 4 em 4 anos, sendo que 2002 a 2006 foi o período de pico responsável por este grupo ser o de mais altas SMRs (veja no mapa da Figura (5.12)). Nos demais anos do período em estudo, as microrregiões deste grupo tiveram SMR em torno de 1, caracterizando que os óbitos registrados corresponderam ao esperado para a região.

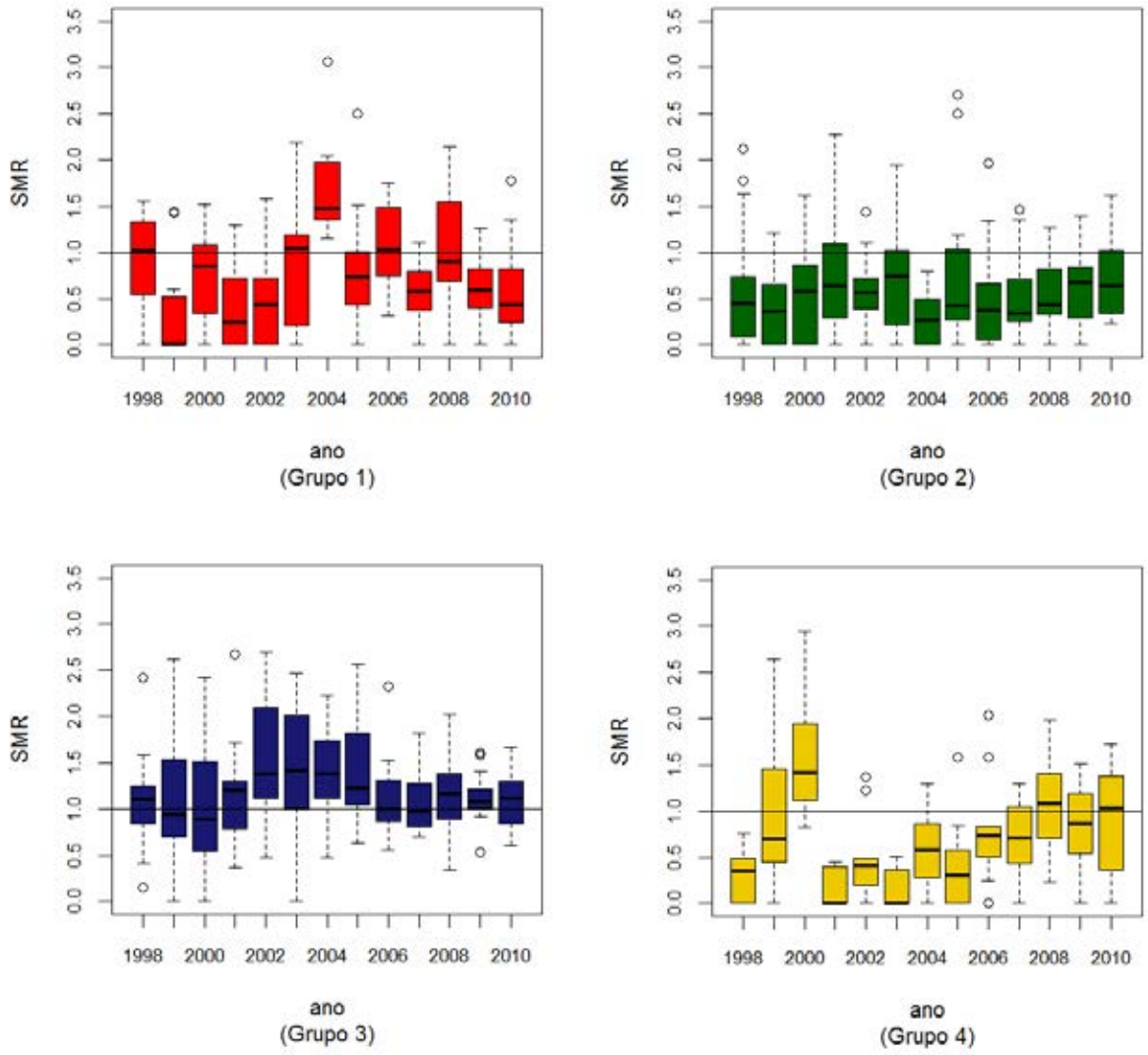

Figura 5.10: *Boxplots dos grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer feminino de mama de 1998 a 2010.*

O grupo 4 (amarelo) teve crescimento na mortalidade nos 3 primeiros anos do período. Em seguida, sofreu queda nas SMRs, que cresceram lentamente, até se estabilizar em torno de 1 de 2008 a 2010. O grupo parece ter tendência de comportamento de 5 em 5 anos.

O grupo 5 (microrregião de Barretos) (Figura 5.11), como de se esperar, registrou alto número de mortes, atingindo o pico do período em 2004, ano em que o número de óbitos chegou a ser 6 vezes maior que o esperado. Apesar disso, é notável que os valores tenham se tornado mais baixos e sem grandes saltos ao fim do período, que, como mencionado, pode se dever não só a uma queda na mortalidade, mas também numa otimização no sistema de informação que coleta os dados.

Ainda é possível verificar o comportamento das SMRs referente aos óbitos por câncer de mama para cada ano do período. A Figura 5.12 dispõe os mapas.

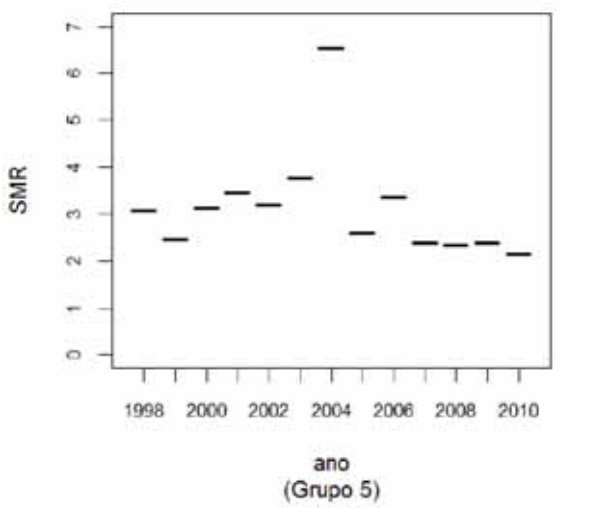

Figura 5.11: *Boxplots do grupo 5 (microrregião de Barretos) da análise de agrupamentos das SMRs referentes aos óbitos por câncer feminino de mama de 1998 a 2010.*

Apesar de ser possível formar grupos bastante específicos para essa doença, ao atentar para o mapa da Figura 5.8, é fácil ver que, de modo geral, os grupos se distribuíram de modo homogêneo pelo estado, sem a formação de áreas com grandes aglomerados de microrregiões pertencentes a um mesmo grupo. Verifica-se, portanto, grande variabilidade espacial para a mortalidade por câncer de mama, talvez devido à forte influência do fator hereditário na manifestação dessa doença.

Um exame nos mapas da Figura 5.12 indica que a interpretação dos grupos formados na análise de agrupamentos parece razoável, já que o comportamento espacial das SMRs nos mapas, como um todo, é bastante semelhante ao dos grupos no mapa da Figura 5.8.

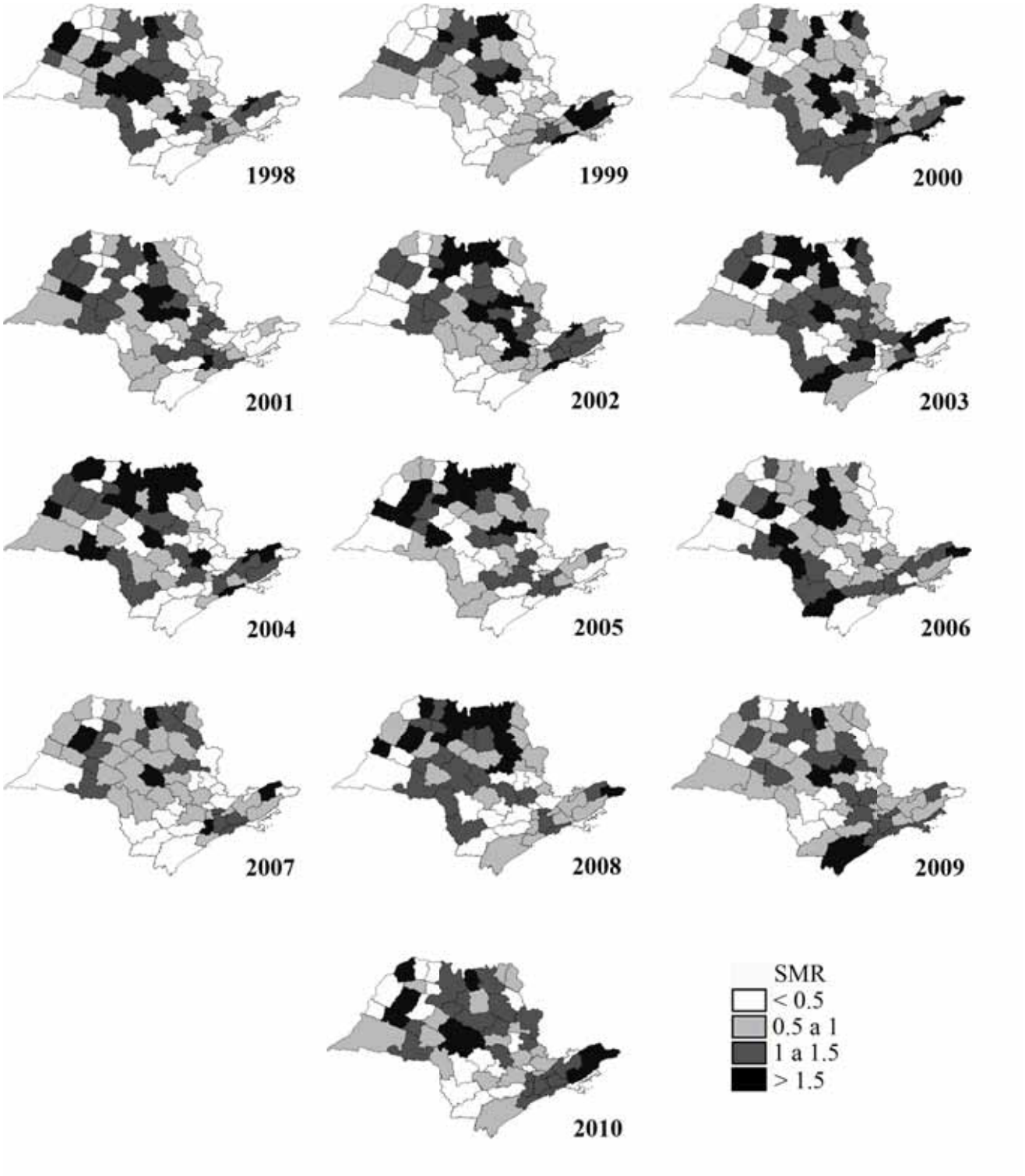

Figura 5.12: *Razões de Mortalidade Padronizadas (SMRs) para a mortalidade por câncer feminino de mama nas microrregiões do estado de São Paulo, de 1998 a 2010.*

5.1.3 Câncer de lábios, cavidade oral e faringe

Sobre a doença: A Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde - Décima Revisão (CID-10) agrupa as neoplasias malignas nos lábios, cavidade oral e faringe em uma única categoria. Esta inclui os tumores malignos em toda glândula e tecido da boca (exceto a pele do lábio), até a faringe. Segundo o INCA, os principais fatores de risco para o câncer da cavidade oral são o tabagismo, o alcoolismo e as infecções pelo HPV (do inglês, *human papiloma virus*, vírus do papiloma humano), sendo que o hábito de fumar e beber aumenta em 30 vezes o risco para o desenvolvimento deste tipo de câncer. O Instituto aponta que 42% dos óbitos por essa neoplasia se devem ao fumo, enquanto 16% ao alcoolismo. A detecção precoce por inspeção visual pode descobrir anormalidades pré-malignas do câncer da cavidade oral que, quando diagnosticado precocemente, apresenta bom prognóstico.

A Figura 5.13 leva a atenção novamente à dificuldade de se interpretar os valores das SMRs na região em estudo levando em conta todas as microrregiões. Torna-se difícil verificar comportamentos específicos e, assim como no caso do câncer feminino de mama, o fato de a maior parte das caixas ficar abaixo da linha demarcadora para $SMR = 1$ pode levar à conclusão de que os óbitos por essa doença foram menores do que o esperado na região em estudo para o período considerado, quando, na realidade, isso não é verdade, como mostrarão os resultados da análise de agrupamento. A quantidade de *outliers* também dificulta a identificação de padrões.

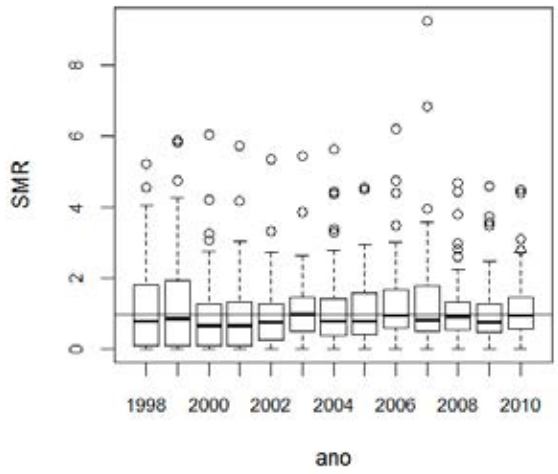

Figura 5.13: *Boxplots da distribuição das SMRs em todas as microrregiões do estado de São Paulo referentes aos óbitos por câncer de lábios, cavidade oral e faringe, de 1998 a 2010.*

A distribuição espacial dos grupos de microrregiões resultantes da análise de agrupamentos encontra-se na Figura 5.14.

Figura 5.14: *Microrregiões do estado de São Paulo segundo os grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de lábios, cavidade oral e faringe, de 1998 a 2010.*

Os dendrogramas da Figura 5.16 especificam as regiões pertencentes a cada grupo, e a hierarquia na divisão de alguns deles pela análise. O que se nota é que, os grupos 3 (azul) e 4 (amarelo) foram considerados semelhantes de alguma forma, e o grupo 5 (rosa) semelhante a ambos. O grupo 2 (verde) se destacou como o mais particular, além de ser o mais volumoso.

Os boxplots da Figura 5.15 podem ajudar a elucidar o significado de cada grupo.

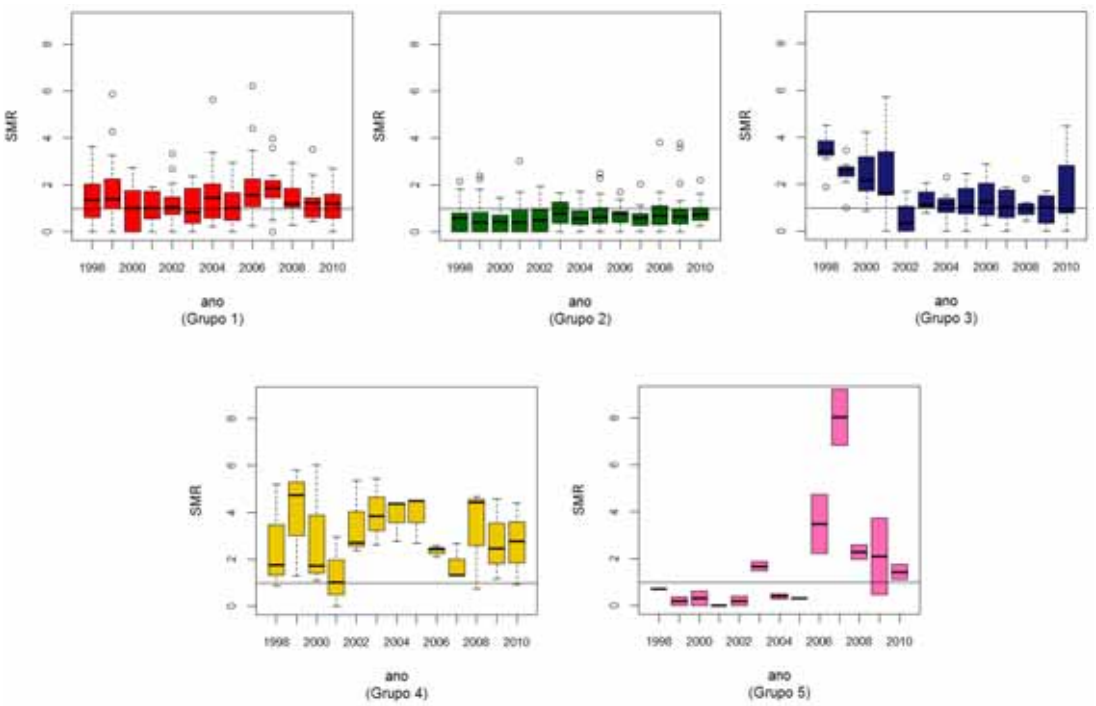

Figura 5.15: *Boxplots dos grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de lábios, cavidade oral e faringe, de 1998 a 2010.*

O grupo 2 (verde), predominante no mapa, é o que contém as microrregiões com número de óbitos inferior ao esperado para praticamente todo o período em estudo. No mapa, observa-se que tais microrregiões estão por todo o estado. Ao grupo 1 (vermelho) pertencem as microrregiões

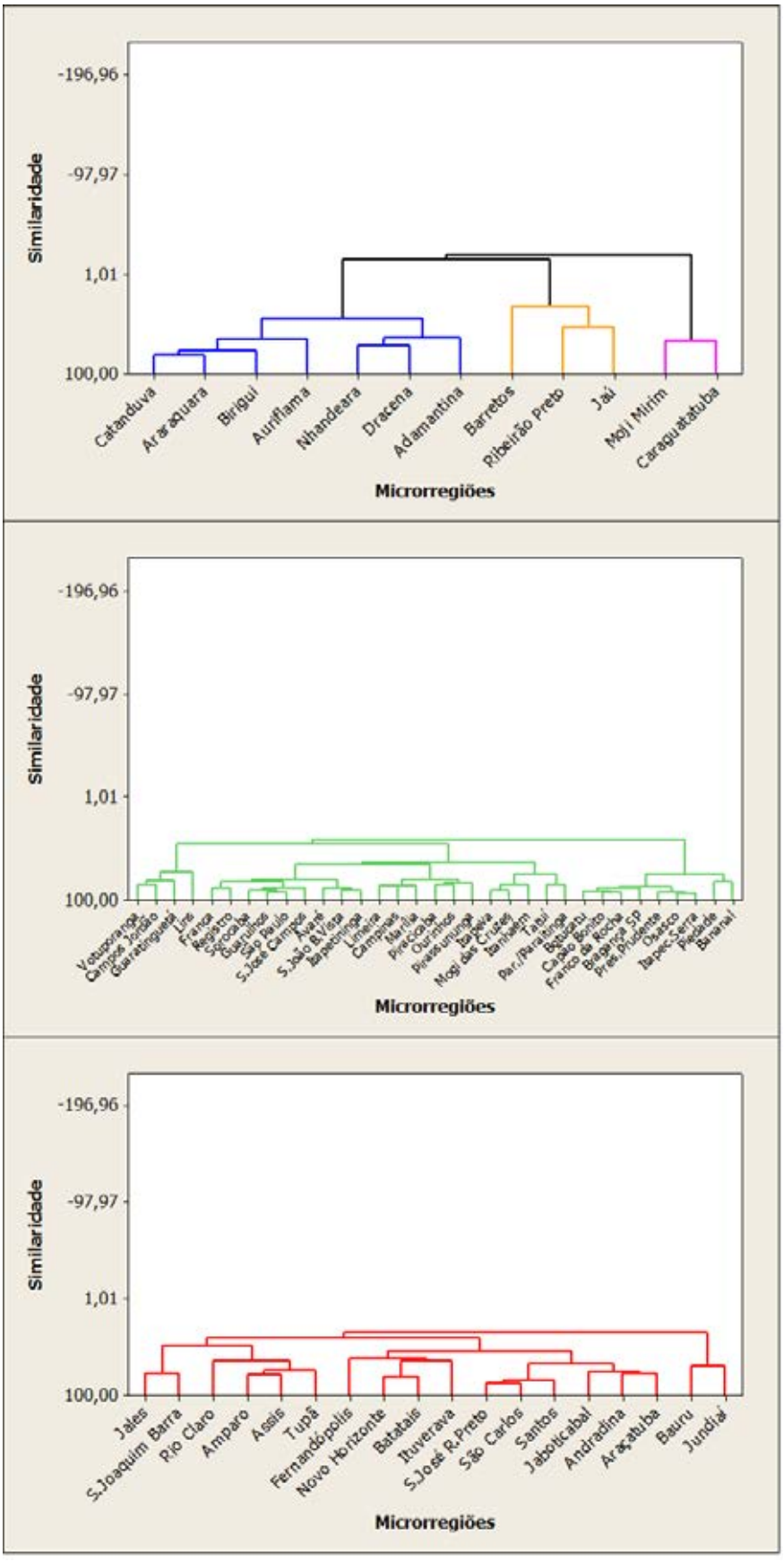

Figura 5.16: *Dendrograma do agrupamento das SMRs referentes aos óbitos por câncer de lábios, cavidade oral e faringe, de 1998 a 2010.*

cujo número de óbitos observados se mostrou próximo ao esperado, em quase todos os anos do período, exceto pela presença de alguns *outliers* de microrregiões que chegaram a ter até 5 vezes mais mortes que o esperado para essa doença em algum momento.

O grupo 3 (azul) teve valores altos para a SMR, assim como os grupos 4 (amarelo) e 5 (rosa), registrou mortalidades mais altas que o esperado, porém, decrescentes. Esses três grupos possuem poucas microrregiões, o que talvez explique parte da alta variabilidade observada em alguns deles (como no 4). Enquanto o grupo 3 teve mortalidade alta, porém decrescente, o grupo 4 teve altas SMRs em todo o período, o que se justifica uma vez que se verifica que duas de suas microrregiões - Barretos e Jaú - possuem hospitais de referência no tratamento de câncer. Com isso, recebe destaque a microrregião de Ribeirão Preto, que se alocou num grupo de SMRs tão altas. No mapa da Figura 5.14 é visível a proximidade desses três grupos, que, concentrados no centro e norte do estado, indicam que essas regiões são as mais afetadas em decorrência da mortalidade por essa neoplasia. Os mapas da Figura 5.17 confirmam essa distribuição espacial. Observou-se esse mesmo padrão de ocorrência das SMRs para o câncer de traqueias, brônquios e pulmão. E, levando em conta que, de acordo com o SILVA *et al.* (2011), 90% dos casos de câncer de pulmão no Brasil são decorrentes do tabagismo, esta variável também deve ser importante em explicar parte dos óbitos por câncer de lábios, cavidade oral e faringe, e, portanto, a similaridade na distribuição espacial dos riscos associados a tais doenças.

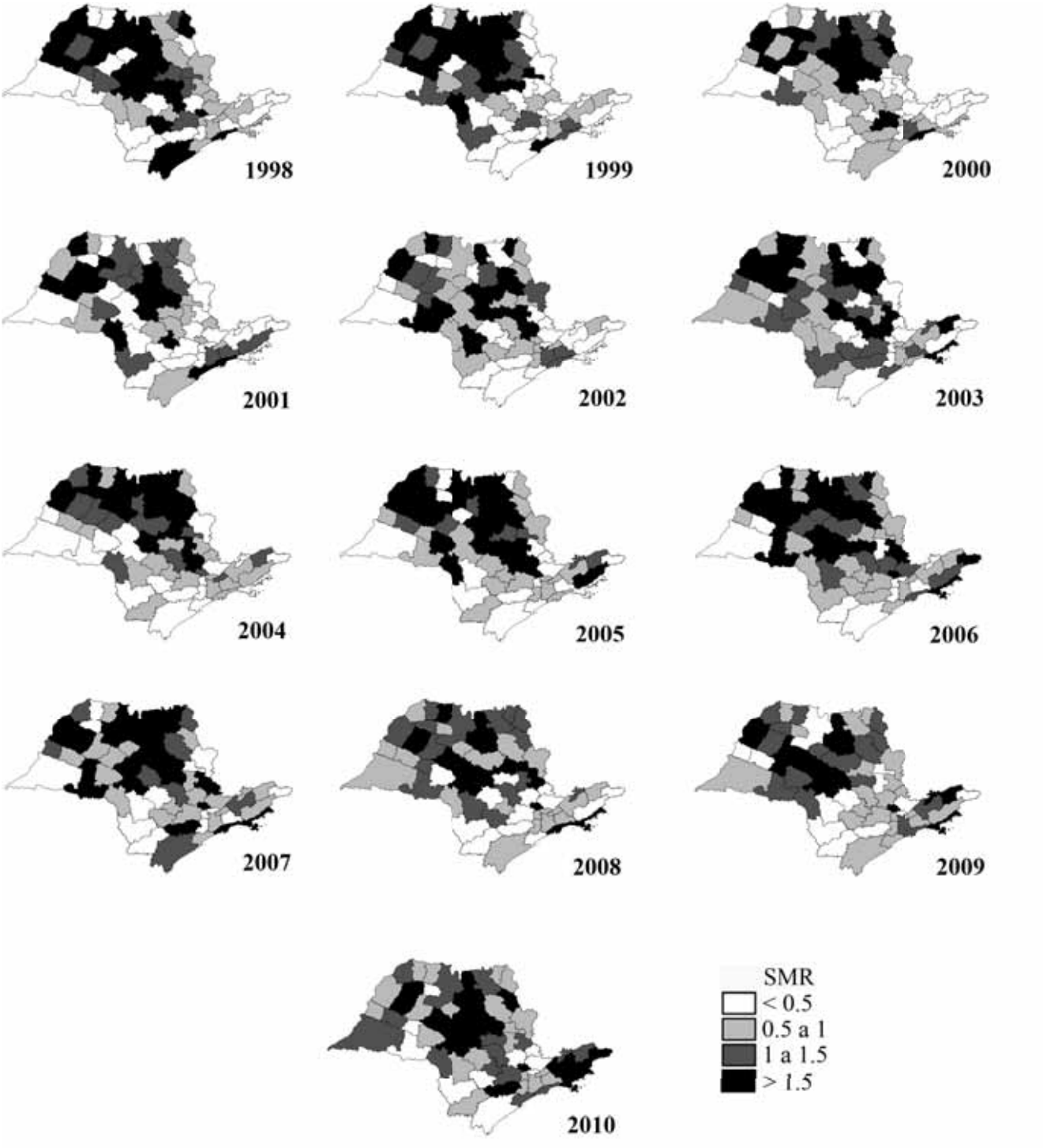

Figura 5.17: *Razões de Mortalidade Padronizadas (SMRs) para a mortalidade por câncer de lábios, cavidade oral e faringe nas microrregiões do estado de São Paulo, de 1998 a 2010.*

5.1.4 Câncer de estômago

Sobre a doença: Também denominado câncer gástrico, os tumores do estômago tem seu pico de incidência na população masculina, por volta dos 70 anos. De acordo com o INCA, cerca de 65% dos pacientes diagnosticados com câncer de estômago têm mais de 50 anos. No Brasil, esses tumores aparecem em terceiro lugar na incidência entre homens e em quinto, entre as mulheres. A nível mundial, o maior número de casos ocorre no Japão, onde são encontrados 780 doentes por 100.000 habitantes. Configura-se como a segunda causa de morte por câncer no mundo, sendo os mais afetados países em desenvolvimento. Apesar disso, as taxas de incidência tem decrescido na maioria dos países. O INCA assegura que, boa parte disso se deve ao aumento do uso de refrigeradores para uma melhor conservação alimentar, aliado a modificações no hábito alimentar da população (aumento da ingestão de frutas, legumes e verduras frescas). Essa mudança no padrão alimentar, junto com melhorias no saneamento básico, também explica a redução na prevalência de infecções pela *Helicobacter pylori* (H. pylori), responsável por 63% dos casos de câncer gástrico. O câncer do estômago é um tipo de tumor que não possui um bom prognóstico, apresentando sobrevida relativa considerada baixa, de apenas cinco anos.

Na Figura 5.18 estão dispostos os boxplots das SMRs referentes aos óbitos por câncer de estômago no estado de São Paulo, para cada ano do período em estudo. Observe que o número de óbitos por essa neoplasia chegou a ser 4 vezes maior que o esperado em algumas microrregiões. A elevada quantidade de outliers indica a existência de muitas microrregiões com alta mortalidade.

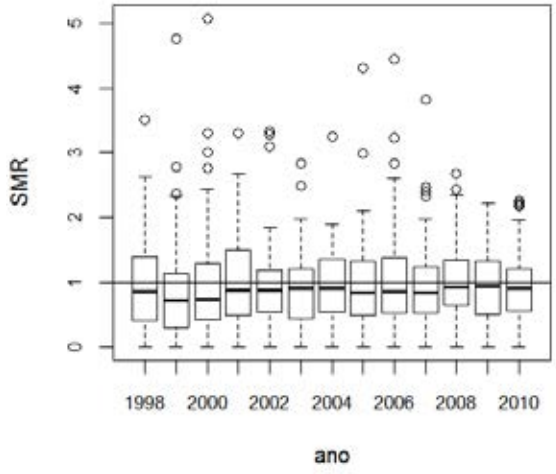

Figura 5.18: *Boxplots da distribuição das SMRs em todas as microrregiões do estado de São Paulo referentes aos óbitos por câncer de estômago, de 1998 a 2010.*

No mapa da Figura 5.20, é notável a predominância dos grupos 1 (vermelho) e 3 (azul), e, atentando a seus respectivos boxplots na Figura 5.21 isso se torna compreensível, já que eles tem comportamentos bastante semelhantes, a não ser pela locação das SMRs, que, no primeiro grupo concentrou-se em torno de 1, enquanto que no terceiro, ficaram abaixo. Isso revela que a maior parte das microrregiões do estado, para a maior parte do período em estudo, tiveram mortalidade compatível ou abaixo do esperado.

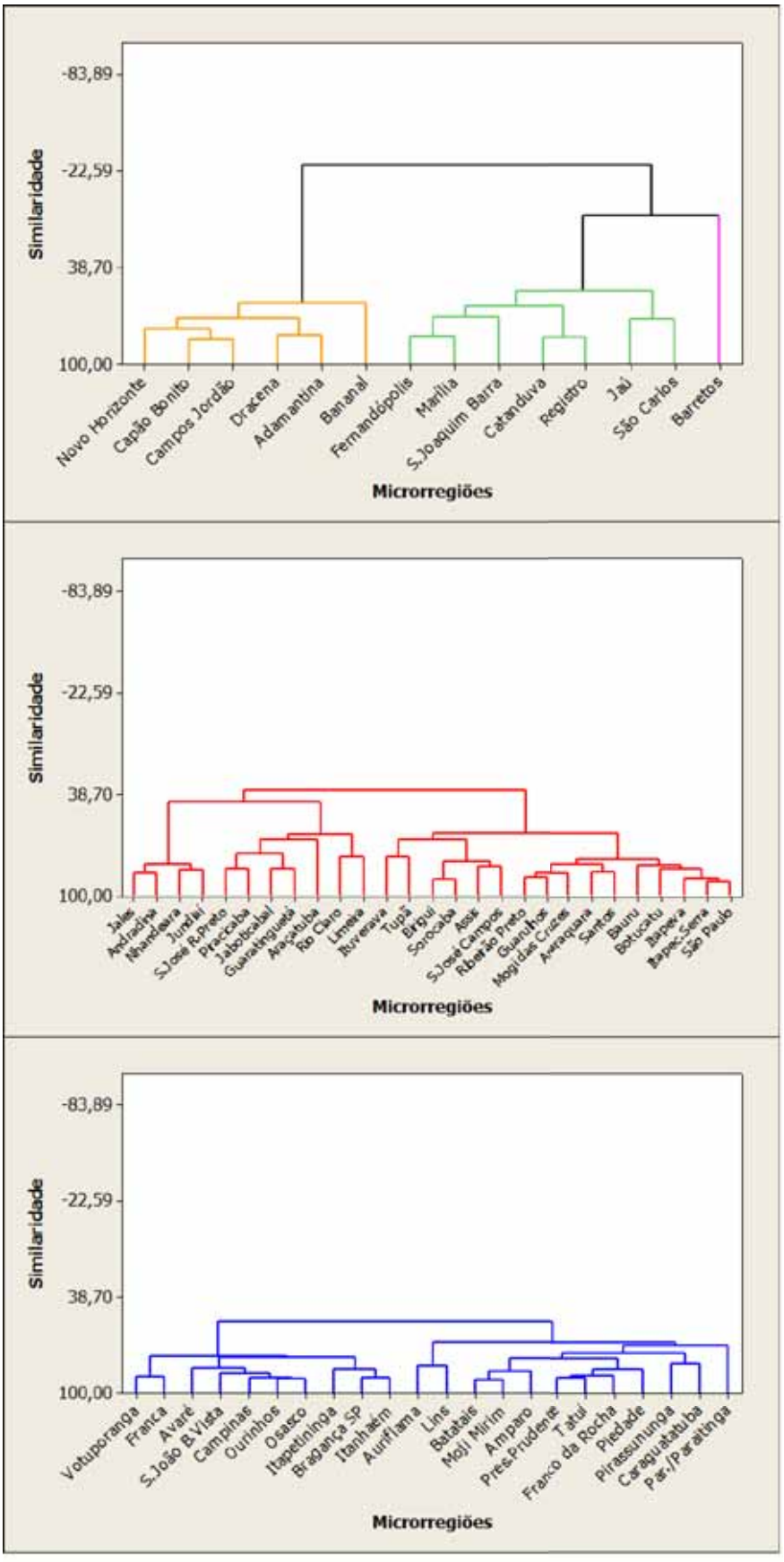

Figura 5.19: *Dendrograma do agrupamento das SMRs referentes aos óbitos por câncer de estômago, de 1998 a 2010.*

No grupo 2 (verde) foram alocadas as microrregiões com os maiores valores para a SMR, geralmente com maior número de óbitos do que o esperado para sua estrutura demográfica. Ao observar o mapa, vemos que essas microrregiões se distribuem sem nenhum padrão aparente, tal qual as do grupo 4 (amarelo), com grande variabilidade nas SMRs e quedas ou crescimentos bruscos em seus valores no decorrer do período.

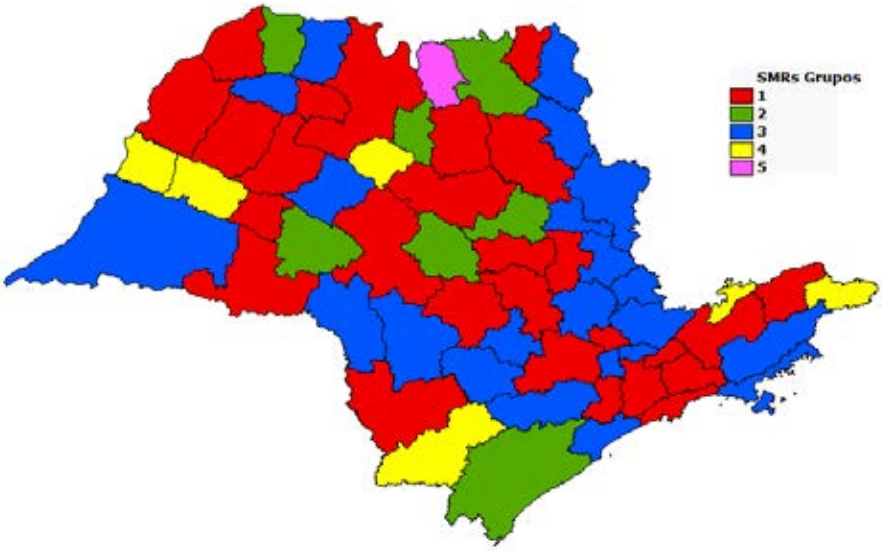

Figura 5.20: *Microrregiões do estado de São Paulo segundo os grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de estômago, de 1998 a 2010.*

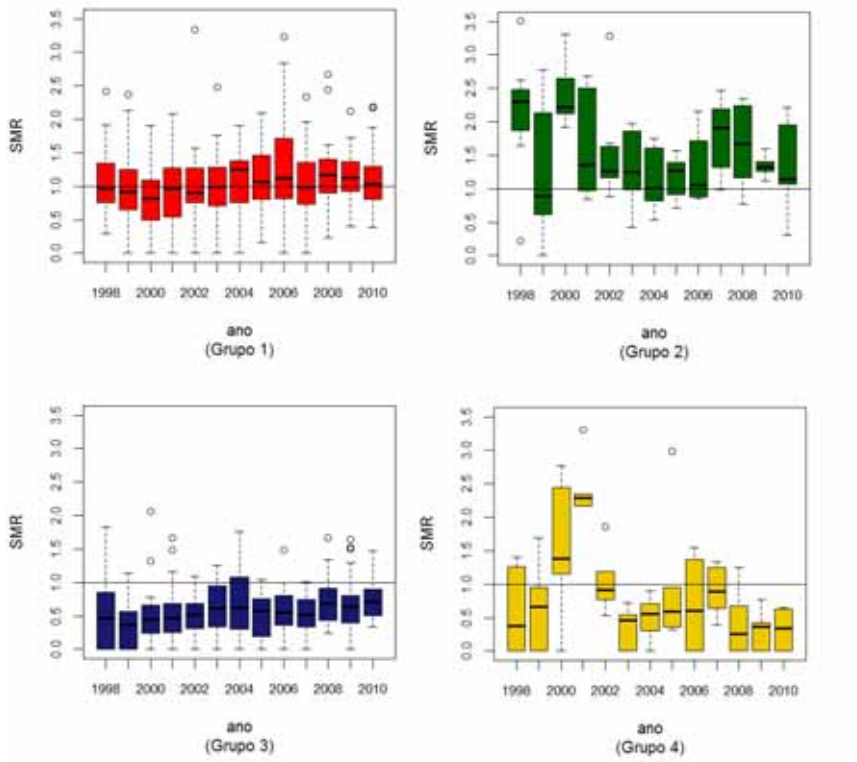

Figura 5.21: *Boxplots dos grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de estômago, de 1998 a 2010.*

Também é possível acompanhar as SMRs para Barretos, dispostas nos boxplots da seguinte figura. Os valores altos não surpreendem, mas a semelhança com os respectivos boxplots para as demais doenças, em que se observa queda e estabilidade nas SMRs para o fim do período, aponta para uma melhoria no sistema de coleta de dados.

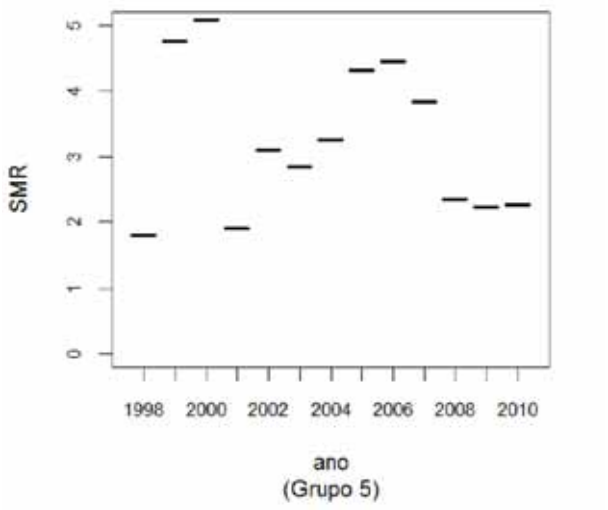

Figura 5.22: *Boxplots dos grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de estômago, de 1998 a 2010.*

Na sequência, os mapas individuais para cada ano do período revelam as particularidades de cada um deles na Figura 5.23. Novamente, as regiões centro e norte aparecem com destaque para altos valores da razão de mortalidade padronizada, mesmo que de maneira suave, o que talvez explique isso não ter ficado explícito na análise de agrupamentos.

Como a análise exploratória neste trabalho se concentra nos valores para as SMRs, é importante lembrar que altos valores para essa estimativa não necessariamente estão associados a altos valores absolutos de óbitos. Assim, uma microrregião que apresentou um número absoluto de óbitos menor do que outra, pode ter um valor para a SMR maior. Isso acontece, por exemplo, em regiões onde a estrutura da população é majoritariamente masculina e/ou idosa. Já foi citado que áreas com tais características possuem maior incidência de câncer, assim como isto também se dá em populações com maior densidade demográfica, por motivos óbvios. No entanto, neste caso, o que está sendo considerado é a busca da resposta à seguinte pergunta: Estão ocorrendo mais óbitos nessa população do que o esperado, levando em conta sua estrutura demográfica? Se a resposta for sim, então cabe às autoridades responsáveis pela saúde pública implementar medidas mais eficientes de diagnóstico, prevenção e tratamento dessas doenças, além de prosseguir investigando fatores etiológicos que possam ser evitáveis ou reduzidos. Por outro lado, se a resposta for não, isso quer dizer que os óbitos estão ocorrendo dentro do que é esperado para a estrutura demográfica da microrregião.

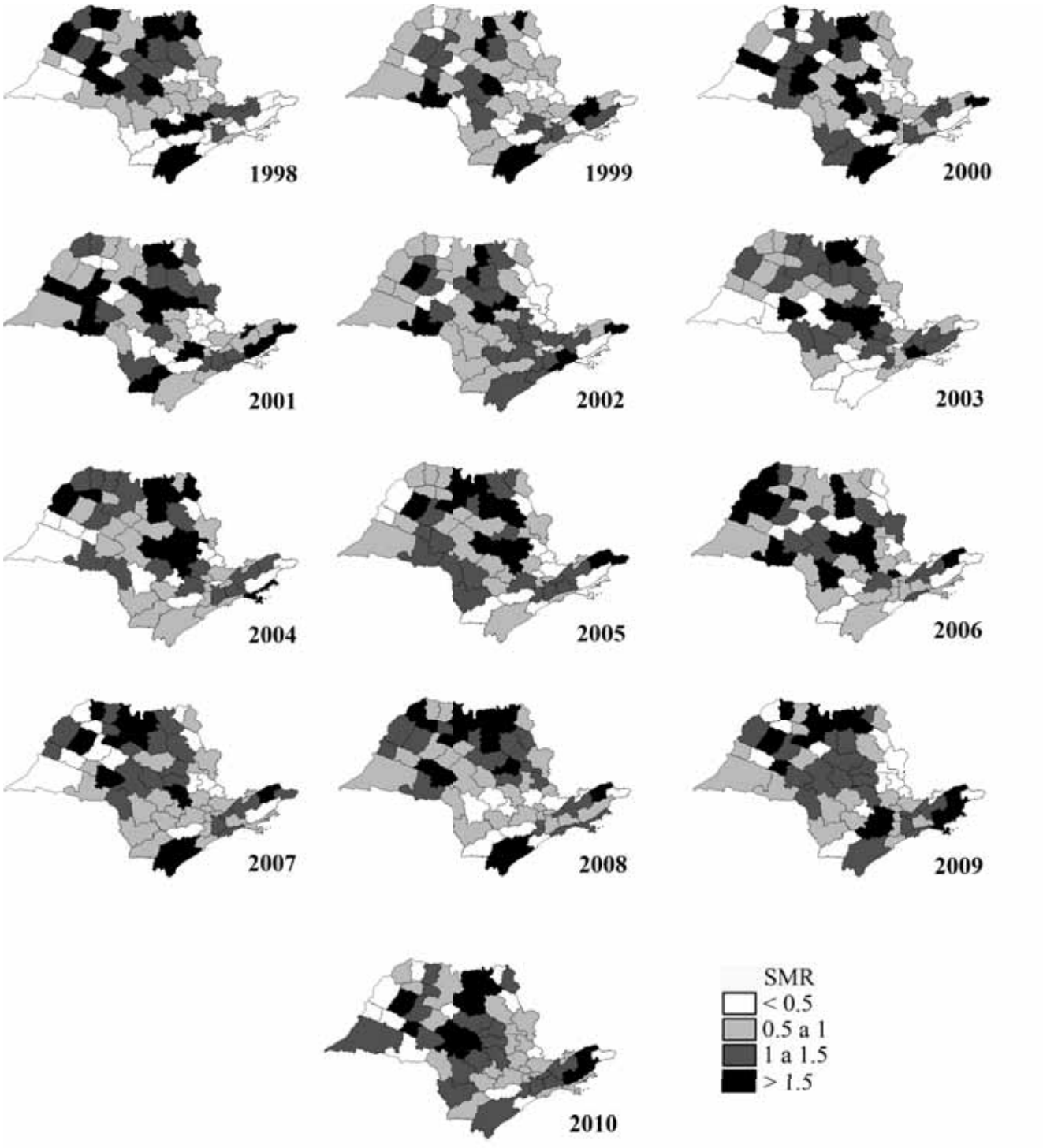

Figura 5.23: *Razões de Mortalidade Padronizadas (SMRs) para a mortalidade por câncer de estômago nas microrregiões do estado de São Paulo, de 1998 a 2010.*

5.1.5 Câncer de cólon

Sobre a doença: O Instituto Nacional do Câncer classifica o câncer de cólon como o terceiro tipo de câncer mais comum entre os homens, e o segundo para as mulheres, com 60% dos casos localizados em regiões mais desenvolvidas. Os padrões geográficos são bem semelhantes em relação ao sexo, embora o masculino apresente maior incidência na maioria das populações. Essa neoplasia é considerada de bom prognóstico se a doença for diagnosticada em estágio inicial. Seu desenvolvimento, assim como o de várias formas comuns de câncer é resultado da interação entre fatores hereditários e ambientais, sendo que, deste último, o mais notável é a dieta. O consumo excessivo de carne vermelha, embutidos e bebidas alcoólicas, o tabagismo e distúrbios de peso favorecem o desenvolvimento dessa doença. Mas os fatores de risco mais relevantes são a história familiar e a predisposição genética ao desenvolvimento de doenças crônicas do intestino. A idade também é considerada um fator de risco, uma vez que tanto a incidência como a mortalidade aumentam com a idade. A história natural dessa neoplasia propicia condições ideais à sua detecção precoce. A pesquisa de sangue oculto nas fezes e métodos endoscópicos são considerados meios de detecção precoce eficientes para esse câncer, pois são capazes de diagnosticar pólipos adenomatosos colorretais (precursores do câncer do cólon e reto), bem como tumores em estágios bem iniciais. Mas, mesmo em países com maiores recursos, a relação custo-benefício em investimentos para estratégias apropriadas de prevenção e detecção precoce do câncer do cólon e reto tem impossibilitado a implantação de rastreamento populacional.

Iniciando a análise para os óbitos por essa doença no estado de São Paulo, considere a Figura 5.24 com informação de todas as microrregiões, para cada ano do período de 1998 a 2010. Com base nessa figura, pode-se dizer que as microrregiões tiveram número de óbitos por câncer de cólon bem próximo do esperado, sendo 1998 a 2001 o sub-período com os menores, mas ao mesmo tempo também os maiores, números de mortes registradas, devido a quantidade de *outliers*, que continua sendo um problema na análise deste gráfico.

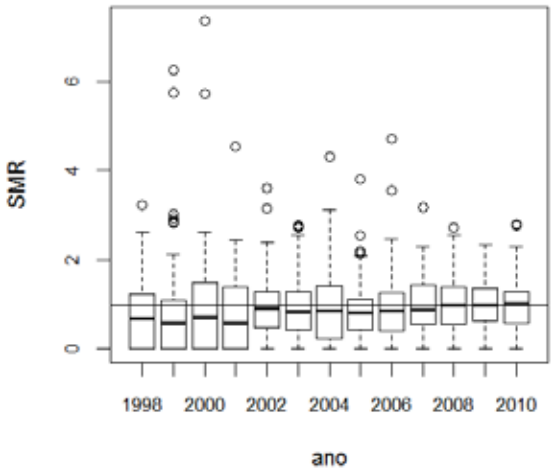

Figura 5.24: *Boxplots da distribuição das SMRs em todas as microrregiões do estado de São Paulo referentes aos óbitos por câncer de cólon, de 1998 a 2010.*

De acordo com as figuras 5.26 e 5.27, os grupos 2 (verde) e 4 (amarelo) foram os que tiveram as maiores SMRs, atentando para que, no grupo 2, apesar disso, as microrregiões tiveram SMR próxima de 1 na maior parte do período. Já no grupo 4, isso não aconteceu, uma vez que a variância dessas razões nas microrregiões deste grupo se mostrou-se alta, fazendo com que essa distribuição oscilasse em boa parte do início do período em diante. Esses grupos apareceram notoriamente no centro e norte do estado, regiões de maior risco, em geral, para a marioria dos cânceres considerados neste trabalho. Os grupos 1 (vermelho) e 3 (azul) tiveram as menores SMRs, entretanto, é válido ressaltar o comportamento temporal crescente dessas razões no grupo 3, que visivelmente sofreu mudanças nas SMRs em sub-períodos de 4 ou 5 anos.

A microrregião de Barretos (Figura 5.29) se comportou novamente como um caso anômalo, porém, com decrescimento evidente nas razões de mortalidade padronizadas.

Nos mapas individuais das SMRs para cada ano do período, pode-se confirmar as conclusões tiradas com base na análise de agrupamentos. Note que os mapas foram ficando mais escuros no decorrer do período, e aparentemente mais "preenchidos", ou seja, com menos microrregiões com pequenas SMRs e mais regiões com altas SMRs. Ao comparar esses mapas com o mapa do agrupamento na Figura 5.26, é possível identificar que isso acontece devido à mudanças principalmente nas microrregiões pertencentes aos grupos 2 (verde) e 3 (azul), cujo comportamento temporal, como já mencionado, é bem evidente em seu boxplot da Figura 5.27.

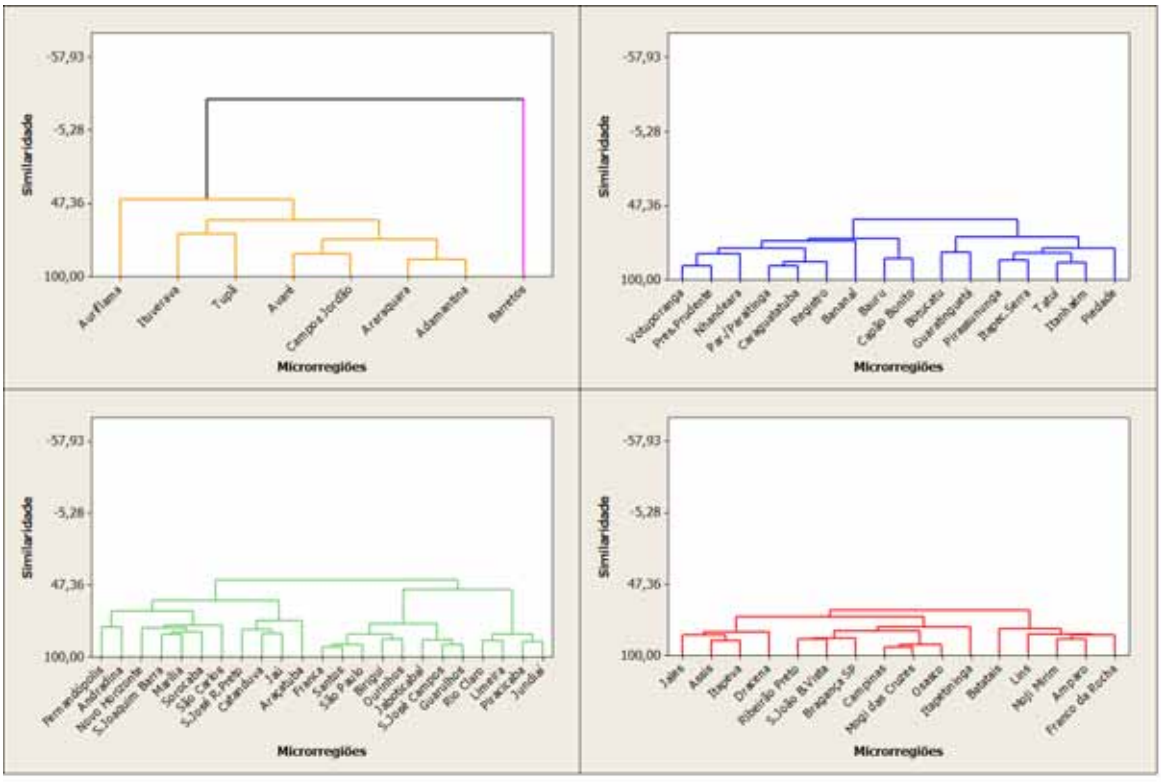

Figura 5.25: *Dendrograma do agrupamento das SMRs referentes aos óbitos por câncer de cólon, de 1998 a 2010.*

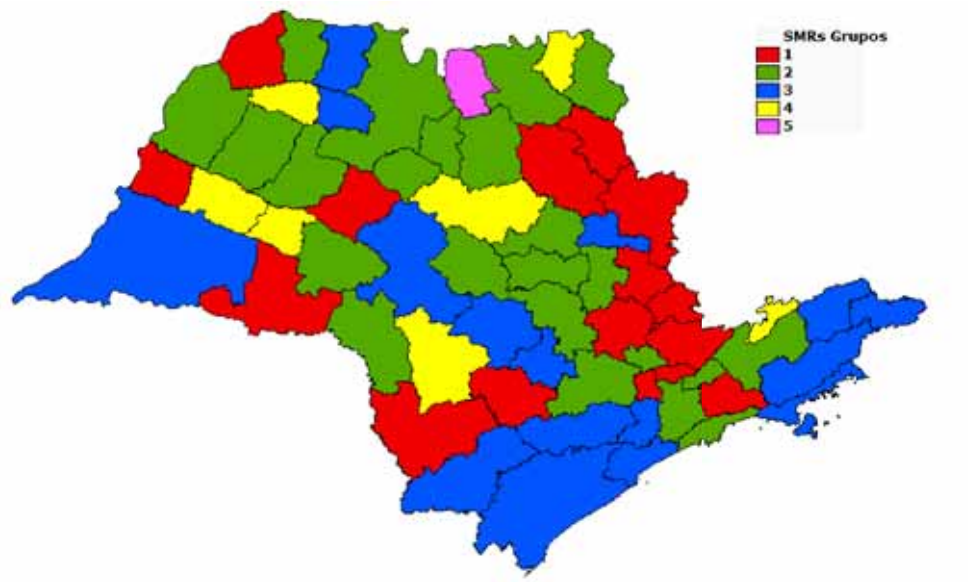

Figura 5.26: *Microrregiões do estado de São Paulo segundo os grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de cólon, de 1998 a 2010.*

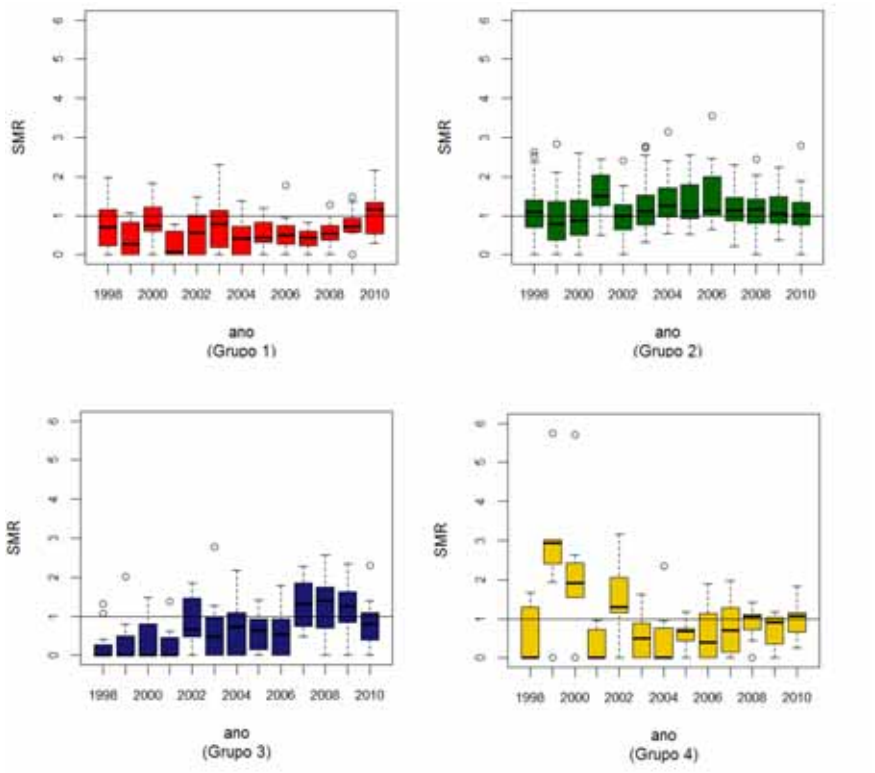

Figura 5.27: *Boxplots dos grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de cólon, de 1998 a 2010.*

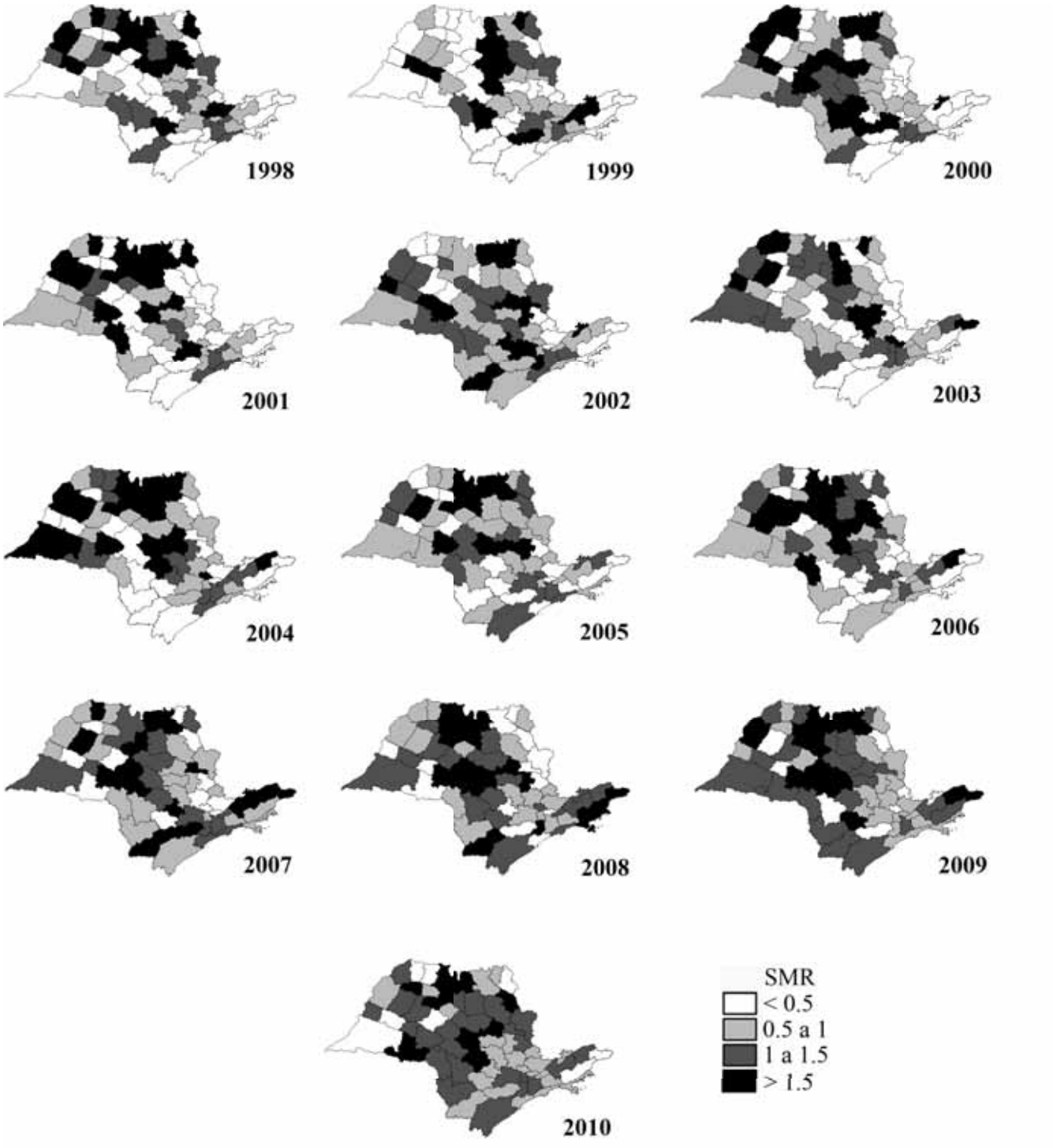

Figura 5.28: *Razões de Mortalidade Padronizadas (SMRs) para a mortalidade por câncer de cólon nas microrregiões do estado de São Paulo, de 1998 a 2010.*

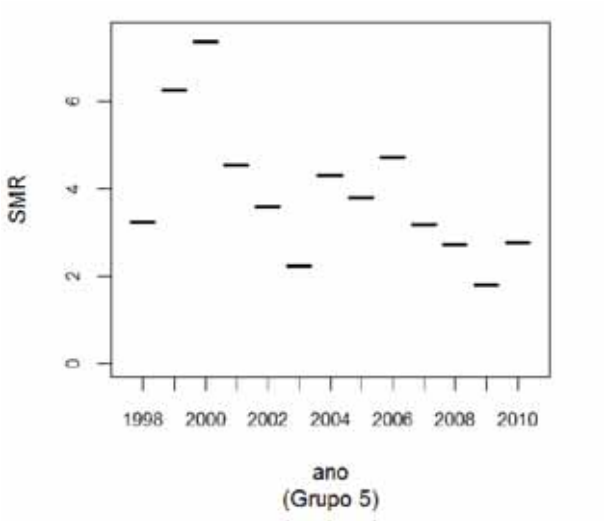

Figura 5.29: *Boxplots dos grupos da análise de agrupamentos das SMRs referentes aos óbitos por câncer de cólon, de 1998 a 2010.*

5.2 Resumo da Análise Exploratória

Embora não seja conclusiva, através do que fora apresentado nas seções anteriores, fica evidente que a análise exploratória contribui muito para um conhecimento prévio dos dados. E, resumindo os pontos mais importantes da análise feita, conclui-se o seguinte para cada doença:

- **Câncer de traqueia, brônquios e pulmão**
  Os altos valores para as SMRs se concentraram no centro e norte do estado. Destaque pode ser dado ao grupo azul da análise de agrupamentos, em que chamam atenção as microrregiões de Araraquara, São Carlos, Jaboticabal, Catanduva, São Joaquim da Barra e Ituverava. A exceção é a microrregião de Jaú, que possui altas taxas devido a conter um hospital de referência.

- **Câncer feminino de mama**
  Esse tipo de doença manifestou maior aleatoriedade na distribuição espacial das razões de mortalidade, de modo que a interpretação da ocorrência dos óbitos nas microrregiões não se mostrou tão clara quanto para a doença anterior. Ainda assim, podem ser destacadas as microrregiões do grupo azul, para o sub-período de 2002 a 2005. Microrregiões com altos valores para os óbitos foram Araçatuba, São José do Rio Preto, Assis e Marília. Também se destacaram nos mapas duas faixas de microrregiões, uma que vai de São Joaquim da Barra até Sorocaba, e outra de Santos a Guaratinguetá, que se localizam, respectivamente, do norte para o centro, e no leste do estado.

- **Câncer de lábios cavidade oral e faringe**
  A visualização da ocorrência de óbitos importantes nesta doença pôde ser melhor observada através do mapa da SMR para o ano de 2010, pois as razões de mortalidade encontram-se

mais suaves neste mapa e resumem bem a informação produzida pela análise de agrupamentos. Numa visão geral, o grupo azul recebe destaque pelas ocorrências dos óbitos entre 1998 e 2001, na faixa de microrregiões que inclui Adamantina, Birigui, Auriflama e Nhandeara, que ficam no oeste do estado. Outra região de microrregiões que se destacaram foi a que se localiza no centro-norte da região, e vai de Barretos até Jaú, passando por Lins.

- **Câncer de estômago**
  Recebem destaque neste tipo de doença as microrregiões do grupo verde: Registro, São Carlos, Marília, Fernandópolis, Catanduva e São Joaquim da Barra, que, embora aparentem ter uma distribuição espacial aleatória, se alocaram adjacentes ou próximas ao grupo vermelho, que, por ter uma grande quantidade de *outliers*, possui microrregiões que tiveram altos valores para a SMR em vários anos do período.

- **Câncer de cólon**
  A distribuição espacial das SMRs para esta doença se mostrou bastante semelhante à da doença anterior, câncer de estômago. Novamente, um aglomerado de microrregiões no centro e norte do estado se mostraram importantes. A análise de agrupamentos não se mostrou muito elucidativa neste caso, uma vez que a variabilidade entre os grupos foi relativamente pequena, não permitindo evidenciar claramente diferenças entre suas microrregiões.

Com respeito à intensidade das SMRs para as doenças, uma análise visual dos mapas apresentados indica que a ordem de importância (das maiores taxas para as menores) foi a seguinte:

Tabela 5.1: *Ranking das doenças de acordo com os resultados da análise exploratória dos dados*

Posição	Doença
1º	Câncer de lábios, cavidade oral e faringe
2º	Câncer de traqueia, brônquios e pulmão
3º	Câncer de mama
4º	Câncer de estômago
5º	Câncer de cólon

Capítulo 6

Aplicação dos modelos hierárquicos Bayesianos em dados de área multivariados

O Capítulo anterior forneceu uma visão geral sobre o comportamento dos dados e as regiões do estado com maiores riscos de óbitos, baseado nas SMRs. A partir das suposições que a análise exploratória proporcionou, o próximo passo é compreender melhor os dados a partir da construção de um modelo apropriado.

Neste Capítulo encontram-se os resultados da aplicação de modelos semelhantes aos considerados no Capítulo 2, levando em conta primeiramente apenas o domínio espacial, e posteriormente, também o domínio temporal. São apresentadas as regiões de maior risco a posteriori para cada um dos modelos, e, por fim, uma comparação entre eles é realizada através do DIC. Antes, porém, é válido descrever os procedimentos de inferência utilizados para a obtenção de tais resultados.

6.1 Procedimentos de Inferência

A partir da especificação do modelo e da escolha das distribuições a priori apropriadas, o objetivo do pesquisador é a obtenção das distribuições a posteriori dos parâmetros de interesse. Tais distribuições podem ser obtidas pela abordagem Bayesiana através do Teorema de Bayes, que diz o seguinte:

Seja φ *o parâmetro de interesse num modelo. Após observar uma amostra* x *de um vetor aleatório* X *relacionado com* φ*, o conhecimento a respeito desse parâmetro pode ser atualizado através da expressão*

$$p(\varphi|x) = \frac{p(\varphi,x)}{p(x)} = \frac{p(x|\varphi)p(\varphi)}{p(x)} = \frac{p(x|\varphi)p(\varphi)}{\int p(\varphi,x)d\varphi} \propto p(x|\varphi)p(\varphi) \tag{6.1}$$

O objetivo deste tópico é descrever os procedimentos de inferência para os modelos utilizados na aplicação em dados de câncer, que são adaptações dos modelos abordados do Capítulo 2. São considerados dois modelos: o primeiro trata-se de um modelo similar ao descrito em (2.2), e o

segundo de um modelo similar ao em (2.7), porém, ambos em uma versão multivariada e sem a presença de covariáveis. Para $i = 1,...,N$ áreas, $t = 1,...,T$ tempos, e $k = 1,...,K$ doenças, os modelos são, respectivamente:

- Modelo 1: Modelo Hierárquico Bayesiano

$$\begin{aligned} Y_{ik}|R_{ik} &\sim Poisson(E_{ik}R_{ik}), \\ log(R_{ik}) &= \alpha_k + \phi_{ik}, \\ \alpha_k &\sim U(-\infty,+\infty), \\ \boldsymbol{\phi}_k &\sim MVCAR(1,\boldsymbol{\Lambda}). \end{aligned}$$

- Modelo 2: Modelo Hierárquico Bayesiano Dinâmico

$$\begin{aligned} Y_{itk}|R_{itk} &\sim Poisson(E_{itk}R_{itk}), \\ log(R_{itk}) &= \alpha_k + \theta_{kt} + \phi_{ik}, \\ \alpha_k &\sim U(-\infty,+\infty), \\ \boldsymbol{\phi}_k &\sim MVCAR(1,\boldsymbol{\Lambda}), \\ \theta_{kt} &= \theta_{k,t-1} + \omega_{kt}, \\ \text{com } \theta_{k0} &\sim N(0,\boldsymbol{W}_K), \\ \omega_t &\sim N(0,\boldsymbol{W}_K). \end{aligned}$$

Em ambos os modelos a matriz $\boldsymbol{\Lambda}$ contém os parâmetros de precisão para os efeitos espaciais, e assume-se para ela distribuição a priori $Wishart(\boldsymbol{R},n)$, em que $n = K = 5$ é escolhido, de modo que a priori seja pouco informativa (Shaddick e Wakefield (2002)). A média a priori da $Wishart(\boldsymbol{R},n)$ é igual a $n\boldsymbol{R}$, o que sugere que uma escolha razoável para $\boldsymbol{R}^{-1}$ pode ser $n\Sigma_0$, em que Σ_0 é a crença a priori acerca da matriz de covariâncias.

Com respeito à matriz $\boldsymbol{W}_K$ especificada para $\omega_t = (\omega_{1t},...,\omega_{Kt})'$ no segundo modelo, esta contém as variâncias $\sigma^2_{\omega k}$, que permitem que diferentes doenças tenham diferentes quantidades de dependência temporal, e $K(K-1)/2$ termos de covariância refletindo a dependência entre cada doença, condicional aos valores dos tempos anteriores.

Agora, para a descrição da distribuição a priori atribuída aos efeitos temporais θ_{kt}, por simplicidade de notação considere uma doença genérica. Seu vetor de efeitos aleatórios espaciais é $\boldsymbol{\theta} = (\theta_1,...,\theta_T)'$, cujo respectivo termo de variância é σ^2_ω. Esta distribuição pode ser escrita como

$$p(\boldsymbol{\theta}|\sigma^2_\omega) \propto \prod_{t=2}^{T} p(\theta_t|\theta_{t-1},\sigma^2_\omega)$$

$$\propto \exp\left\{-\frac{1}{2\sigma_\omega^2}\sum_{t=2}^{T}\left(\theta_t-\theta_{t-1}^2\right)\right\}$$
$$\propto \exp\left\{-\frac{1}{2\sigma_\omega^2}\sum_{t=1}^{T} n_t\theta_t\left(\theta_t-\bar{\theta}_t^2\right)\right\}$$

em que n_t indica o número de vizinhos de θ_t e $\bar{\theta}_t$ a média destes, isto é, de θ_{t-1} e θ_{t+1}. Observe que esta distribuição pode ser expressa como

$$p(\theta_t|\theta_{-t},\sigma_\omega^2) \sim \begin{cases} N\left(\theta_{t+1},\sigma_\omega^2\right), & \text{t = 1,} \\ N\left(\dfrac{\theta_{t-1}+\theta_{t+1}}{2},\dfrac{\sigma_\omega^2}{2}\right), & \text{t = 2,...,}T-1, \\ N\left(\theta_{t-1},\sigma_\omega^2\right), & \text{t = }T. \end{cases} \tag{6.2}$$

em que θ_{-t} representa o vetor de todos os elementos de $\boldsymbol{\theta}$, exceto θ_t.

Isso mostra que, em uma dimensão, o passeio aleatório Gaussiano se reduz à distribuição CAR intrínseca (ver Fahrmeir e Lang (2001)). Dessa forma, a especificação em 6.2 é equivalente a

$$(\theta_t|\theta_{-t},\sigma_\omega^2) \sim N\left(\sum_j C_{tj}\theta_j,\sigma_\omega^2 M_{tt}\right), \text{ para } t=1,...,T,$$

em que $C_{tj}=\dfrac{W_{tj}}{W_{t+}}$, $W_{t+}=\sum_j W_{tj}$ e $W_{tj}=1$ se $j=(t-1)$ ou $j=(t+1)$ e 0 caso contrário. Também, $M_{tt}=\dfrac{1}{W_{t+}}$. Por isso, no OpenBUGS, pode-se ajustar como distribuição a priori um passeio aleatório de ordem um (denotada por *RW(1)*) através da distribuição *"car.normal"*, ou, analogamente, da distribuição *"mv.car"* para dados multivariados, que é o caso dos dados de aplicação deste trabalho. Neste contexto, Shaddick e Wakefield (2002) utilizaram uma modelagem espaço-temporal para quatro poluentes medidos diariamente em oito pontos de monitorização, na cidade de Londres, ao longo de quatro anos. A modelagem foi conduzida de modo a investigar o efeito da poluição do ar na saúde, e para os efeitos aleatórios temporais dos poluentes foi atribuída como distribuição a priori um passeio aleatório, sendo que a implementação de parte deste modelo encontra-se disponível no módulo GeoBUGS, do OpenBUGS.

Considere $\boldsymbol{\varphi}$ como sendo a coleção de todos os parâmetros desconhecidos do modelo e $\boldsymbol{Y}$ a matriz de observações para a variável $\boldsymbol{Y}$ em cada caso. Em ambos os modelos $N=63$, microrregiões, e $K=5$ doenças. Para o segundo modelo $T=13$ anos de um período. Assim, a dimensão da matriz de observações $\boldsymbol{Y}$ no primeiro modelo é 63×5 e no segundo $63\times5\times13$. A diferença básica entre eles é a inclusão do domínio temporal no segundo modelo. Segue a descrição da inferência para um deles.

- Inferência sob o Modelo 1: Aqui, $\boldsymbol{\varphi}=(\boldsymbol{\alpha},\boldsymbol{\phi},\boldsymbol{\Lambda})$, sendo que $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5)'$ e $\boldsymbol{\phi}=$

$(\boldsymbol{\phi}_1', \boldsymbol{\phi}_2', \boldsymbol{\phi}_3', \boldsymbol{\phi}_4', \boldsymbol{\phi}_5')$. Pelo teorema de Bayes,

$$\begin{aligned} p(\boldsymbol{\varphi}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\varphi})p(\boldsymbol{\varphi}) \\ p(\boldsymbol{\varphi}|\boldsymbol{y}) &\propto \prod_{i=1}^{N}\prod_{k=1}^{K} \exp\left[y_{ik}\log(E_{ik}R_{ik}) - E_{ik}R_{ik} - \log(y_{ik}!)\right] \prod_{k=1}^{K} p(\alpha_k, \phi_k, \boldsymbol{\Lambda}) \\ &\propto \prod_{i=1}^{N}\prod_{k=1}^{K} \exp\left[y_{ik}\log(E_{ik}R_{ik}) - E_{ik}R_{ik} - \log(y_{ik}!)\right] \left[\prod_{k=1}^{K} p(\alpha_k)p(\phi_k|\boldsymbol{\Lambda})\right] p(\boldsymbol{\Lambda}). \end{aligned}$$

- Inferência sob o Modelo 2: Aqui, $\boldsymbol{\varphi} = (\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda})$, sendo que $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)'$, $\boldsymbol{\theta} = (\theta_{1,1}, ..., \theta_{5,13})$ e $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \boldsymbol{\phi}_2', \boldsymbol{\phi}_3', \boldsymbol{\phi}_4', \boldsymbol{\phi}_5')$. Pelo teorema de Bayes,

$$\begin{aligned} p(\boldsymbol{\varphi}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\varphi})p(\boldsymbol{\varphi}) \\ p(\boldsymbol{\varphi}|\boldsymbol{y}) &\propto \prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{k=1}^{K} \exp\left[y_{itk}\log(E_{itk}R_{itk}) - E_{itk}R_{itk} - \log(y_{itk}!)\right] \prod_{t=1}^{T}\prod_{k=1}^{K} p(\alpha_k, \theta_{kt}, \phi_k, W_k, \boldsymbol{\Lambda}) \\ &\propto \prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{k=1}^{K} \exp\left[y_{itk}\log(E_{itk}R_{itk}) - E_{itk}R_{itk} - \log(y_{itk}!)\right] \prod_{k=1}^{K} p(\alpha_k) \\ &\quad \left[\prod_{t=1}^{T}\prod_{k=1}^{K} p(\theta_{kt}|\theta_{k,t-1}, W_k)p(\phi_k|\boldsymbol{\Lambda})\right] p(\boldsymbol{\Lambda}). \end{aligned}$$

As distribuições a posteriori sob os modelos 1 e 2 apresentados possuem forma complexa e desconhecida. Assim, é necessária a utilização de métodos computacionalmente intensivos para fazer inferência a respeito dos parâmetros desconhecidos destes modelos. Neste trabalho é utilizado o método de MCMC, apresentado no Capítulo 4. O *software* utilizado foi o OpenBUGS. Este pacote estatístico já tem implementado internamente as rotinas para estimar os parâmetros via MCMC, cabendo ao usuário a especificação do modelo, das distribuições a priori e de valores iniciais para os hiperparâmetros. A partir de tais definições, e da construção das distribuições condicionais completas, o amostrador de Gibbs implementado no OpenBUGS permite o uso de diversas rotinas para amostrar de forma eficiente as distribuições condicionais. Caso não seja possível construir as distribuições condicionais completas, o *software* utiliza o algoritmo de Metropolis-Hastings com a distribuição proposta sendo Gaussiana e centrada no valor atual do parâmetro. Nos Apêndices A e B encontram-se, respectivamente, os códigos utilizados no OpenBUGS para realizar inferência acerca dos modelos 1 e 2 acima. É importante ressaltar que no OpenBUGS está disponível o código de um modelo para o mapeamento de duas doenças no oeste de Yorkshire, Reino Unido: câncer de cavidade oral e câncer de pulmão (Thomas *et al.* (2004)). Este código foi ampliado para implementar os modelos 1 e 2 considerados aqui.

Os dados de aplicação são referentes aos óbitos por 5 tipos de doenças: câncer de traqueia, brônquios e pulmão; câncer feminino de mama; câncer de estômago; câncer de lábios, cavidade oral e faringe; e câncer de cólon nas 63 microrregiões do estado de São Paulo, no decorrer do

período de 1998 a 2010. No modelo 1, que não leva em conta os espaços de tempo, o dado consiste na soma dos óbitos para todo o período em cada microrregião; no modelo 2 consiste na quantidade anual de óbitos para cada microrregião. Assim, no primeiro modelo Y é uma matriz composta de $63 \times 5 = 315$ valores, enquanto no segundo modelo Y é tridimensional, com $63 \times 5 \times 13 = 4095$ valores. Em ambos os casos também é necessário fornecer ao OpenBUGS os respectivos valores esperados para Y, que corresponde à mesma quantidade de valores deste vetor.

Para a estimação, foram realizadas 15000 iterações, sendo descartadas as 5000 iterações iniciais e armazenadas as 10000 posteriores, com um salto de 10 observações para melhorar a convergência. Esta foi monitorada através da análise da trajetória da cadeia, dos gráficos das autocorrelações, e de funções disponíveis no pacote CODA, que encontram-se no Apêndice C apenas para alguns dos parâmetros monitorados, uma vez que os demais possuem gráficos semelhantes e, portanto, levam à indicação de convergência de modo semelhante.

6.2 Aplicação do Modelo Hierárquico Bayesiano

Através do GeoBUGS, é possível mapear qualquer variável especificada para o modelo, de modo a visualizar sua dinâmica espacial. Aqui, no entanto, como o interesse principal é compreender a distribuição espacial dos riscos de cada doença, foram construídos os mapas para a média a posteriori do risco relativo associado a cada doença. Para averiguar a localização das microrregiões no estado de São Paulo mencionadas no decorrer desta seção, recorra ao mapa do Apêndice E. Antes de analisar os riscos relativos, porém, considere um resumo dos resultados obtidos a posteriori na Tabela 6.1.

Tabela 6.1: *Estimativas a posteriori dos parâmetros, erros padrão, e intervalos de credibilidade de 95%*

Parâmetro	Média	Erro Padrão	ICr(95%)
α_1	-0,1769	0,01939	(-0,2194; -0,1388)
α_2	-0,2439	0,02277	(-0,2890; -0,1977)
α_3	-0,0057	0,02211	(-0,0489; -0,0377)
α_4	-0,0973	0,00171	(-0,1319; -0,0637)
α_5	-0,0978	0,02033	(-0,1377; -0,0585)
σ_{ϕ_1}	0,9432	0,0937	(0,7802; 1,1470)
σ_{ϕ_2}	0,804	0,0865	(0,6533; 1,0991)
σ_{ϕ_3}	0,9581	0,0973	(0,7887; 1,1640)
σ_{ϕ_4}	0,6734	0,0715	(0,5481; 0,8281)
σ_{ϕ_5}	0,6699	0,0740	(0,5402; 0,8307)

Com respeito ao intercepto α_k para as doenças, atente para que o cálculo do risco relativo se

dá tal que $RR_{ik} = \exp(\alpha_k + \phi_{ik})$, de modo que $\exp(\alpha_1) = 0,84$, $\exp(\alpha_2) = 0,78$, $\exp(\alpha_3) = 0,99$, $\exp(\alpha_4) = 0,91$ e $\exp(\alpha_5) = 0,91$ é o quanto cada intercepto aumenta no risco relativo de cada doença. Sendo que o desejável é que os riscos não ultrapassem 1, o que indicaria que os óbitos estão ocorrendo de acordo com o esperado, conclui-se que estes termos são importantes para o risco e, portanto, relevantes ao modelo. Observe que tais termos são significativos, uma vez que seus intervalos de credibilidade não contém o valor 0.

Em relação aos termos de desvio dos efeitos σ_{ϕ_k}, os maiores foram observados para o câncer de lábios, cavidade oral e faringe ($\sigma_{\phi_3} = 0,9581$) e para câncer de traqueia, brônquios e pulmão ($\sigma_{\phi_1} = 0,9432$), respectivamente. Essas últimas estimativas são obtidas das colunas da matriz de covariâncias da distribuição MVCAR, que, embora sejam calculadas de modo conjunto, podem ser isoladas para a interpretação individual dos efeitos para cada doença. Assim, as doenças que tiveram maior variabilidade devido à sua estrutura espacial foram câncer de lábios, cavidade oral e faringe e câncer de traqueia, brônquios e pulmão.

6.2.1 Câncer de traqueia, brônquios e pulmão

Como se pode ver na Figura 6.1, o mapa indica que o risco de se morrer devido à câncer de traqueia, brônquios e pulmão no estado de São Paulo é maior no centro e no norte, pois é nessas regiões que se concentram as microrregiões pertencentes às classes de riscos mais altos do mapa, o que é coerente com os resultados obtidos na análise exploratória.

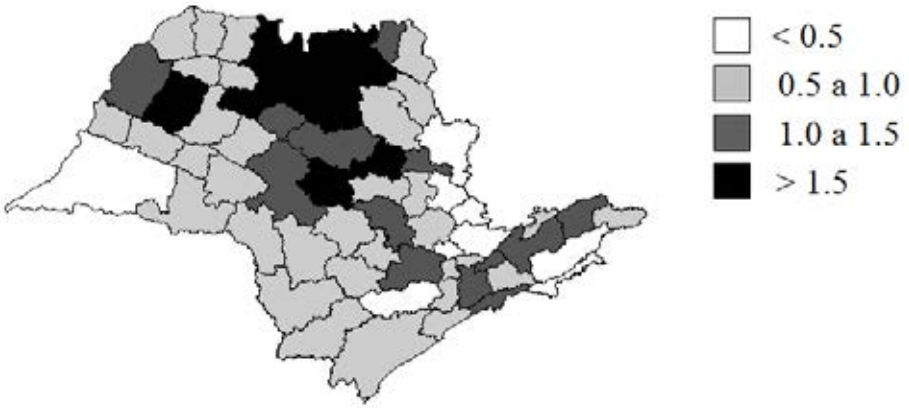

Figura 6.1: *Risco a posteriori obtido para o modelo referente aos óbitos por câncer de traqueia, brônquios e pulmão nas microrregiões do estado de São Paulo, de 1998 a 2010.*

Com exceção das microrregiões que contém hospitais de referência no tratamento do câncer (onde os riscos tendem a ser altos), destacam-se como tendo valores altos para o risco de mortalidade por câncer de traqueia, brônquios e pulmão a microrregião de Araçatuba, uma faixa no norte do estado que vai da microrregião de São José do Rio Preto até a de Ituverava, e uma faixa que vai das microrregiões de Santos até Guaratinguetá. Ao todo, 21 microrregiões apresentaram risco a posteriori maior que 1, e, destas, 9 tiveram as estimativas acima de 1,5.

Em comparação com as SMRs obtidas para a mesma doença, apresentadas no Capítulo anterior, percebe-se uma suavização nas estimativas obtidas pelo modelo bayesiano, o que facilita a compreensão dos riscos na região de estudo, bem como uma melhor identificação das microrregiões e de áreas com maior risco de óbito.

A Figura 6.2 apresenta os boxplots para as distribuições a posteriori dos riscos relativos associados a cada microrregião em estudo.

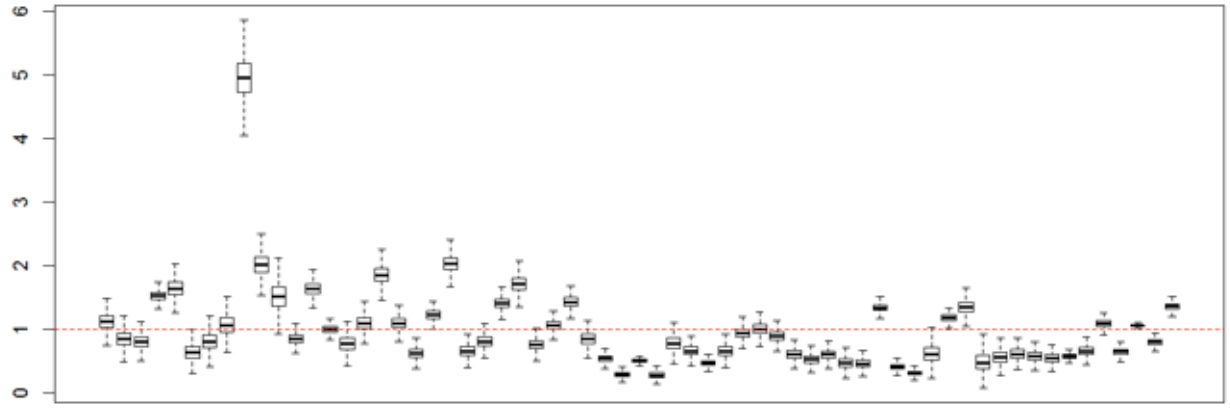

Figura 6.2: *Boxplots para as distribuições a posteriori dos riscos relativos associados ao câncer de traqueia, brônquios e pulmão, para cada microrregião do estado de São Paulo.*

6.2.2 Câncer de mama feminino

No caso desta doença, o resultado do modelo aponta para a existência de maior risco de óbito também no centro e norte do estado, e em algumas nos extremos da região, como mostra o mapa da Figura 6.3. Observe que esse padrão espacial foi identificado na análise exploratória. No entanto, este mapa é bem mais informativo do que o resultante da análise de agrupamentos, na Figura 5.8, cuja dispersão dos grupos dificultou um pouco a compreensão da dinâmica espacial.

Nota-se que a distribuição espacial dos riscos aqui é semelhante à do câncer de traqueia, brônquios e pulmão. No entanto, estas diferem na intensidade, sendo que o risco para o câncer de mama é bem menor, além de que nesta doença as estimativas do risco parecem estar distribuídas de modo mais aleatório. Naturalmente, o fator hereditário está fortemente ligado à manifestação desta doença, o que explica o fato da distribuição espacial de seus óbitos não estar tão bem definida quanto nas demais doenças consideradas aqui, que possivelmente sofrem maior interferência de fatores etiológicos encontrados no ambiente.

Ao todo, 19 microrregiões tiveram risco acima de 1, sendo que apenas Jaú e Barretos apresentaram estimativa maior que 1,5, risco este que não deve ser levado tão seriamente em conta, uma vez que ambas microrregiões possuem hospital de referência ao tratamento de câncer. Destacam-se as microrregiões de Assis e Marília, que tiveram risco maior que o esperado, bem como uma faixa de microrregiões que vai de Campinas até Santos, e as microrregiões de Araçatuba e Votuporanga também apresentaram riscos maior que o esperado.

A Figura 6.4 apresenta os boxplots para as distribuições a posteriori dos riscos relativos associados a cada microrregião em estudo.

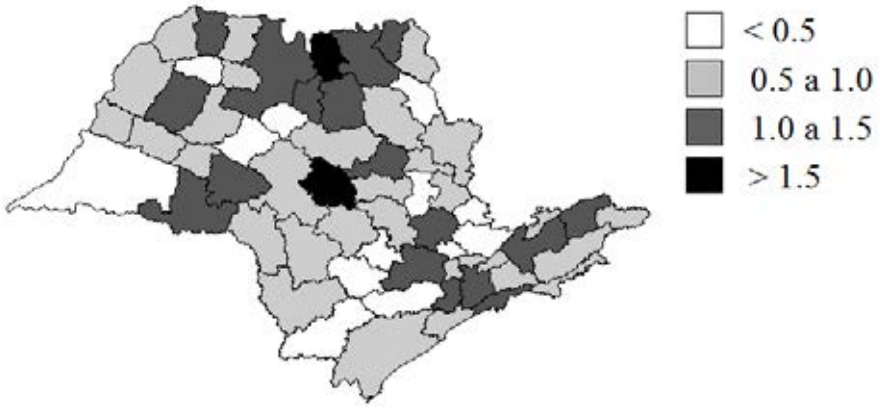

Figura 6.3: *Risco a posteriori obtido para o modelo referente aos óbitos por câncer feminino de mama nas microrregiões do estado de São Paulo, de 1998 a 2010.*

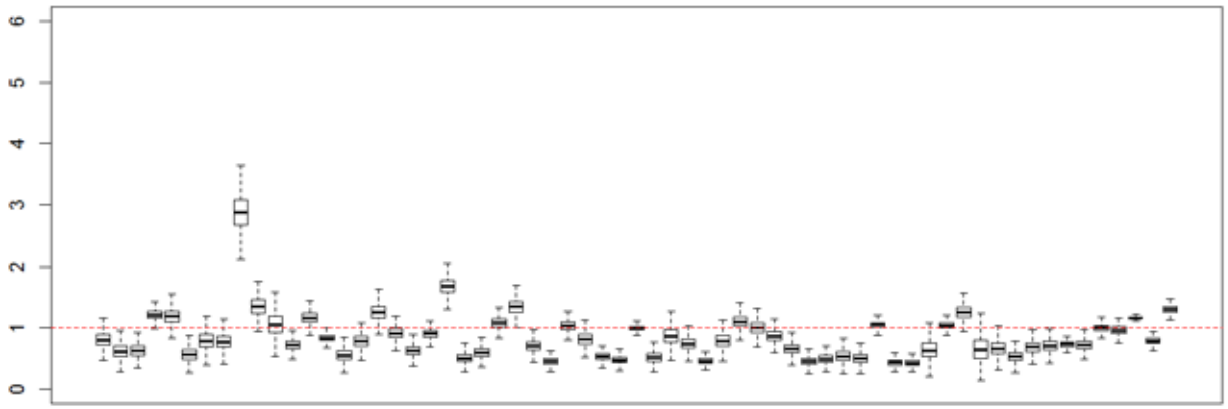

Figura 6.4: *Boxplots para as distribuições a posteriori dos riscos relativos associados ao câncer feminino de mama, para cada microrregião do estado de São Paulo.*

6.2.3 Câncer de lábios, cavidade oral e faringe

Note que para essa doença, a dinâmica espacial dos riscos a posteriori obtida é bastante semelhante à do câncer de traqueia, brônquios e pulmão, porém, com intensidade ainda maior do que aquela. Analisando o mapa a seguir, é fácil observar a presença de um grande grupo de microrregiões, presentes no centro e norte do estado com alto risco de óbito por câncer de lábios, cavidade oral e faringe. Dentro deste grande grupo, destacam-se ainda dois subgrupos mais agravantes, de microrregiões com risco mais de 50% maior do que o esperado. Merece atenção especial o grupo a esquerda do mapa, do qual fazem parte as microrregiões de Andradina, Araçatuba, Birigui, Auriflama e Fernandópolis, que para as demais doenças em estudo não tiveram riscos tão altos. Isso também se aplica às microrregiões de Jundiaí, Moji Mirim e Caraguatatuba, que isoladamente

apresentaram riscos elevados. Ao todo, 29 microrregiões tiveram risco maior que o esperado, das quais, 18 apresentaram risco maior que 1,5.

A Figura 6.6 apresenta os boxplots para as distribuições a posteriori dos riscos relativos associados a cada microrregião em estudo.

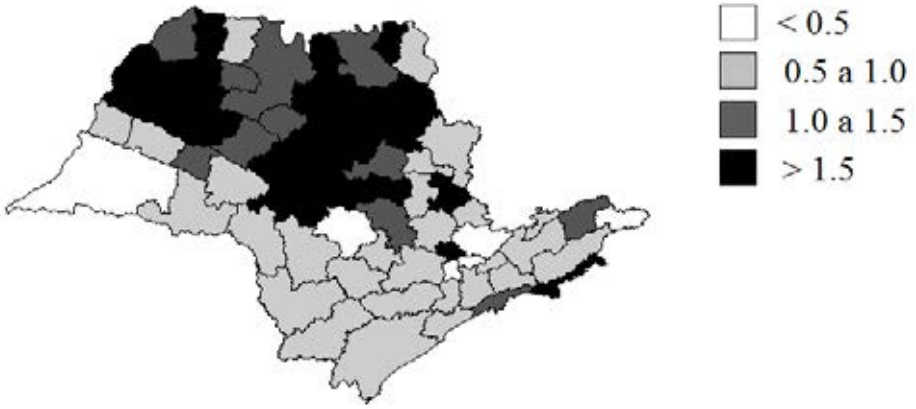

Figura 6.5: *Risco a posteriori obtido para o modelo referente aos óbitos por câncer de lábios, cavidade oral e faringe nas microrregiões do estado de São Paulo, de 1998 a 2010.*

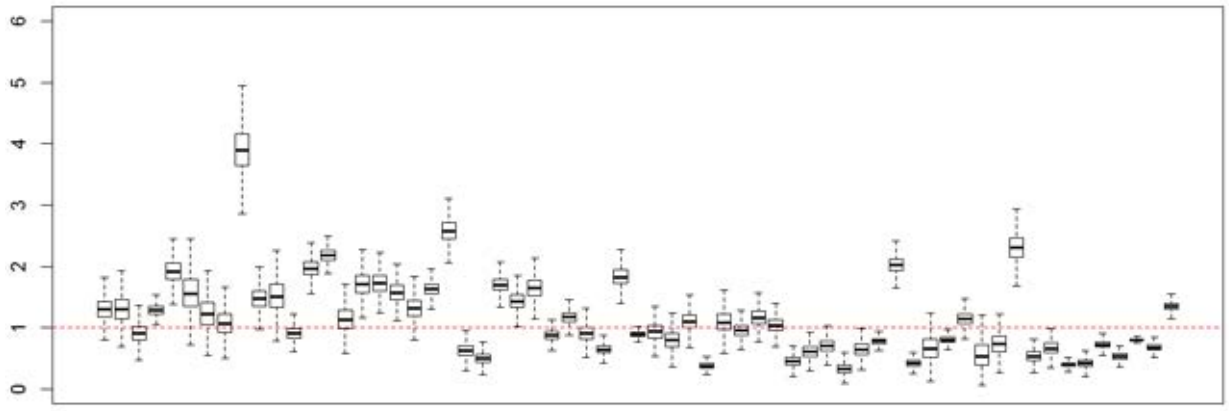

Figura 6.6: *Boxplots para as distribuições a posteriori dos riscos relativos associados ao câncer de lábios, cavidade oral e faringe, para cada microrregião do estado de São Paulo.*

6.2.4 Câncer de estômago

Como se pode ver no mapa da Figura 6.7, a distribuição espacial do risco para esta doença abrange altos riscos numa faixa que vai do norte ao sul do estado, sendo que das microrregiões que encontram-se na faixa de maiores riscos no mapa, destaca-se a de Guaratinguetá, uma vez que as demais microrregiões pertencentes a esta classe possuem hospitais de referência no tratamento do câncer (Barretos e Jaú) ou fazem fronteira com microrregiões com tal característica (São Carlos e São Joaquim da Barra). Esta distribuição espacial não se parece com a de nenhuma das outras doenças consideradas até aqui. Numa análise geral, 22 microrregiões apresentaram risco a posteriori maior que o esperado para sua estrutura demográfica e características de sua população, das quais 6 tiveram estimativa maior que 1,5.

A Figura 6.8 apresenta os boxplots para as distribuições a posteriori dos riscos relativos associados a cada microrregião em estudo.

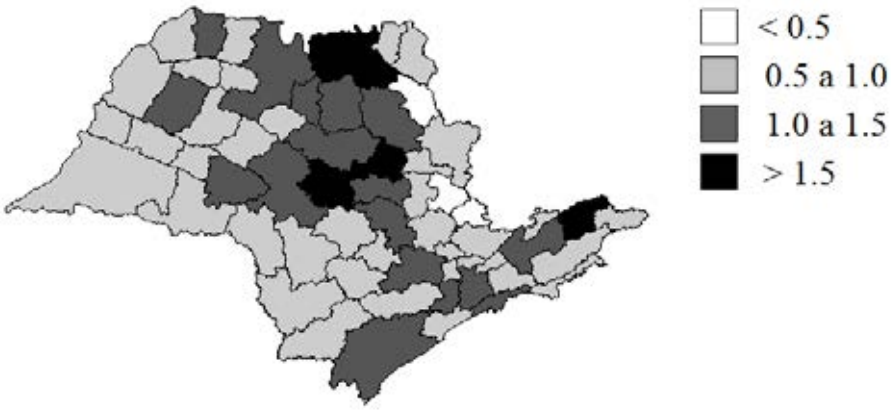

Figura 6.7: *Risco a posteriori obtido para o modelo referente aos óbitos por câncer de estômago nas microrregiões do estado de São Paulo, de 1998 a 2010.*

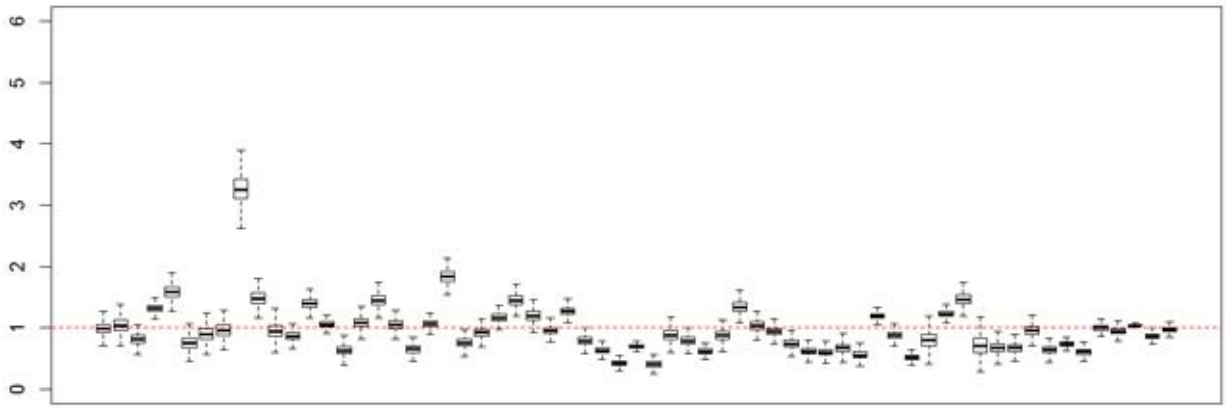

Figura 6.8: *Boxplots para as distribuições a posteriori dos riscos relativos associados ao câncer de estômago, para cada microrregião do estado de São Paulo.*

6.2.5 Câncer de Cólon

Ao comparar este mapa com o mapas das figura 6.7 é possível notar que a distribuição espacial do risco por câncer de cólon possui a área de maior risco quase que idêntica à obtida para o câncer de estômago, além de que as microrregiões pertencentes à última classe do mapa são as mesmas para as duas doenças. Isto é, além do centro e norte do estado apresentar altos riscos de óbito, microrregiões do sul e das extremidades do estado também estiveram na classe de riscos maiores que o esperado. A diferença entre as duas doenças consiste basicamente num conjunto de aproximadamente 5 microrregiões no noroeste do estado que tiveram risco maior que o esperado para esta doença, e para câncer de estômago não. Desta forma, espera-se que os óbitos por esses dois tipos de cânceres estejam correlacionados no espaço, ou seja, ocorrendo de forma semelhante no decorrer do estado, devido a algum fator de risco em comum entre as doenças.

A Figura 6.10 apresenta os boxplots para as distribuições a posteriori dos riscos relativos associados a essa doença, para cada microrregião em estudo.

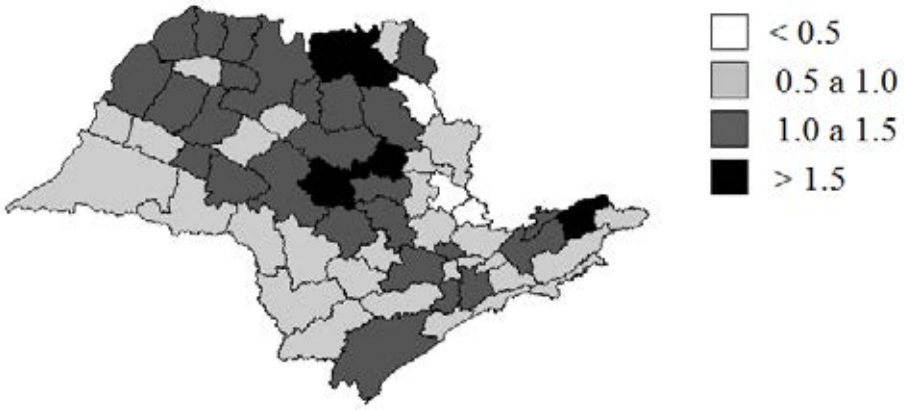

Figura 6.9: *Risco a posteriori obtido para o modelo 1 referente aos óbitos por câncer de cólon nas microrregiões do estado de São Paulo, de 1998 a 2010.*

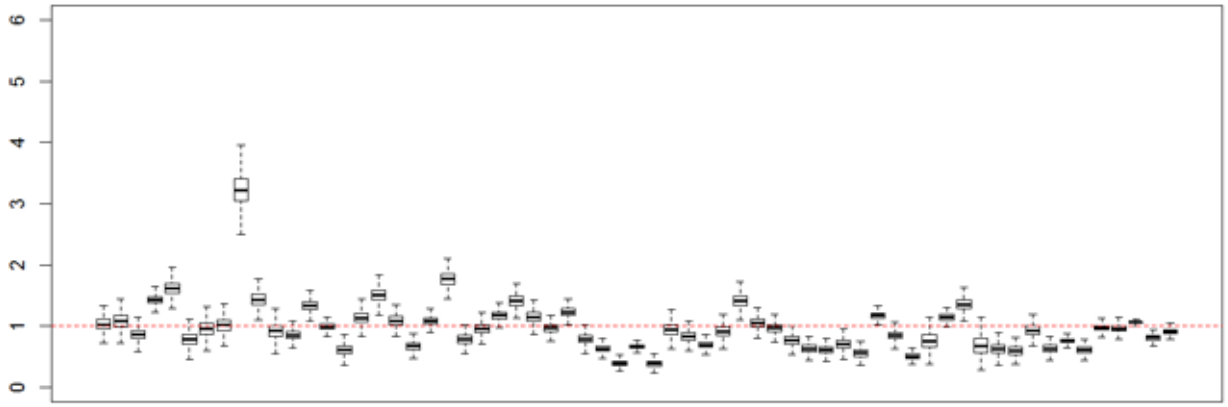

Figura 6.10: *Boxplots para as distribuições a posteriori dos riscos relativos associados ao câncer de cólon, para cada microrregião do estado de São Paulo.*

Considere agora como a ocorrência dos óbitos decorrentes das doenças em estudo estão corre-

lacionados no decorrer do estado de São Paulo. Apenas para resumir, 31 microrregiões apresentaram risco relativo de óbito para câncer de cólon acima do esperado para sua estrutura demográfica, das quais 6 tiveram estimativa acima de 1,5.

Levando em conta os resultados para cada doença, é estabelecido o seguinte *ranking*, segundo a importância dos riscos, devido a quantidade de microrregiões com valores maiores que 1 e que 1,5 para o risco relativo. Observe que esse *ranking* difere do estabelecido na análise exploratória, pois os cânceres de cólon e de mama trocam as posições, sendo que da análise realizada aqui fica evidente que a colocação do câncer de cólon acima do de mama é muito mais condizente com a informação fornecida pelos dados.

Tabela 6.2: *Ranking das doenças de acordo com os resultados do modelo hierárquico bayesiano*

Posição	Doença	RR > 1	RR > 1,5
1º	Câncer de lábios, cavidade oral e faringe	29	18
2º	Câncer de traqueia, brônquios e pulmão	21	9
3º	Câncer de cólon	31	6
4º	Câncer de estômago	22	6
5º	Câncer de mama	19	2

6.2.6 Correlação a posteriori para as doenças

Como mencionado, um dos atrativos do uso da priori MVCAR é a possibilidade de modelar a correlação entre as variáveis, neste caso, os riscos das doenças. Seja Υ_1 a matriz de correlação para os riscos deste modelo, obtida a partir da respectiva matriz $\mathbf{\Lambda}$ do modelo. Os valores obtidos pelo modelo para esta matriz foram

$$\Upsilon_1 = \begin{bmatrix} 1,0000 & 0,7704 & 0,3093 & 0,8392 & 0,8348 \\ 0,7704 & 1,0000 & 0,2985 & 0,7362 & 0,7213 \\ 0,3093 & 0,2985 & 1,0000 & 0,3804 & 0,3190 \\ 0,8392 & 0,7362 & 0,3804 & 1,0000 & 0,9775 \\ 0,8348 & 0,7213 & 0,319 & 0,9775 & 1,0000 \end{bmatrix}$$

Nota-se forte correlação entre câncer de traqueia, brônquios e pulmão *versus* câncer de estômago (0,8392); câncer de traqueia, brônquios e pulmão *versus* câncer de cólon (0,8348), e câncer de estômago *versus* câncer de cólon (0,9775), sendo que este último resultado não surpreende, dada a semelhança notada entre os mapas para tais doenças, como mencionado no tópico anterior. Estes resultados são intrigantes, levando à hipótese de existência de fatores de risco em comum para tais doenças. Uma perspectiva futura para continuidade deste trabalho é inserir covariáveis no modelo, com o objetivo de identificar características das regiões com maiores riscos.

De acordo com o artigo de Guerra *et al.* (2005), todos os tipos de câncer estudados neste trabalho tem sua manifestação associada à exposição a um grande número de fatores de riscos ambientais relacionados ao processo de industrialização - agentes químicos, físicos e biológicos - e de exposição a outros fatores relacionados às disparidades sociais, o que explicaria parte das correlações observadas entre estas doenças. Segundo os autores, o tabagismo, por exemplo, contribui não somente para o aumento da ocorrência de câncer de traqueia, brônquios e pulmão no país, mas também para a incidência de outros tipos de câncer, tais como câncer de estômago e câncer de lábios, cavidade oral e faringe, principalmente se associado a consumo de álcool e precárias condições de saúde, outros fatores de risco muito comuns no Brasil.

Por outro lado, tanto câncer de estômago, como câncer de mama, e de cólon relacionam-se a hábitos dietéticos, e a um status sócio-econômico elevado, observado principalmente na região sudeste do país, indicando a possível importância de uma variável como o IDH (Índice de Desenvolvimento Humano) em explicar parte da dependência entre tais doenças.

Ainda segundo os autores, no Brasil, o aumento de doenças relacionadas ao hábito do fumo pode ser explicado, em parte, pela aceleração no consumo do tabaco no decorrer dos anos e a difusão do tabagismo na população feminina. Sendo esta uma das principais causas associadas aos óbitos pelos cânceres estudados neste trabalho, um interesse que pode surgir acerca dos dados diz respeito à sua ocorrência no decorrer do tempo. Uma análise desta natureza pode elucidar quais momentos no decorrer de um período em estudo se mostraram determinantes na manifestação de altos riscos para determinada doença. Assim, na próxima seção são apresentados os resultados da aplicação do Modelo Hierárquico Bayesiano Dinâmico, como instrumento para estudar os dados ao longo do tempo.

6.3 Aplicação do Modelo Hierárquico Bayesiano Dinâmico

A Tabela 6.3 apresenta os resultados obtidos para um conjunto de parâmetros do modelo 2. Nota-se que os valores obtidos para os interceptos foram muito próximos aos obtidos no modelo anterior, o que, consequentemente, leva às mesmas interpretações. A variabilidade dos efeitos aleatórios espaciais, monitorada através dos valores de σ_{ϕ_k} , se mostrou maior também para câncer de lábios, cavidade oral e faringe (0,9567), seguido pelo câncer de traqueia, brônquios e pulmão (0,942), sendo que estes, bem como os demais valores de σ_{ϕ_k} também ficaram muito parecidos com os obtidos no modelo 1.

Note que o valor obtido para este parâmetro referente a câncer de cólon foi o menor entre as doenças (0,6679), indicando que a variabilidade dos óbitos se deve menos à sua estrutura espacial nesta doença do que nas outras.

Com respeito aos efeitos temporais, que é o que difere este modelo do anterior, todos os valores para σ_{θ_k} ficaram em torno de $0,07$, e se mostraram significativos. No entanto, nenhum dos valores obtidos para os próprios θ_k (ver Apêndice D) foi significativo para explicar os riscos relativos. Este fato, somado à semelhança entre os demais parâmetros monitorados nos dois modelos

Tabela 6.3: *Estimativas a posteriori dos parâmetros, erros padrão, e intervalos de credibilidade de 95%*

Parâmetro	Média	Erro Padrão	ICr(95%)
α_1	-0,1776	0,01953	(-0,2165; -0,1396)
α_2	-0,2450	0,02239	(-0,2892; -0,2016)
α_3	-0,0060	0,02223	(-0,0489; -0,0369)
α_4	-0,0983	0,01769	(-0,1336; -0,0642)
α_5	-0,0979	0,02105	(-0,1397; -0,0572)
σ_{ϕ_1}	0,9420	0,0937	(0,7779; 1,1460)
σ_{ϕ_2}	0,8043	0,0865	(0,6514; 0,9930)
σ_{ϕ_3}	0,9567	0,0981	(0,7842; 1,1700)
σ_{ϕ_4}	0,6749	0,0728	(0,5489; 0,8321)
σ_{ϕ_5}	0,6679	0,0743	(0,5385; 0,8269)
$\sigma_{\theta_{1\cdot}}$	0,07714	0,0177	(0,05124; 0,1192)
$\sigma_{\theta_{2\cdot}}$	0,07859	0,0177	(0,05222; 0,1199)
$\sigma_{\theta_{3\cdot}}$	0,07982	0,0184	(0,05291; 0,1240)
$\sigma_{\theta_{4\cdot}}$	0,07747	0,0175	(0,05152; 0,1186)
$\sigma_{\theta_{5\cdot}}$	0,08084	0,0189	(0,05280; 0,1260)

indica que a inclusão de um efeito aleatório temporal distribuído segundo um passeio aleatório de ordem um não é importante para explicar os riscos, o que leva à conclusão de que ou o período de tempo considerado neste estudo não é grande o suficiente para que a variação temporal possa ser capturada desta forma, ou os óbitos em questão variam segundo outro tipo de distribuição.

Referente às correlações entre as doenças obtidas neste modelo, considere a matriz Υ_2. Observe que os valores das correlações não diferem significativamente dos obtidos para o modelo anterior. Isto é mais um indicativo de que este modelo não está capturando mais informação a respeito da ocorrência simultânea dos óbitos pelas doenças em estudo. No entanto, ainda existe a vantagem de que ele possibilita obter as estimativas do risco em cada nível de tempo pertencente ao período em estudo, isto é, torna possível observar o risco de morte no estado de São Paulo em cada ano do período, o que pode ser visto nas Figuras 6.11 a 6.15.

Note que agora é possível compreender quando determinadas microrregiões tiveram riscos altos, já que para algumas delas isso não aconteceu em todo o período. Por exemplo, na Figura 6.11 é possível observar que a microrregião de Araraquara teve risco bastante alto apenas em 2003 e 2004, mantendo estabilidade nos demais anos do período. Assim, os altos riscos foram atípicos nesse local, não devendo ser de séria preocupação sua ocorrência; exceto se manifestada novamente.

De modo geral, os riscos se comportaram da mesma maneira que no modelo anterior, apenas

neste caso tem-se a vantagem de analisá-los individualmente ano a ano, como já mencionado. Em relação às SMRs, é visível, através dessas Figuras, que as estimativas dos riscos obtidas por este modelo são mais suaves e de maior facilidade de interpretação do que as estimativas clássicas apresentadas no capítulo anterior.

Novamente, é possível estabelecer um *ranking* de gravidade das doenças, com base nos riscos obtidos a posteriori.

Tabela 6.4: *Ranking das doenças de acordo com os resultados do modelo hierárquico bayesiano dinâmico*

Posição	Doença
1º	Câncer de lábios, cavidade oral e faringe
2º	Câncer de traqueia, brônquios e pulmão
3º	Câncer de cólon
4º	Câncer de estômago
5º	Câncer de mama

Observe que este *ranking* coincide com o que fora obtido no modelo anterior (Tabela 6.2), reforçando a conclusão de que ambos os modelos capturam a estrutura dos riscos relativos de modo similar.

$$\Upsilon_2 = \begin{bmatrix} 1,0000 & 0,7677 & 0,3089 & 0,8384 & 0,8356 \\ 0,7677 & 1,0000 & 0,2977 & 0,7358 & 0,7195 \\ 0,3089 & 0,2977 & 1,0000 & 0,3757 & 0,3198 \\ 0,8384 & 0,7358 & 0,3757 & 1,0000 & 0,9779 \\ 0,8356 & 0,7195 & 0,3198 & 0,9779 & 1,0000 \end{bmatrix}$$

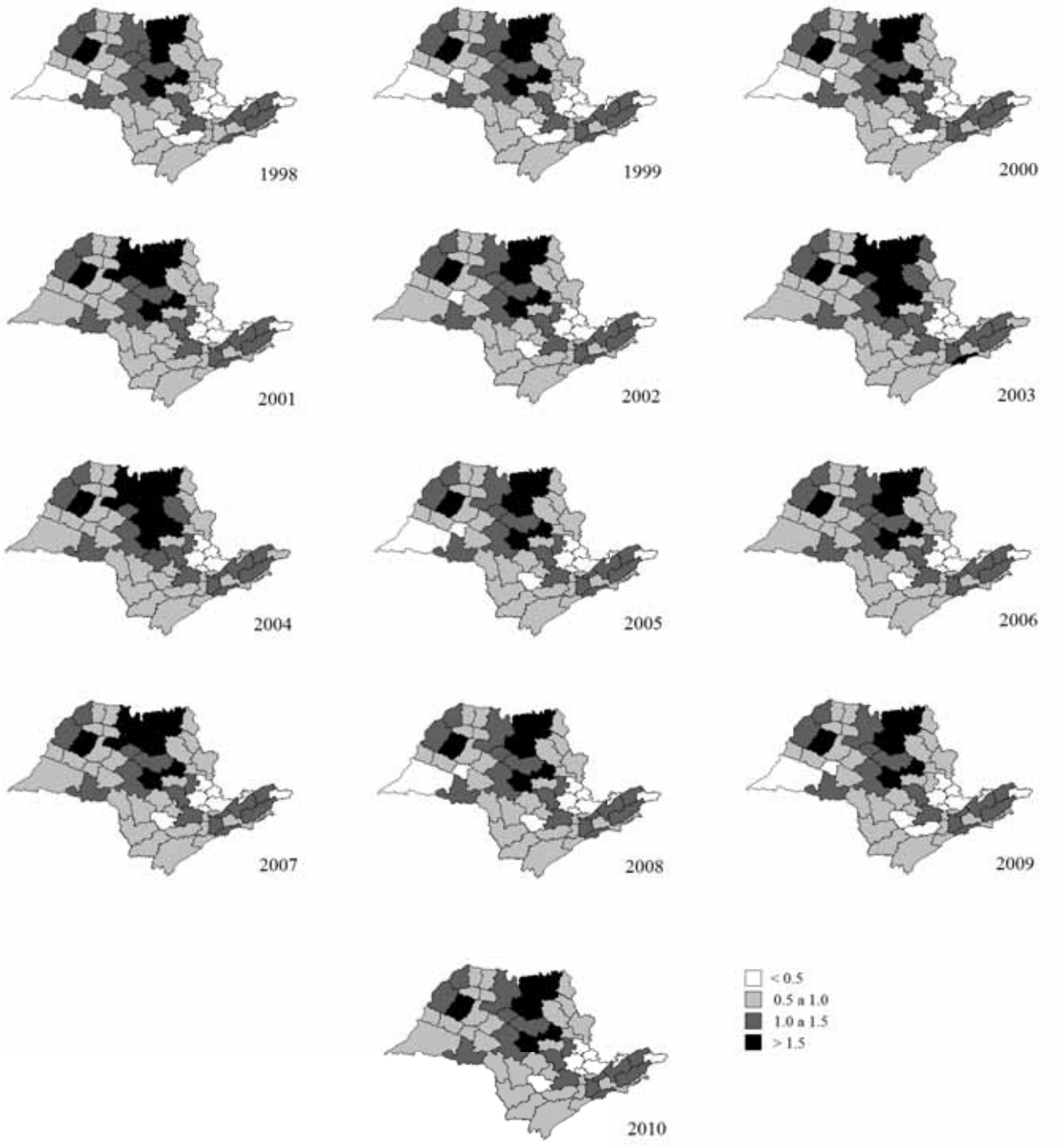

Figura 6.11: *Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de traqueia, brônquios e pulmão nas microrregiões do estado de São Paulo, de 1998 a 2010.*

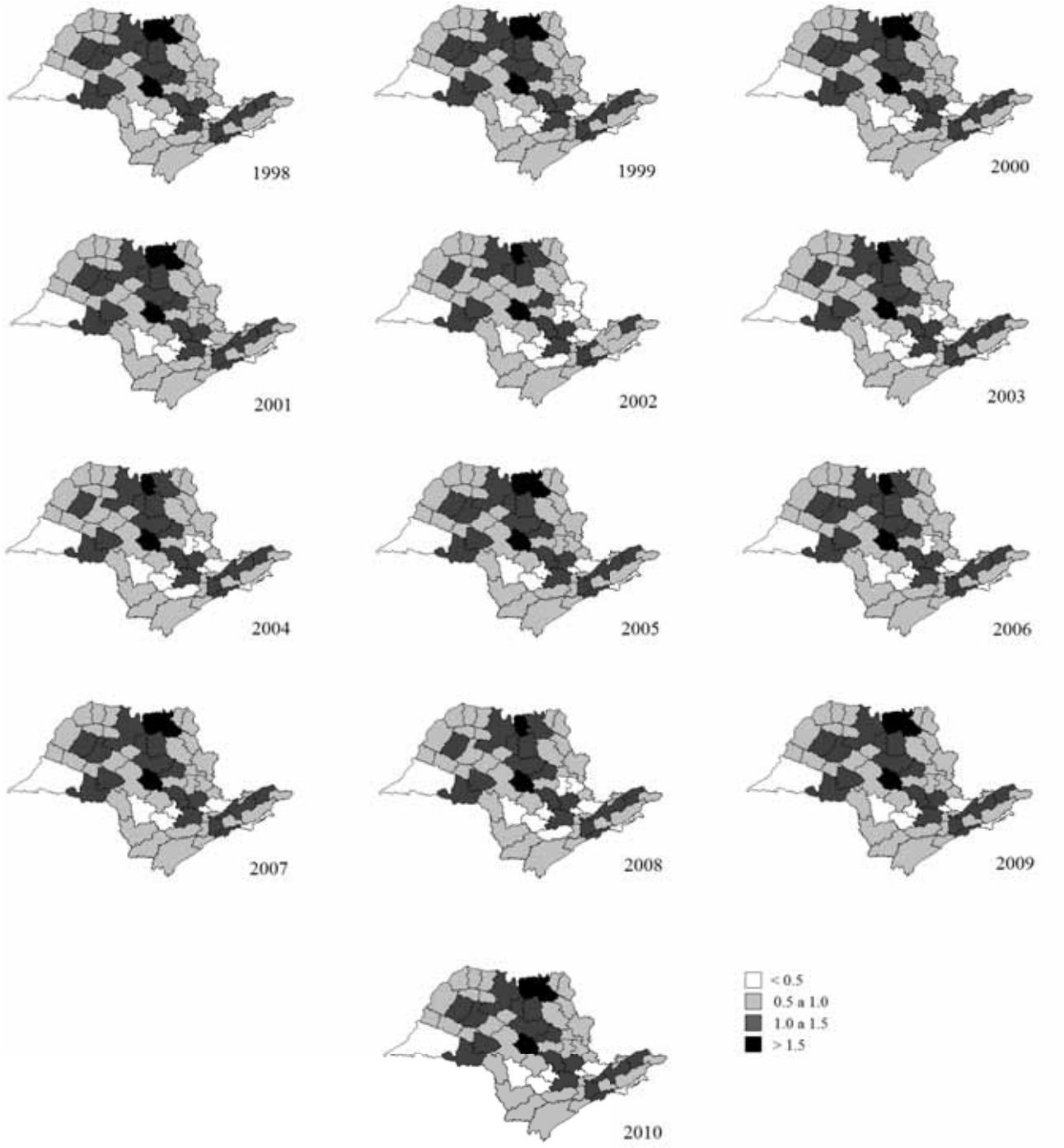

Figura 6.12: *Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de mama feminino nas microrregiões do estado de São Paulo, de 1998 a 2010.*

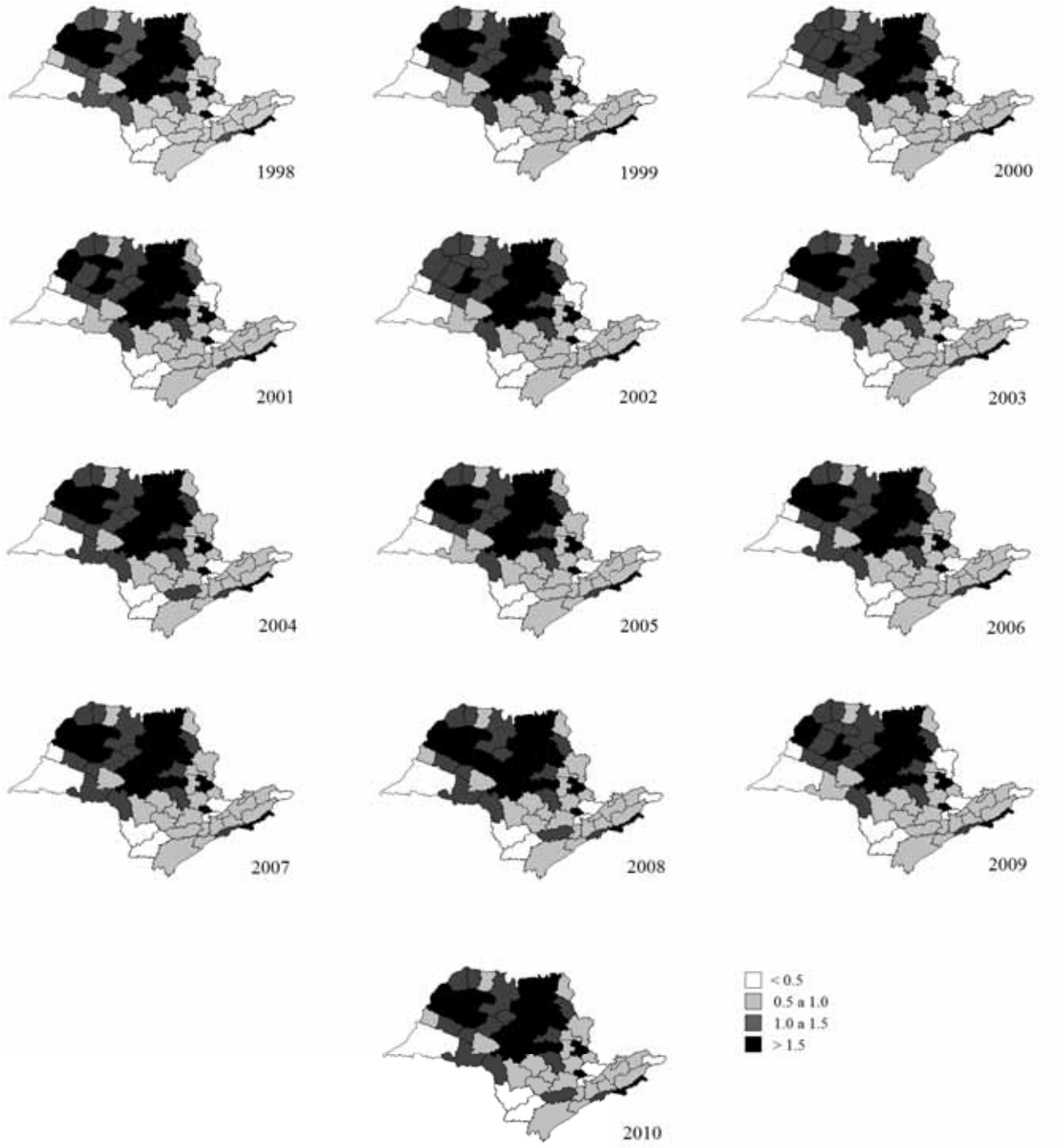

Figura 6.13: *Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de lábios, cavidade oral e faringe nas microrregiões do estado de São Paulo, de 1998 a 2010.*

Figura 6.14: *Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de estômago nas microrregiões do estado de São Paulo, de 1998 a 2010.*

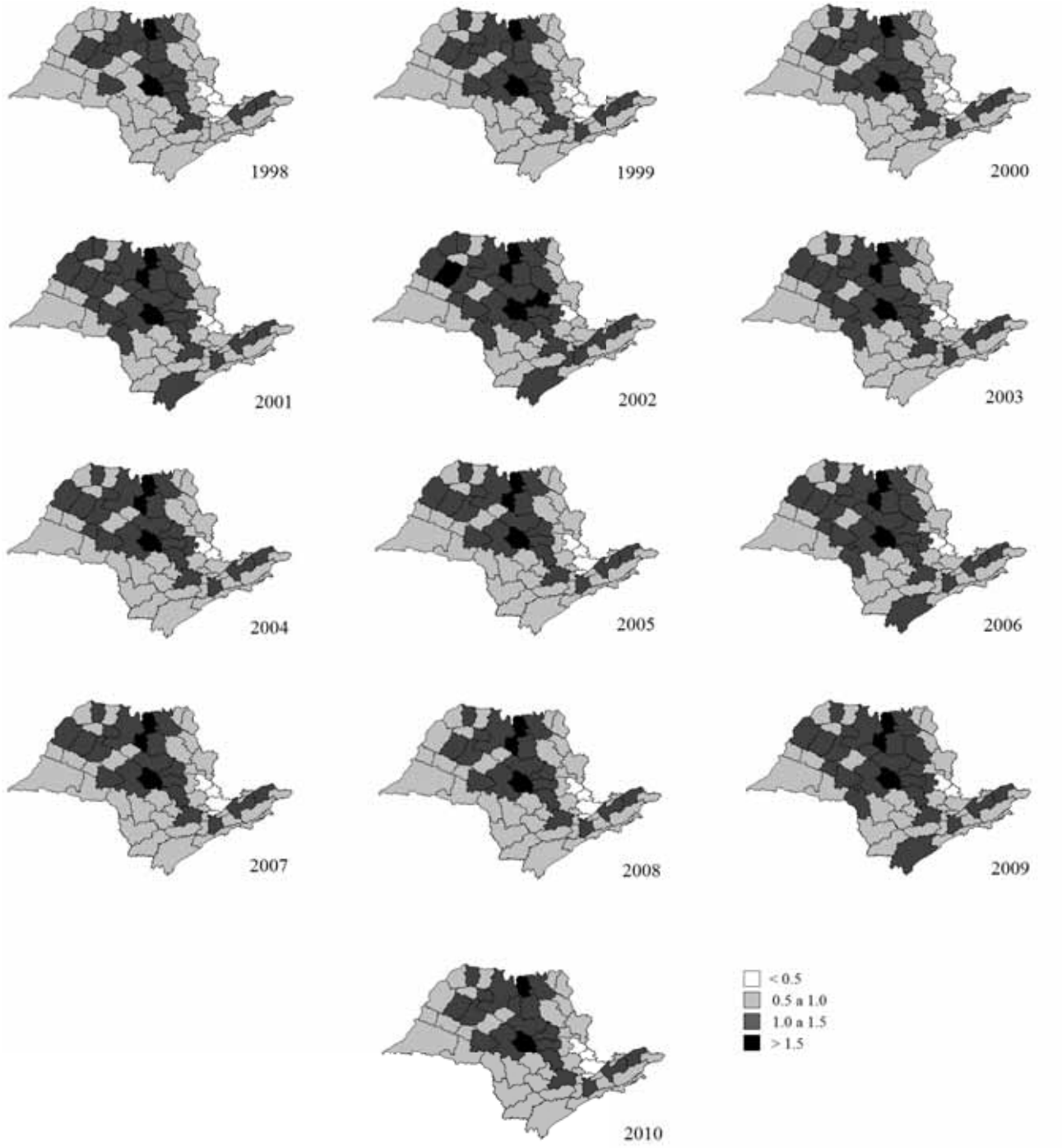

Figura 6.15: *Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de cólon nas microrregiões do estado de São Paulo, de 1998 a 2010.*

6.4 Comparação dos modelos através do Critério DIC (Deviance Information Criterion)

O critério DIC (Deviance Information Criterion (Spiegelhalter *et al.* (2002))) é uma generalização dos critérios AIC (Akaike information criterion) e BIC (Bayesian information criterion). É particularmente útil em problemas de seleção de modelos bayesianos nos quais as distribuições a posteriori são obtidas através de simulação via MCMC. Assim como os critérios AIC e BIC, consiste em uma aproximação assintótica e só é válido quando a distribuição a posteriori é aproximadamente Normal multivariada. Gelman *et al.* (2013b) apresentam o estado da arte no contexto de seleção bayesiana de modelos e concluem que os critérios propostos até o momento não são suficientemente eficientes. Comparam resultados para o AIC, *Deviance* e WAIC (Watanabe AIC) em três exemplos. Embora o DIC seja criticado por não ser adequado para casos em que a distribuição a posteriori é assimétrica, pois seu cálculo envolve a média a posteriori, está é a medida utilizada neste trabalho.

Para o cálculo do DIC, defina o desvio $D(\varphi) = -2\log(p(\boldsymbol{y}|\varphi)) + c$, em que $\boldsymbol{y}$ é o vetor de dados, φ é o vetor de parâmetros de interesse sob determinado modelo, $p(\boldsymbol{y}|\varphi)$ é a função de verossimilhança, e c é uma constante cancelada nos cálculos que comparam diferentes modelos e, portanto, não precisa ser conhecida. A esperança $\bar{D} = E_{\varphi|\boldsymbol{y}}[D]$ é uma medida de quão bem o modelo se ajusta aos dados. Quanto maior esta for, pior o ajuste.

Existem dois cálculos utilizados comumente para encontrar a quantidade efetiva de parâmetros do modelo. O primeiro, conforme descrito em Spiegelhalter *et al.* (2002) é $p_D = \bar{D} - D(\bar{\varphi})$, em que $\bar{\varphi}$ é a esperança de φ. O segundo, tal como descrito em Gelman *et al.* (2013a) é $p_D = p_V = \frac{1}{2}\widehat{var}(D(\varphi))$. Quanto maior o número efetivo de parâmetros é, mais fácil é para o modelo ajustar os dados, com isso o desvio tem de ser penalizado.

O DIC é então calculado como

$$DIC = p_D + \bar{D},$$

ou, equivalentemente como

$$DIC = D(\bar{\varphi}) + 2p_D.$$

A ideia é que os modelos com menor DIC devem ser preferidos à modelos com maiores valores para essa estatística. Os modelos são penalizados tanto pelo valor de $\bar{D}$, o que favorece um bom ajuste, como também (em comum com AIC e BIC) pelo número efetivo de parâmetros p_D. Uma vez que $\bar{D}$ diminui à medida que o número de parâmetros em um modelo aumenta, p_D compensa este efeito, favorecendo modelos com um número menor de parâmetros.

A vantagem do DIC em relação a outros critérios para a seleção de um modelo é que este é facilmente calculado a partir das amostras geradas por uma simulação de MCMC, ao passo que o AIC e o BIC exigem o cálculo do máximo da verossimilhança sobre φ, que não está prontamente

disponível a partir da simulação MCMC. Por outro lado, para calcular DIC basta simplesmente calcular $\bar{D}$ como sendo a média de $D(\varphi)$ sobre as amostras de φ, e $D(\bar{\varphi})$ como o valor de D avaliado na média das amostras de φ. Por fim, o DIC segue a partir dessas aproximações.

Os valores para o DIC fornecidos pelo OpenBUGS para os modelos estudados neste trabalho foram de 2186 para o modelo 1 e 19510 para o modelo 2, induzindo a que deve-se preferir o modelo 1 ao modelo 2 para ajustar os dados de câncer considerados. A discrepância entre tais valores da estatística evidencia que o modelo 2, devido à inclusão dos efeitos temporais θ_{kt}, é penalizado por um excesso de parâmetros que não melhoram significativamente o conhecimento acerca dos riscos. Apesar disso, a discussão realizada na seção anterior é válida.

Capítulo 7

Conclusões e perspectivas futuras

Como já mencionado, a proposta de trabalho para esta dissertação consistiu no estudo da classe de modelos hierárquicos dinâmicos aplicados a dados de área multivariados. Isto é, a proposta envolveu o estudo de modelos que permitem incorporar em sua estrutura as dimensões espaço e tempo. Neste contexto, este relatório apresenta uma introdução sobre os modelos clássico de riscos relativos, de Poisson com efeitos aleatórios associados à estrutura espacial e com evolução temporal dos parâmetros. Para o procedimento de inferência, no modelo clássico de riscos relativos utiliza-se estimadores de máxima verossimilhança e para o modelo de Poisson a abordagem Bayesiana. Neste último caso, a distribuição a posteriori conjunta não apresenta forma fechada e métodos numéricos são necessários. Nesta etapa do trabalho utilizou-se o amostrador de Gibbs implementado através do Software OpenBUGS.

Os dados escolhidos para a aplicação foram os óbitos pelos cinco cânceres de maior letalidade nas microrregiões do estado de São Paulo, registrados para o período de 1998 a 2010. Inicialmente, apresentou-se uma análise exploratória dos dados, consistindo no mapeamento das estimativas de máxima verossimilhança obtidas através do modelo clássico de riscos relativos (SMRs) e numa análise de agrupamento das microrregiões de acordo com estas. Ao agrupar as SMRs, percebeu-se a presença de grupos de microrregiões com comportamentos particulares. Observando seu comportamento no tempo, ficou evidente que, para algumas doenças, e alguns grupos de microrregiões, houve tendência temporal na ocorrência dos óbitos no decorrer do período em estudo.

Além disso, aplicou-se o modelo de Poisson nos dados agrupados para todo o período, especificando como distribuição a priori para os efeitos aleatórios o modelo CAR intrínseco, uma alternativa da classe de modelos condicionais autorregressivos (CAR) proposta por Besag (1974), que inclui também o Modelo de convolução, sendo que este último, embora tenha sido apresentado no Capítulo 3, não foi implementado devido ao fato de não ser considerado superior ao CAR intrínseco para a estimação dos riscos. Quanto a estes, ao longo do texto foram consideradas suas características e atrativos. Como os dados utilizados na aplicação são multivariados, a versão multivariada do modelo ICAR (Besag e Kooperberg (1995)), o MVCAR, foi aplicada para modelar os efeitos aleatórios espaciais dos dados. Os resultados evidenciaram forte autocorrelação entre os efeitos espaciais de câncer de pulmão e câncer de estômago, câncer de pulmão e câncer de cólon;

e câncer de estômago e câncer de cólon. Numa segunda etapa do trabalho foi aplicado o Modelo Hierárquico Bayesiano Dinâmico, que possibilitou aprofundar o conhecimento dos riscos relativos das doenças em estudo, com respeito à sua ocorrência no decorrer do tempo. De um modo geral, as estimativas obtidas não diferiram significativamente das obtidas para o modelo sem a dimensão temporal. Com isso, apesar de a proposta para este trabalho poder ser considerada cumprida, uma perspectiva futura para aprofundar o conhecimento nos modelos em estudo é trabalhar com a inclusão de covariáveis neste modelo, bem como nas implicações teóricas a que tal mudança leva, e nas vantagens de tal abordagem com respeito à capturar a estrutura espacial comum entre as doenças.

Referências Bibliográficas

Assunção e Krainski (2009) Renato Assunção e Elias Krainski. Neighborhood dependence in bayesian spatial models. *Biometrical Journal*, 51(5):851–869. Citado na pág. 21

Assunção e Castro (2004) Renato M Assunção e Mônica SM Castro. Multiple cancer sites incidence rates estimation using a multivariate bayesian model. *International journal of epidemiology*, 33(3):508–516. Citado na pág. 6

Besag (1974) Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, páginas 192–236. Citado na pág. 9, 14, 17, 18, 79

Besag e Kooperberg (1995) Julian Besag e Charles Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746. Citado na pág. 18, 20, 79

Besag *et al.* **(1991)** Julian Besag, Jeremy York e Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1): 1–20. Citado na pág. 8, 14, 19, 20

Best *et al.* **(2005)** Nicky Best, Sylvia Richardson e Andrew Thomson. A comparison of bayesian spatial models for disease mapping. *Statistical methods in medical research*, 14(1):35–59. Citado na pág. 8

Best *et al.* **(1999)** Nicola G Best, Katja Ickstadt e Robert L Wolpert. Spatial poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American statistical association*, 95(452):1076–1088. Citado na pág. 6

Carlin e Banerjee (2003) Bradley P Carlin e Sudipto Banerjee. Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian statistics*, 7:45–63. Citado na pág. 21

Clayton e Kaldor (1987) David Clayton e John Kaldor. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, páginas 671–681. Citado na pág. 6

Congdon (2007) Peter Congdon. *Bayesian statistical modelling*, volume 704. Wiley. com. Citado na pág. 6, 20

Cressie (1993) Noel AC Cressie. *Statistics for Spatial Data, revised edition*. Wiley, New York. Citado na pág. 15, 16

DATASUS (Visitada em junho/2013) DATASUS. Sistema de informaÃ§ao de mortalidade (sim). URL http://200.214.130.44/sim/default.asp. Citado na pág. 29

Eberly *et al.* **(2000)** Lynn E Eberly, Bradley P Carlin *et al.* Identifiability and convergence issues for markov chain monte carlo fitting of spatial models. *Statistics in Medicine*, 19(1718):2279–2294. Citado na pág. 20, 21

Fahrmeir e Lang (2001) Ludwig Fahrmeir e Stefan Lang. Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220. Citado na pág. 59

Gelfand e Smith (1990) Alan E Gelfand e Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409. Citado na pág. 27

Gelfand e Vounatsou (2003) Alan E Gelfand e Penelope Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15. Citado na pág. 22

Gelman *et al.* **(2013a)** Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari e Donald B Rubin. *Bayesian data analysis*. CRC press. Citado na pág. 77

Gelman *et al.* **(2013b)** Andrew Gelman, Jessica Hwang e Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, páginas 1–20. Citado na pág. 77

Geman e Geman (1984) Stuart Geman e Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741. Citado na pág. 25

Guerra *et al.* **(2005)** Maximiliano Ribeiro Guerra, CV de M Gallo, GAS Mendonça e GA Silva. Risco de câncer no brasil: tendências e estudos epidemiológicos mais recentes. *Rev bras cancerol*, 51(3):227–34. Citado na pág. 69

Harrison e Stevens (1976) P Jeffrey Harrison e Colin F Stevens. Bayesian forecasting. *Journal of the Royal Statistical Society. Series B (Methodological)*, páginas 205–247. Citado na pág. 10

Hastings (1970) W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109. Citado na pág. 24

Hilbe (2011) Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press. Citado na pág. 6

Kramer e Williamson (2013) Michael R. Kramer e Rebecca Williamson. Multivariate bayesian spatial model of preterm birth and cardiovascular disease among georgia women: Evidence for life course social determinants of health. *Spatial and Spatio-temporal Epidemiology*, 6(0):25 – 35. Citado na pág. 21

Künsch (1987) Hans R Künsch. Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika*, 74(3):517–524. Citado na pág. 18

Lang e Brezger (2000) Stefan Lang e Andreas Brezger. Bayesx-software for bayesian inference based on markov chain monte carlo simulation techniques. Citado na pág. 9

Lawson (2008) Andrew B Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*, volume 32. CRC Press. Citado na pág. 7, 8, 22

Lunn *et al.* **(2009)** David Lunn, David Spiegelhalter, Andrew Thomas e Nicky Best. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067. Citado na pág. 1

Lunn *et al.* **(2000)** David J Lunn, Andrew Thomas, Nicky Best e David Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337. Citado na pág. 9

Metropolis *et al.* **(1953)** Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller e Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087. Citado na pág. 24

Mollié (1996) Annie Mollié. Bayesian mapping of disease. *Markov chain Monte Carlo in practice*, 1:359–379. Citado na pág. 13

Ripley (2005) Brian D Ripley. *Spatial statistics*, volume 575. Wiley. com. Citado na pág. 15

Rodrigues e Assunção (2012) Erica Castilho Rodrigues e R Assunção. Bayesian spatial models with a mixture neighborhood structure. *Journal of Multivariate Analysis*, 109:88–102. Citado na pág. 21

Shaddick e Wakefield (2002) Gavin Shaddick e Jon Wakefield. Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):351–372. Citado na pág. 58, 59

SILVA *et al.* **(2011)** APR de SILVA, CP Noronha, JLO Silva *et al.* Estimativa 2012: incidência de câncer no brasil. *Rio de Janeiro: Instituto Nacional de Câncer José Alencar Gomes da Silva*. Citado na pág. 44

Song *et al.* **(2006)** J.J. Song, M. Ghosh, S. Miaou e B. Mallick. Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, 97(1):246 – 273. Citado na pág. 7, 21

Spiegelhalter *et al.* **(2002)** David J Spiegelhalter, Nicola G Best, Bradley P Carlin e Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. Citado na pág. 77

Stern e Cressie (2000) Hal S Stern e Noel Cressie. Posterior predictive model checks for disease mapping models. *Statistics in medicine*, 19(17-18):2377–2397. Citado na pág. 3, 19

Thomas *et al.* **(2004)** Andrew Thomas, Nicky Best, Dave Lunn, Richard Arnold e David Spiegelhalter. Geobugs user manual. *Á< www. mrc-bsu. cam. ac. uk/bugs/winbugs/geobugs. shtml*. Citado na pág. 8, 60

Waller *et al.* **(1997)** Lance A Waller, Bradley P Carlin, Hong Xia e Alan E Gelfand. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92 (438):607–617. Citado na pág. 7

West e Harrison (1997) Mike West e Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, New York. Citado na pág. 10

West *et al.* **(1985)** Mike West, P Jeff Harrison e Helio S Migon. Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389): 73–83. Citado na pág. 10

Xia *et al.* **(1997)** Hong Xia, BRADLEY P Carlin e Lance A Waller. Hierarchical models for mapping ohio lung cancer rates. *Environmetrics*, 8(2):107–120. Citado na pág. 7

Xie e Carlin (2006) Yang Xie e Bradley P. Carlin. Measures of bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10):3458 – 3477. Citado na pág. 21, 22

Apêndice A - Código do OpenBUGS para aplicação do Modelo Hierárquico Bayesiano

```
model {
for (i in 1 : Nareas) {
for (k in 1 : Ndiseases) {
Y[i, k] ~ dpois(mu[i, k])
log(mu[i, k]) <- log(E[i, k]) + alpha[k] + S[k, i]
RR1[i] <- exp(alpha[1] + S[1, i])
RR2[i] <- exp(alpha[2] + S[2, i])
RR3[i] <- exp(alpha[3] + S[3, i])
RR4[i] <- exp(alpha[4] + S[4, i])
RR5[i] <- exp(alpha[5] + S[5, i])     }
S[1:Ndiseases, 1 : Nareas] ~ mv.car(adj[], weights[], num[], omega[ , ])
for (i in 1:sumNumNeigh) {weights[i] <- 1 }
# Outras prioris
for (k in 1 : Ndiseases) {alpha[k] ~ dflat()}
omega[1 : Ndiseases, 1 : Ndiseases] ~ dwish(R[ , ], Ndiseases)
sigma2[1 : Ndiseases, 1 : Ndiseases] <- inverse(omega[ , ])
# Quantidades de interesse
sigma[1] <- sqrt(sigma2[1, 1])
sigma[2] <- sqrt(sigma2[2, 2])
sigma[3] <- sqrt(sigma2[3, 3])
sigma[4] <- sqrt(sigma2[4, 4])
sigma[5] <- sqrt(sigma2[5, 5])
corr12 <- sigma2[1, 2] / (sigma[1] * sigma[2])
corr13 <- sigma2[1, 3] / (sigma[1] * sigma[3])
corr14 <- sigma2[1, 4] / (sigma[1] * sigma[4])
corr15 <- sigma2[1, 5] / (sigma[1] * sigma[5])
corr23 <- sigma2[2, 3] / (sigma[2] * sigma[3])
corr24 <- sigma2[2, 4] / (sigma[2] * sigma[4])
corr25 <- sigma2[2, 5] / (sigma[2] * sigma[5])
corr34 <- sigma2[3, 4] / (sigma[3] * sigma[4])
corr35 <- sigma2[3, 5] / (sigma[3] * sigma[5])
corr45 <- sigma2[4, 5] / (sigma[4] * sigma[5])
mean1 <-  mean(S[1,])
mean2 <- mean(S[2,])
mean3 <- mean(S[3,])
mean4 <- mean(S[4,])
mean5 <- mean(S[5,]) }
```

Apêndice B - Código do OpenBUGS para aplicação do Modelo Hierárquico Bayesiano Dinâmico

```
model {
for (i in 1 : Nareas) {
for (k in 1 : Ndiseases) {
for (t in 1:T) {
Y[k,i, t] ~ dpois(mu[k,i, t])
log(mu[k,i, t]) <- log(E[k,i, t]) + alpha[k] + theta[k,t] + S[k, i]
RR[k,i, t]<- exp(alpha[k] + theta[k,t] + S[k, i])
}}}
# Distribuições a priori
for (k in 1 : Ndiseases) {
alpha[k] ~ dflat()}
S[1:Ndiseases, 1:Nareas] ~ mv.car(adj[], weights[], num[], omega[ , ])
for (i in 1:sumNumNeigh) {
weights[i] <- 1 }
omega[1 : Ndiseases, 1 : Ndiseases] ~ dwish(R[ , ], Ndiseases)
sigma2[1 : Ndiseases, 1 : Ndiseases] <- inverse(omega[ , ])
theta[1:Ndiseases,1:T] ~ mv.car(adjt[], weightst[], numt[], omegat[ , ])
# Especificar a matrizes de peso e de adjacência correspondentes a priori RW(1)
for(t in 1:1) {
weightst[t] <- 1;
adjt[t] <- t+1;
numt[t] <- 1
}
for(t in 2:(T-1)) {
weightst[2+(t-2)*2] <- 1;
adjt[2+(t-2)*2] <- t-1
weightst[3+(t-2)*2] <- 1;
adjt[3+(t-2)*2] <- t+1;
numt[t] <- 2
}
for(t in T:T) {
weightst[(T-2)*2 + 2] <- 1;
adjt[(T-2)*2 + 2] <- t-1;
numt[t] <- 1
}
omegat[1 : Ndiseases, 1 : Ndiseases] ~ dwish(Rt[ , ], Ndiseases)
```

```
sigmat2[1 : Ndiseases, 1 : Ndiseases] <- inverse(omegat[ , ])
# Valores de interesse
for(k in 1:Ndiseases) {
sigma[k] <- sqrt(sigma2[k, k])
}
for (j in 1 : Ndiseases) {
for(k in 1:Ndiseases) {
corr[j,k]<-sigma2[j,k] / (sigma[j] * sigma[k])}}
for (k in 1:Ndiseases) {
meanS[k] <-  mean(S[k,])}
for(k in 1:Ndiseases) {
sigmat[k] <- sqrt(sigmat2[k, k])}
for (j in 1 : Ndiseases) {
for(k in 1:Ndiseases) {
corrt[j,k]<-sigmat2[j,k] / (sigmat[j] * sigmat[k])}}
for(k in 1:Ndiseases) {
meant[k] <- mean(theta[k,])}}
```

Apêndice C - Gráficos para análise de convergência dos modelos

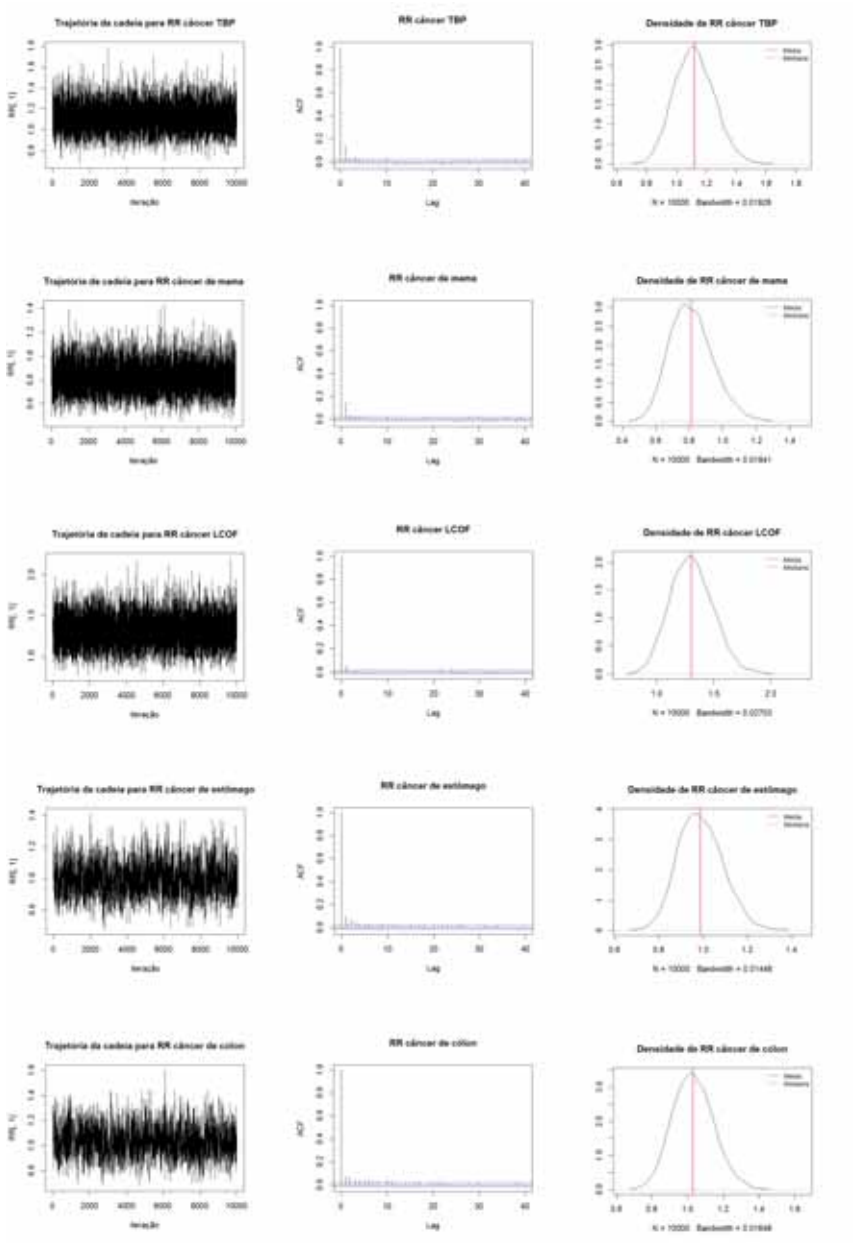

Figura 1: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para o Risco Relativo obtido para o modelo 1 referente aos óbitos por cada doença em estudo segundo as microrregiões do estado de São Paulo, de 1998 a 2010.*

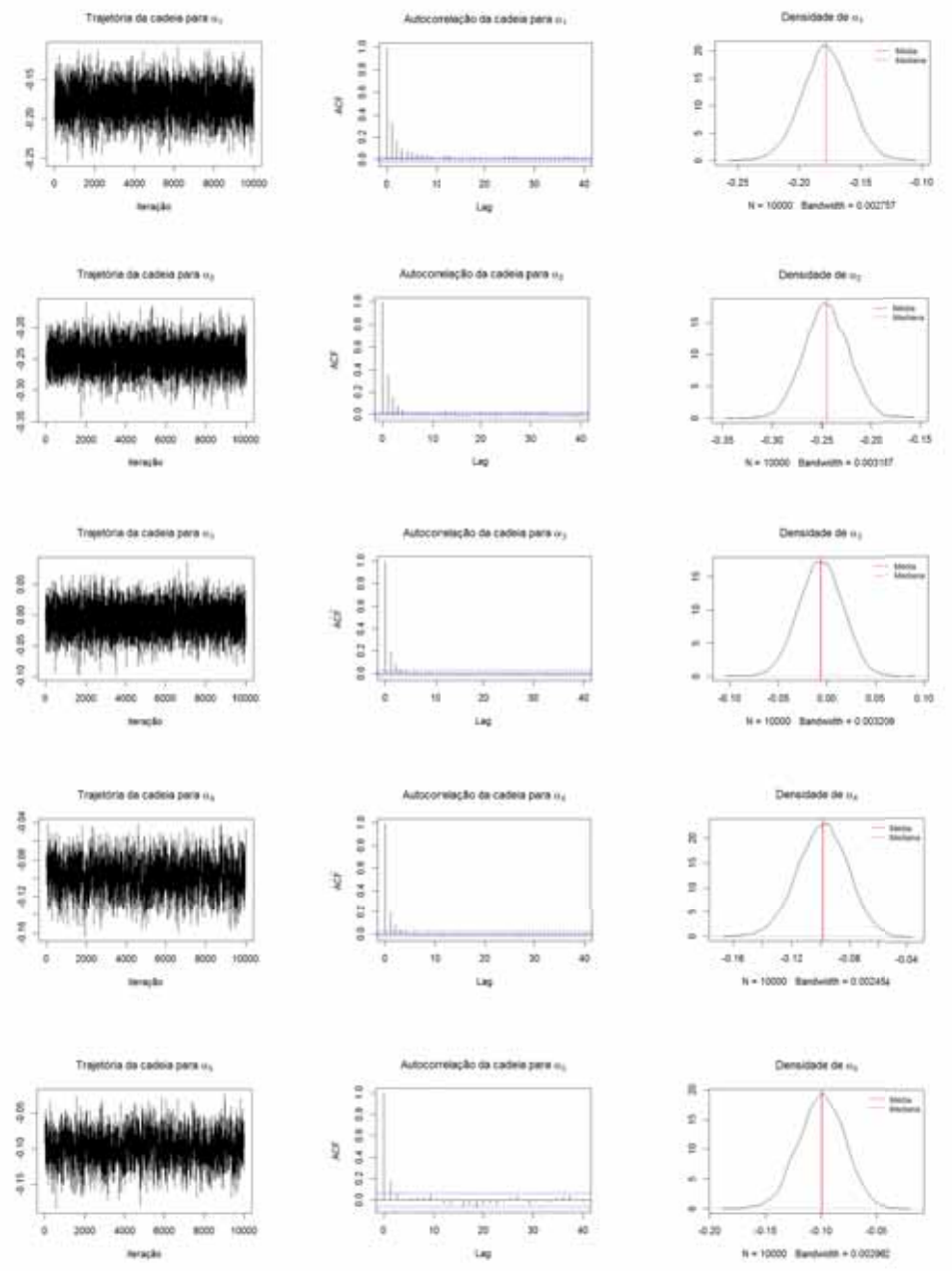

Figura 2: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para o parâmetro de intercepto a posteriori obtido para o modelo 2 referente aos óbitos por cada doença em estudo segundo as microrregiões do estado de São Paulo, de 1998 a 2010.*

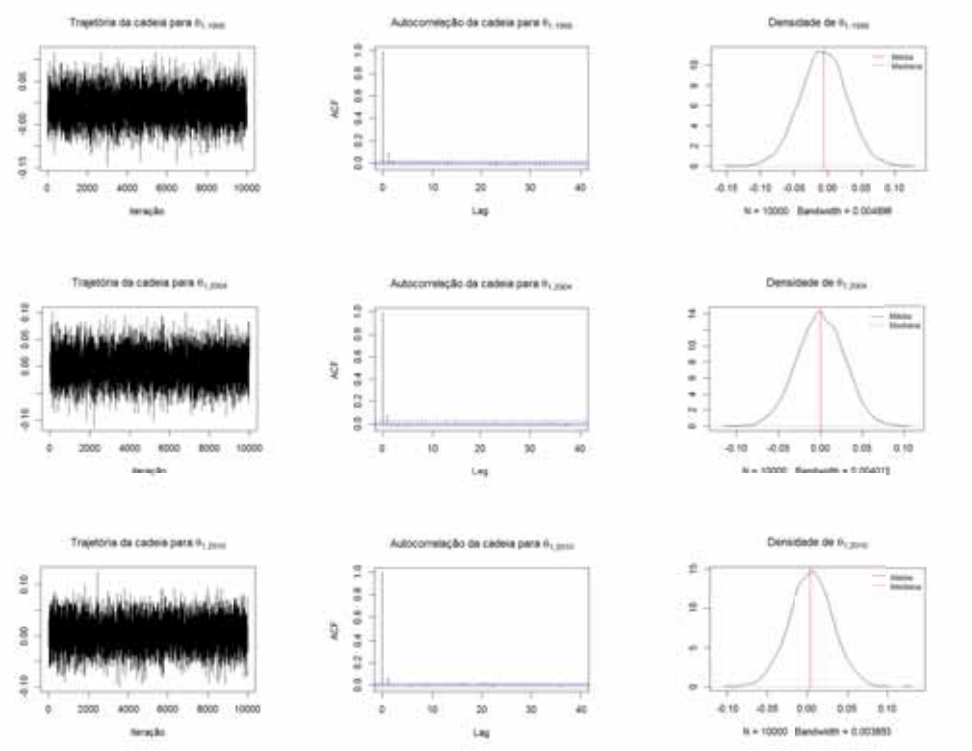

Figura 3: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para* θ*, obtido para o modelo 2 referente aos óbitos por câncer de traqueia, brônquios e pulmão segundo as microrregiões do estado de São Paulo, de 1998 a 2010.*

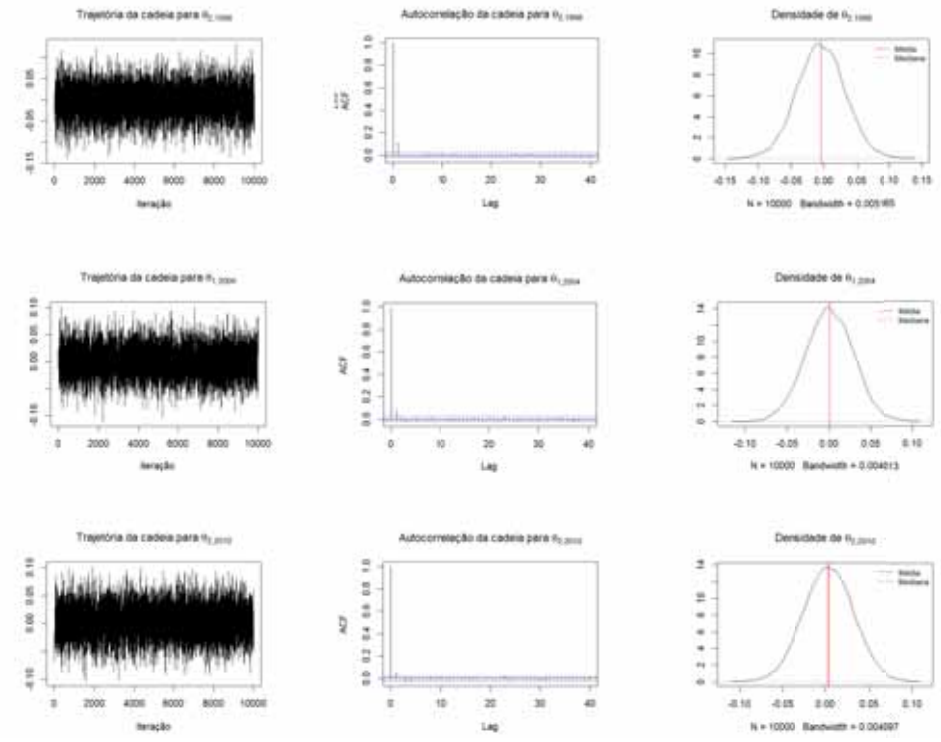

Figura 4: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para* θ*, obtido para o modelo 2 referente aos óbitos por câncer de mama feminino segundo as microrregiões do estado de São Paulo, de 1998 a 2010.*

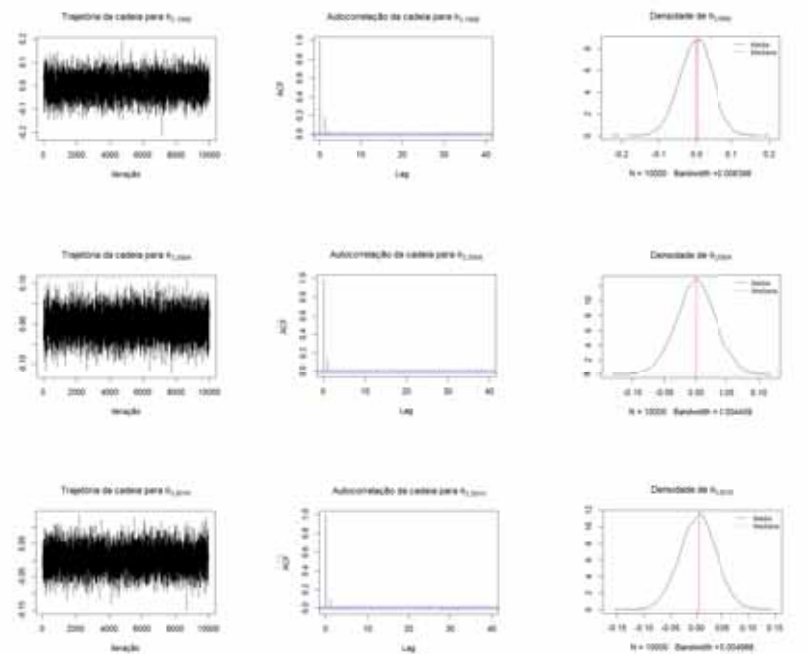

Figura 5: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para* θ*, obtido para o modelo 2 referente aos óbitos por câncer de lábios, cavidade oral e faringe as microrregiões do estado de São Paulo, de 1998 a 2010.*

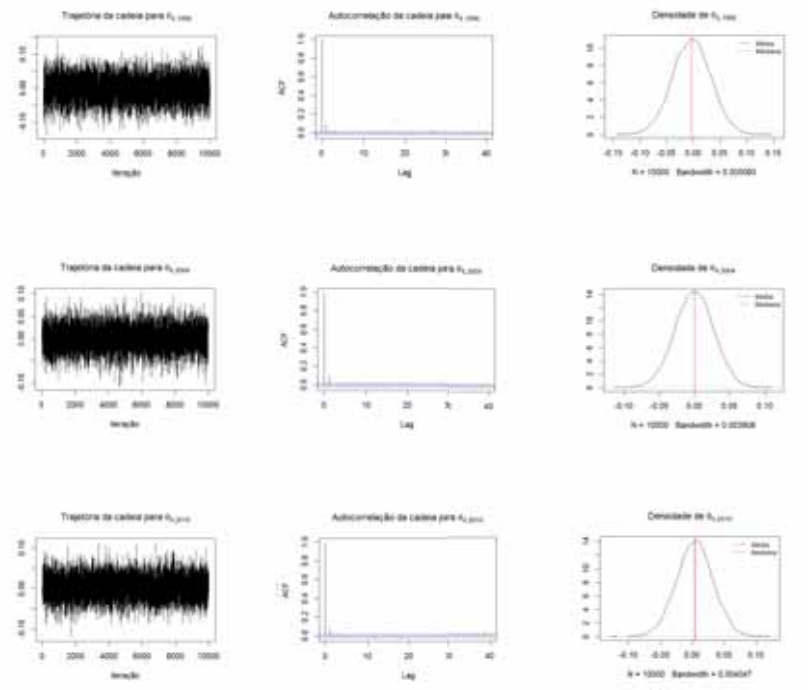

Figura 6: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para* θ*, obtido para o modelo 2 referente aos óbitos por câncer de estômago segundo as microrregiões do estado de São Paulo, de 1998 a 2010.*

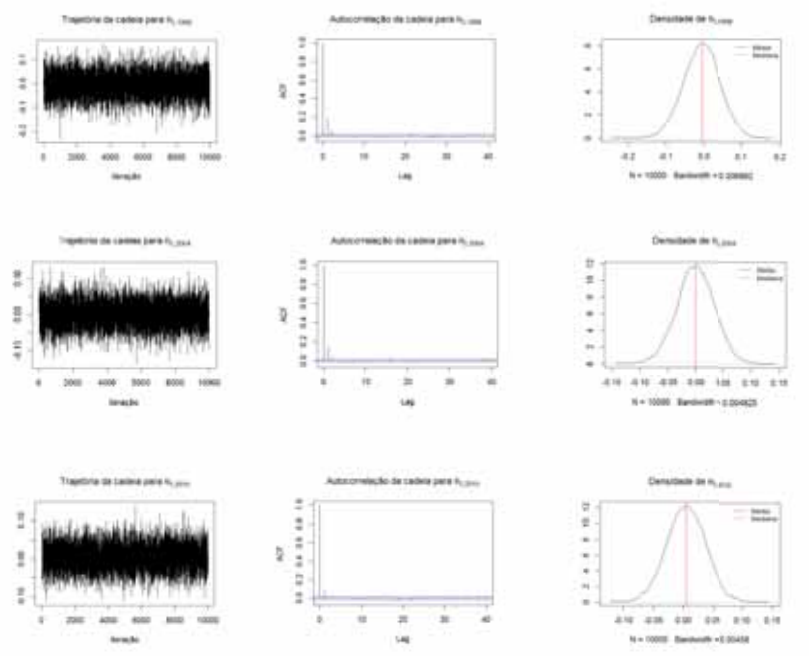

Figura 7: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para* θ*, obtido para o modelo 2 referente aos óbitos por câncer de cólon segundo as microrregiões do estado de São Paulo, de 1998 a 2010.*

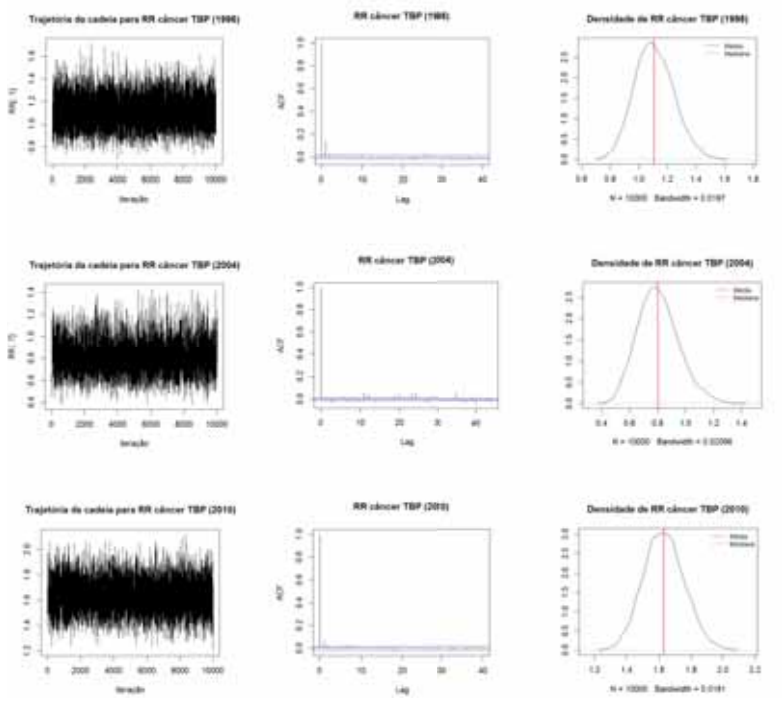

Figura 8: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para o Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de traqueia, brônquios e pulmão segundo as microrregiões do estado de São Paulo, para três anos do período.*

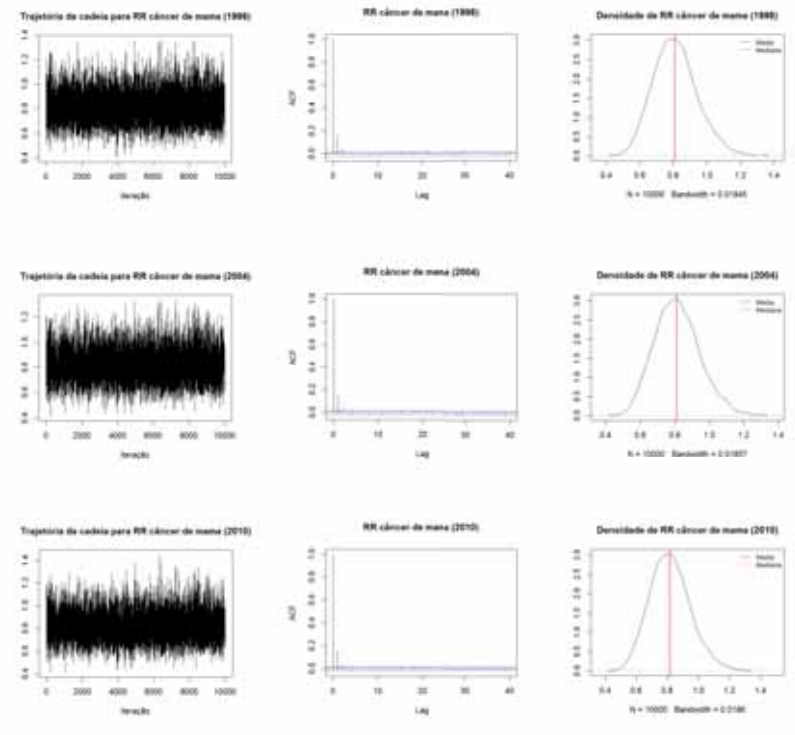

Figura 9: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para o Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de mama feminino segundo as microrregiões do estado de São Paulo, para três anos do período.*

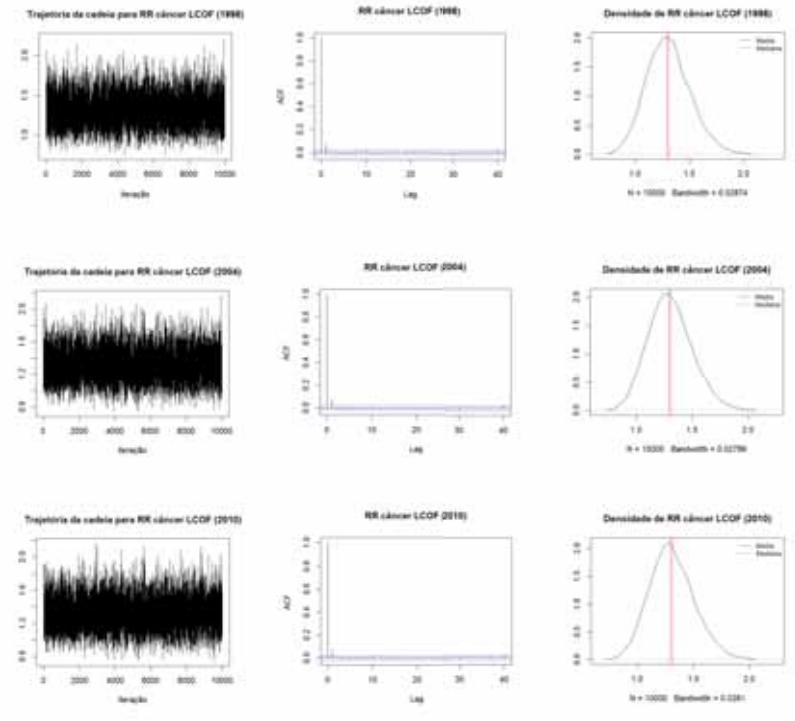

Figura 10: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para o Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de lábios, cavidade oral e faringe segundo as microrregiões do estado de São Paulo, para três anos do período.*

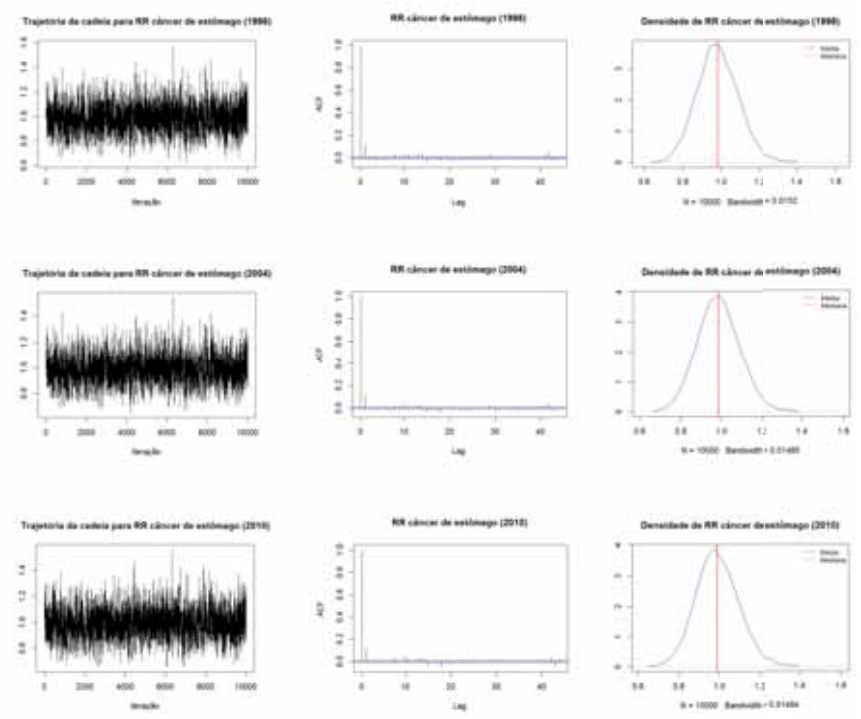

Figura 11: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para o Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de estômago segundo as microrregiões do estado de São Paulo, para três anos do período.*

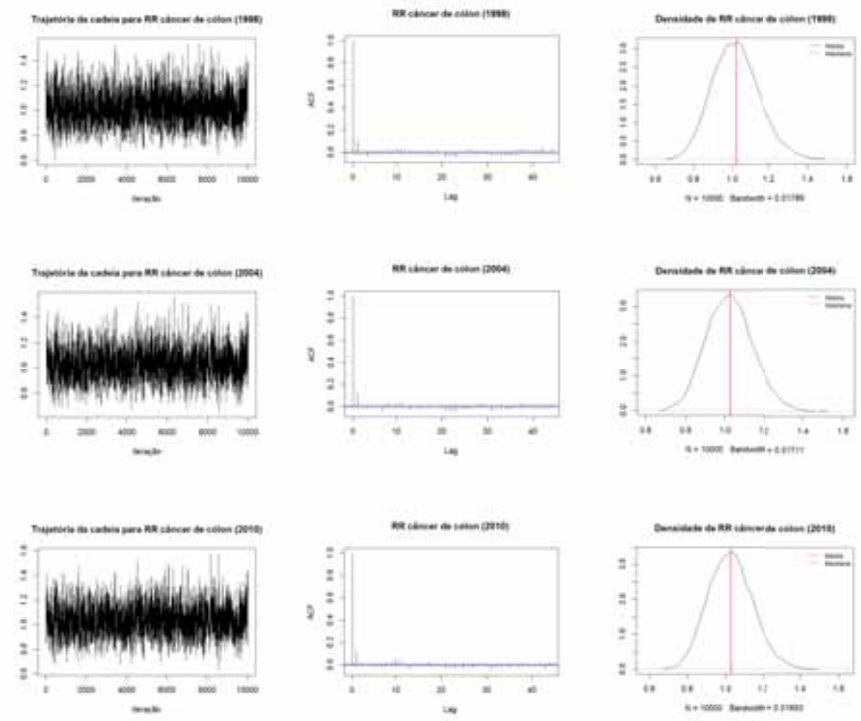

Figura 12: *Gráficos da trajetória, autocorrelação, e densidade a posteriori da cadeia para o Risco a posteriori obtido para o modelo 2 referente aos óbitos por câncer de cólon segundo as microrregiões do estado de São Paulo, para três anos do período.*

Apêndice D - Estimativas dos efeitos temporais do modelo dinâmico para cada doença

Tabela 1: *Estimativas dos parâmetros, erros padrões, e intervalos de credibilidade de 95% para os efeitos temporais do modelo 2 referente à câncer de traqueia, brônquios e pulmão*

Parâmetro	Média	Erro Padrão	ICr(95%)
$\theta_{1,1}$	-0.00658	0.03442	(-0.07436, 0.06058)
$\theta_{1,2}$	-0.004817	0.03382	(-0.07219, 0.06075)
$\theta_{1,3}$	-0.002662	0.03211	(-0.06527, 0.06058)
$\theta_{1,4}$	-0.001656	0.03109	(-0.06282, 0.05927)
$\theta_{1,5}$	-2.81E-01	0.02907	(-0.05692, 0.05712)
$\theta_{1,6}$	6.19E-01	0.03011	(-0.05873, 0.05893)
$\theta_{1,7}$	3.04E-01	0.02813	(-0.05520, 0.05532)
$\theta_{1,8}$	7.16E-01	0.02736	(-0.05352, 0.05482)
$\theta_{1,9}$	0.001727	0.02703	(-0.05045, 0.05430)
$\theta_{1,10}$	0.003061	0.02687	(-0.05029, 0.05639)
$\theta_{1,11}$	0.002471	0.02727	(-0.05227, 0.05564)
$\theta_{1,12}$	0.002646	0.02561	(-0.04783, 0.05298)
$\theta_{1,13}$	0.004452	0.02732	(-0.05019, 0.05849)

Tabela 2: *Estimativas dos parâmetros, erros padrões, e intervalos de credibilidade de 95% para os efeitos temporais do modelo 2 referente à câncer de mama feminino*

Parâmetro	Média	Erro Padrão	ICr(95%)
$\theta_{2,1}$	-0.004763	0.03635	(-0.07611, 0.06698)
$\theta_{2,2}$	-0.003798	0.0349	(-0.07308, 0.06395)
$\theta_{2,3}$	-0.003181	0.0327	(-0.06888, 0.06133)
$\theta_{2,4}$	-0.002319	0.03323	(-0.06680, 0.06239)
$\theta_{2,5}$	-9.93E-01	0.03418	(-0.06866, 0.06635)
$\theta_{2,6}$	-0.001243	0.03584	(-0.07344, 0.06910)
$\theta_{2,7}$	-5.41E-01	0.03337	(-0.06533, 0.06425)
$\theta_{2,8}$	0.001287	0.0306	(-0.05956, 0.06126)
$\theta_{2,9}$	0.002156	0.02937	(-0.05506, 0.05914)
$\theta_{2,10}$	0.003566	0.02863	(-0.05266, 0.05993)
$\theta_{2,11}$	0.003551	0.02867	(-0.05246, 0.06000)
$\theta_{2,12}$	0.003234	0.0277	(-0.05064, 0.05779)
$\theta_{2,13}$	0.003044	0.02872	(-0.05345, 0.05903)

Tabela 3: *Estimativas dos parâmetros, erros padrões, e intervalos de credibilidade de 95% para os efeitos temporais do modelo 2 referente à câncer de lábio, cavidade oral e faringe*

Parâmetro	Média	Erro Padrão	ICr(95%)
$\theta_{3,1}$	-1.64E-01	0.04485	(-0.08856 0.08746)
$\theta_{3,2}$	-5.78E-01	0.03989	(-0.07857 0.07657)
$\theta_{3,3}$	0.001101	0.03776	(-0.07255 0.07438)
$\theta_{3,4}$	5.98E-04	0.03648	(-0.07029 0.07288)
$\theta_{3,5}$	7.61E-01	0.03424	(-0.06795 0.06794)
$\theta_{3,6}$	4.69E-01	0.03318	(-0.06608 0.06538)
$\theta_{3,7}$	-0.001282	0.03192	(-0.06416 0.06182)
$\theta_{3,8}$	-1.38E-01	0.03223	(-0.06369 0.06295)
$\theta_{3,9}$	-4.41E-01	0.02952	(-0.05810 0.05717)
$\theta_{3,10}$	4.70E-01	0.02924	(-0.05633 0.05705)
$\theta_{3,11}$	-0.002196	0.03212	(-0.06638 0.06078)
$\theta_{3,12}$	-0.001938	0.03186	(-0.06415 0.05986)
$\theta_{3,13}$	0.003338	0.03504	(-0.06507 0.07229)

Tabela 4: *Estimativas dos parâmetros, erros padrões, e intervalos de credibilidade de 95% para os efeitos temporais do modelo 2 referente à câncer de estômago*

Parâmetro	Média	Erro Padrão	ICr(95%)
$\theta_{4,1}$	-0.003098	0.0357	(-0.07399, 0.06685)
$\theta_{4,2}$	-0.002974	0.03469	(-0.07246, 0.06446)
$\theta_{4,3}$	-0.001789	0.03264	(-0.06446, 0.06185)
$\theta_{4,4}$	-0.001411	0.03189	(-0.06417, 0.06058)
$\theta_{4,5}$	-7.87E-01	0.02969	(-0.05992, 0.05828)
$\theta_{4,6}$	-3.56E-01	0.02885	(-0.05669, 0.05612)
$\theta_{4,7}$	-3.61E-01	0.02739	(-0.05444, 0.05231)
$\theta_{4,8}$	9.00E-01	0.02796	(-0.05476, 0.05562)
$\theta_{4,9}$	9.55E-04	0.02757	(-0.05348, 0.05523)
$\theta_{4,10}$	0.001722	0.02693	(-0.05214, 0.05349)
$\theta_{4,11}$	9.77E-01	0.02881	(-0.05548, 0.05724)
$\theta_{4,12}$	0.002516	0.02731	(-0.05098, 0.05606)
$\theta_{4,13}$	0.003706	0.02848	(-0.05297, 0.05858)

Tabela 5: *Estimativas dos parâmetros, erros padrões, e intervalos de credibilidade de 95% para os efeitos temporais do modelo 2 referente à câncer de cólon*

Parâmetro	Média	Erro Padrão	ICr(95%)
$\theta_{5,1}$	-0.004759	0.04869	(-0.10110, 0.09020)
$\theta_{5,2}$	-0.003865	0.04116	(-0.08498, 0.07785)
$\theta_{5,3}$	-0.002746	0.03988	(-0.08106, 0.07482)
$\theta_{5,4}$	-0.00156	0.03887	(-0.07896, 0.07475)
$\theta_{5,5}$	-7.62E-01	0.03654	(-0.07280, 0.07078)
$\theta_{5,6}$	3.92E-01	0.03424	(-0.06695, 0.06763)
$\theta_{5,7}$	4.89E-01	0.03394	(-0.06566, 0.06603)
$\theta_{5,8}$	9.83E-01	0.03296	(-0.06390, 0.06620)
$\theta_{5,9}$	0.001335	0.03283	(-0.06381, 0.06542)
$\theta_{5,10}$	0.002110	0.03104	(-0.05898, 0.06311)
$\theta_{5,11}$	0.001411	0.03153	(-0.06037, 0.06398)
$\theta_{5,12}$	0.002349	0.03053	(-0.05725, 0.06262)
$\theta_{5,13}$	0.004623	0.0321	(-0.05817, 0.06713)

Apêndice E - Mapa do estado de São Paulo segundo as microrregiões do IBGE

Figura 13: *Mapa do Estado de São Paulo segundo microrregiões.*