UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Estatística Espacial em Dados de Área: Uma Modelagem Inteiramente Bayesiana para o Mapeamento de Doenças Aplicada à Dados Relacionados com a Natalidade em Mulheres Jovens de Porto Alegre

Autor: Rafael Bassegio Caumo
Orientador: Prof. Dra. Patrícia Klarmann Ziegelmann

Porto Alegre, 13 de Dezembro de 2006.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Estatística Espacial em Dados de Área: Uma Modelagem Inteiramente Bayesiana para o Mapeamento de Doenças Aplicada à Dados Relacionados com a Natalidade em Mulheres Jovens de Porto Alegre

Autor: Rafael Bassegio Caumo

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Prof. Dra. Patrícia Klarmann Ziegelmann (Orientador).
Valéria Dozolina Sartori Bassani (Estatística convidada da PMPA).

Porto Alegre, 13 de Dezembro de 2006.

Agradecimentos

Gostaria de agradecer a todos que de alguma forma me apoiaram ou estiveram ao meu lado, direta ou indiretamente, nesta longínqua caminhada. Sinto-me na obrigação de citar todas estas insubstituíveis pessoas nesta monografia.

Agradeço à todos os meus familiares, em especial ao meu pai, Jatir, e à minha mãe, Santina, assim como ao meu irmão, Lucas. Por estarem sempre ao meu lado, assim como por toda a base e o sustento de vital importância fornecido.

Agradeço também a todos os amigos e colegas. Em especial aos parceiros de futebol do Vira, assim como aos de banda Estela, àqueles das trips pelo surf, aos colegas de curso, aos amigos de infância e PSP e aos companheiros de colorado. Estão todos muito bem guardados, cada um com muita consideração.

Agradeço à minha namorada, Mariana, e família. Sempre presente, me deu muita força pra continuar. Devo muito deste longo percurso percorrido a ela que, tanto nos momentos de felicidade quanto nos momentos desgastantes de estudos, trabalho e sobrecargas, se apresentou sempre companheira e disposta para o que fosse necessário.

Agradeço aos professores do Departamento de Estatística, em especial à minha orientadora Patrícia K. Ziegelmann por toda atenção e dedicação dispensada, contribuindo extremamente para a evolução deste trabalho. Foi um prazer poder trabalharmos juntos.

Agradeço ao pessoal da PED/FEE, em especial ao Jéferson. Mais do que um chefe, se demonstrou um co-orientador tanto para minha formação profissional quanto acadêmica.

Agradeço ao pessoal da equipe técnica do Observatório da cidade de Porto Alegre/ Prefeitura Municipal de Porto Alegre, em especial à Valéria pela disponibilidade de tempo despendido e contribuição através do fornecimento dos dados, assim como à Denise pela produção dos mapas utilizados neste trabalho.

O Intermédio Entre o Universo Cotidiano e as Leis Determinísticas

A Estatística, na condição de ciência matemática, é aquela que trata de compilar o universo das facticidades do mundo cotidiano para a dimensão dos números, leis e funções matemáticas. Porém, como é de consentimento comum, os fatos cotidianos não podem ser expressos por leis determinísticas. Em outras palavras, ainda não é da capacidade de nós, seres humanos sob constantes transformações, predizer eventos com absoluta certeza ou afirmar verdades únicas.

É, portanto, que a sociedade atual vem acolhendo o conceito das probabilidades, olímpico sustento de toda a teoria Estatística.

Trata-se esta de uma visão moderna, onde os eventos passam ser interpretados da forma mais crivelmente apropriada possível. Aceitando as idéias de probabilidade, distribuição de probabilidade, variabilidade, esperança, entre outras.

É atribuída, então, com orgulho, ao Estatístico, a sublime missão de transcrever numericamente o fato de acordo com sua essência cotidiana. Eis que o profissional da área passa a considerar o conceito da variabilidade em todas as leis e funções matemáticas.

Não é difícil de concordar que o ser humano não é capaz de prever com absoluta certeza os eventos cotidianos, pois a variabilidade está praticamente sempre presente. Em contrapartida, com base em uma teoria consistente e robusta, já mundialmente adotada pela comunidade científica e por todos os órgãos e instituições com conhecimento avançado, a Estatística assume o posto da ciência que visa o aperfeiçoamento do domínio do ser humano sobre os eventos. Através do trabalho acerca da compreensão da essência dos fatos e acerca da manipulação da variabilidade (objetivando, é claro, redução da incerteza). E, ao Estatístico, surge a competência de manejar esta ciência e seus correspondentes mecanismos de quantificação do observável.

Resumo

Estatística Espacial em Dados de Área: Uma Modelagem Inteiramente Bayesiana para o Mapeamento de Doenças Aplicada à Dados Relacionados com a Natalidade em Mulheres Jovens de Porto Alegre

Autor:

Rafael Bassegio Caumo

Orientadora:

Patrícia Klarmann Ziegelmann

Este trabalho tem foco principal na apresentação de um processo de modelagem que vêm sendo considerado como o atualmente mais evoluído para a geração de estimativas que superam problemas intrínsecos à modelagem clássica utilizáveis no mapeamento de doenças, área da Epidemiologia Espacial em expansão e de muita utilidade pública: a modelagem inteiramente Bayesiana. Como procedimento ilustrativo da modelagem, realizou-se a aplicação prática de cinco modelos inteiramente Bayesianos – não estruturado, espacialmente estruturado, convolução sem covariáveis, convolução com uma covariável e convolução com duas covariáveis – no mapeamento de um risco relativo associado à natalidade de mulheres com idade inferior à 20 anos de idade por bairros de Porto Alegre para o ano de 2005. Os dados foram obtidos junto à equipe técnica do Observatório da cidade de Porto Alegre/Prefeitura Municipal de Porto Alegre. Anteriormente à apresentação dos conceitos de mapeamento de doenças e modelo inteiramente Bayesiano, este trabalho faz uma revisão bibliográfica a respeito de tópicos que fornecem sustentabilidade para uma boa compreensão por parte do leitor. Dentre os assuntos revisados: estatística Bayesiana e metodologias MCMC, estatística espacial e análise de dados de área, epidemiologia espacial e o mapeamento de doenças. Para a produção deste, utilizou-se do Software WinBUGS 1.4, que também aparece introduzido em uma seção desta monografia.

Palavras-Chave: modelo inteiramente Bayesiano; não-estruturado; espacialmente estruturado; convolução; modelo Bayesiano hierárquico; mapeamento de doenças; epidemiologia espacial; estatística espacial; análise de dados de área; estatística Bayesiana; metodologia MCMC; natalidade em mulheres jovens de Porto Alegre; risco relativo; problema das pequenas áreas; WinBUGS.

Índice

Capítulo 1

Introdução

O Mapeamento de Doenças corresponde a uma área da Epidemiologia Espacial que, por sua vez, se trata de um ramo da Estatística Espacial. Esta área da epidemiologia espacial vem assumindo o posto de ferramenta imprescindível e de alto grau de importância para a análise da saúde pública regional, sustentando e direcionando tomadas de decisões e – para quando se aplicar – a fomentação de políticas públicas. Principalmente por seus objetivos principais de descrever e compreender a disposição da doença ou fenômeno ao longo do espaço geográfico através de mapas limpos, ou seja, livres de quaisquer ruídos aleatórios e perturbações externas que acrescentem à variabilidade populacional e degredem a verdadeira distribuição espacial do fenômeno, além de sugerir hipóteses específicas que relacionem a incidência do fenômeno com informações adicionais incluídas na análise.

A epidemiologia espacial nada mais é do que a junção entre os conceitos de epidemiologia com estatística espacial. Para o primeiro, considera-se que é a ciência que estuda quantitativamente fenômenos de saúde e doenças, buscando compreender ao máximo os desfechos e fatores explicativos. E, por estatística espacial, entende-se a área da Estatística que trata de compreender a distribuição espacial de dados oriundos de fenômenos ocorridos no espaço geográfico. Ou seja, estuda métodos científicos para a coleta, descrição, visualização e análise de dados que possuem coordenadas geográficas.

Assim como se pode trabalhar com diversas metodologias intrínsecas ao Mapeamento de Doenças, à Epidemiologia e à Estatística Espacial, sob perspectiva Ortodoxa (Visão Clássica) da ciência Estatística, pode se optar, sabendo que em muitos casos esta se faz expressivamente mais conveniente ou terminantemente necessária, por abordar as mesmas questões sob a ótica da corrente Bayesiana do pensamento estatístico.

A estatística Bayesiana adota uma visão diferente da clássica, principalmente no que se relaciona à quantidade desconhecida de interesse (parâmetro). Para os Bayesianos, os parâmetros para os quais se deseja inferir são quantidades, por desconhecidas, aleatórias, diferentemente da visão clássica que os considera quantidades desconhecidas, porém fixas.

Baseado nestas condições, o estatístico clássico leva em consideração para o processo inferencial apenas a informação captada pelos dados que possui, enquanto o Bayesiano atribui distribuição de probabilidade à priori para representar a incerteza que possui, que convém com a aleatoriedade do parâmetro de interesse, e a combina com a informação carregada pelos dados observacionais ou experimentais.

A estatística Bayesiana já foi, por muito tempo, atravancada pela indisponibilidade de tecnologia avançada por sofrer de um problema frequentemente presente quando tratamos com aplicações práticas. Este é muito conhecido como o problema geral da estatística Bayesiana e compreende que o tratamento analítico para a obtenção da distribuição à posteriori de interesse é extremamente complicado e muitas vezes impossível, fazendo com que o processo inferencial Bayesiana não possa ser realizado.

Atualmente, entretanto, em virtude do avanço computacional e da disponibilidade de tecnologias avançadas, este problema foi superado. O grande responsável pela superação de tal problema atende pelo nome de Markov Chain Monte Carlo (MCMC - Monte Carlo via Cadeias de Markov) e se trata essencialmente de um procedimento computacional que permite uma boa aproximação para a posteriori desconhecida de interesse.

Sendo assim, a estatística Bayesiana rompeu suas limitações e vem, a cada dia mais, se tornando extremamente popular. Além disso, vem sendo ampliada e desenvolvida para os mais diversos campos de estudo.

1.1. Objetivos

Considerando todas as informações anteriores e que este trabalho tem foco principal na exposição da metodologia estatística (e não na aplicação prática), relata-se o principal objetivo deste trabalho: apresentar um modelo inteiramente Bayesiano para o mapeamento de doenças.

Como um dos objetivos secundários, deseja-se expor conceitos básicos de estatística Bayesiana, metodologias MCMC, estatística espacial, análise de dados de área, epidemiologia espacial e mapeamento de doenças sob a ótica clássica e Bayesiana.

Outro objetivo secundário consiste na ilustração do mapeamento inteiramente Bayesiano através do tratamento com dados reais.

1.2. Justificativas

O grande interesse do autor pela estatística espacial e pela inferência Bayesiana conduziram-no a produzir este trabalho e a explorar e apresentar ao máximo possível, considerando a limitação de tempo, aspectos relacionados a estes temas.

Especificamente a respeito da metodologia inteiramente Bayesiana, foi escolhida, primeiramente, por tratar de uma técnica intrínseca a um procedimento de vital importância social – mapeamento de doenças – e que supera limitações presentes na abordagem clássica do mapeamento de doenças.

1.3. Metodologia

Para atingir os objetivos de interesse – primeiramente – utilizou-se de revisão bibliográfica para a exposição dos conceitos necessários para uma boa compreensão da estrutura geral que acerca o mapeamento de doenças.

Em um segundo momento, realizou-se uma aplicação prática do conceito principal exposto na revisão bibliográfica – modelagem inteiramente Bayesiana para o mapeamento de doenças – a partir de dados reais que tratam da natalidade de mulheres jovens de Porto Alegre em 2005, obtidos junto à Prefeitura Municipal de Porto Alegre.

1.4. Estrutura

A revisão conceitual bibliográfica está compreendida em três capítulos. O Capítulo 2, inicialmente, traça um paralelo entre a estatística Bayesiana e a estatística clássica. Apresenta também conceitos gerais da estatística Bayesiana e aborda alguns tópicos computacionais Bayesianos avançados, como a metodologia MCMC e o software Winbugs.

O Capítulo 3 apresenta ao leitor fundamentos da estatística espacial, da tipologia dos dados com os quais a mesma trabalha e sobre a análise de dados de área. Além disso, aborda brevemente o conceito de epidemiologia espacial e suas ramificações.

O Capítulo 4 – último da revisão bibliográfica – introduz o leitor ao mapeamento de doenças, passando desde a modelagem clássica e suas limitações até a modelagem Bayesiana hierárquica e inteiramente Bayesiana.

Fechando o trabalho, o Capítulo 5 expõe a análise prática realizada com dados reais obtidos junto à Prefeitura Municipal de Porto Alegre. Este capítulo poderia ter sido extremamente mais aprofundado, considerando a possibilidade de tópicos a serem explorados dentro do contexto prático abordado. Apesar de interessante, uma completa e detalhada análise sobre algum problema real é anteparada pela limitação de tempo para a produção de uma monografia.

Capítulo 2

Estatística Bayesiana

Apesar de que, em situações práticas, a estatística Bayesiana tem aparecido com destaque somente nos últimos anos, sua história é muito antiga e tem origem associada ao trabalho original do reverendo Thomas Bayes (1763). Porém, a complexidade do processo analítico de muitos problemas cujo enfoque Bayesiano lhes é atribuido, fizeram com que a aplicação prática dos mesmos estagnasse a espera de um período em que máquinas – na forma atual de computadores – e seus recursos técnicos eletrônicos avançados pudessem novamente abrir as portas para a propagação deste ramo da estatística.

Atualmente, a estatística Bayesiana, além de apresentar uma linha de pensamento diferenciada, surge com alternativas e propostas para diversos paradigmas e entraves intrínsecos à perspectiva Ortodoxa (Clássica) da estatística.

Neste Capítulo, em um primeiro momento (Seção 2.1), traçar-se-á um paralelo comparativo entre as duas abordagens da Estatística. Já na Seção 2.2, serão apresentados os fundamentos básicos da estatística Bayesiana. As Seções 2.3 e 2.4 apresentarão ao leitor tópicos avançados do ramo Bayesiano – na forma de, respectivamente às seções, metodologia avançada (MCMC) e software computacional – que contribuem para atual bom momento de glórias e avanços da abordagem Bayesiana. O software em questão corresponde ao WinBUGS, que permite a implementação de modelos Bayesianos através do uso de MCMC.

Quanto à referências que tratam da estatística Bayesiana, aquelas consultadas para a construção deste texto foram Box e Tiao (1992) e Paulino (2003). Referências para MCMC podem ser encontradas na seção 2.3.

2.1. Inferência Clássica X Inferência Bayesiana

A filosofia da escola Clássica é fundamentada no crédulo de que toda a informação a respeito do fenômeno aleatório que se deseja compreender é oriunda dos dados a serem obtidos pelo pesquisador, ou seja, da amostra. Sendo assim, toda e qualquer inferência

(realização de generalizações para uma população) é embasada apenas naquilo que se observou – seja através da obtenção de uma amostra ou da realização de um experimento – ou naquilo que poderia ser observado caso o experimento fosse repetido ou uma nova amostra fosse observada.

Em complemento à ideologia Clássica apresentada no parágrafo anterior, o Estatístico Clássico é aquele que considera que os parâmetros desconhecidos para os quais se deseja inferir são quantidades fixas. Em um exemplo prático, de acordo com o pensamento clássico, pode-se imaginar que o efeito médio que a aplicação de certo tratamento em pacientes gravemente doentes tem sobre o tempo de sobrevida dos mesmos é uma quantia específica, fixa, porém desconhecida.

A escola Bayesiana, em contrapartida, acredita que os parâmetros de interesse, por serem desconhecidos, se tratam de quantidades aleatórias. Desta forma, a especificação de distribuições de probabilidade para os parâmetros é natural. Em exemplo análogo ao anteriormente apresentado, o pensamento Bayesiano considera que o efeito médio é dotado de variabilidade, propondo que, a incerteza que se tem a seu respeito deva ser representada por uma distribuição de probabilidade. Essa distribuição é de caráter subjetivista, diferentemente do caráter frequentista clássico. É interessante proceder a associação deste pensamento à dados que envolvem opinião pública, doenças, entre outros.

Por outro ponto de vista, então, os Bayesianos consideram que não só os dados obtidos são importantes para o processo inferencial, como também a informação, com correspondente grau de incerteza (precisão), que o pesquisador carrega a respeito daquilo que está estudando. Sendo assim, é aberto espaço para que o conhecimento e a experiência do pesquisador no assunto possam contribuir com as inferências a serem realizadas. Este conhecimento do pesquisador sobre os parâmetros de interesse é dito “conhecimento à priori” e será representado pela “distribuição à priori”.

A informação à priori (distribuição à priori) será combinada com a informação presente nos dados obtidos (função de verossimilhança) de modo que aquele que carregar um maior grau de certeza seja mais valorizado. O resultado final desta combinação de fontes de informação é representado pela distribuição à posteriori e será a base para toda e qualquer inferência que se fizer interessante.

Muitos consideram a Estatística Clássica como um caso particular da Estatística Bayesiana pelo fato de que, se o pesquisador não possuir conhecimento algum sobre aquilo que vier a estudar, tratará de considerar uma distribuição à priori não informativa (que não adiciona informação) no estudo e, consequentemente, as inferências serão realizadas apenas com base nos dados.

Porém, apesar de que os resultados da inferência Bayesiana, sob especificação de uma priori não informativa, são iguais aos encontrados pela inferência clássica, é importante ressaltar que a interpretação dos mesmos é diferenciada devido aos seus métodos de obtenção. Além disso, vale destacar que a estatística Bayesiana se apresenta muito mais flexível e abrangente, no sentido que lida melhor com o relaxamento de suposições que, principalmente em modelos mais complexos, são estabelecidas pela abordagem clássica com intuito de permitir a continuidade da análise.

Um exemplo da flexibilidade da abordagem Bayesiana é a incorporação de um componente espacialmente estruturado no modelo para o mapeamento de doenças, detalhadamente apresentado no Capítulo 4.

2.2. Fundamentos da Estatística Bayesiana

Como já mencionado na seção anterior, distribuição à priori é um modelo probabilístico que representa o conhecimento assumido pelo pesquisador – previamente à observação dos dados – em relação a um parâmetro desconhecido θ. Sendo assim, ao considerar a quantidade de interesse θ, deve-se assumir uma distribuição $\pi(\theta)$ à priori para a mesma.

Após a escolha de uma priori adequada, é realizado um processo observacional ou experimental para a obtenção de dados (obtenção de uma amostra). A informação contida nos dados fornecerá a verossimilhança, através de uma função $l(x \mid \theta)$, para cada um dos possíveis valores de θ.

A combinação da informação à priori com a informação obtida através das observações é realizada através do Teorema de Bayes, resultando na distribuição à posteriori. A obtenção da posteriori, para o caso de θ contínuo, é dada por

$$p(\theta \mid x) = \frac{p(\theta, x)}{\int p(\theta, x) d\theta} = \frac{p(\theta) p(x \mid \theta)}{\int p(\theta) p(x \mid \theta) d\theta} = \frac{p(\theta) p(x \mid \theta)}{p(x)}, \tag{2.1}$$

ou, seguindo a terminologia deste trabalho,

$$\pi(\theta \mid x) = \frac{\pi(\theta) l(x \mid \theta)}{p(x)}. \tag{2.2}$$

Para o caso de θ discreto, o procedimento de obtenção da posteriori segue diretamente da substituição da integral de 2.2 por um somatório em θ. Sem perda de generalidade, no decorrer deste trabalho pensar-se-á sempre em θ contínuo.

Previamente à observação de uma amostra, $p(x)$ representa a distribuição marginal da variável aleatória X, ou seja, a distribuição da v.a. X condicionada em θ e ponderada pela priori $\pi(\theta)$. Esta função é chamada de função preditiva à priori e é muito utilizada para se checar a adequação do modelo através de predições via $p(x)$, anteriores à observação de x. A idéia consiste em calcular o valor da preditiva para o ponto observado na amostra. Se o valor calculado for baixo, deve-se questionar o modelo, visto que valores baixos representam que – sob o modelo construído – o observado pela amostra recebia pouca probabilidade de ser verificado. Maiores detalhes a respeito deste tópico podem ser encontrados Gelman (1997).

Em 2.2, visto que a amostra já foi observada, $p(x)$ passa a ser uma constante em relação à θ. A esta é atribuído o nome de constante normalizadora, baseado no fato de que é responsável por transformar o numerador de 2.2 em uma distribuição de probabilidade. É por isso que muitas vezes a posteriori é apresentada na forma da proporcionalidade

$$\pi(\theta \mid x) \propto \pi(\theta) l(x \mid \theta).$$

No caso da observação de uma segunda amostra, para que a informação sobre o parâmetro de interesse seja atualizada por estes novos dados, basta considerar a distribuição à posteriori obtida com a primeira amostra como priori a ser combinada com a segunda amostra. Sendo assim, da mesma forma, através do Teorema de Bayes, podemos atualizar a informação e obter uma segunda distribuição à posteriori.

A posteriori é, portanto, uma distribuição de probabilidade para θ que representa todo o conhecimento que se tem sobre o parâmetro de interesse. Corresponde ao principal resultado da inferência Bayesiana, visto que toda inferência que se deseja realizar a respeito de θ pode ser retirada de tal distribuição.

Na estatística Bayesiana, então, logicamente, toda inferência é baseada na distribuição à posteriori que, por sua vez, nem sempre é derivável, principalmente por dificuldades com o cálculo da preditiva $p(x)$ (visto que, geralmente, se trata de uma integral múltipla que depende do número de parâmetros desconhecidos no modelo), fazendo com que a posteriori não possa ser analiticamente obtida. Fica caracterizado então o problema geral da inferência Bayesiana, com solução proposta na próxima seção.

O estimador de Bayes é dado por $E(\theta \mid x) = \int \theta\pi(\theta \mid x)d\theta$. É com base neste e na moda da distribuição à posteriori, conhecida por estimador de máxima verossimilhança generalizado, que usualmente as estimações pontuais são realizadas. Apesar de não muito utilizada, a mediana da distribuição à posteriori também pode funcionar como uma estimativa pontual.

As inferências intervalares – na abordagem bayesiana denominadas intervalos de credibilidade ou intervalos de probabilidade à posteriori – também derivam, logicamente, da posteriori e são usualmente obtidas através de delimitações nos quantis. Outra possibilidade para geração dos intervalos de credibilidade atende por HDR ("Highest Density Region") e consiste em inserir no intervalo os valores de θ mais prováveis, não se restringindo a atribuir mesma probabilidade a ambas as caudas da distribuição à posteriori. Consequentemente, o HDR consiste no intervalo de credibilidade que, para uma probabilidade fixada, apresenta menor amplitude dentre todos os intervalos de credibilidade com esta probabilidade.

A interpretação de um intervalo de credibilidade é diferenciada da relacionada aos intervalos de confiança da estatística clássica. Na clássica, interpretamos que o intervalo de confiança obtido contém o verdadeiro valor do parâmetro com determinada confiança, pois considera que o parâmetro desconhecido é fixo. Aleatório, por conseguinte, são os possíveis intervalos de confiança. Ou seja, antes de se observar a amostra, têm-se determinada probabilidade de vir a se obter um intervalo de confiança que contenha o verdadeiro valor do parâmetro.

Já a interpretação Bayesiana afirma que o verdadeiro valor do parâmetro está contido no intervalo de credibilidade – que representa um intervalo de probabilidade – especificado com probabilidade pré-determinada. Considerando que, agora, o parâmetro é que é suposto aleatório.

Quanto à distribuição à priori, esta é encarregada de incorporar a informação alheia aos dados no modelo, representando o conhecimento prévio do pesquisador em relação ao parâmetro desconhecido. A influência da informação à priori nos resultados será baixa para os casos em que a distribuição à priori apresentar baixa precisão (alta incerteza do pesquisador) ou que os dados carregarem muita informação (amostra grande, baixa incerteza). Logo, à medida que o tamanho da amostra aumenta, a contribuição da informação à priori na posteriori diminui.

Para os casos em que o pesquisador não possuir ou não desejar adicionar conhecimento anterior sobre o parâmetro de interesse, têm-se como alternativa as prioris não informativas. Estas procedem de forma a atribuir mesma probabilidade à priori para qualquer possível valor do parâmetro. Tais prioris podem ser tanto impróprias quanto próprias. Um exemplo de priori imprópria é a distribuição uniforme no intervalo $(-\infty,+\infty)$. Ou seja, trata-se de uma não informação perfeita – pois atribui mesma probabilidade para todo o espaço do parâmetro de interesse – porém não corresponde a uma distribuição de probabilidade por não satisfazer as propriedades necessárias para isso. Prioris próprias são todas as que podem ser classificadas como uma distribuição de probabilidade.

Nos casos de priori não informativa, a posteriori refletirá apenas a informação contida nos dados e, consequentemente, os resultados corresponderão aos mesmos do enfoque clássico, quando neste o problema for tratável. Eis porque alguns estatísticos defendem a idéia de que a estatística clássica nada mais é do que um caso particular da estatística Bayesiana.

Assim como as prioris não informativas, as prioris conjugadas correspondem a um conceito de prioris muito utilizadas na prática. Estas correspondem a prioris que, ao serem combinadas com certa função de verossimilhança, resultam em uma distribuição à posteriori da mesma família, facilitando o processo analítico. As prioris conjugadas são abrangentes e flexíveis, permitindo fácil manuseio, inclusive com a possibilidade de serem não informativas.

Pode se pensar em dividir a especificação da distribuição a priori em estágios. Além de que, em muitos casos, pode facilitar a especificação ou permitir inserir alguma estruturação na priori, esta abordagem é usual para determinadas situações experimentais. Supondo que θ é o parâmetro de interesse e ϕ é o parâmetro da priori (hiperparâmetro) de

θ, a distribuição à priori marginal de interesse do parâmetro θ pode ser apresentada, com base na especificação hierárquica, por

$$\pi(\theta) = \int \pi(\theta, \phi) d\phi = \int \pi(\theta \mid \phi) \pi(\phi) d\phi,$$

onde $\pi(\phi)$ corresponde à hiperpriori – priori para o hiperparâmetro ϕ – e representa segundo estágio desta hierarquia. Ou seja, quando se deseja atribuir hiperprioris para os hiperparâmetros, cria-se uma estrutura de hierárquica. Apesar de que não haja limitações quanto ao número de estágios, a dificuldade de interpretação dos hiperparâmetros dos estágios mais altos faz com que os mesmos tenham a si atribuídos distribuições não informativas.

Outro resultado interessante da abordagem Bayesiana é a distribuição preditiva para uma observação futura $\tilde{X}$, chamada de distribuição preditiva à posteriori e denotada por $p(\tilde{X} \mid x)$. Esta pode ser obtida através de

$$p(\tilde{X} \mid x) = \int f(\tilde{X} \mid \theta) \pi(\theta \mid x) d\theta.$$

A preditiva à posteriori consiste em uma distribuição de probabilidade de onde se obtém informações a respeito do esperado para uma observação futura $\tilde{X}$, considerando o atual conhecimento sobre o parâmetro desconhecido θ. Sendo assim, é desta distribuição que são derivadas predições sobre a quantidade a ser futuramente re-observada $\tilde{X}$.

A seção seguinte abordará um tratamento usual empregado no combate ao problema geral da inferência Bayesiana (já introduzido nesta seção), que consiste na impossibilidade de derivação da distribuição à posteriori, principalmente para os casos de modelos mais avançados.

2.3. Markov Chain Monte Carlo (MCMC)

É para os casos em que, através do processo analítico ou até mesmo por aproximação numérica, não se consegue obter a distribuição à posteriori de interesse que a metodologia de integração Monte Carlo via cadeias de Markov (MCMC) surge como solução. Trata-se essencialmente de um procedimento de simulação estocástica que permite uma boa aproximação para a posteriori desconhecida, de muita utilidade sobre a ótica Bayesiana

para problemas práticos, principalmente por não restringir o número de parâmetros no modelo e possuir um algoritmo de simulação relativamente simples.

No contexto inferencial clássico, a simulação estocástica é um procedimento rotineiramente utilizado para a estimação de parâmetros que não podem ou apresentam extrema complexidade para ser analiticamente estimados. Ou seja, é frequentemente utilizada para estimação de momentos de funções de variáveis aleatórias (produto, soma, etc.). O procedimento consiste em simular das distribuições dos parâmetros de interesse e remodelar o resultado da simulação, forçando a amostra final a se adequar a uma amostra oriunda da distribuição da função da variável aleatória em questão.

Para os casos em que não se conhece a distribuição de onde se precisa simular – como é o caso de diversas posterioris da inferência Bayesiana aplicada – a simulação estocástica surge na forma de MCMC, permitindo assim o processo inferencial. Resumidamente, os métodos MCMC possibilitam a geração de amostras de distribuições à posteriori com função densidade de probabilidade desconhecida ou demasiadamente complicada.

Neste trabalho serão apresentados dois dos principais métodos MCMC. O primeiro, algoritmo de Metropolis-Hastings, proposto por Metropolis (1953) e extendido por Hastings (1970), é o caso mais geral. O segundo – proposta de Geman e Geman (1984) – corresponde ao amostrador de Gibbs (Gibbs Sampling) e representa um caso particular do algoritmo de Metropolis-Hastings. Como referência para simulação estocástica e MCMC, cita-se tanto Gamerman e Lopes (2006) quanto Gilks, Richardson e Spiegelhalter (1996), que também apresenta diversas diferentes aplicações práticas dos métodos MCMC.

2.3.1. O Algoritmo de Metropolis-Hastings

A idéia geral consiste na geração de valores a partir de uma distribuição $q(.|.)$, chamada de distribuição proposta, que pode ser arbitrariamente escolhida. Os valores gerados $\boldsymbol{\theta}^*$ são candidatos à formadores de uma amostra oriunda da distribuição à posteriori de interesse $\pi(.)$, onde $\boldsymbol{\theta}$ pode ser escalar (um parâmetro de interesse) ou corresponder à um vetor (mais de um parâmetro de interesse). Os candidatos $\boldsymbol{\theta}^*$ farão parte da amostra de $\pi(\boldsymbol{\theta})$ com probabilidade específica que garante a equivalência entre a amostra resultante dos valores candidatos selecionados com uma amostra derivada da

distribuição de interesse $\pi(.)$. À esta probabilidade atribui-se o nome de probabilidade de aceitação. Resumidamente, trata-se de um procedimento que garante a possibilidade de amostramos de distribuições não conhecidas.

Mais especificamente, constrói-se uma cadeia de Markov onde, a cada tempo $t \geq 0$, gera-se um candidato para o próximo estado, $\boldsymbol{\theta}^*$, a partir de uma distribuição proposta $q(.|\boldsymbol{\theta}_t)$. O valor gerado dependerá dos estados anteriores através de, no máximo, o atual valor da cadeia e será "aceito" e adicionado à cadeia de forma que $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^*$ com probabilidade de aceitação dada por

$$\alpha(\boldsymbol{\theta}_t, \boldsymbol{\theta}^*) = \min\left(1; \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_t)q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_t)}\right). \tag{2.3}$$

O mecanismo funciona com intuito de rejeitar com maior probabilidade os pontos simulados que forem muito discrepantes em relação aos esperados pela distribuição $\pi(.)$. Percebe-se, também de forma intuitiva, que quando a distribuição proposta e a distribuição de interesse forem similares, tal probabilidade tende a ser próxima de um, aumentando a possibilidade de aceitação do valor gerado.

Considerando o caso específico de um parâmetro de interesse, então, interpreta-se o vetor $\underset{\sim}{\theta}$ de valores gerados como uma cadeia de Markov pelo fato de que cada valor θ_t depende no máximo do seu passado imediatamente anterior. Assim como $q(.|.)$ corresponde a matriz Kernel de transição, suposta homogênea em t (não depende de t).

O algoritmo de Metropolis-Hastings é formalizado pelos seguintes passos:

1. Define-se $t = 0$ e escolhe-se um valor arbitrário para θ_0 (valor inicial);

2. Gera-se um valor θ^* a partir de $q(.|\theta_t)$;

3. Gera-se um valor U de uma distribuição $Uniforme(0,1)$;

4. Se $U \leq \alpha(\theta_t, \theta^*)$, faz-se $\theta_{t+1} = \theta^*$. Caso contrário, $\theta_{t+1} = \theta_t$;

5. Incrementa-se t;

6. Retorna-se ao passo 2, continuando as iterações até um tamanho n suficientemente grande.

Sob certas condições de regularidade, a cadeia de Markov "esquece" os estados iniciais e converge para a distribuição estacionária $\pi(.)$. Sendo assim, para o processo inferencial, desconsideram-se os m primeiros valores gerados de $\underset{\sim}{\theta}$ (desconsidera-se o período de burn-in, ou seja, aquecimento). Obtém-se então, após a verificação da convergência da cadeia, em $\underset{\sim}{\theta} = (\theta_{m+1}, \theta_{m+1}, ..., \theta_n)$, uma amostra não independente da distribuição $\pi(.)$. A dependência se faz presente pelo fato de que os novos valores são gerados a partir de uma distribuição de probabilidade condicionada ao valor do estado imediatamente anterior.

Para reduzir a dependência, um procedimento usual é de considerar um intervalo que, no inglês, é conhecido como "thin interval". Este consiste em aproveitar valores espaçados da cadeia de acordo com um intervalo pré-fixado. A idéia que fundamenta este conceito acredita que valores mais distantes no tempo estão menos autocorrelacionados.

O leitor pode ainda estar questionando o fato de utilizarmos uma probabilidade de aceitação que depende da distribuição desconhecida de interesse $\pi(.)$. Note, porém, que a probabilidade de aceitação depende de $\pi(.)$ somente através do quociente $\pi(\theta^*)/\pi(\theta_t)$. Por definição, e considerando θ escalar, a posteriori é dada por

$$\pi(.) = \pi(\theta \mid x) = \frac{\pi(\theta) l(x \mid \theta)}{p(x)}, \tag{2.4}$$

onde $p(x)$ é uma constante de normalização que não depende de θ. Desta forma, pode-se perceber que, tanto para $\theta = \theta^*$ quanto para $\theta = \theta_t$, o denominador de (2.4) não se altera. Logo

$$\alpha(\theta_t, \theta^*) = \min\left\{1; \frac{\pi(\theta = \theta^*) l(\underset{\sim}{x} \mid \theta = \theta^*) q(\theta_t \mid \theta^*)}{\pi(\theta = \theta_t) l(\underset{\sim}{x} \mid \theta = \theta_t) q(\theta^* \mid \theta_t)}\right\}$$

e o processo depende apenas da priori $\pi(\theta)$ e da verossimilhança $l(\underset{\sim}{x} \mid \theta)$.

Teremos então, por fim, uma amostra da distribuição de interesse de onde, através da integração de Monte Carlo, podemos obter estimativas de momentos da posteriori desconhecida de interesse. Além disso, pode-se fazer uso dos valores amostrados para obtenção de uma estimativa da própria posteriori $\pi(\theta \mid x)$. Uma técnica muito utilizada para esta finalidade corresponde ao método não-paramétrico denominado "Kernel smoothing".

2.3.2. *Gibbs Sampling*

O amostrador de Gibbs (Gibbs Sampling) é um caso particular do algoritmo de Metropolis-Hastings, adequado para ocasiões em que $\boldsymbol{\theta}$ é um vetor de dimensão p, tal que $p \geq 2$. Logo, compartilha dos mesmos princípios básicos já apresentados na Subseção 2.3.1.

Para a aplicação desta metodologia se faz necessário que a posteriori conjunta – mesmo que desconhecida – possa, pelo menos, ser traduzida em posterioris condicionais completas.

Por exemplo, se estamos interessados em uma distribuição que compreende três parâmetros, sendo $\pi(.) = \pi(\theta^{(1)}, \theta^{(2)}, \theta^{(3)})$, necessitaríamos conhecer ou $\pi(\theta^{(1)} \mid \theta^{(2)}\theta^{(3)})$, $\pi(\theta^{(2)} \mid \theta^{(1)}\theta^{(3)})$ e $\pi(\theta^{(3)} \mid \theta^{(1)}\theta^{(2)})$, ou quaisquer outras possibilidades de condicionais completas, tal qual $\pi(\theta^{(1)} \mid \theta^{(2)}\theta^{(3)})$ e $\pi(\theta^{(2)}\theta^{(3)} \mid \theta^{(1)})$, para que se faça possível a implementação do amostrador de Gibbs.

Visto que o amostrador de Gibbs corresponde a um caso particular do algoritmo de Metropolis-Hastings com distribuições propostas iguais às condicionais completas, pode-se demonstrar que a probabilidade de aceitação de cada valor gerado é igual a 1, fazendo com que a cadeia sempre aceite os novos candidatos. Isso faz com que, na prática, o amostrador de Gibbs seja mais eficiente e rápido.

O amostrador de Gibbs é formalizado pelos seguintes passos:

1. Define-se $t = 0$ e escolhem-se valores iniciais $\boldsymbol{\theta}_0 = (\theta_0^{(1)}, \theta_0^{(2)}, ..., \theta_0^{(p)})$;

2. Obtém-se o próximo estado da cadeia, $\boldsymbol{\theta}_{t+1} = (\theta_{t+1}^{(1)}, \theta_{t+1}^{(2)}, ..., \theta_{t+1}^{(p)})$, com base nas gerações sucessivas oriundas das condicionais completas, tal que

$$\begin{aligned}
\theta_{t+1}^{(1)} &= \pi(\theta^{(1)} \mid \theta_t^{(2)}, \theta_t^{(3)}, ..., \theta_t^{(p)}) \\
\theta_{t+1}^{(2)} &= \pi(\theta^{(2)} \mid \theta_{t+1}^{(1)}, \theta_t^{(3)}, ..., \theta_t^{(p)}) \\
\theta_{t+1}^{(3)} &= \pi(\theta^{(3)} \mid \theta_{t+1}^{(1)}, \theta_{t+1}^{(2)}, \theta_t^{(4)}, ..., \theta_t^{(p)}); \\
&\vdots \\
\theta_{t+1}^{(p)} &= \pi(\theta^{(p)} \mid \theta_{t+1}^{(1)}, ..., \theta_{t+1}^{(p-1)})
\end{aligned}$$

3. Incrementa-se t;

4. Retorna-se ao passo 2, procedendo as iterações até tamanho n suficientemente grande e satisfatório.

A determinação da amostra final da distribuição $\pi(.)$ de interesse segue os mesmos princípios dos já apresentados na Seção 2.2.

2.3.3. Algoritmos Híbridos

Os algoritmos híbridos consideram misturas de metodologias MCMC, no sentido que, se a busca por todas as condicionais completas para a utilização do Gibbs Sampling não for atendida, pode-se considerar apenas as condicionais completas que se conhece. Para as outras, desconhecidas, se faz uso do algoritmo geral de Metropolis-Hastings.

2.3.4. O Processo Inferencial

As inferências, como tradicionalmente na estatística Bayesiana, derivam única e exclusivamente da posteriori. Logo, qualquer estimativa que se desejar obter deve derivar da amostra simulada, visto que não possuímos a forma funcional da posteriori. O desconhecimento da posteriori e a possibilidade de inferirmos unicamente através da amostra simulada implica na inacessibilidade à estimação da moda da distribuição à posteriori, assim como, paralelamente, não se torna possível a obtenção de intervalos HDR de credibilidade.

2.4. O Software WinBUGS

O WinBUGS (Win Bayesian inference Using Gibbs Sampling) é um software livre[1], em ascensão na comunidade Estatística, que permite fácil manuseio de diversas técnicas avançadas Bayesianas. Está sendo desenvolvido na Unidade de Bioestatística do Medical Research Council em conjunto com o Imperial College, ambos na Inglaterra, sendo implementado por Thomas et al. (1992). Consiste em um conjunto de ferramentas que permite a especificação de diversos tipos de modelos, estimando parâmetros – sob abordagem Bayesiana – via metodologias MCMC.

O resultado do procedimento interno do WinBUGS consiste essencialmente em uma amostra dos parâmetros requisitados, de acordo com os dados de entrada e o modelo

[1] O download pode ser realizado através do website http://www.mrc-bsu.cam.ac.uk/bugs/

especificado. São dessas amostras que o processo inferencial deriva, com algumas estimativas e medidas podendo ser imediatamente acessadas através de menus e botões.

A linguagem do sistema, assim como a linguagem de entrada e saída dos dados, equivale à linguagem do sistema $S+$ e, portanto, qualquer outro interesse por parte do usuário que não estiver disponível nos menus e botões do programa pode ser requisitado através da especificação de sintaxes.

Em seu procedimento interno de obtenção das amostras oriundas da distribuição à posteriori de interesse, ao utilizar o algoritmo de Metropolis-Hastings, o software assume uma distribuição Normal centrada no atual valor da cadeia como distribuição proposta $q(.|.)$. Os valores iniciais para os parâmetros de interesse podem tanto ser especificados pelo usuário quanto gerados automaticamente pelo programa.

Para o caso do amostrador de Gibbs, o WinBUGS amostra diretamente após um auto reconhecimento de funções condicionais completas conjugadas. Caso o programa não consiga detectar tais condicionais completas conjugadas, passa a trabalhar com uma dentre duas rotinas possíveis. A primeira delas, ARS (Adaptative Rejection Sampling), pode ser utilizada para amostrar de forma eficiente de qualquer distribuição condicional com função de densidade log-côncava. A segunda, ARMS (Adaptative Rejection Metropolis Sampling), generaliza a rotina anterior para os demais casos de distribuições condicionais.

O WinBUGS, além de extrema facilidade de uso, acompanha um manual explicativo para o usuário, com diversos exemplos aplicados, permitindo que, mesmo aqueles que não tem intimidade com a inferência Bayesiana possam fazer uso do mesmo e gerar resultados. A recomendação é que o usuário tenha noção e reflita sobre o que e como o software está agindo. O mau uso do WinBUGS pode vir a fornecer tanto bons resultados para o problema errado quando maus resultados para o problema certo.

Capítulo 3

Estatística Espacial

O desenvolvimento de técnicas e modelos estatísticos que trabalham de acordo com uma perspectiva geográfica (ou espacial) é relativamente recente. O interesse da percepção espacial dos fenômenos surgiu principalmente no campo da Geologia e outras ciências do solo, do processamento de imagens, da epidemiologia, das ciências rurais, da ecologia, florestal, das ciências ambientais, entre outros. Atualmente, porém, o estudo espacial dos fenômenos vem se expandindo para todas as áreas em que o mesmo possa ser tratável e, consequentemente, diversos pesquisadores, de acordo com suas correspondentes áreas de interesse, vêm contribuindo para a propagação da Estatística Espacial.

Este capítulo inicia – Seção 3.1 – com a apresentação de uma visão inicial e geral sobre a estatística espacial. A tipologia dos dados utilizáveis em estatística espacial é considerada na Seção 3.2. A seção 3.3, dada a amplitude de metodologias da estatística espacial e o foco específico deste trabalho no mapeamento de doenças, compreende a apresentação mais aprofundada dos dados de área. Por fim, a Seção 3.4 apresenta ao leitor um ramo – e suas correspondentes ramificações – da estatística espacial que atualmente vem sendo muito difundido: A Epidemiologia Espacial.

Algumas das principais referências de estatística espacial compreendem Bailey e Gatrell (1995), Ripley (1981) e Cressie (1993). Quanto à produção nacional, cita-se o bom trabalho de Druck, Carvalho, Câmara e Monteiro (2004). Para epidemiologia espacial, considera-se Elliott, Wakefield, Best e Briggs (2001) e Lawson (2001).

3.1. Introdução à Estatística Espacial

A Estatística Espacial é a área da Estatística que trata de compreender a distribuição espacial de dados oriundos de fenômenos ocorridos no espaço geográfico. Ou seja, estuda métodos científicos para a coleta, descrição, visualização e análise de dados que possuem coordenadas geográficas. Quando, na análise em questão, deseja-se analisar o fenômeno no espaço ao longo do tempo, passamos a tratar de um problema espaço-temporal.

Logo, para que um problema seja de Estatística Espacial, os dados obrigatoriamente devem possuir um índice que faz referencia à uma localização geográfica. Ou seja, a referência geográfica é explicitamente utilizada na modelagem.

A percepção visual da distribuição espacial dos dados é bastante eficaz no sentido de traduzir os padrões existentes com considerações objetivas, assim como na percepção da associação com as possíveis causas, direcionando e sustentando as tomadas de decisões.

Segundo Bailey & Gatrell (1995):

> Análise de Dados Espaciais trata das análises onde dados observáveis são obtidos a partir de algum processo operando no espaço e para os quais se utilizam métodos para descrever ou explicar o comportamento deste processo e sua possível relação com outros fenômenos espaciais. Desta forma, o objetivo da Análise de Dados Espaciais é de aumentar a compreensão básica do processo, assim como buscar evidências em relação às hipóteses estabelecidas ou ainda predizer valores em áreas onde as observações não foram feitas.

A grande evolução computacional das técnicas de mapeamento e da acessibilidade aos Sistemas de Informação Geográfica (SIG) são os principais responsáveis pelo avanço da estatística espacial e pela possibilidade da realização de modelagens sofisticadas dentro deste contexto.

Um Sistema de Informação Geográfica é um conjunto de equipamentos e programas computacionais que possibilitam a integração de mapas e gráficos com um banco de dados, além de ferramentas capazes de coletar, armazenar, manejar, analisar e visualizar informações georeferenciadas. Possibilitando a visualização espacial de variáveis, permitindo fácil detecção do padrão espacial.

3.2. Tipologia dos Dados Espaciais

Na análise espacial, três tipos de dados georeferenciados são principalmente considerados. A diferenciação entre os tipos de dados diz respeito a sua natureza estocástica. Consequentemente, diferentes metodologias estatísticas são empregadas na análise de cada um dos tipos.

Nesta monografia, em virtude da aplicação prática abordada e das limitações deste tipo de trabalho, apenas a análise dos dados de área será discutida de forma mais

aprofundada. Porém, uma introdução a respeito dos três tipos de dados georeferenciados é apresentada na seqüência.

3.2.1. Processos Pontuais

São aqueles que identificam eventos ou fenômenos como pontos localizados no espaço. Neste caso, o interesse principal consiste nas coordenadas geográficas que representam a localização exata dos acontecimentos. Na prática, dados relacionados à crimes são um exemplo dos frequentemente estudados. O foco é a detecção de padrões e fontes de influência – se aleatórios ou não – para distribuição espacial dos pontos.

3.2.2. Dados de Área

São dados usualmente obtidos através de levantamentos populacionais, tais quais censos, estatísticas de saúde, cadastramentos populacionais, entre outros, agregados por áreas de uma região. Em outras palavras, representar-se-á cada uma das áreas do mapa por uma quantia para cada uma das variáveis do estudo. As áreas são subdivisões do mapa – é tradicional chamar todo o espectro geográfico por região – com, supostamente, homogeneidade interna, usualmente delimitadas por polígonos fechados. Na prática, porém, as áreas constituem partições de caráter administrativo, político ou geofísico. Uma discussão mais profunda sobre dados de área aparece na Seção 3.3 desta monografia.

3.2.3. Superfícies Contínuas

Esta seção admite outras nomenclaturas e engloba subdivisões muito encontradas na literatura. Tais quais geoestatística e superfícies aleatórias. Os dados se tratam de amostras de campo, regular ou irregularmente distribuídos. Em situações cotidianas, normalmente as amostras são derivadas de estações fixas de monitoração, coleta ou medição de certa variável de interesse. O objetivo da análise deste tipo de dado é de modelar uma superfície espacial que represente o comportamento da variável em estudo no espaço geográfico. Para isso, através da modelagem, trata-se de expandir os resultados amostrados nas estações de coleta para as demais regiões que não tiveram informações coletadas. O principal resultado das análises deste tipo de dado compreende mapas geológicos e topográficos. Como exemplo, poderíamos pensar na modelagem da superfície da qualidade do ar em algum estudo sobre poluição.

3.2.4. Dados de Interação Espacial

Estes dados, cuja localização é considerada fixa tal qual em superfícies contínuas, correspondem a um par ordenado que indicam ponto de saída (origem) e ponto de chegada (destino). Através destes, se torna viável compreender o comportamento dos fluxos – identificando acessibilidade questões como acessibilidade e atratividade – através da modelagem, que pode inclusive permitir predizer efeitos oriundos de alterações no cenário dos mesmos. Como exemplo, cita-se estudos migratórios que podem estar associados a um planejamento comercial, de transportes e de saúde.

3.3. Análise de Dados de Área

Nos dados de área, considera-se uma região dividida em áreas contíguas, disjuntas e bem definidas. Para cada uma das áreas da região é associada uma quantia, ou indicador, de acordo com a característica que se está estudando. Em outras palavras, podemos dizer que na análise de dados de área lidamos com eventos agregados em espaços delimitados por polígonos fechados. Ou seja, os valores (indicadores) associados às áreas não estão relacionados à localização específica pontual de um evento no espaço, mas correspondem a uma quantidade que representa um padrão global para a área em relação à variável em estudo. Na prática, as divisões geográficas que resultam nas áreas são – normalmente – de caráter político, administrativo e geofísico, geralmente caracterizadas por bairros, municípios, setores censitários.

Muitas vezes, principalmente sob perspectiva epidemiológica, ambiental e sócio-econômica, e de acordo com a abordagem prioritária desta monografia, os dados são apresentados na forma de contagens de ocorrências de um evento. Faz-se, todavia, a necessidade de alguma padronização na quantia bruta da contagem observada, visto que as áreas habitualmente apresentam populações em risco de tamanhos diferentes. A maneira mais comum de tratar este problema é através do cálculo de taxas de incidência, riscos relativos e proporções. Tais são usualmente obtidos através de levantamentos populacionais, como censos, estatísticas de saúde, cadastramentos populacionais, entre outros.

O possível questionamento do leitor quanto à prática da modelagem estatística para dados derivados de levantamentos populacionais tem resposta presente no próximo

capítulo – mais especificamente na Seção 4.2 –, quando a análise de dados de área é abordada no contexto do mapeamento de doenças. Confesso que no período inicial da produção desta monografia indaguei por algum tempo este questionamento, visto que a literatura se demonstra ainda limitada e em processo de desenvolvimento, dado o volume de assuntos associados tanto a estatística espacial quanto a técnicas utilizáveis em dados de área.

O objetivo da análise de dados de área não consiste na predição de valores para áreas não observadas, visto que, na quase totalidade das vezes, todas as áreas apresentam informações disponíveis. Sendo assim, o objetivo principal obedece à identificação de determinado padrão ou configuração espacial no que diz respeito à variável aleatória de interesse, assim como possíveis relações no espaço com covariáveis.

Sendo assim, a abordagem para dados de área apresentada neste trabalho é, portanto, apenas recomendada para dados obtidos através de levantamentos populacionais. Para os casos em que a numeração completa dos eventos não for possível e a amostragem for a única possibilidade, maiores considerações na modelagem são necessárias. Um exemplo de trabalho que considera a análise de dados epidemiológicos amostrados é Nejjari et al. (1993).

Metodologias de amostragem de dados espaciais estão, de maneira satisfatória, apresentadas em Ripley (1981), Cressie (1993) e Thomson (1992). Entretanto, ao considerarmos dados de epidemiologia espacial, nos confrontamos com metodologias ainda não suficientemente desenvolvidas, possivelmente pelo fato de que, nesta área específica, os dados são geralmente derivados de totais enumerações de eventos. Estes derivam de registros oficiais com obtenção rotineira ou estatísticas oficiais governamentais. Sendo assim, a integralidade dos dados espaciais epidemiológicos é então apenas questionada quanto aos erros de registro ou falta de diagnóstico, inerentes à ação do pesquisador.

3.3.1. Representação Gráfica

O procedimento padrão de representação gráfica para dados de área corresponde ao Mapa com Padrão de Cores (Choropleth Map), que apresenta as áreas da região coloridas de acordo com uma escala discreta associada aos valores correspondentes de cada área. Bailey e Gatrell (1995) apresentam uma sugestão para o cálculo no número de classes da

escala de cores. As classes poderiam ser definidas por intervalos iguais, através dos quantis, com base em desvios padrões ou até com freqüências pré-fixadas nas caudas (trimmed).

Na prática, normalmente o pesquisador já conhece os valores críticos ou intervalos de interesse para o fenômeno que está estudando. Consequentemente, talvez o mais interessante a se fazer não seja aplicar fórmulas para determinar o número nem como serão definidas as classes da escala de cores, mas sim considerar o interesse do pesquisador ou do especialista.

3.3.2. Autocorrelação Espacial

Intuitivamente, podemos acreditar que áreas próximas, dependendo daquilo que se estuda, tendem a apresentar valores mais similares (relação direta) ou dissimilares (relação inversa). Esta idéia de dependência espacial está associada ao conceito estatístico de autocorrelação espacial, onde cálculo da autocorrelação espacial obedece à maneira de quantificação da dependência espacial. O termo autocorrelação assume o prefixo "auto" por fazer referência a uma mesma variável aleatória, apenas considerando correlação entre diferentes localizações.

Moran (1950) e Geary (1954) apresentam índices que, assim como o variograma, correspondem a ferramentas utilizáveis na quantificação da magnitude desta autocorrelação. Tais índices, porém, carregam limitações ao considerar que a variável aleatória de interesse é identicamente distribuída nas áreas. Para – principalmente – dados epidemiológicos, onde na maioria das vezes consideramos taxas ou riscos, é muito difícil que esta suposição seja satisfeita pelo fato de que a distribuição deste tipo de dado depende do tamanho da população em risco, propondo distribuições de probabilidades diferentes para a variável aleatória associada a cada área. Ou seja, áreas com tamanhos diferentes (contingente de população em risco diferentes), em que taxas, riscos ou proporções estão sendo consideradas, apresentarão variabilidades diferentes, acarretando, consequentemente, a não aplicabilidade dos índices de Moran e Geary. Assunção e Reis (1999) propuseram um índice que mede a autocorrelação espacial para dados epidemiológicos.

Nesta monografia, não se fará necessário abordar nenhum destes índices, pois não compete ao mapeamento de doenças se fixar no estudo e na quantificação da dependência

espacial entre as áreas. Para o foco deste trabalho, é suficiente apenas saber com quais outras cada uma das áreas se comunica. Uma melhor apresentação sobre as relações necessárias entre as áreas para a modelagem utilizada no mapeamento de doenças será apresentada no Capítulo 4.

3.4. Epidemiologia Espacial

A etimologia da palavra epidemiologia apresenta "o estudo das doenças" como significado para a mesma. Mais do que isto, a epidemiologia é a ciência que estuda quantitativamente fenômenos de saúde e doenças, buscando compreender ao máximo os desfechos e fatores explicativos.

Quanto ao conceito de epidemiologia espacial, ramo ou subárea da estatística espacial, seu significado lógico se origina da fusão dos conceitos de estatística espacial e epidemiologia, como não poderia deixar de ser. Nas palavras de Lawson (2001): "*A epidemiologia espacial corresponde à análise da distribuição espacial (geográfica) da incidência de doenças*".

Dentro da epidemiologia espacial, uma das maneiras – utilizada por Elliott, Wakefield, Best & Briggs (2001) – de categorização da natureza ou tipo do estudo e, consequentemente, dos métodos de análise a serem empregados, efetua tal distinção de acordo com as quatro tipologias apresentadas na seqüência.

3.4.1. Mapeamento de Doenças

O mapeamento de doenças é realizado com intuito de descrever e sumarizar por completo a distribuição e a variação da doença no mapa. Os modelos aqui aplicados simplesmente fornecem estimativas para o parâmetro de interesse associado à doença em estudo, de maneira satisfatória para a geração de mapas informativos, ou seja, que possam ser traduzidos em considerações objetivas. Sendo assim, através das estimativas calculadas, deseja-se unicamente levantar informações sobre a etiologia da doença capazes de auxiliar em tomadas de decisões – por exemplo, para a implementação de políticas públicas –, sustentar hipóteses e sugerir estudos futuros. O próximo capítulo abordará de forma mais aprofundada este tipo de estudo.

3.4.2. Estudo de Correlação Ecológica

Esta tipologia compreende outras diversas variações em sua nomenclatura. Análise ecológica, regressão ecológica e estudo de correlação geográfica são alguns dos exemplos. Este tipo de estudo apresenta modelos muito similares aos modelos utilizados para o mapeamento de doenças e, sob a mesma perspectiva estatística, fornecem a distribuição da doença no mapa. Entretanto, apresenta a grande diferença em relação ao mapeamento de doenças no que se refere aos interesses e objetivos do estudo. Aqui, prioriza-se compreender e estabelecer a relação entre a distribuição espacial da incidência da doença e variáveis explicativas, diferentemente do foco descritivo do mapeamento de doenças. Tais variáveis podem ser, por exemplo, de âmbito ambiental ou social.

3.4.3. Estudo de Origens

São estudos apropriados para quando se suspeita de que pontos (indústrias, por exemplo) ou linhas (rodovias, por exemplo) no espaço geográfico representem a origem ou fonte do acréscimo no risco/taxa de incidência da doença ou apresentam potencial perigo ambiental.

3.4.4. Cluster de Doenças

Este consiste na detecção de clusters (conglomerados), caracterizados por região geográfica com elevada incidência da doença, sem que haja hipóteses etiológicas pré-definidas. Ou seja, preocupa-se com a análise de agregações atípicas de doenças.

Capítulo 4

Mapeamento de Doenças

O mapeamento de doenças vem apresentando considerável progresso no desenvolvimento de suas metodologias. A preocupação com a qualidade dos métodos surge ao classificarmos a técnica como ferramenta imprescindível e de alto grau de importância para a análise da saúde pública regional. Além disso, vem ostentando o posto de uma das principais aplicações no campo dos dados de área, fazendo com que, por conseguinte, sejam assumidas as características de um estudo em dados de área – em especial no que refere ao trabalho com dados oriundos de levantamentos populacionais – apresentadas na Seção 3.3.

Neste capítulo, a idéia fundamental do mapeamento de doenças é apresentada na Seção 4.1, seguida pelos princípios da modelagem estatística – passando pelo conceito de que a verdadeira taxa de incidência, risco relativo ou proporção, de cada área, em um determinado espaço de tempo, é vista como fruto de uma variável aleatória – e pela abordagem clássica do mapeamento de riscos relativos (Seções 4.2 e 4.3, respectivamente). A Seção 4.4 apresenta o problema das pequenas áreas, classificado como uma das grandes justificativas para utilização da modelagem Bayesiana apresentada, na forma de modelo hierárquico, na Seção 4.5. No entanto, o foco deste trabalho é abordado na Seção 4.6 e consiste na apresentação do modelo inteiramente Bayesiano para produção de estimativas melhores e mais realistas para os riscos das áreas em análise, com intuito de superar problemas que serão apresentados na seção 4.4 deste trabalho. A Seção 4.7 faz referência à especificação da relação entre as áreas de uma região através de uma matriz de pesos. O capítulo finaliza com a Seção 4.8, que trata brevemente a respeito de critérios de seleção e comparação entre diferentes modelos utilizados na geração de estimativas a serem mapeadas.

Como referência para mapeamento de doenças, além de Lawson (2001) e Elliot (2001), já citados no capítulo anterior, sugere-se Mollié (1996) e a produção nacional de Assunção (2001). Em relação a artigos e aplicações práticas, cita-se Assunção (1998), Richardson(2004) e Ehlers (2006).

4.1. Introdução ao Mapeamento de Doenças

Como já foi dito na Subseção 3.4.1, corresponde a um dos ramos da epidemiologia espacial, assim como a análise ecológica, estudo de origens e cluster de doenças. Esta técnica apresenta grande interesse e utilidade dentro da sociedade, visto que, através dela, podemos identificar áreas de risco elevado ou de desempenho admirável, fornecendo sustentabilidade para implementação de políticas públicas, amparando hipóteses e sugerindo diretrizes para novos estudos.

Nas palavras de Lawson (2000):

> O mapeamento de doenças pode ser utilizado para se avaliar a necessidade de alocação de recursos para a saúde de acordo com a variação geográfica, assim como pode ser útil em estudos de investigação da relação da incidência da doença com variáveis explicativas.

Para a primeira das utilidades citadas por Lawson (2000), considera-se que a proposta do mapeamento é de produzir mapas "limpos". Ou seja, livres de quaisquer ruídos aleatórios e perturbações externas que acrescentem à variabilidade populacional e degredem a verdadeira distribuição espacial do fenômeno. A segunda das utilidades consiste em apresentar hipóteses específicas que relacionem a incidência a ser estimada com informações adicionais incluídas na análise (covariáveis, por exemplo).

Na maioria das vezes, o mapeamento é constituído por taxas de incidência, riscos relativos ou proporções, associados às áreas do mapa. Ou seja, para cada uma das áreas da região sob análise, um valor de taxa, risco ou proporção, é atribuído. A próxima seção introduz o pensamento da modelagem estatística na geração de estimativas para os tais riscos relativos, taxas de incidência ou proporções.

4.2. O Modelo Estatístico

Em termos estatísticos, o objetivo do mapeamento de doenças é de desenvolver metodologias para criar, através de estimativas geradas por um modelo proposto, um mapa que descreva de forma correta a distribuição espacial dos riscos nas áreas de uma região. Há de se ressaltar que a modelagem não se aplica apenas para o contexto em que o que se deseja mapear são doenças ou variáveis epidemiológicas, mas também à outras situações onde o interesse consiste em variáveis sociais, econômicas, entre outras.

Um questionamento natural pode transparecer ao leitor quanto à utilização de modelos estatísticos para dados de levantamentos populacionais, visto que estes normalmente são interpretados como quantidades fixas. Na abordagem epidemiológica (assim como em quaisquer outras que o leitor puder associar o seguinte raciocínio), porém, número observado de eventos (contagem) é visto como uma realização de uma variável aleatória, mesmo que proceda de um levantamento populacional. O argumento de defesa consiste na idéia de que o número de ocorrências de um evento é fruto de ações, decisões e acontecimentos aleatórios e que, se pudéssemos transcorrer no tempo, haveria a possibilidade de que a contagem antes verificada não fosse a mesma.

Sendo o número de eventos uma variável aleatória, nada mais natural que se desenvolvam modelos estatísticos para a mesma. As mais diversas modelagens têm sido empregadas, desenvolvidas e discutidas, com intuito fornecer estimativas robustas a respeito do fenômeno em estudo.

Formalizando, considere Y_i, para $(i = 1,...,n)$, a variável aleatória número de eventos, onde i é o índice que faz referência à área geográfica e o conjunto das n áreas é chamado de região. Em uma primeira etapa, é necessário atribuir uma distribuição de probabilidade para Y_i. Normalmente, quando se trabalha com contagem de eventos e, principalmente, se os eventos forem raros, a distribuição empregada é a de Poisson. Ou seja,

$$Y_i \mid E_i, \lambda_i \sim Poisson(E_i \lambda_i), \; i = 1,...,n. \tag{4.1}$$

E, consequentemente, a função de verossimilhança associada é

$$f(y_i \mid \lambda_i) = \frac{e^{-E_i \lambda_i} (E_i \lambda_i)^{y_i}}{y_i!}.$$

A quantidade E_i corresponde ao número esperado de eventos na área i caso o risco nessa área seja igual ao risco total na região, ou seja, a suposição é de risco constante em toda região. Não é, porém, considerada como parâmetro desconhecido, mas sim conhecida, tal que, usualmente utiliza-se

$$E_i = rN_i = \frac{\sum_i y_i}{\sum_i N_i} N_i, \tag{4.2}$$

onde r é a taxa de incidência global observada e N_i é o tamanho da população em risco na área i.

O cálculo do valor esperado E_i pode também ser realizado fazendo-se uma padronização indireta. Neste caso, consideram-se classes (de acordo com covariáveis como idade, sexo, entre outras) e a suposição é de que o risco seja constante dentro de cada área em toda região, porém não necessariamente igual entre as classes. Como exemplo, poderíamos pensar em dividir uma população em risco por faixas etárias (classes) e calcular um E_i para cada faixa etária, considerando que o risco do fenômeno sobre as diferentes faixas não é o mesmo (veja Assunção (2001), página 35).

A quantidade λ_i é o parâmetro de interesse e representa o verdadeiro risco da área i, relativo ao global. Por risco relativo, entende-se a razão entre a probabilidade de ocorrência do evento de dois grupos de indivíduos, no caso, entre a dos indivíduos de certa área i e a dos indivíduos da região inteira.

Para o caso de eventos não raros, a suposição de distribuição de *Poisson* para Y_i em (4.1) não é adequada. A alternativa natural é a do emprego da distribuição *Binomial*. Esta situação, porém, não será apresentada neste trabalho.

4.3. Modelagem Clássica para Riscos Relativos

Esta consiste na forma clássica e mais simples do mapeamento de doenças. Aqui, simples estimativas dos riscos relativos de interesse são atribuídos às áreas do mapa. Utilizando o método de máxima verossimilhança, pode-se mostrar que a estimativa para o verdadeiro risco na área i é dada pela SMR_i (*Standard Mortality Ratio*), onde

$$SMR_i = \hat{\lambda}_i = \frac{\frac{y_i}{N_i}}{\frac{\sum_i y_i}{\sum_i N_i}} = \frac{\frac{y_i}{N_i}}{r} = \frac{y_i}{rN_i} = \frac{y_i}{E_i}, \tag{4.3}$$

com erro padrão dado por $s_i = \sqrt{y_i}/E_i$.

Três importantes considerações a respeito da modelagem clássica contribuem para que esta seja gradativamente, na prática, substituída pela modelagem Bayesiana. A primeira delas está apresentada na Seção 4.4 e profere que as estimativas desta forma geradas para os riscos frequentemente são perturbadas pelo problema das pequenas áreas. Uma segunda consideração é que, se, além de um tamanho de área pequeno, o evento for

raro, podemos constatar um problema de variabilidade superior àquela assumida pelo modelo de Poisson. Por fim, a modelagem clássica também não leva em consideração a possibilidade de autocorrelação espacial, apresentada na Seção 3.3.2.

4.4. O Problema das Pequenas Áreas

A análise de dados de área prioriza o trabalho com áreas pequenas, visando a precisão mais próxima possível da pontual. Porém, pode acarretar em um problema muito comum quando trabalhamos com taxas, riscos e proporções, popularmente conhecido como o problema das pequenas áreas. Especificamente, este não está relacionado a um espaço geográfico pequeno, mas sim à áreas com pequena população em risco. Nestas situações, a diferença de uma unidade na contagem dos eventos, que pode inclusive estar relacionada a erros substanciais de registro ou à falta de diagnóstico, pode alterar bruscamente um risco relativo, taxa ou proporção.

Exemplo: Considere uma região composta por três áreas tal que:

Área A = 1 evento em uma população em risco de tamanho 20;

Área B = 1 evento em uma população em risco de tamanho 200;

Área C = 1 evento em uma população em risco de tamanho 2000.

Com a utilização de (4.3), calcula-se que $SMR_A = SMR_B = SMR_C = 1$. E, ao considerar um evento a mais nessa região, pode-se verificar que:

- Caso o evento adicional tenha ocorrido na área A, então

  $SMR_A \cong 1,98$, enquanto $SMR_B = SMR_C \cong 0,9911$;

- Caso o evento adicional tenha ocorrido na área B, então

  $SMR_B \cong 1,0901$, enquanto $SMR_A = SMR_C \cong 0,9911$;

- Caso o evento adicional tenha ocorrido na área C, então

  $SMR_C \cong 1,0010$, enquanto $SMR_A = SMR_B \cong 0,9911$.

Ao nos depararmos com uma informação do tipo $SMR_A \cong 1,98$, se cruamente analisada, estaria representando um quadro de calamidade pública. Este exemplo intuitivamente indica o quanto a confiabilidade dos riscos relativos associadas a

populações pequenas é baixa, fruto de alta flutuação aleatória. Ou seja, uma diferença casual e atípica na ocorrência de poucos casos do evento – que faria com que, na prática, a nova SMR_A tivesse seu valor, do ponto de vista interpretativo, desprezado – poderia, então, estar mascarando ou alterando, de forma não correspondente ao comportamento realista, o risco relativo associado às áreas.

Modelos estatísticos mais sofisticados são implementados com objetivo de permitir inclusive a análise e a interpretação dos riscos associadas às pequenas áreas, através de estimativas mais realistas e próximas de um risco que realmente represente a situação em que se encontram as áreas da região. Ou seja, para populações em risco reduzidas onde taxas, riscos ou proporções, estão sendo analisados, um tratamento adicional nos dados deve ser considerado. Como solução, principalmente três idéias são propostas na literatura.

4.4.1. Agregar Áreas

Consiste no ato de agregar pequenas áreas próximas, com intuito de aumentar o contingente populacional. Esta sugestão, além de permitir que áreas com comportamento diferente sejam unidas, dissimulando o real valor da variável em questão para a nova área gerada, confronta com o princípio maior de identificação da distribuição espacial com maior precisão possível, ou seja, em áreas mais restritas. Em outras palavras, perdemos a informação localizada.

4.4.2. Mapa de Probabilidade

Proposto por Choynowski (1959), se trata de um mapa temático onde os indicadores relacionados a cada uma das áreas são substituídos pela probabilidade de se obter uma contagem que é mais extrema do que aquela observada, sob hipótese nula de risco constante na região (ou p-valor de qualquer outro teste de interesse). Sendo assim, quando um resultado observado já corresponder a um valor extremo, muito alto ou muito baixo, em relação ao esperado sob hipótese de risco constante em toda a região do mapa (uniformidade), a probabilidade associada à área será baixa. Para populações pequenas, o mapa de probabilidade é satisfatório, além de substituir as taxas por uma medida que leva em conta a natureza estocástica da contagem, considerando sua variabilidade. Para populações grandes, porém, pode apresentar frequentemente valores associados às áreas próximos a zero, quando de fato a distância entre observado e o esperado foi relativamente

baixa. Outros problemas e inconveniências dos mapas de probabilidades estão apresentados em Assunção (2001) – veja página 40 –, e fazem com que esta técnica perca espaço para o desenvolvimento de metodologias Bayesianas.

4.4.3. Modelagem Bayesiana

Esta vem sendo muito difundida, se caracterizando por uma metodologia de estimação para o mapeamento de doenças que funciona como alternativa às duas outras soluções propostas anteriormente no combate ao problema das pequenas áreas. A modelagem Bayesiana tem como principal fundamento a idéia de incluir uma distribuição à priori para os riscos relativos, possibilitando que não só a informação da própria área seja utilizada na construção da estimativa do risco, como também as informações derivadas de outras áreas da região, incorporando o conceito de autocorrelação espacial apresentado na Subseção 3.3.2. Além disso, a atribuição de uma priori que considere o efeito espacial permite estabelecer uma conexão entra as áreas – não presente na modelagem clássica – que pode incorporar a idéia intuitiva de regularidade perante as áreas de uma mesma região ocasionada por ações públicas ou características políticas comuns à esta determinada região. Sendo assim, procura uma melhor estimativa para a taxa ou risco associado às áreas da região em estudo. A modelagem Bayesiana, para que realmente corresponda às características acima relatadas, necessita assumir uma estrutura hierárquica, como apresentado na Seção 4.5.

4.5. Modelagem Bayesiana Hierárquica para Riscos Relativos

Por modelagem Bayesiana, como já foi mencionado, entende-se aquela que atribui informações à priori para os riscos relativos das áreas. Consequentemente, as inferências são derivadas da distribuição à posteriori resultante da combinação entre a informação da priori com a oriunda dos dados.

Sendo assim, ao considerarmos uma distribuição à priori para os riscos relativos, já adentramos em um caso de modelagem hierárquica[2], tal que o primeiro estágio da hierarquia está associado à verossimilhança $l(\boldsymbol{y} \mid \boldsymbol{\lambda})$, o segundo à priori $\pi(\boldsymbol{\lambda} \mid \boldsymbol{\gamma})$ condicionada aos hiperparâmetros.

Com fins de formalização, temos que:

$\boldsymbol{y} = (y_1, ..., y_n)$ é o vetor das contagens dos eventos;

$\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$ é o vetor de parâmetros (riscos relativos) a serem estimados;

n é o número de áreas da região.

Considera-se padrão o procedimento de atribuir uma distribuição de Poisson para cada Y_i, como já foi visto na Seção 4.2, supondo independência entre as contagens das áreas. A hipótese de independência é bastante realista para quando o evento for não contagioso. Sendo assim, a função de verossimilhança associada ao vetor $\boldsymbol{y}$ corresponde a

$$l(\boldsymbol{y} \mid \boldsymbol{\lambda}) = \prod_{i=1}^{n} \frac{e^{-E_i \lambda_i} (E_i \lambda_i)^{y_i}}{y_i !}.$$

Quanto à determinação das prioris, percebe-se que, na prática, prioris especificadas com todos os hiperparâmetros (parâmetros da priori) conhecidos são raramente utilizadas. A abordagem usual considera que as distribuições à priori para os riscos têm estrutura composta por hiperparâmetros desconhecidos.

Então, posteriormente à definição de prioris para os riscos relativos, é preciso que os hiperparâmetros sejam estabelecidos. A especificação à respeito dos hiperparâmetros deve ser realizada subjetivamente de modo a retratar o conhecimento do pesquisador a respeito do risco relativo em questão. Duas maneiras de especificações dos hiperparâmetros acabam por subdividir esta modelagem Bayesiana em duas correntes: a modelagem Bayesiana empírica e a modelagem inteiramente Bayesiana.

---

[2] Na literatura, porém, a definição do termo modelagem Bayesiana hierárquica, quando tratamos do mapeamento de doenças, apresenta-se duvidoso e muitas vezes confundido com a modelagem inteiramente Bayesiana (veja Subseção 4.5.2), visto que esta compreende uma modelagem com três estágios de hierarquia.

4.5.1. Modelagem Bayesiana Empírica

A possibilidade da modelagem Bayesiana empírica, não profundamente discutida neste trabalho, assume hiperparâmetros desconhecidos, porém utiliza estimativas pontuais de máxima verossimilhança para tais, como uma forma de aproximação. Esta modelagem, no entanto, pode subestimar a variabilidade dos riscos estimados (além de não lidar bem com a variabilidade das estimativas produzidas), essencialmente quando a priori condicionada aos hiperparâmetros apresentar alta dispersão, além do que não permite generalizações para situações mais complexas, como casos de análise espaço-temporal. Ou seja, a estimação Bayesiana empírica ignora a variabilidade introduzida pelos hiperparâmetros justamente por não considerar a incerteza na estimação dos mesmos (considera-os fixos).

A idéia básica consiste na utilização dos próprios dados da amostra para estimar os valores dos hiperparâmetros e então utilizar estas estimativas como valores fixos para os mesmos.

Na seqüência deste trabalho, não trataremos mais a respeito da estimação bayesiana empírica. Como principal referência, cita-se Marshall (1991).

4.5.2. Modelagem Inteiramente Bayesiana

A modelagem inteiramente Bayesiana considera que, por assumir que os hiperparâmetros são quantidades aleatórias, hiperprioris devem ser especificadas aos mesmos. A literatura classifica o método inteiramente Bayesiano como preferível, principalmente por considerar toda a variabilidade através da imposição de distribuições de probabilidade para os hiperparâmetros e por possuir um processo inferencial mais rico, apesar de mais sofisticado.

É por isso que a modelagem inteiramente Bayesiana para os riscos das áreas em análise não só se apresenta superior à metodologia de estimação Bayesiana empírica como supera contras e desvantagens do agregar de áreas e dos mapas de probabilidades no combate ao problema das pequenas áreas, reduzindo consideravelmente o problema da variabilidade dos riscos estimados.

A especificação de hiperprioris se torna possível através da inserção de uma estrutura hierárquica na modelagem. Além disso, especificação inteiramente Bayesiana

compreende, pelo menos, três estágios de hierarquia, sendo o primeiro associado à verossimilhança $l(\boldsymbol{y} \mid \boldsymbol{\lambda})$, o segundo à priori $\pi(\boldsymbol{\lambda} \mid \boldsymbol{\gamma})$ – condicionada aos hiperparâmetros – e o terceiro à hiperpriori $\pi(\boldsymbol{\gamma})$.

Então, ao considerarmos que os hiperparâmetros dos riscos são quantidades aleatórias, caracterizamos:

$\boldsymbol{\gamma}$ é o vetor de parâmetros da priori (ou hiperparâmetros);

$\pi(\boldsymbol{\lambda} \mid \boldsymbol{\gamma})$ é a priori do vetor de parâmetros $\boldsymbol{\lambda}$ condicionada aos hiperparâmetros;

$\pi(\boldsymbol{\gamma})$ é a hiperpriori associada ao vetor de hiperparâmetros.

Consequentemente, a posteriori conjunta para todos os parâmetros desconhecidos do modelo resulta em

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\gamma} \mid y) = \frac{l(\boldsymbol{y} \mid \boldsymbol{\lambda})\pi(\boldsymbol{\lambda} \mid \boldsymbol{\gamma})\pi(\boldsymbol{\gamma})}{\int l(\boldsymbol{y} \mid \boldsymbol{\lambda})\pi(\boldsymbol{\lambda}, \boldsymbol{\gamma}) d\boldsymbol{\lambda} d\boldsymbol{\gamma}},$$

e a posteriori marginal de interesse é dada por

$$\pi(\boldsymbol{\lambda} \mid y) = \int \pi(\boldsymbol{\lambda}, \boldsymbol{\gamma} \mid y) d\boldsymbol{\gamma}.$$

Ao tratar com modelagens Bayesianas, é muito comum se deparar com integrais que não permitem a obtenção da solução por meios analíticos e, algumas vezes, até por integração numérica. Eis quando se faz necessária a utilização de métodos computacionalmente intensivos. Mais especificamente, métodos MCMC.

Popularmente, assim como neste trabalho, todavia, não opta-se por modelar o risco relativo λ_i, mas sim $\eta_i = \log(\lambda_i)$. A justificativa é baseada no fato de que as prioris que para $\log(\lambda_i)$ podem ser especificadas facilitam o trabalho com a modelagem (Mollié (1996)). Mais especificamente, enquanto a priori conjugada para λ_i se trata de uma distribuição Gama, para η_i torna-se adequada a atribuição de uma distribuição Normal. Na seqüência do trabalho, então consideraremos

$$\boldsymbol{\eta} = (\log(\lambda_1), ..., \log(\lambda_n)). \tag{4.4}$$

4.6. A Especificação do Modelo Inteiramente Bayesiano

Já foi apresentada na Subseção 4.5.2 a situação que caracteriza a modelagem inteiramente Bayesiana. Esta seção, contudo, tratará de abordar desde a escolha das prioris até a escolha das hiperprioris, mesmo sabendo que apenas a segunda caracteriza a modelagem inteiramente Bayesiana. Optou-se, porém, por tratar nesta seção a respeito da especificação das prioris pela conveniência da associação entre a estruturação da priori com a especificação das hiperprioris.

Quanto à informação à priori, representada pela distribuição à priori, será responsável por especificar a variabilidade dos riscos relativos ao longo do mapa. Em outras palavras, esta reflete o conhecimento que temos sobre a variação dos riscos no espaço. Consequentemente, é através da modelagem da distribuição à priori que podemos introduzir a dependência espacial entre os riscos.

4.6.1. Prioris para Riscos Não-Estruturados

Quando falamos em riscos não estruturados, nos referimos a um modelo de risco que não considera uma estrutura que permita a inserção de efeito espacial. Este é o modelo mais simples e admite que o logaritmo natural do risco possa ser visualizado de tal forma que

$$\eta_i = \mu + u_i \,, \tag{4.5}$$

onde μ representa uma média global para os $\eta_i's$ e u_i obedece a efeitos aleatórios que correspondem à heterogeneidade não estruturada presente nos dados. Uma suposição natural é considerar os efeitos aleatórios $u_i's$ das áreas independentes e identicamente distribuídos, tal que

$$u_i \sim N(0, \sigma^2) \,. \tag{4.6}$$

Para este caso, então, têm-se que a priori empregada ao logaritmo dos riscos corresponde à

$$(\eta_i \mid \boldsymbol{\gamma}) \sim N(\mu, \sigma^2) \,, \tag{4.7}$$

onde $\boldsymbol{\gamma} = (\mu, \sigma^2)$.

Sob hipótese de independência entre os logaritmos naturais dos riscos das áreas, a posteriori conjunta é dada por

$$\pi(\boldsymbol{\eta} \mid \boldsymbol{\gamma}) = \pi(\boldsymbol{\eta} \mid \mu, \sigma^2) = \prod_{i=1}^{n} \pi(\eta_i \mid \mu, \sigma^2). \tag{4.8}$$

Alternativamente à estrutura apresentada em (4.5), pode-se permitir que a observação do comportamento de uma covariável contribua na explicação do risco de cada área. Neste caso, diferentes médias estarão associadas às áreas com diferentes valores para as covariáveis. A idéia consiste em considerar que a média de cada área é estimada através de uma regressão linear com a covariável. Ou seja,

$$\eta_i = \alpha_0 + \alpha_1 z_i + u_i, \tag{4.9}$$

onde:

α_0 representa a media geral de η;

α_1 representa o efeito médio global de z em η;

z_i é o valor observado da covariável Z na área i.

Esta nova alternativa mantém a suposição de homogeneidade das variâncias. Ou seja, continua-se assumindo que variabilidade dos efeitos aleatórios u_i é comum para todas as áreas. Conclusivamente, a priori para o logaritmo do risco pode ser expressa por

$$(\eta_i \mid \boldsymbol{\gamma}) \sim N(\mu_i, \sigma^2) = N(\alpha_0 + \alpha_1 z_i, \sigma^2). \tag{4.10}$$

E, sob hipótese de independência entre os logaritmos naturais dos riscos das áreas,

$$\pi(\eta \mid \boldsymbol{\gamma}) = \prod_{i=1}^{n} \pi(\eta_i \mid \mu_i, \sigma^2). \tag{4.11}$$

Até este momento, apenas considerando a estrutura da distribuição à priori, não se adentrou na modelagem inteiramente Bayesiana, mas sim em uma metodologia Bayesiana hierárquica em dois estágios que pode também ser empregada no caso da modelagem Bayesiana empírica.

A modelagem inteiramente Bayesiana exige especificação de hiperprioris para os hiperparâmetros α_0, α_1 e σ^2. Usualmente, as hiperprioris especificadas são conjugadas, visando facilidades analíticas, e, para α_0 e α_1, atribui-se, respectivamente, $N(0, \sigma^2_{\alpha_0})$ e $N(0, \sigma^2_{\alpha_1})$. Se a idéia for utilizar uma priori não informativa, sugere-se tanto a atribuição de

altos valores para $\sigma^2_{\alpha_0}$ e $\sigma^2_{\alpha_1}$ nas Normais apresentadas anteriormente quanto uma distribuição uniforme $U(-\infty,+\infty)$ para α_0 e α_1. A segunda corresponde a uma priori imprópria, mas conduz a uma posteriori própria. Para $\tau^2 = 1/\sigma^2$, atribui-se uma priori conjugada $Gama(a,b)$, com a e b fixados.

O modelo com uma variável pode facilmente ser ampliado para um maior número de covariáveis associadas aos riscos das áreas. Quando isso ocorrer, e se assumirmos que os efeitos aleatórios das áreas são independentes, considera-se que o vetor $\boldsymbol{\alpha}$ tem distribuição Normal multivariada com vetor de médias 0 e matriz de covariâncias $\Sigma = \mathbf{I}\sigma^2_{\alpha_j}$, onde j é o índice que faz referência à covariável e $\mathbf{I}$ corresponde à matriz identidade.

Considerando o vetor de quantidades desconhecidas $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \boldsymbol{u}, \sigma^2)$, temos que sua posteriori é dada por

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{n} l(y_i \mid \alpha_0, \alpha_1, u_i)\pi(\alpha_0)\pi(\alpha_1)\pi(u_i \mid \sigma^2)\pi(\sigma^2). \qquad (4.12)$$

Tal posteriori, devido à considerável complexidade e às prioris que usualmente são empregadas, não apresenta forma fechada, criando a necessidade de se utilizar metodologias MCMC para o processo inferencial, com intuito já apresentado na seção 2.3.

4.6.2. Prioris para Riscos Espacialmente Estruturados

Já foi mencionada a naturalidade da idéia de que áreas geográficas próximas tendem a apresentar correlação entre os riscos. Então, um modelo para o risco que considera a estrutura espacial local existente será o associado ao tratarmos de riscos espacialmente estruturados.

A informação à priori será a responsável pela inserção do conhecimento a respeito do efeito espacial no modelo, através da imposição de determinada estrutura de covariância. Besag et al. (1974) introduziu a idéia dos modelos autoregressivos condicionais (Conditional Autoregression (CAR)), que permitem variação espacialmente estruturada. Autoregressivo, nesse contexto, faz referência à utilização da informação da mesma variável de interesse, porém correspondente às outras áreas do mapa na construção da estimativa para determinada área i. Uma classe muito conveniente, na prática, é a dos modelos de campos aleatórios de Markov (Markov Random Field, MRF). Tais consideram

que o risco a área i tem correlação unicamente com seus vizinhos imediatos δ_i, ou seja, apenas com as áreas que fazem fronteira com a área i. Mais especificamente, a distribuição à priori para o risco da área i, condicionada ao risco de todas as outras áreas do mapa, depende apenas dos riscos dos vizinhos imediatos δ_i de i.

Usualmente trabalha-se com os modelos MRF Gaussianos, que especificam distribuição Normal com média dependente dos valores η_{δ_i} para priori condicional de η_i. Estes assumem variância condicional constante. Entretanto, para mapas irregulares, onde o número de vizinhos varia de área para área, é apropriada a utilização dos modelos autoregressivos Gaussianos intrínsecos, que serão o foco daqui pra frente. Para estes modelos, a variância condicional à priori de η_i, dado todos os logaritmos naturais dos riscos das outras áreas do mapa, é inversamente proporcional ao número de vizinhos da área i.

Para começarmos a introduzir a modelagem, consideramos η_i na seguinte forma:

$$\eta_i = \log(\lambda_i) = \mu_i + b_i . \tag{4.13}$$

Onde μ_i é o mesmo já apresentado na decomposição de η_i da Seção 4.6.1 e b_i são os efeitos aleatórios que representam a componente espacialmente estruturada, diferentemente dos efeitos aleatórios não-estruturados $u_i's$.

Então, enquanto μ_i pode ser comum às áreas do mapa ou estimado por uma regressão com uma ou mais covariáveis, b_i tem distribuição condicional aos efeitos espaciais das outras áreas, segundo o modelo CAR gaussiano intrínseco, dada por

$$\pi(b_i \mid b_j, j \neq i) \sim N\left(\frac{\sum_{j \in \delta_i} w_{ij} b_j}{\sum_{j \in \delta_i} w_{ij}}, \frac{\sigma^2}{\sum_{j \in \delta_i} w_{ij}} \right), \tag{4.14}$$

onde w_{ij} representa o peso (em outras palavras, a influência) do vizinho j na área i, em discussão na Seção 4.7. Percebe-se então que quanto maior o número de vizinhos da área i, menor é a variabilidade que vamos atribuir à priori na estimação de η_i.

Logo, a distribuição conjunta condicional à priori, sob hipótese de independência à priori entre os efeitos aleatórios espaciais b_i, é dada por

$$\pi(\boldsymbol{b} \mid \sigma^2) \propto \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j<i} w_{ij} (b_i - b_j)^2 \right\}. \tag{4.15}$$

Esta priori possui média zero e matriz inversa de covariância com elementos da diagonal principal iguais a w_{i+}/σ^2 e demais elementos iguais a $-w_{ij}/\sigma^2$, onde $w_{i+} = \sum_{j \in \delta_i} w_{ij}$. Além disso, é uma priori imprópria, mas que pode ter este problema corrigido através da imposição de $\sum_i b_i = 0$.

Usualmente, os pesos w_{ij} são especificados de forma binária, de maneira que se j é vizinho de i, $w_{ij} = w_{ji} = 1$. Caso contrário, $w_{ij} = w_{ji} = 0$. Neste caso, a distribuição condicional de b_i se reduz a

$$\pi(b_i \mid b_j, j \neq i) \sim N\left(\frac{\sum_{j \in \delta_i} b_j}{n_i}, \frac{\sigma^2}{n_i} \right), \tag{4.16}$$

onde n_i corresponde ao número de áreas vizinhas à i.

Considerando o vetor de quantidades desconhecidas $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \boldsymbol{b}, \sigma^2)$, temos que sua posteriori é dada por

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{n} l(y_i \mid \alpha_0, \alpha_1, b_i) \pi(\alpha_0) \pi(\alpha_1) \pi(b_i \mid \sigma^2) \pi(\sigma^2). \tag{4.17}$$

E, assim como no caso anterior, necessitar-se-á recorrer à metodologias MCMC para a realização do processo inferencial.

4.6.3. Prioris de Convolução

Como, na prática, não se tem muito clara a idéia de como escolher entre priori estruturada ou não-estruturada, a solução proposta por Besag (1989), Besag e Mollié (1989) e Besag et al (1991) foi de trabalhar com uma distribuição intermediária composta por uma parte estruturada e outra não-estruturada. Esta se chama priori Gaussiana de convolução e propõe que

$$\eta_i = \mu_i + u_i + b_i, \tag{4.18}$$

onde os componentes são supostos independentes e correspondem aos mesmos anteriormente especificados. Ou seja, u_i representa a heterogeneidade não-estruturada enquanto b_i a variação espacialmente estruturada.

Então, u_i, à priori, sob hipótese de independência dos efeitos aleatórios, tem distribuição $N(0,\sigma_u^2)$ e b_i é modelado por um autoregressivo Gaussiano intrínseco com variâncias condicionais proporcionais a σ_b^2. Onde σ_u^2 e σ_b^2 são hiperparâmetros.

Como são independentes, a soma $u_i + b_i$ tem variância igual a soma das variâncias. Ou seja, igual à $\sigma_u^2 + \sigma_b^2 / w_{i+}$. Quando a razão σ_b^2 / σ_u^2 for pequena, significa que a heterogeneidade não estruturada é mais importante, enquanto se esta razão for pequena, podemos dizer que a variação espacialmente estruturada é quem domina.

4.7. Especificação da Matriz de Pesos w:

A matriz de pesos é uma matriz simétrica, $n \times n$, onde nela ficar-se-á explicitada a relação que as áreas têm, pelo fato de estarmos trabalhando com os modelos MRF (veja Subseção 4.6.2), apenas com seus vizinhos. Ou seja, é esta matriz que vai especificar a influências dos vizinhos na área i, promovendo a estruturação espacial.

Esta seção tem por objetivo apenas apresentar três dos possíveis métodos de escolha da matriz de pesos w.

A matriz binária, mais simples e comumente utilizada na literatura, será a utilizada neste trabalho e considera apenas o fato das áreas fazerem fronteira ou não, assumindo $w_{ij} = 1$ caso seja verdade, ou $w_{ij} = 0$, caso contrário.

Uma segunda opção para matriz w é aquela que considera maior peso ou, em outras palavras, maior influência para tamanhos de fronteira maiores. Ou seja, será baseada nos tamanhos das fronteiras. Por exemplo, w_{ij} pode ser igual ao tamanho, em km, da fronteira entre as áreas i e j. Esta opção, porém, não é tão utilizada na prática pela dificuldade, muitas vezes, de se medir a fronteira.

A terceira e última opção para matriz w apresentada nesta monografia é uma extensão da segunda alternativa e considera não só o tamanho das fronteiras como a presença de barreiras naturais. Por exemplo, dependendo da variável de estudo, pode ser

razoável considerar que duas áreas não sejam tão correlacionadas quando entre elas existe uma montanha, até mesmo um rio ou qualquer outra característica geográfica que possa interferir. Como referência, Mollié (1996).

4.8. Critérios para Comparação e Seleção de Modelos

O critério de comparação e seleção entre modelos com diferentes níveis de complexidade frequentemente encontrado na literatura relacionada e mais popular nas aplicações práticas corresponde ao Critério de Informação da Deviance (Deviance Information Criterion, DIC). Proposto por Spiegelhalter, Best, Carlin, and Linde (2002), trata-se de uma generalização do Critério de Informação de Akaike (AIC) e é dado por

$$DIC(M_k) = \bar{D}_k + p_k = 2\bar{D}_k - D(\bar{\theta}_k), \tag{4.19}$$

onde:

θ_k é o vetor de parâmetros do modelo k;

$\bar{\theta}_k$ é a média à posteriori de θ_k;

D é a função deviance, que mede o ajuste do modelo;

$\bar{D}_k = E_{\theta_k|y}(D(\theta_k))$, representando a média à posteriori da deviance;

$D(\bar{\theta}_k)$ é uma estimativa pontual da deviance obtida através da utilização da média à *posteriori* de θ na expressão do desvio que pode se alterar dependendo da parametrização especificadas nas prioris do modelo;

$p_k = E_{\theta_k|y}(D(\theta_k)) - D(E_{\theta_k|y}(\theta_k)) = \bar{D}_k - D(\bar{\theta}_k)$ é um termo penalizador que mede a complexidade do modelo;

$D(\theta_k) = -2\log l(y \mid \theta_k, k) + 2\log f(y)$ é a posteriori da deviance;

onde:

$l(y \mid \theta_k, k)$ é a função de verossimilhança;

$f(y)$ é uma função de padronização dos dados.

Visto que este critério considera a média à posteriori como estimativa para os parâmetros do modelo, não seria recomendada a utilização do *DIC* para distribuições multimodais ou com grande assimetria.

A estatística *DIC* indica melhor ajuste do modelo relativo ao número de parâmetros quanto menor for seu valor. Ou seja, aquele modelo que obtiver o menor *DIC* calculado é considerado como aquele que melhor pode predizer um novo conjunto de dados com a mesma estrutura dos dados observados. Tal estatística é de fácil e instantânea obtenção através do software WinBugs e obedece à utilizada na seleção entre os modelos neste trabalho. Maiores informações a respeito da *DIC* podem ser encontradas em Spiegelhalter *et al*. (2002).

Uma medida de adequação dos modelos também encontrada na literatura é a da Validação Cruzada. Por não ser discutida neste trabalho, fica como referência Gelfand (1996).

Capítulo 5

Mapeando Riscos Associados à Natalidade em Mulheres Jovens de Porto Alegre

5.1. Introdução

A aplicação prática desta monografia consiste na utilização de técnicas já apresentadas anteriormente em dados reais da cidade de Porto Alegre. Mais especificamente, trataremos de modelar e mapear o risco RR para cada bairro, com $SMR's$ dadas por

$$SMR_i = \frac{\dfrac{\text{total de nascimentos em mulheres abaixo de vinte anos de idade em 2005 no bairro i}}{\text{total de nascimentos em 2005 no bairro i}}}{\dfrac{\text{total de nascimentos em mulheres abaixo de vinte anos de idade em 2005}}{\text{total de nascimentos em 2005}}},$$

com objetivo de identificar áreas geográficas – bairros, para a análise em questão – com comportamento preocupante, visando sustentação para políticas públicas e tomadas de decisões. Afinal, é de consentimento geral que altas taxas de natalidade em mulheres com menos de vinte anos representam, no mínimo, um quadro a ser estudado. O mapeamento será feito com base em estimativas para os $RR's$ oriundas de modelos inteiramente Bayesianos, sabidas as qualidades de tais.

O estudo foi conduzido com a utilização dos softwares WinBugs 1.4.1 – implementação dos modelos e geração das estimativas – e MapInfo 8.0 – produção dos mapas –, além de dados obtidos junto à Prefeitura Municipal de Porto Alegre.

Os dados de natalidade em questão são referentes ao ano de 2005. A disponibilidade é pelo número de nascidos vivos e número de nascidos vivos originários de mães com menos de 20 anos, segregados pelos bairros de Porto Alegre. Especificamente, a razão entre o número de nascidos vivos naturais de mães com menos de vinte anos de idade e o total de nascidos vivos não corresponde à medida tradicional da taxa de natalidade. Em contrapartida, fornece, de modo suficientemente interessante para a detecção das áreas preocupantes, uma medida de intensidade ou representatividade da natalidade em

mulheres com menos de 20 anos de idade em relação à natalidade total, de modo que consideremos

$$Risco_i = \frac{\text{n° de nascimentos em mulheres com menos de vinte anos no bairro i em 2005}}{\text{n° de nascimentos no bairro i em 2005}}$$

Em outras palavras, podemos interpretar a razão em questão como uma medida da natalidade em mulheres jovens relativa à natalidade total. Além disso, segue que, quando tratarmos em risco na seqüência desse trabalho, estaremos nos referindo à razão acima apresentada. De modo que, se um risco é preocupante e alto para determinado bairro, estamos tratando de um bairro com grande proporção de nascimentos oriundos de mulheres com menos de vinte anos de idade em relação ao total de nascimentos. Consequentemente, o risco relativo do bairro i é a razão entre o risco do bairro i e o risco de Porto Alegre.

Também existe a disponibilidade por dados referentes a outras variáveis que, se detectadas como explicativas para o interesse do estudo, poderão vir a ser incluídas na modelagem como covariáveis, com intuito de auxiliar a aperfeiçoar a geração das estimativas para os riscos relativos de cada bairro. Estas correspondem a oito variáveis de caráter econômico e educacional extraídas dos dados levantados no censo demográfico realizado em 2000 pelo IBGE.

Tanto os dados relativos à natalidade quanto os candidatos à covariáveis derivam de levantamentos populacionais. Os primeiros fazem referência à registros oficiais de saúde do município de Porto Alegre enquanto os outros obedecem ao censo demográfico de 2000, considerando que o interesse neste estudo não é de traçar uma regressão ou algum modelo de previsão para os riscos dos bairros de Porto Alegre, mas sim de identificar, através da distribuição espacial dos riscos, áreas com determinado perfil, em certo instante do tempo, com base em estimativas dos riscos dos bairros melhoradas pela modelagem inteiramente bayesiana. Julgou-se que, mesmo que defasadas, se as candidatas à covariáveis apresentarem correlação ou, consequentemente, poder de explicação para os riscos, estas poderiam vir a ser incluídas na modelagem com intuito de melhorar as estimativas geradas. Outro argumento que reforça a defesa para a utilização das variáveis defasadas desrespeita ao pequeno intervalo de defasagem das variáveis. Para variáveis econômicas e educacionais tais quais, respectivamente, renda média e taxa de analfabetismo, percebe-se

que 5 anos não é período suficiente para gerar grandes alterações nas variáveis a nível de bairro.

A divisão em bairros utilizada para Porto Alegre está baseada na regionalização aprovada pelo Conselho do Orçamento Participativo em 1997, diferente – porém com substancial semelhança – da oficial divisão de bairros da cidade. A tabela 5.1 exemplifica o banco de dados disponível.

Bairro	ID	y	total	z1	z2	z3	z4	z5	z6	z7	z8
Agronomia	1	68	249	4,23	3,00	41,07	5,81	7,55	22,82	6,12	7,52
Anchieta	2	9	20	4,92	90,30	40,61	9,70	14,44	31,67	5,88	11,82
Arquipélago	3	41	156	3,15	96,52	56,56	3,48	14,19	35,36	4,71	11,52
Auxiliadora	4	4	83	19,85	40,91	4,14	59,09	0,23	1,84	12,75	0,60
Azenha	5	19	128	11,50	64,77	12,64	35,23	1,57	5,87	10,65	1,67
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Vila Nova	82	69	531	6,12	85,75	26,78	14,25	3,67	13,44	7,88	3,59

Tabela 5.1: Parte do Banco de Dados Utilizado

Onde:

- *ID* é a identificação numérica dos bairros de Porto Alegre, totalizando 82 bairros;
- *y* é o número de nascidos vivos naturais de mulheres com menos de 20 anos em 2005;
- *total* é o número total de nascimentos em 2005;
- $z1$ é a rendia média dos responsáveis, exclusive aqueles sem rendimentos, em salários mínimos, em 2000;
- $z2$ é a porcentagem de responsáveis por domicílio com renda até 10 salários mínimos, em 2000;
- $z3$ é a porcentagem de responsáveis por domicílio com renda até 2 salários mínimos, em 2000;
- $z4$ é a porcentagem de responsáveis por domicílio com renda superior a 10 salários mínimos, em 2000;
- $z5$ é a porcentagem de responsáveis por domicílio analfabetos, em 2000;
- $z6$ é a porcentagem de responsáveis por domicílio com menos de quatro anos de estudo, fazendo referência ao analfabetismo funcional, em 2000;

- $z7$ é a escolaridade média, em anos de estudo, dos responsáveis por domicílios em 2000;
- $z8$ é a taxa de analfabetismo na população de 15 anos e mais, em 2000.

5.2. Modelagem Estatística

A modelagem estatística propõe que o número de nascimentos em mulheres com menos de vinte anos no bairro i segue uma distribuição

$$Y_i \mid E_i, \lambda_i \sim Poisson(E_i \lambda_i).$$

Onde E_i foi calculado através da maneira usual, como apresentado em (4.2), e representa o número esperado de nascimentos oriundos de mulheres jovens para o bairro i no ano de 2005, sob hipótese de que o risco associado à este evento é constante para os bairros e igual ao risco global de Porto Alegre. Não se fez necessário a utilização de padronização indireta principalmente porque estamos trabalhando com uma faixa etária específica e somente com indivíduos do sexo feminino.

5.3. Mapa com Padrão de Cores

O mapeamento dos riscos relativos estimados pelos modelos propostos (um via abordagem clássica e cinco sob a ótica Bayesiana) para os bairros foi realizado com base em cinco classes intervalares definidas pelo especialista (leia-se equipe do Observatório de Porto Alegre/Prefeitura Municipal de Porto Alegre).

As classes estão de acordo com uma escala de tons de vermelho, de modo que quanto mais escuro for tom da cor, maior é o risco relativo estimado associado à área. A Tabela 5.2 relaciona as cinco classes com o respectivo intervalo associado.

Classe	Cor	Intervalo de $\hat{RR}$
1		0,00 |- 0,51
2		0,51 |- 1,02
3		1,02 |- 1,53
4		1,53 |- 2,04
5		2,04 |-| 2,561

Tabela 5.2 - Escala de Cores Utilizada no Mapeamento dos Riscos Relativos Estimados

Na prática, poderíamos interpretar que bairros com riscos relativos estimados associados à classe 5 apresentam situação de extrema preocupação, se caracterizando por áreas que necessitam de políticas públicas imediatas. Bairros que apresentarem estimativa do risco correspondente à classe 4 também apresentam um quadro muito preocupante e requerível de políticas direcionadas.

Infelizmente, o pouco tempo disponível para o estudo da aplicação, assim como a dependência da equipe do Observatório de Porto Alegre/Prefeitura Municipal de Porto Alegre para a produção dos mapas, fez com que novas classes intervalares ficassem impossibilitadas de serem utilizadas.

5.4. Análise Clássica

Preliminarmente à implementação da modelagem Bayesiana, construiu-se as estimativas clássicas – representadas pela SMR – para os riscos relativos dos bairros, com objetivo de justificar a posterior utilização da abordagem Bayesiana.

A Figura 5.1 apresenta gráficos das SMR's e dos erros padrões das estimativas do risco relativo das áreas, ambos versus o tamanho das populações em risco.

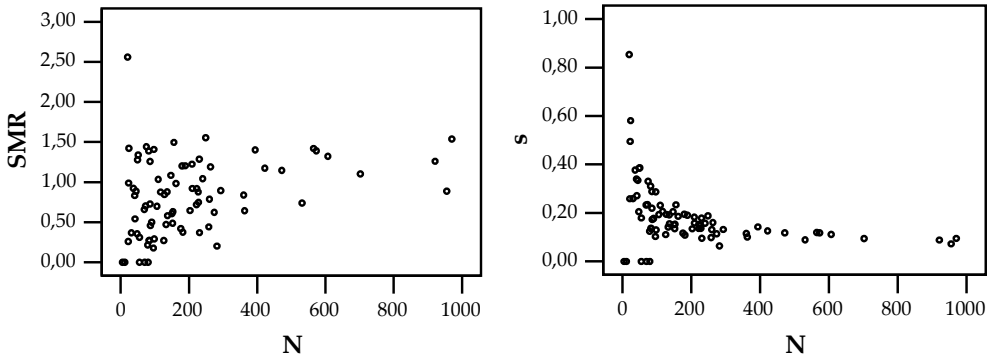

Figura 5.1 - Gráfico SMR's x População em Risco e Gráfico Erro Padrão Estimado x População em Risco

Pode-se claramente perceber a ocorrência do fenômeno apresentado na Seção 4.4 desta monografia, episódio freqüente quando empregamos a modelagem clássica de riscos relativos em áreas com populações em risco reduzidas. A alta flutuação aleatória verificada nas estimativas e a falta de consideração com a influência espacial sobre as áreas servem como argumentos de sustento para o emprego da modelagem Bayesiana.

A Figura 5.2 apresenta o mapa resultante da atribuição das estimativas pontuais da modelagem clássica (SMR's) às correspondentes bairros de Porto Alegre.

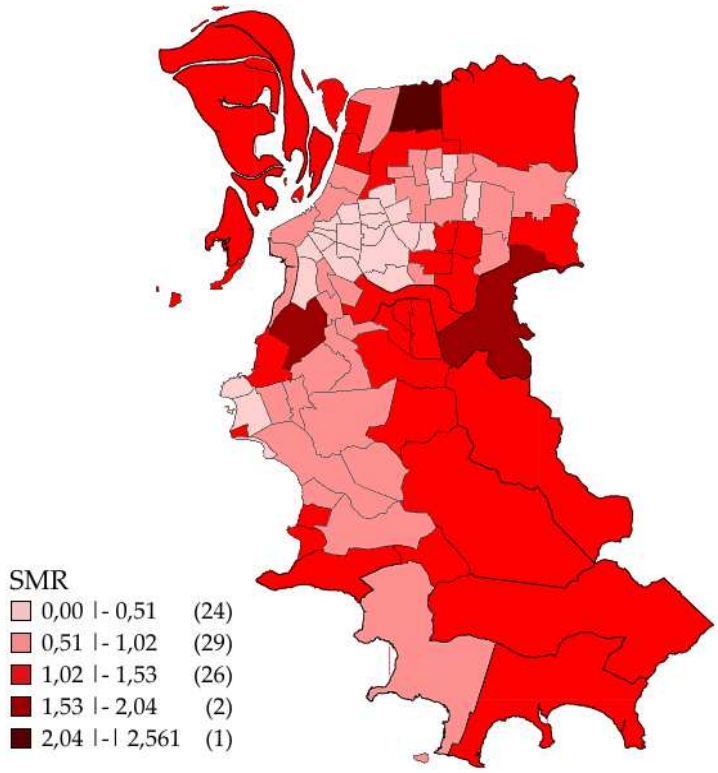

Figura 5.2 - Mapa com Estimativas Clássicas (SMR) para os Riscos Relativos dos Bairros

Uma análise comparativa entre os mapas e estimativas geradas pela abordagem clássica e pelos modelos Bayesianos está apresentada e mais detalhadamente discutida na Seção 5.6.

5.5. Análise Bayesiana

Justificada a utilização da modelagem Bayesiana, testou-se a implementação de cinco diferentes modelos inteiramente Bayesianos, tal que:

- Modelo 1 – Modelo para riscos não-estruturados sem covariável;
- Modelo 2 – Modelo para riscos espacialmente estruturados sem covariáveis;
- Modelo 3 – Modelo de convolução sem covariáveis;
- Modelo 4 – Modelo de convolução com uma covariável;
- Modelo 5 – Modelo de convolução com duas covariáveis.

Para os modelos que admitem estrutura espacial, consideraram-se vizinhos os bairros i e j que fazem fronteira terrestre (veja Seção 4.7). De tal forma que $w_{ij} = w_{ji} = 1$ se i e j forem vizinhos e $w_{ij} = w_{ji} = 0$ caso contrário. É importante considerar, porém, que o bairro Arquipélago foi interpretado como sem vizinhos, visto que o mesmo não satisfez o critério de vizinhança pré-definido.

Com exceção da priori determinada para o efeito espacial, que segue o modelo CAR gaussiano intrínseco, todas as outras prioris empregadas nos modelos construídos são não informativas, essencialmente devido ao desconhecimento que se tem a respeito da distribuição espacial do risco relativo em análise.

A escolha das covariáveis não foi realizada de maneira aprofundada e minuciosa, visto que a prioridade do mapeamento de doenças não consiste especificamente na compreensão e no estabelecimento da relação entre a distribuição espacial da incidência da doença e variáveis explicativas. Então, sabendo que nosso objetivo é apenas de retratar o quadro da natalidade em mulheres jovens de Porto Alegre para o ano de 2005, dispensamos não muito tempo para a escolha das covariáveis inseridas nos modelos 4 e 5. A escolha foi baseada no conhecimento teórico prévio de que o rendimento familiar e a educação dos chefes de família estão correlacionados às taxas de natalidade.

Em um primeiro momento, optou se então – para o modelo 4 – por inserir uma dentre as possíveis covariáveis. Esta foi escolhida de acordo com uma análise do coeficiente de correlação de Pearson entre as variáveis do banco e a SMR calculada, onde aquela variável que teve maior correlação detectada para com a SMR foi inserida no modelo. Este critério é fundamentado pela idéia de que a variável mais correlacionada com o risco possui maior poder de explicação em relação ao fenômeno de estudo, ou seja, ao nascimento de crianças em mulheres jovens, contribuindo para uma melhor estimativa do risco de interesse. A Tabela 5.3 contém as correlações entre as variáveis disponíveis e a SMR calculada.

	z1	z2	z3	z4	z5	z6	z7	z8
SMR	-0,754	0,707	0,831	-0,828	0,846	0,845	-0,862	0,788

Tabela 5.3: Correlações de Pearson Entre as Variáveis e a SMR

Logo, o modelo 4 considerou a contribuição da escolaridade média, em anos de estudo, dos responsáveis por domicílios no ano de 2000 (z7) na construção das estimativas dos riscos relativos.

O modelo 5, por sua vez, considerou também a informação da variável z3 (porcentagem de responsáveis por domicílio com renda até 2 salários mínimos, em 2000) para a construção das estimativas. Esta escolha considerou que a inserção de duas variáveis de mesma tipologia (educação e rendimento) no modelo poderia apresentar um caso de multicolinearidade. É por isso que se optou por utilizar uma variável (a mais correlacionada com a SMR) de cada um dos dois gêneros.

5.5.1. Implementação via WinBUGS

O WinBugs é um software que permite fácil implementação dos modelos apresentados. Especificamente, quanto aos dados, necessita-se somente determinar a contagem do evento – número de nascimentos observados em mulheres com menos de vinte anos em 2005 para os bairros de Porto Alegre –, o valor esperado da contagem para os bairros – de acordo com o apresentado na Seção 4.2 – e, se for o caso, os dados associados às covariáveis. Além disso, especifica-se a função de verossimilhança associada aos dados, a estrutura do risco, assim como e as prioris e hiperprioris necessárias, como já apresentado na Seção 4.6.

Visto que a amostra à posteriori gerada e, consequentemente, as estimativas derivam de metodologias MCMC, faz-se necessária a determinação de valores iniciais para a cadeia em simulação dos parâmetros desconhecidos.

As prioris, hiperprioris e valores iniciais utilizados para os cinco modelos foram pré-determinados de acordo com aqueles que vem sendo utilizados a nível de literatura.

Para todos os modelos implementados, fez-se uso de semente igual à 123 e gerou-se 205000 simulações. As 5000 primeiras foram descartadas como período de burn-in. Dentre as 200000 restantes, aplicou-se um *thin interval* de 40, de modo que obtivemos uma amostra final de 5000 simulações.

5.5.2. Teste de Convergência

Para as prioris e hiperprioris atribuídas, testou-se a imposição de diferentes valores iniciais para as cadeias dos parâmetros, com intuito de verificar se a convergência estava sendo atingida em diferentes inicializações e se, independentemente dos valores iniciais, as estimativas a posteriori se mantinham similares.

Logo nos primeiros testes, confirmou-se a lógica iterativa das metodologias MCMC ao verificarmos que, mesmo que o software exija, nem todos os parâmetros necessitariam ser inicializados. Pois, após alterarmos o valor inicial de alguns parâmetros, suas correspondentes cadeias simuladas não se alteraram. Presenciamos este fato ao alterarmos os valores iniciais de α_0 e da precisão τ_u dos efeitos aleatórios u_i.

Quanto à precisão τ_b dos efeitos espaciais b_i, mesmo mudanças bruscas (como, por exemplo, de 0,5 para 100) fizeram com que, tanto as estimativas dos riscos relativos de interesse quanto às estimativas para o efeito aleatório e o para o efeito espacial que influência as áreas, se verificassem alterações nas estimativas apenas, em média, na terceira casa após a vírgula. Ou seja, a convergência foi obtida.

Ao alterarmos o valor inicial de α_1 (inclusive de zero para 100), novamente detectou-se obtenção da convergência e mínimas alterações nas estimativas finais, mais discretas ainda que às mudanças percebidas ao alterarmos o valor inicial de τ_b, tanto nas estimativas do risco relativo quanto do próprio α_1.

Para mudanças bruscas nos valores iniciais de u_i e b_i (de um para 100), verifica-se que as cadeias do risco relativo e dos correspondentes efeitos retardam a o alcance da convergência. Porém, a partir do momento que é atingida, a cadeia segue um comportamento de acordo com o esperado.

Para todas as tentativas, se considerarmos o processo depois de obtida a convergência, a estatística DIC permaneceu-se praticamente inalterada e com mudança mínima intrínseca a flutuação aleatória da própria.

5.5.3. Análise de Sensibilidade

A análise de sensibilidade corresponde a uma prática que tem por objetivo verificar a suscetibilidade do modelo perante diferentes especificações de prioris e parâmetros das prioris. Para maiores informações, sugere-se Gelman (1997).

Neste caso, foi aplicada e consistiu em substituir tais prioris e hiperprioris empregadas e que seguem a sugestão literária por outras quaisquer, a livre arbítrio, com objetivo de visualizar possíveis alterações e o comportamento das mesmas no que se refere aos resultados à posteriori gerados pelo software.

Ao substituirmos a priori *dflat*() atribuída para α_0 por uma priori também não informativa $N(0, e^{-5})$, percebeu-se que a estimativa pontual de α_0 sofreu alterações apenas na quarta casa após a vírgula, enquanto os riscos relativos estimados não sofreram alterações. Alterações na priori atribuída para α_1 também resultaram em pouquíssimas alterações na estimativa final dos riscos relativos.

Ao modificarmos às prioris de τ_u e τ_b de $Gama(0,5;0,0005)$ para $Gama(0,0005;0,0005)$, as estimativas dos riscos relativos apresentaram algumas mudanças na segunda casa após a vírgula, enquanto as estimativas para os efeitos aleatórios e espaciais apresentaram mudanças inclusive na primeira casa após a vírgula.

5.5.4. Modelo 1

Para o modelo 1 – não estruturado – implementado, considerou-se, à priori:

$$u_i \sim Normal(0, 1/\tau^2);$$

$$\mu \sim dflat();$$

$$\tau \sim Gamma(0,5;0,0005).$$

Onde a função intrinseca do WinBUGS *dflat*() corresponde a uma distribuição imprópria não informativa que não compromete o andamento do processo e, mesmo assim, conduz a uma posteriori própria. Os valores iniciais atribuídos foram $\tau = 0,5$, $\mu = 0$ e $u_i = 1$, para $i = 1,...,n$.

A Figura 5.3 representa o mapeamento com as médias dos riscos relativos de interesse à posteriori, por bairros.

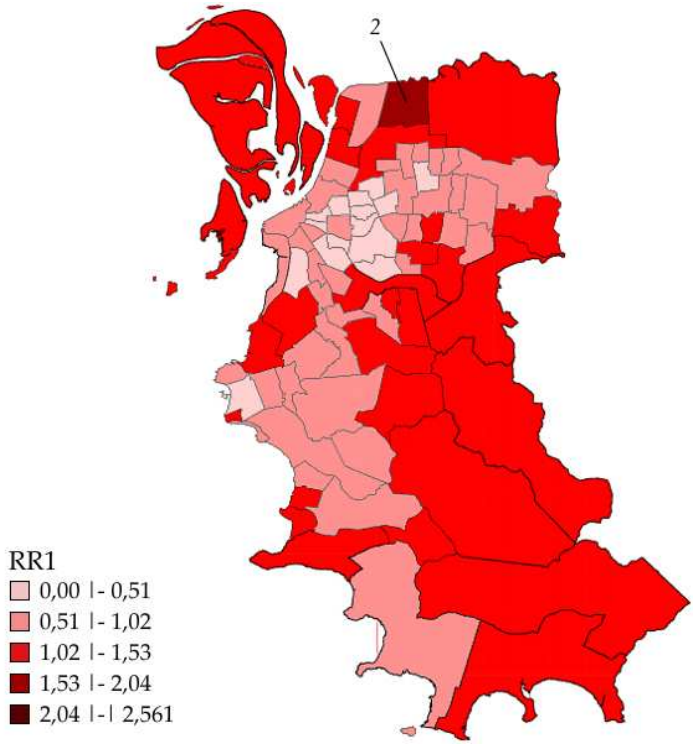

Figura 5.3 - Mapa com Estimativas pelo Modelo 1 (RR1) para os Riscos Relativos dos Bairros

Percebe-se que, ao considerarmos uma modelagem que não recebe a contribuição de covariáveis nem de efeito espacial para a obtenção de estimativas para o verdadeiro valor do risco relativo associado aos bairros, apenas o bairro Anchieta (2) apresentou situação muito preocupante, com um risco relativo estimado em 1,721. Logo, estima-se, pelo modelo 1, que o bairro Anchieta (2) apresenta o risco para o fenômeno em questão aproximadamente 72% superior ao risco de Porto Alegre.

5.5.5. Modelo 2

O modelo espacialmente estruturado teve especificações à priori dadas por:

$$b_i \mid b_j, j \neq i \sim N\left(\frac{\sum_{j \in \delta_i} b_j}{n_i}, \frac{1}{n_i \tau^2}\right);$$

$$\mu \sim dflat();$$

$$\tau \sim Gamma(0,5;0,0005).$$

Os valores iniciais corresponderam à $\tau = 0,5$, $\mu = 0$ e $b_i = 1$, para $i = 1,...,n$.

O mapeamento resultante com as médias dos riscos relativos de interesse à posteriori, por bairros, está apresentado na Figura 5.4.

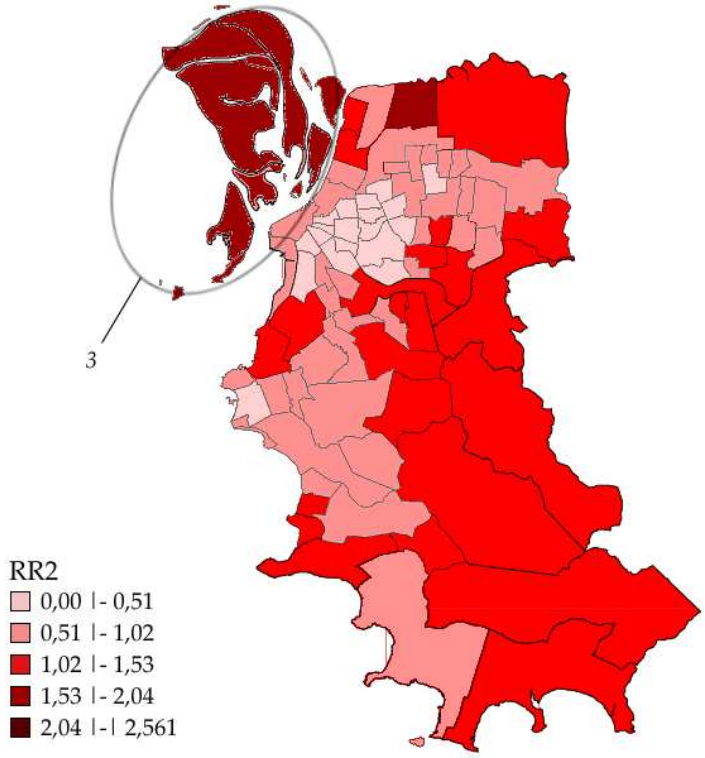

Figura 5.4 - Mapa com Estimativas pelo Modelo 2 (RR2) para os Riscos Relativos dos Bairros

Então, ao considerarmos, além dos dados observados por bairro, uma estrutura que admite apenas a influência espacial que os bairros vizinhos determinam, os riscos relativos estimados pelo modelo indicam os bairros Anchieta e Arquipélago (3) como aqueles de pior situação em Porto Alegre. Para o bairro Arquipélago (3), por exemplo, estima-se um risco aproximadamente 92% superior ao de Porto Alegre.

5.5.6. Modelo 3

O modelo 3 corresponde ao de convolução e, à priori:

$$u_i \sim Normal(0, 1/\tau_u^2);$$

$$b_i \mid b_j, j \neq i \sim N\left(\frac{\sum_{j\in\delta_i} b_j}{n_i}, \frac{1}{n_i \tau_b^2}\right);$$

$$\mu \sim dflat();$$

$$\tau_u \sim Gamma(0,5; 0,0005);$$

$$\tau_b \sim Gamma(0,5;0,0005)\,.$$

Com valores iniciais dados por $\tau_u = \tau_b = 0,5$, $\mu = 0$ e $u_i = b_i = 1$, para $i = 1,...,n$.

Segue o mapeamento com as médias dos riscos relativos de interesse à posteriori, apresentado na Figura 5.5.

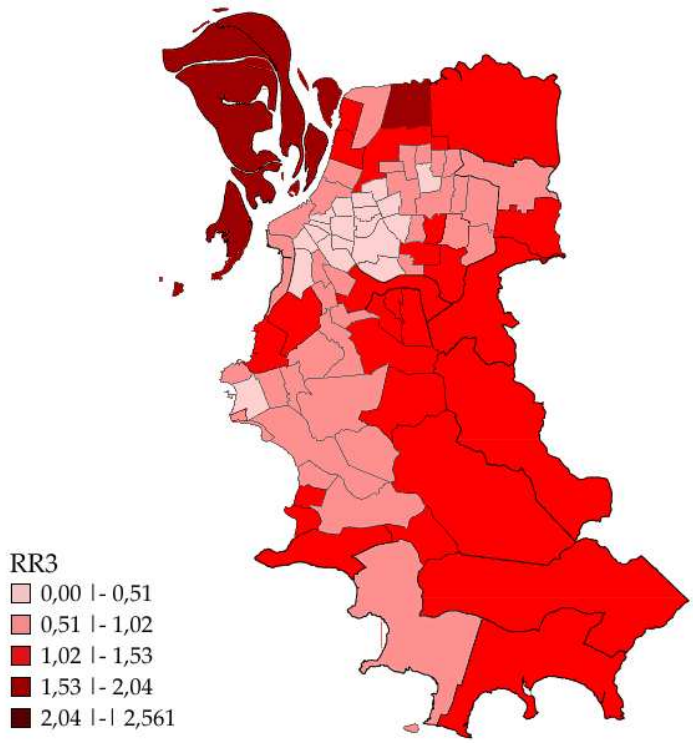

Figura 5.5 - Mapa com Estimativas pelo Modelo 3 (RR3) para os Riscos Relativos dos Bairros

Ao considerarmos tanto o efeito aleatório quanto o efeito espacial na modelagem, o bairro Anchieta e Arquipélago mais uma vez obtiveram estimativas para seu verdadeiro risco relativo alarmantes. Desta vez, estima-se que o bairro Anchieta tem risco aproximadamente 79% superior ao risco de Porto Alegre, enquanto para o Arquipélago estimou-se 63%.

5.5.7. Modelo 4

Este modelo considera a variável *z7* como covariável e assume as mesmas imposições à priori do modelo 3, com exceção de:

$$\alpha_0 \sim dflat()\,;$$

$$\alpha_1 \sim Normal(0; e^{-5})\,.$$

Os valores iniciais foram $\tau_u = \tau_b = 0,5$, $\alpha_0 = \alpha_1 = 0$ e $u_i = b_i = 1$, para $i = 1,...,n$.

A Figura 5.6 representa o mapeamento da média à posteriori do risco relativo.

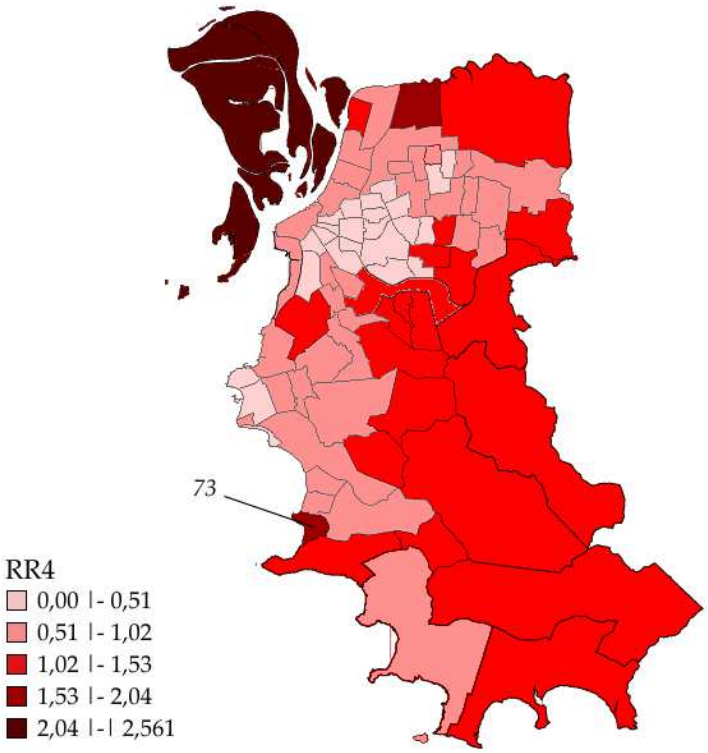

Figura 5.6 - Mapa com Estimativas pelo Modelo 4 (RR4) para os Riscos Relativos dos Bairros

Ao anexarmos a contribuição da covariável z7 (escolaridade média, em anos de estudo, dos responsáveis por domicílios em 2000) para a estimativa do risco relativo dos bairros no modelo de convolução, verificou-se – principalmente – três bairros em situação preocupante; Anchieta, Arquipélago e Serraria (73). Entre estes, o Arquipélago apresentou um quadro de extrema preocupação, com risco estimado em aproximadamente 109% superior ao riso de Porto Alegre.

5.5.8. Modelo 5

O modelo 5 é uma extensão do modelo 4 e, além de $z3$ como covariável, considera:

$$\alpha_2 \sim Normal(0; e^{-5}) .$$

Os valores iniciais utilizados foram $\tau_u = \tau_b = 0,5$, $\alpha_0 = \alpha_1 = \alpha_2 = 0$ e $u_i = b_i = 1$, para $i = 1,...,n$.

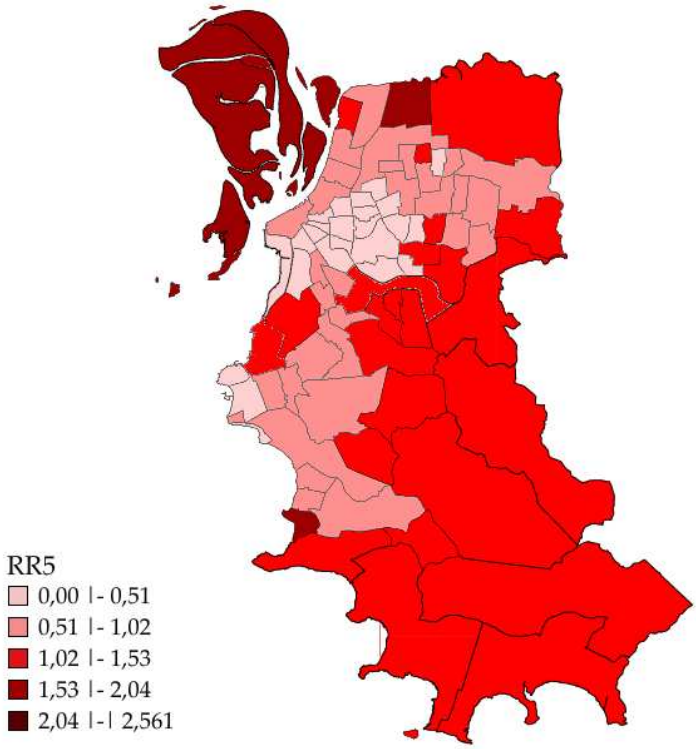

Figura 5.7 - Mapa com Estimativas pelo Modelo 5 (RR5) para os Riscos Relativos dos Bairros

Após a inserção de mais uma covariável uma variável no modelo de convolução (z3 – porcentagem de responsáveis por domicílio com renda até 2 salários mínimos, em 2000), visando uma melhor estimativa para os verdadeiros riscos relativos, estimou-se que os bairros Arquipélago, Anchieta e Serraria possuem riscos aproximadamente iguais à, respectivamente, 102%, 68% e 54% superiores ao risco de Porto Alegre.

Para o modelo 5, trataremos de apresentar algumas questões referentes ao comportamento das cadeias simuladas para que o leitor possa tomar conhecimento da forma que os resultados se apresentaram, com objetivo de visualizar itens associados à convergência das cadeias, forma das posterioris e medidas resumo em geral.

A figura 5.8 apresenta o comportamento da cadeia completa de simulações para σ_u e σ_b.

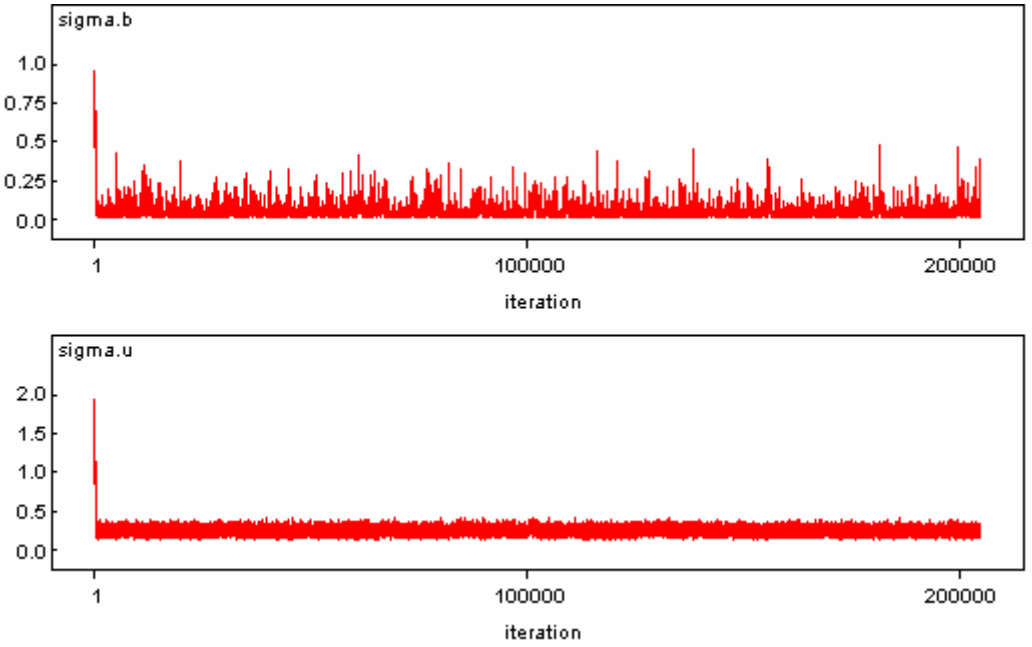

Figura 5.8 – Trajetórias Completas da Simulação de σ_b e σ_u

Figura 5.9 expõe os gráficos das posterioris – estimadas não parametricamente pelo método de suavização de Kernel – correspondentes à amostra final de tamanho 5000 das cadeias anteriormente apresentadas em conjunto com as posterioris derivadas da amostra final, também estimadas por Kernel, de outros hiperparâmetros.

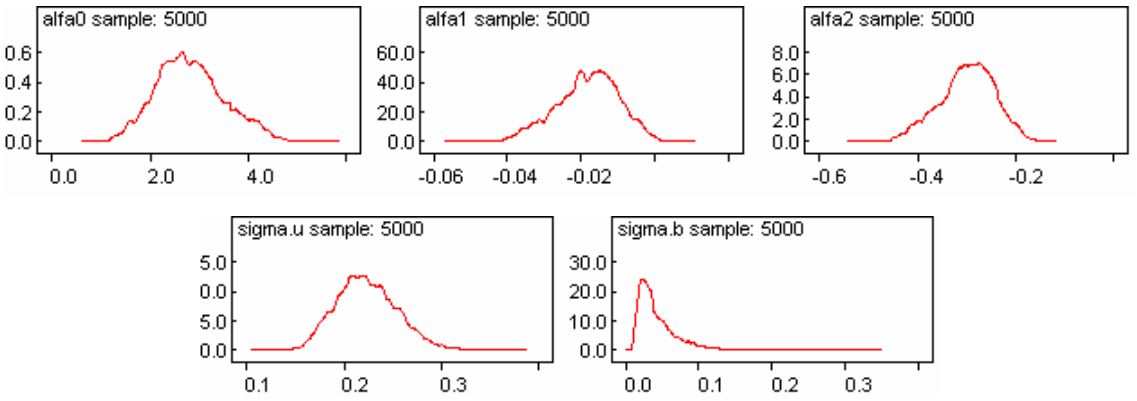

Figura 5.9 – Posterioris de α_0, α_1, α_2, σ_u e σ_b Estimadas Não Parametricamente por Kernel

A tabela 5.4 apresenta medidas resumo à posteriori dos hiperparâmetros.

	Média	Desvio Padrão	Mediana	L.I. IC (95%)	L.S. I.C. (95%)
alfa0	2,849	0,732	2,790	1,568	4,420
alfa1	-0,018	0,009	-0,018	-0,037	-0,003
alfa2	-0,300	0,059	-0,295	-0,427	-0,195
sigma.b	0,048	0,038	0,036	0,013	0,155
sigma.u	0,224	0,033	0,222	0,166	0,295

Tabela 5.4 – Medidas Resumo à Posteriori de Hiperparâmetros

As tabelas 5.5 e 5.6 apresentam, respectivamente, medidas resumo à posteriori dos riscos relativos e dos efeitos aleatórios e espaciais dos bairros Agronomia (1), Anchieta (2) Arquipélago (3), Bom Fim (10), Centro (17), Jardim Lindóia (39), Sarandi (72) e Teresópolis (74). A escolha destes bairros foi feita de maneira arbitrária apenas como ilustração.

	Média	Desvio Padrão	Mediana	L.I. IC (95%)	L.S. I.C. (95%)
RR5[1]	1,492	0,165	1,485	1,194	1,839
RR5[2]	1,679	0,356	1,635	1,100	2,473
RR5[3]	2,022	0,261	2,010	1,554	2,573
RR5[10]	0,302	0,069	0,298	0,182	0,449
RR5[17]	0,553	0,076	0,550	0,417	0,713
RR5[39]	0,477	0,094	0,467	0,315	0,684
RR5[72]	1,147	0,066	1,147	1,020	1,280
RR5[74]	0,638	0,106	0,630	0,450	0,869

Tabela 5.5 - Medidas Resumo à Posteriori do Risco Relativo de Algumas Áreas

	Média	Desvio Padrão	Mediana	L.I. IC (95%)	L.S. I.C. (95%)
u5[1]	0,131	0,116	0,131	-0,096	0,359
u5[2]	0,150	0,201	0,146	-0,236	0,549
u5[3]	-0,701	0,143	-0,697	-0,986	-0,427
u5[10]	-0,188	0,210	-0,178	-0,624	0,198
u5[17]	0,291	0,149	0,293	0,004	0,580
u5[39]	0,024	0,191	0,023	-0,350	0,399
u5[72]	-0,121	0,087	-0,119	-0,300	0,045
u5[74]	-0,004	0,165	-0,003	-0,339	0,312
b5[1]	0,003	0,027	0,002	-0,047	0,067
b5[2]	0,007	0,045	0,003	-0,070	0,114
b5[4]	-0,002	0,032	-0,001	-0,069	0,062
b5[10]	-0,840	0,032	0,208	-0,066	0,059
b5[17]	0,006	0,032	0,003	-0,048	0,083
b5[39]	0,390	0,034	0,413	-0,074	0,071
b5[72]	-0,875	0,036	-0,419	-0,076	0,077
b5[74]	0,003	0,030	0,001	-0,051	0,071

Tabela 5.6 - Medidas Resumo à Posteriori dos Efeitos Aleatórios e Espaciais de Algumas Áreas

As figura 5.10 apresenta a trajetória simulada referente a amostra final do risco relativo dos bairros Agronomia(1) e Anchieta(2).

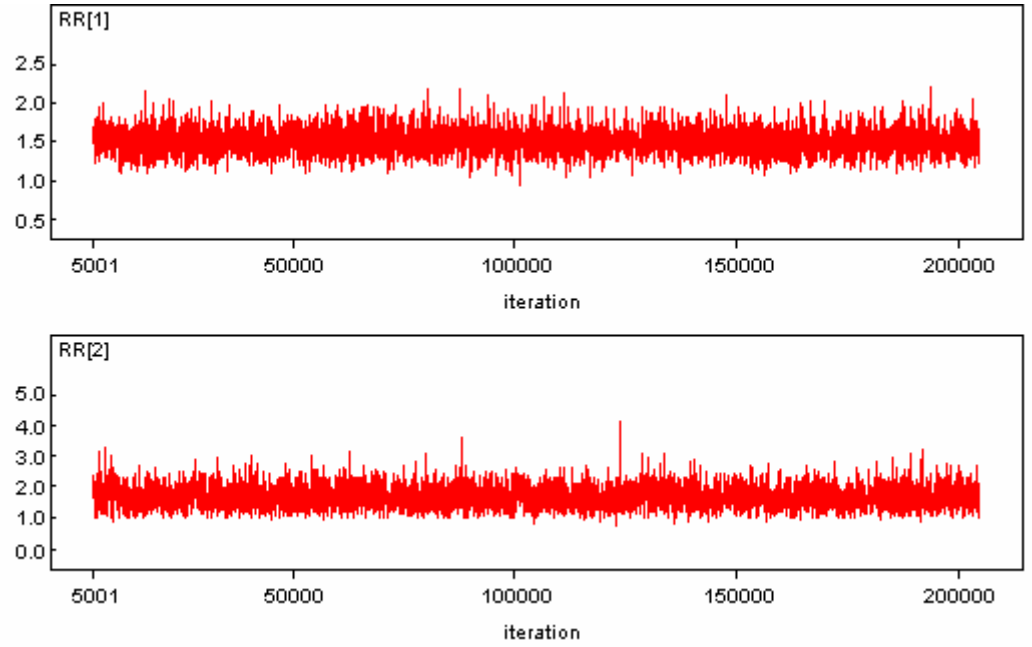

Figura 5.10 – Trajetórias que Compõem a Amostra Final para o RR dos Bairros Agronomia e Anchieta

As figura 5.11 apresenta a trajetória simulada referente a amostra final do risco relativo dos bairros Arquipélago(3) e Bom Fim(10).

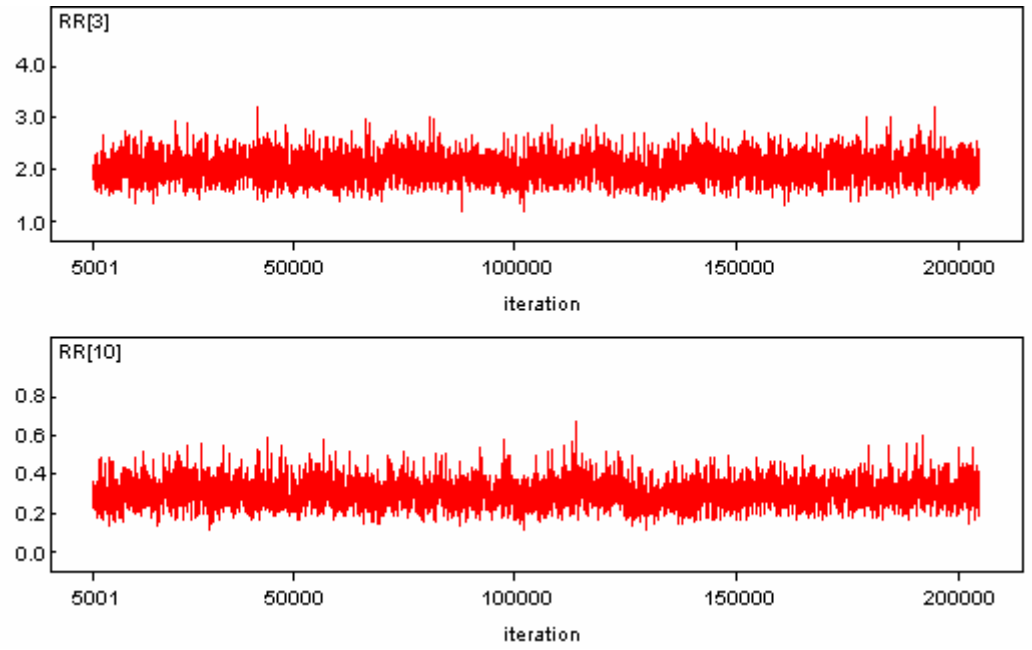

Figura 5.11 – Trajetórias que Compõem a Amostra Final para o RR dos Bairros Arquipélago e Bom Fim

As figura 5.12 apresenta a trajetória simulada referente a amostra final do risco relativo dos bairros Centro(17) e Jardim Lindóia(39).

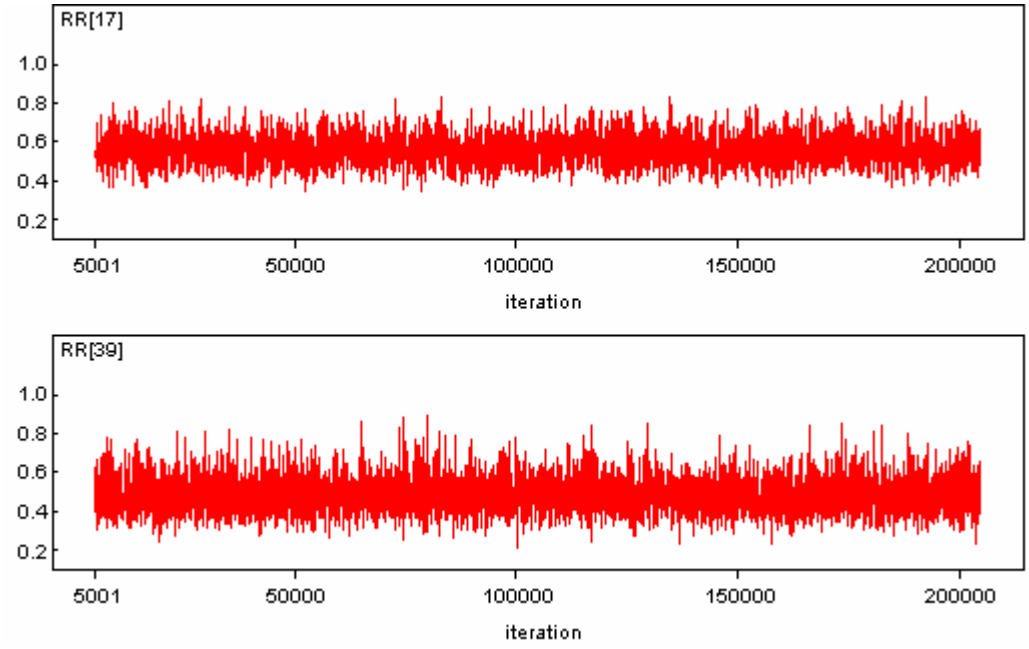

Figura 5.12 – Trajetórias que Compõem a Amostra Final para o RR dos Bairros Centro e Jd. Lindóia

As figura 5.13 apresenta a trajetória simulada referente a amostra final do risco relativo dos bairros Sarandi(72) e Teresópolis(74).

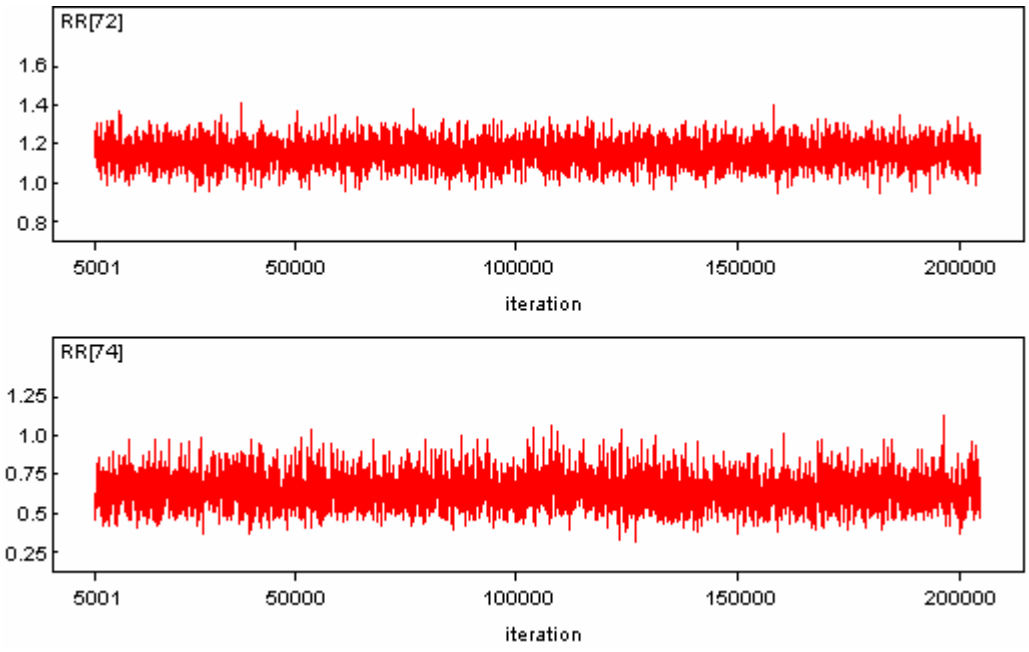

Figura 5.13 – Trajetórias que Compõem a Amostra Final para o RR dos Bairros Sarandi e Teresópolis

A figura 5.14 apresenta os gráficos das posterioris – estimadas não parametricamente pelo método de suavização de Kernel – correspondentes à amostra final de tamanho 5000 das cadeias de simulação do risco relativo dos bairros Agronomia(1), Anchieta(2), Arquipélago(3) e Bom Fim(10).

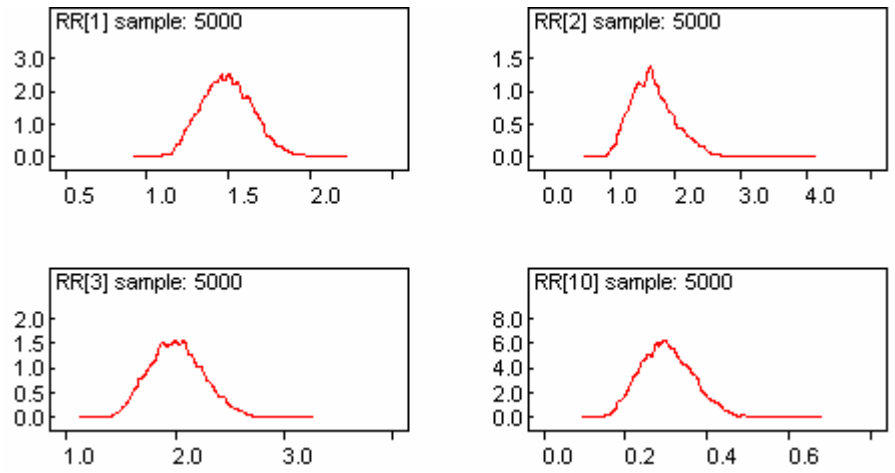

Figura 5.14 – Posterioris para o RR de Alguns Bairros Estimadas Não Parametricamente por Kernel

A figura 5.15, por sua vez, faz referência aos bairros Centro (17), Jardim Lindóia (39), Sarandi(72) e Teresópolis(74).

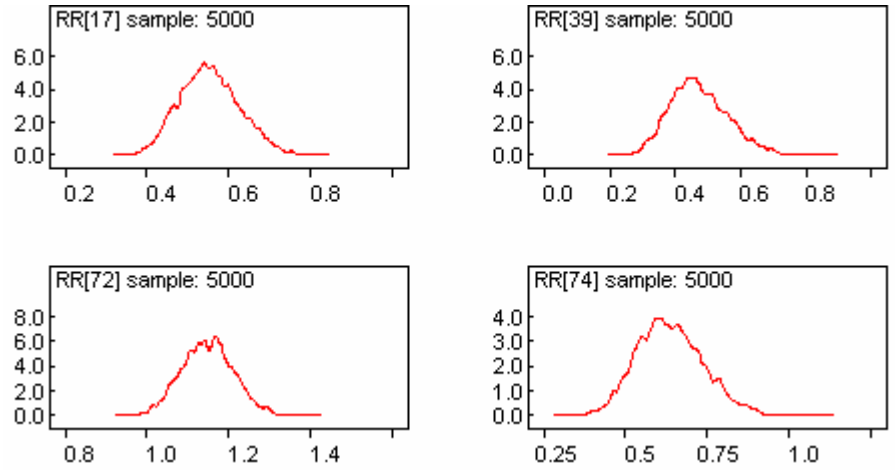

Figura 5.15 – Posterioris para o RR de Alguns Bairros Estimadas Não Parametricamente por Kernel

A Figura 5.16 apresenta um gráfico com as médias à posteriori dos riscos relativos de todos os bairros de Porto Alegre, assim como os intervalos de credibilidade (com 95% de probabilidade) correspondentes. A linha vermelha do gráfico indica a média geral dos riscos relativos ($\overline{RR5} \cong 0,827$).

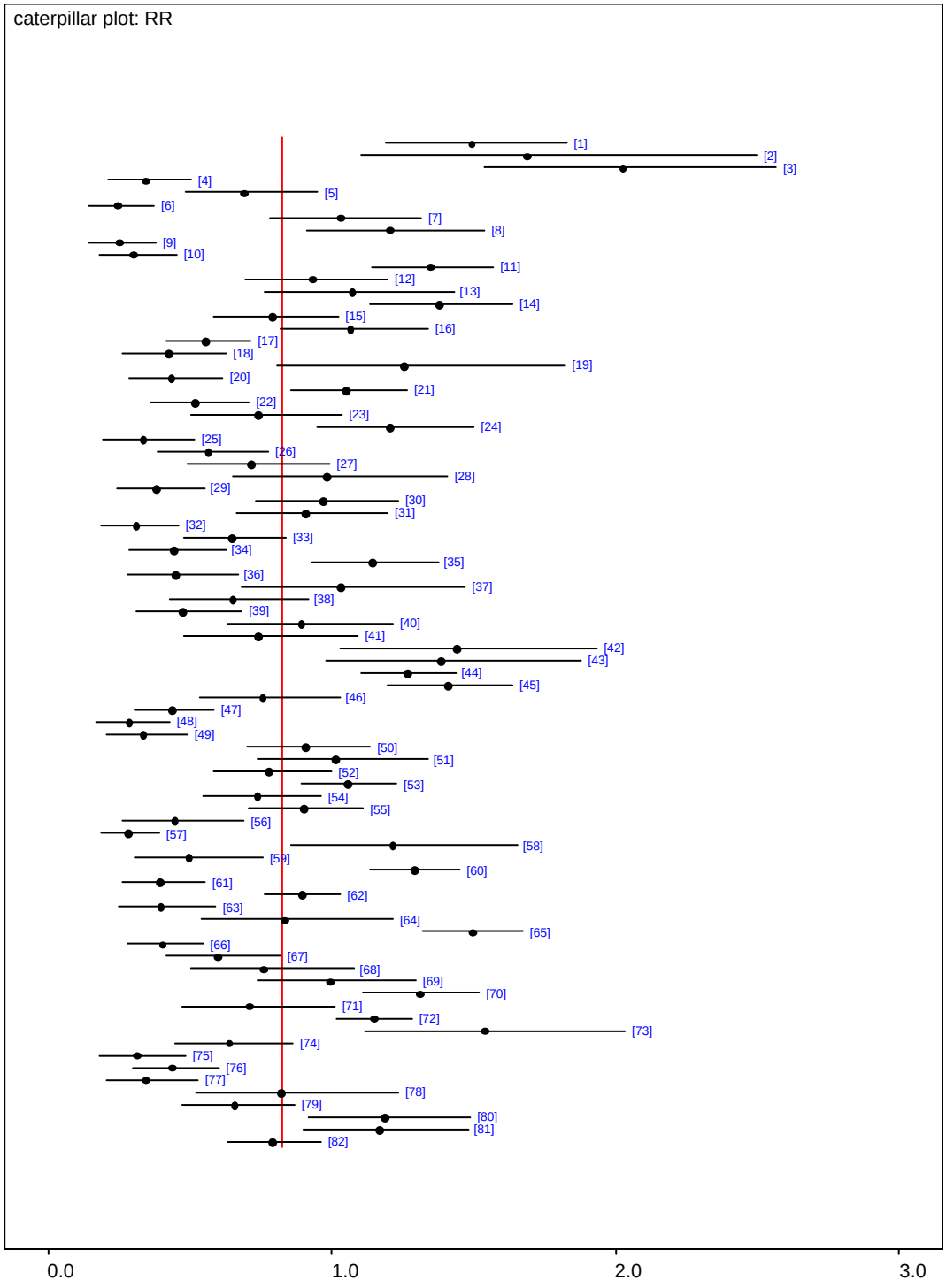

Figura 5.16 – Médias e IC 95% à Posteriori dos Riscos Relativos dos Bairros de Porto Alegre

5.6. Análise Comparativa

Anteriormente à escolha dentre o melhor dos cinco modelos empregados, apresenta-se a Figura 5.16, com o mapeamento resultante por todos os processos de estimação.

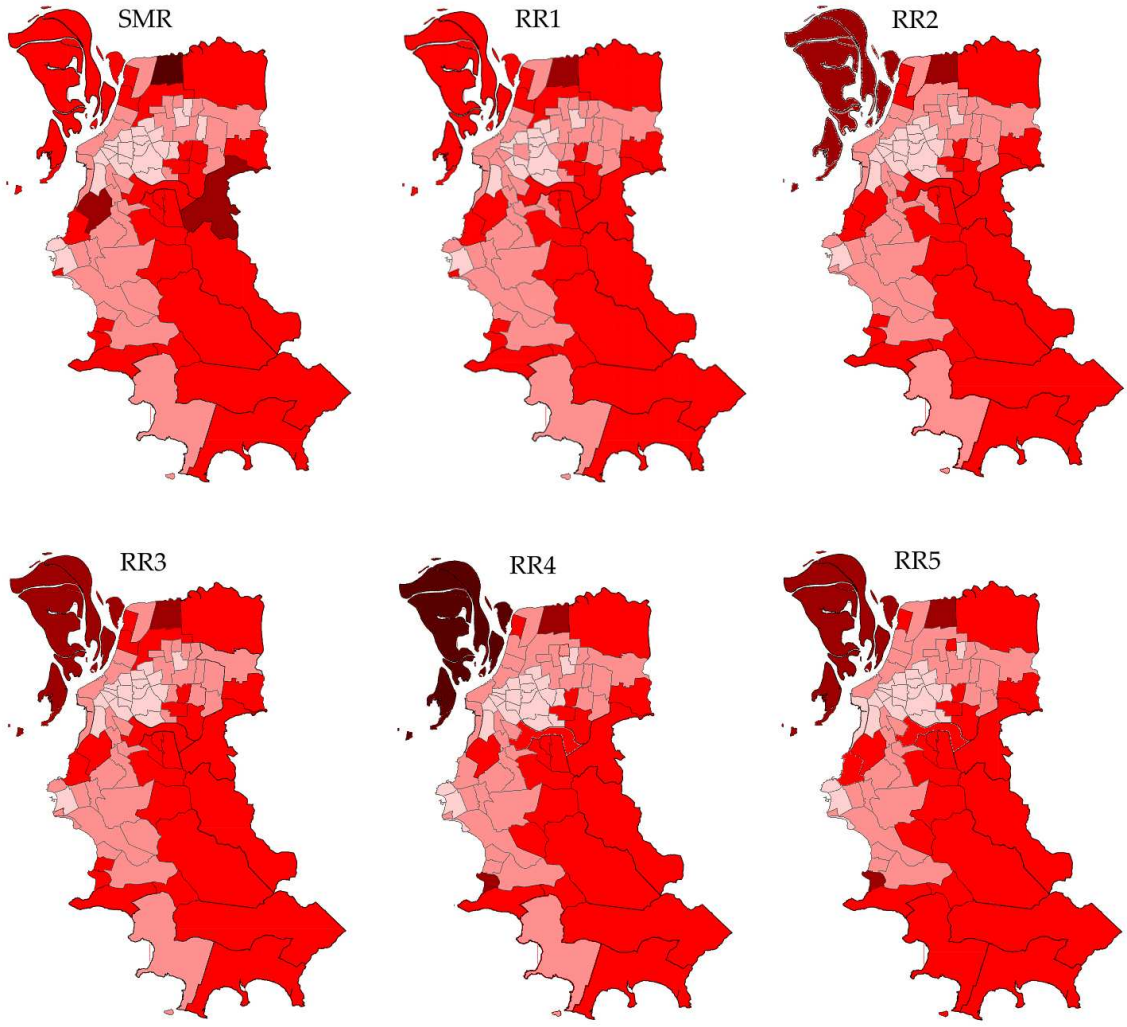

Figura 5.17 – Mapas com Diferentes Estimativas para o Risco Relativo dos Bairros de Porto Alegre

As estimativas para os riscos relativos de todos os bairros de Porto Alegre segundo a abordagem clássica e todos os modelos inteiramente Bayesianos estão apresentadas em uma tabela no Apêndice F deste trabalho.

O bairro Anchieta pode ser analisado para ilustrarmos as conseqüências da modelagem inteiramente Bayesiana em riscos relativos atribuídos à pequenas áreas. Para este bairro, verificou-se, em 2005, a ocorrência de 9 nascimentos oriundos de mulheres jovens em um total de 20 crianças nascidas, enquanto o esperado para este número, sob hipótese de risco constante ao longo de Porto Alegre, era de aproximadamente 4 nascimentos.

Sendo assim, tanto o modelo que considerou exclusivamente a inserção de um componente de aleatoriedade para o risco relativo do bairro Anchieta quanto os modelos que consideram apenas o efeito espacial, ambos os efeitos ou covariáveis, produziram

estimativas menos preocupantes para o risco relativo da área. Pois, justamente por considerarem a possibilidade do efeito casual, assim como a autocorrelação espacial e a informação contida em covariáveis altamente correlacionadas com o risco relativo do bairro, os modelos atribuíram o alto risco relativo observado à um fenômeno casual e geraram estimativas mais realistas de acordo com um cenário que considera muito mais informações do que apenas o número de nascidos vivos em mulheres jovens e o número de nascidos vivos total do bairro Anchieta. Considerando o modelo 5, a estimativa para o risco relativo do bairro Anchieta reduziu dos iniciais 2,56 (estimativa clássica – SMR) para aproximadamente 1,68.

O bairro Arquipélago, se analisado pela estimativa clássica, demonstra menor risco relativo do que o próprio bairro Anchieta (aproximadamente 1,50 contra 2,56). De acordo com o modelo 5, porém, que considera um efeito aleatório e duas covariáveis para a estimação do verdadeiro risco relativo do bairro (não considera efeito espacial para este caso devido ao fato de que o bairro Arquipélago não possui vizinhos), a situação se inverte. As fontes adicionais de informação contribuíram para que o risco relativo do bairro Arquipélago fosse agora estimado em aproximadamente 2,02, caracterizando uma situação muito pior do que a estimada para o bairro Anchieta.

O bairro Pedra Redonda também apresenta uma situação interessante. Este registrou apenas 5 nascimentos em 2005, sendo que nenhum derivou de mulheres com idade inferior a 20 anos. Sendo assim, a estimativa clássica (SMR) para o risco relativo deste bairro seria 0. Um caso extremo, pois estaríamos estimando pontualmente que não há nascimentos em mulheres jovens neste bairro. O modelo inteiramente Bayesiano 5, porém, estimou que, de acordo com a consideração dos efeitos aleatório e espacial e das covariáveis utilizadas, o risco relativo do bairro Pedra Redonda é de aproximadamente 0,44.

O bairro Jardim Floresta, que, com o nascimento de 7 crianças em mulheres jovens, apresentou SMR de aproximadamente 0,88, teve esta estimativa reduzida de acordo com os modelos que consideram o efeito espacial (possivelmente pelo fato de seus vizinhos apresentarem risco relativo baixo), porém aumentada para aproximadamente 1,03 ao considerarmos as duas variáveis, de acordo com o modelo 5. Ao checarmos os valores das covariáveis associadas ao bairro Jardim Floresta, verificou-se índices muito ruins. Neste bairro, segundo o censo demográfico de 2000, aproximadamente 28% dos chefes de família tem renda de até dois salários mínimos. Além disso, seus chefes de família, também de

acordo com o censo de 2000, têm escolaridade média de aproximadamente 7,6 anos de estudo. O bairro Serraria apresentou a mesma situação descrita para o bairro Jardim Floresta. Com estimativa, porém, pelo modelo 5 de aproximadamente 1,54.

No bairro Jardim Lindóia, por sua vez, observou-se 8 nascimentos em mulheres jovens para o ano de 2005, resultando em uma SMR de 0,5. Ao considerarmos os modelos que admitem a inserção da autocorrelação espacial, o risco relativo para o bairro aumentou, conseqüência da influência de riscos ruins associados aos seus vizinhos. Ao considerarmos os modelos com covariáveis, novamente o risco relativo estimado para o bairro Jardim Lindóia diminuiu, de modo que a estimativa gerada pelo modelo 5 foi de aproximadamente 0,48. O bairro Chácara das Pedras apresentou um comportamento similar ao do Jardim Lindóia ao longo dos 5 modelos especificados.

O bairro Praia de Belas foi outro que melhorou sua estimativa para o risco relativo após a incorporação da informação de covariáveis na modelagem. Com um observado de quatro crianças nascidas de mulheres jovens, este bairro apresentou SMR de aproximadamente 0,99, enquanto a estimativa gerada pelo modelo 5 foi de aproximadamente 0,49. Com situação semelhante à do bairro Praia de Belas conforme os cinco modelos, cita-se a do bairro Guarujá.

A Tabela 5.7 apresenta a ordenação dos 13 bairros de Porto Alegre com estimativas mais baixas (melhores) para o risco relativo, segundo o modelo 5.

Classificação	ID	Bairro	*RR*5
1	6	Bela Vista	0,245
2	9	Boa Vista	0,249
3	57	Petrópolis	0,281
4	48	Moinhos de Vento	0,286
5	10	Bom Fim	0,302
6	32	Independência	0,310
7	75	Três Figueiras	0,314
8	49	Mont' Serrat	0,333
9	25	Farroupilha	0,334
10	77	Vila Assunção	0,342
11	4	Auxiliadora	0,342
12	29	Higienópolis	0,380
13	61	Rio Branco	0,390

Tabela 5.7 – Os Treze Bairros com Menor Risco Relativo Estimado pelo Modelo 5

Para os 13 piores bairros (bairros com risco estimado mais alto de acordo com o modelo 5), a Tabela 5.8 apresenta os resultados.

Classificação	ID	Bairro	*RR*5
82	3	Arquipélago	2,022
81	2	Anchieta	1,679
80	73	Serraria	1,537
79	1	Agronomia	1,492
78	65	Santa Teresa	1,491
77	42	Lageado	1,437
76	45	Mário Quintana	1,407
75	43	Lami	1,380
74	14	Cascata	1,375
73	11	Bom Jesus	1,348
72	70	São José	1,307
71	60	Restinga	1,290
70	44	Lomba do Pinheiro	1,266

Tabela 5.8 - Os Treze Bairros com Maior Risco Relativo Estimado pelo Modelo 5

5.6.1. Comparação Entre Modelos

O critério DIC (veja Seção 4.8), para comparação e seleção entre modelos, calculado para os cinco modelos inteiramente Bayesianos apresentou os resultados que estão explicitados na tabela 5.9.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
DIC	543,628	544,140	530,301	510,019	506,998

Tabela 5.9 – Estatísticas DIC para os Cinco Modelos Bayesianos

Com base neste critério, classifica-se o modelo 5 como melhor modelo para o mapeamento do risco relativo em questão neste estudo.

Capítulo 6

Considerações Finais

Em virtude de algumas limitações, este trabalho não pode ter análises e interpretações da aplicação prática esgotadas ao máximo, infelizmente, visto que muito ainda se poderia fazer. Como entraves, principalmente o pouco tempo disponível para o estudo da aplicação e a dependência da equipe do Observatório de Porto Alegre/Prefeitura Municipal de Porto Alegre para a produção dos mapas.

Em compensação, após a apresentação do previsto conforme a estrutura apresentada na Seção 1.4, pode-se implementar os procedimentos discutidos neste trabalho de maneira satisfatória para a compreensão da importância e utilidade das técnicas.

O mapeamento de doenças segundo um modelo inteiramente Bayesiano permitiu produzir mapas mais informativos no que se refere ao risco em estudo, derivados de estimativas mais realistas (veja Apêndice F), fornecendo uma percepção mais confiável sobre a distribuição espacial da natalidade em mulheres com idade inferior a 20 anos de idade em Porto Alegre.

Os resultados, na percepção do autor, obtidos, seguindo as sugestões de implementação computacional da literatura, foram extremamente satisfatórios. Para uma análise mais minuciosa e útil para a compreensão do quanto a especificação da modelagem – leia-se prioris, valores iniciais, matriz de pesos, etc. – e questões relativas a hipóteses sociais/econômicas/geográficas se relacionam com os efeitos aleatórios e espacias estimados, assim como com os riscos relativos e outras estimativas possíveis, porém, necessitasse despender mais tempo.

Uma análise em conjunto do autor com um especialista no fenômeno em questão, levantando hipóteses, analisando resultados e sugerindo estudos posteriores, pode ser muito proveitosa para a sustentação de políticas públicas e tomadas de decisões. Sendo assim, este trabalho é concluído com a noção de que muito ainda se pode trabalhar em cima do que aqui se apresentou.

Referências Bibliográficas

Assunção, R. M., Barreto, S. M., Guerra, H. L. & Sakurai, E. (1998). Mapas de taxas epidemiológicas: Uma abordagem Bayesiana. *Cadernos de Saúde Pública*, 14, 713-723.

Assunção, R. M. & Reis, E. A. (1999). A new proposal to adjust Moran's I for population density. *Statistics in Medicine*, 18, 2147-2162.

Assunção, R M. (2001). *Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia*. São Carlos: UFSCAR. Disponível em www.est.ufmg.br/~assuncao (acessado em 20/11/2006).

Bailey, T. C. & Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. London: Longman.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 330-418. Reprinted, with biographical note by Bernard, G. A., in Biometrika, 45, 293-315 (1958).

Besag, J. (1974). Spatial interaction and the statistics analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, B, 36, 192-236.

Besag, J. (1989). Towards Bayesian image analysis. *Journal of Applied Statistics*, 16, 395-407.

Besag, J. & Mollié, A. (1989). Bayesian mapping of mortality rates. *Bulletin of the International Statistical Institute*, 53, 127-128.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1-21.

Box, S. E. P. & Tiao, S. C. (1992). *Bayesian Inference in Statistical Analysis*. New York: John Wiley and Sons.

Choynowski, M. (1959). Maps based on probabilities. *Journal of the American Statistical Association*, 54, 385-388.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: John Wiley and Sons, revised edition.

Druck, S., Carvalho, M. S., Câmara, G., Monteiro, A. V. M. (2004). *Análise Espacial de Dados Geográficos.* Brasília: EMBRAPA. Disponível em http://www.dpi.inpe.br/gilberto/livro/analise/ (acessado em 20/11/2006).

Ehlers, R. S., Silva, S. A. & Mota, L. L. M. (2006). Fully Bayesian Spatial Analysis of Homicide Rates. *Journal Estadística*, 58.

Elliott, P., Wakefield, J., Best, N. & Briggs, D. (2001). *Spatial Epidemiology: Methods and Applications*. London: Oxford University Press.

Gamerman, D. & Lopes H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* London: Chapman & Hall, 2nd edition.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5, 115-145.

Gelfand, A. E. (1996). Model determination using sampling-based methods. Em *Monte Carlo Markov Chain in Practice* (Editors Gilks, W. R., Richardson, S. & Spiegelhalter, D. J.).

Gelman, A. (1997). *Bayesian data analysis*. Boca Raton: Chapman & Hall.

Geman, S. & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57, 97-109.

Lawson, A. B. (2001). *Statistical Methods in Spatial Epidemiology*. Chichester: John Wiley.

Marshall, R. J. (1991). Mapping disease and mortality rates using empirical Bayes estimators. *Applied statistics*, 41, 283-294.

Metropolis, N., Rosenbulth, A. W., Rosenbulth, M. N., Teller, A. H. & Teller, E. (1953). Equation of State Calculations by Fast Computing Machine. *Journal of Chemical Physics*, 21, 1089-1091.

Mollié, A. (1996). Bayesian Mapping of Disease. Em *Monte Carlo Markov Chain in Practice* (Editors Gilks, W. R., Richardson, S. & Spiegelhalter, D. J.).

Moran, P. A. P. (1950). The interpretation of statistical maps. *Journal of Royal Statistical Society, series B*, 10, 243-251.

Nejjari, C., Tessier, J. F., Dartigues, J. F., Barberger-Gateau, P., Letenneur, L., Salamon, R. (1993). The Relationship between dyspnoca and main lifetime occupation in the elderly. *International Journal of Epidemiology*, 22, 848-854.

Paulino, C. D., Amaral-Turkman M. A. & Murteira B. (2003). *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian.

Richardson, S., Thomson, A., Best, N. & Elliott, P. (2004). Interpreting posterior relative risk estimates in disease-mapping studies. *Environ Health Perspectives*, 112, 1016-1025.

Ripley, B. D. (1981). *Spatial Statistics*. Chichester: John Wiley.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, B, 64, 583–639.

Thomas, A., Spiegelhalter, D. J. & Gilks, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs Sampling. Em *Bayesian Statistics 4*, 837-842 (Editors Bernardo, J. M., Berger, J. O., Dawid, A. P. & Smith, A. F. M.).

Thomson, S. (1992). *Sampling*. New York: Wiley.

Apêndice A – Código em WinBugs para o Modelo 1

```
model
{
  for (i in 1 : N)
    {
      y[i] ~ dpois(lambda[i])
      log(lambda[i]) <- log(E[i]) + mi + u[i]
      RR[i] <- exp(mi + u[i])
    }
  for (i in 1 : N)
    {
      u[i] ~ dnorm(0, tau)
    }
  mi  ~ dflat()
  tau  ~ dgamma(0.5, 0.0005)
  sigma <- 1/sqrt(tau)
}
```

Apêndice B – Código em WinBugs para o Modelo 2

```
model
{
  for (i in 1 : N)
    {
      y[i] ~ dpois(lambda[i])
      log(lambda[i]) <- log(E[i]) + mi + b[i]
      RR[i] <- exp(mi + b[i])
    }
  b[1:N] ~ car.normal(adj[], weights[], num[], tau)
  for(k in 1:totalvizi)
    {
      weights[k] <- 1
    }
  mi  ~ dflat()
  tau  ~ dgamma(0.5, 0.0005)
  sigma <- 1/sqrt(tau)
}
```

Apêndice C – Código em WinBugs para o Modelo 3

```
model
{
 for (i in 1 : N)
  {
   y[i] ~ dpois(lambda[i])
   log(lambda[i]) <- log(E[i]) + mi + u[i] + b[i]
   RR[i] <- exp(mi + u[i] + b[i])
  }
 for (i in 1 : N)
  {
   u[i] ~ dnorm(0, tau.u)
  }
 b[1:N] ~ car.normal(adj[], weights[], num[], tau.b)
 for(k in 1:totalvizi)
  {
   weights[k] <- 1
  }
 mi  ~ dflat()
 tau.u  ~ dgamma(0.5, 0.0005)
 tau.b  ~ dgamma(0.5, 0.0005)
 sigma.u <- 1/sqrt(tau.u)
 sigma.b <- 1/sqrt(tau.b)
}
```

Apêndice D – Código em WinBugs para o Modelo 4

```
model
{
  for (i in 1 : N)
    {
      y[i] ~ dpois(lambda[i])
      log(lambda[i]) <- log(E[i]) + alfa0 + (alfa1*z7[i]) + u[i] + b[i]
      RR[i] <- exp(alfa0 + (alfa1*z7[i]) + u[i] + b[i])
    }
  for (i in 1 : N)
    {
      u[i] ~ dnorm(0, tau.u)
    }
  b[1:N] ~ car.normal(adj[], weights[], num[], tau.b)
  for(k in 1:totalvizi)
    {
      weights[k] <- 1
    }
  alfa0 ~ dflat()
  alfa1 ~ dnorm(0.0, 1.0E-5)
  tau.u  ~ dgamma(0.5, 0.0005)
  tau.b  ~ dgamma(0.5, 0.0005)
  sigma.u <- 1/sqrt(tau.u)
  sigma.b <- 1/sqrt(tau.b)
}
```

Apêndice E – Código em WinBugs para o Modelo 5

```
model
{
 for (i in 1 : N)
  {
   y[i] ~ dpois(lambda[i])
   log(lambda[i]) <- log(E[i]) + alfa0 + (alfa1*z3[i]) + (alfa2*z7[i]) + u[i] + b[i]
   RR[i] <- exp(alfa0 + (alfa1*z3[i]) + (alfa2*z7[i]) + u[i] + b[i])
  }
 for (i in 1 : N)
  {
   u[i] ~ dnorm(0, tau.u)
  }
 b[1:N] ~ car.normal(adj[], weights[], num[], tau.b)
 for(k in 1:totalvizi)
  {
   weights[k] <- 1
  }
 alfa0 ~ dflat()
 alfa1 ~ dnorm(0.0, 1.0E-5)
 alfa2 ~ dnorm(0.0, 1.0E-5)
 tau.u  ~ dgamma(0.5, 0.0005)
 tau.b  ~ dgamma(0.5, 0.0005)
 sigma.u <- 1/sqrt(tau.u)
 sigma.b <- 1/sqrt(tau.b)
}
```

Apêndice F – Tabela das Estimativas

Na tabela abaixo, seguem as estimativas do risco relativo dos bairros correspondentes à modelagem clássica e aos cinco modelos inteiramente Bayesianos.

ID	Bairro	*SMR*	*RR*1	*RR*2	*RR*3	*RR*4	*RR*5
1	Agronomia	1,554	1,495	1,459	1,487	1,493	1,492
2	Anchieta	2,560	1,721	1,808	1,788	1,644	1,679
3	Arquipélago	1,495	1,402	1,923	1,629	2,091	2,022
4	Auxiliadora	0,274	0,438	0,335	0,356	0,364	0,342
5	Azenha	0,844	0,830	0,724	0,768	0,685	0,689
6	Bela Vista	0,178	0,361	0,259	0,281	0,284	0,245
7	Belém Novo	0,918	0,904	0,930	0,945	1,008	1,031
8	Belém Velho	1,201	1,150	1,139	1,173	1,202	1,206
9	Boa Vista	0,000	0,289	0,248	0,249	0,282	0,249
10	Bom Fim	0,000	0,308	0,258	0,255	0,324	0,302
11	Bom Jesus	1,390	1,368	1,281	1,334	1,362	1,348
12	Camaquã	0,921	0,908	0,845	0,884	0,907	0,929
13	Campo Novo	0,875	0,858	0,863	0,886	1,022	1,072
14	Cascata	1,401	1,368	1,315	1,362	1,390	1,375
15	Cavalhada	0,718	0,723	0,716	0,733	0,769	0,790
16	Cel Aparício Borges	1,043	1,019	1,019	1,036	1,046	1,063
17	Centro	0,643	0,652	0,578	0,610	0,564	0,553
18	Chácara das Pedras	0,356	0,543	0,586	0,547	0,445	0,421
19	Chapéu do Sol	1,339	1,174	1,211	1,242	1,508	1,253
20	Cidade Baixa	0,471	0,533	0,486	0,499	0,454	0,434
21	Cristal	1,145	1,129	1,069	1,107	1,056	1,051
22	Cristo Redentor	0,420	0,481	0,479	0,483	0,505	0,518
23	Espirito Santo	0,728	0,743	0,748	0,777	0,737	0,740
24	Farrapos	1,190	1,154	1,147	1,173	1,235	1,208
25	Farroupilha	0,000	0,600	0,376	0,399	0,370	0,334
26	Floresta	0,581	0,620	0,516	0,547	0,555	0,561
27	Glória	0,698	0,711	0,763	0,751	0,681	0,717
28	Guarujá	1,277	1,116	1,142	1,189	0,936	0,979
29	Higienópolis	0,290	0,431	0,367	0,378	0,394	0,380
30	Hípica	0,877	0,862	0,882	0,894	0,934	0,968
31	Humaitá	0,983	0,956	0,987	0,986	0,872	0,902
32	Independência	0,216	0,408	0,306	0,320	0,338	0,310
33	Ipanema	0,623	0,636	0,649	0,661	0,635	0,645
34	Jardim Botânico	0,271	0,395	0,356	0,368	0,446	0,441
35	Jardim Carvalho	1,173	1,149	1,122	1,148	1,131	1,145
36	Jardim do Salso	0,542	0,653	0,632	0,621	0,469	0,448
37	Jardim Floresta	0,885	0,851	0,769	0,790	0,980	1,034
38	Jardim Itu	0,458	0,556	0,576	0,574	0,617	0,652
39	Jardim Lindóia	0,500	0,579	0,571	0,575	0,473	0,477

40	Jardim Sabará	1,034	0,983	0,967	0,978	0,862	0,893
41	Jardim São Pedro	0,923	0,872	0,794	0,801	0,683	0,737
42	Lageado	1,388	1,251	1,296	1,323	1,455	1,437
43	Lami	1,441	1,287	1,352	1,382	1,370	1,380
44	Lomba do Pinheiro	1,260	1,247	1,226	1,254	1,270	1,266
45	Mário Quintana	1,420	1,396	1,355	1,391	1,434	1,407
46	Medianeira	0,878	0,859	0,824	0,843	0,743	0,754
47	Menino Deus	0,441	0,483	0,500	0,491	0,447	0,434
48	Moinhos de Vento	0,000	0,348	0,249	0,260	0,308	0,286
49	Mont' Serrat	0,271	0,435	0,278	0,313	0,361	0,333
50	Morro Santana	0,893	0,884	0,897	0,906	0,892	0,907
51	Navegantes	1,084	1,036	1,030	1,045	0,964	1,010
52	Nonoai	0,788	0,785	0,785	0,793	0,768	0,776
53	Partenon	1,102	1,089	1,023	1,073	1,052	1,055
54	Passo da Areia	0,749	0,751	0,661	0,701	0,719	0,735
55	Passo das Pedras	0,838	0,832	0,822	0,840	0,891	0,896
56	Pedra Redonda	0,000	0,721	0,615	0,683	0,471	0,443
57	Petrópolis	0,202	0,286	0,304	0,287	0,299	0,281
58	Ponta Grossa	1,257	1,150	1,184	1,210	1,134	1,212
59	Praia de Belas	0,989	0,899	0,822	0,810	0,533	0,494
60	Restinga	1,295	1,283	1,244	1,282	1,295	1,290
61	Rio Branco	0,487	0,542	0,369	0,426	0,421	0,390
62	Rubem Berta	0,888	0,886	0,859	0,882	0,885	0,892
63	Santa Cecilia	0,310	0,498	0,322	0,358	0,414	0,395
64	Santa Maria Goretti	0,833	0,818	0,807	0,793	0,760	0,829
65	Santa Teresa	1,537	1,519	1,459	1,506	1,499	1,491
66	Santana	0,369	0,428	0,392	0,402	0,417	0,402
67	Santo Antônio	0,607	0,637	0,618	0,619	0,582	0,598
68	São Geraldo	0,701	0,727	0,718	0,719	0,697	0,759
69	São João	1,204	1,159	0,977	1,078	0,992	0,991
70	São José	1,322	1,306	1,279	1,305	1,314	1,307
71	São Sebastião	0,660	0,702	0,714	0,707	0,654	0,710
72	Sarandi	1,138	1,131	1,103	1,132	1,142	1,147
73	Serraria	1,408	1,286	1,312	1,334	1,558	1,537
74	Teresópolis	0,632	0,655	0,706	0,689	0,641	0,638
75	Três Figueiras	0,259	0,590	0,438	0,448	0,351	0,314
76	Tristeza	0,375	0,443	0,481	0,471	0,446	0,438
77	Vila Assunção	0,367	0,595	0,588	0,586	0,379	0,342
78	Vila Conceição	1,422	1,111	0,919	1,019	0,915	0,822
79	Vila Ipiranga	0,645	0,664	0,650	0,660	0,641	0,657
80	Vila Jardim	1,286	1,236	1,194	1,226	1,196	1,181
81	Vila João Pessoa	1,225	1,176	1,182	1,192	1,158	1,167
82	Vila Nova	0,739	0,742	0,734	0,747	0,780	0,790