FERNANDA RODRIGUES VARGAS

Estimativas bayesianas da taxa de letalidade do infarto agudo do miocárdio

Orientador: Prof. Dr. Renato Martins Assunção
Coorientador: Prof. Dr. Marcelo Azevedo Costa

Belo Horizonte
03 de maio de 2013

Estimativas bayesianas da taxa de letalidade do infarto agudo do miocárdio

Esta é a versão original da dissertação elaborada pela candidata (Fernanda Rodrigues Vargas), tal como submetida à Comissão Julgadora.

Comissão Julgadora:

- Prof. Dr. Renato Martins Assunção (orientador) - UFMG
- Prof. Dr. Marcelo Azevedo Costa (coorientador) - UFMG
- Prof. Dr. Francisco Louzada Neto - USP
- Prof. Dr. Marcos Oliveira Prates - UFMG

Agradecimentos

Agradeço primeiro a Deus por guiar meu caminho e me dar forças para concluir esta etapa.

Agradeço à minha família por todo carinho e apoio, à minha mãe Suzana e avó Francisca. Ao Carlos pelo seu amor, compreensão e companheirismo em todos os momentos. Às minhas grandes incentivadoras e amigas Inára e Clarissa.

Aos queridos amigos Laura e Rodrigo pela amizade e almoços acolhedores de domingo; Luís Gustavo por dividir seu conhecimento e me ajudar no desenvolvimento do trabalho de pesquisa que muito me acrescentou; Paola pelo incentivo e companhia; Daniela pelas palavras de apoio e carinho.

À querida professora Jandyra Fachel pelo incentivo e exemplo, e ao professor Fernando Pulgati pelo grande apoio e estímulo.

Aos professores Renato Assunção e Marcelo Azevedo pela orientação e oportunidade de trabalhar no desenvolvimento das atividades de pesquisa no Laboratório de Estatística Espacial, com as quais adquiri muita experiência e conhecimento.

Ao professor Marcos Prates pela ajuda e contribuições dadas a este trabalho.

Às secretárias de pós-graduação, Rose e Rogéria, pelas palavras de incentivo e apoio, e à Maiza pelas conversas acompanhadas pelo seu bom café.

Resumo

A análise da distribuição geográfica da incidência de uma doença e sua relação com fatores de risco são fontes de informações relevantes nos estudos epidemiológicos e de saúde pública, porque sugerem hipóteses que conduzem à investigação e monitoramento das possíveis causas da doença. No mapeamento de doenças os métodos Bayesianos são amplamente usados, principalmente pela possibilidade de adotar uma estrutura hierárquica para a modelagem dos dados. O propósito desta dissertação é estimar a letalidade por infarto agudo do miocárdio (IAM) nas microrregiões brasileiras. Como os infartos são poucos nas microrregiões pequenas a letalidade acaba sendo mal estimada nestes casos. Assim, procuramos estudar conjuntamente outro fenômeno associado com a letalidade, a internação por IAM. Este fenômeno é mais comum e isto permite que o modelo Bayesiano use esta informação para estimar melhor a letalidade. Deste modo, neste trabalho analisamos conjuntamente os dados de internação e letalidade por IAM das microrregiões brasileiras. Para isso utilizou-se o modelo componente compartilhado proposto por Knorr-Held e Best (2001) especificado por uma nova formulação aqui proposta, avaliada e validada através de um estudo de simulação. Usamos o método INLA (*Integrated Nested Laplace Approximations*) para estimar o modelo. As estimativas suavizadas para a letalidade diminuíram o efeito das flutuações aleatórias não associadas ao risco, tornando perceptíveis as microrregiões com estimativas acima da média nacional. Além disso, verificamos que a relação entre a letalidade e a taxa de internação não é linear, havendo uma dependência da variabilidade entre a letalidade e o logaritmo da taxa de internação.

Palavras-chave: Epidemiologia espacial, estatística espacial, mapeamento de doenças, MCMC, INLA.

Abstract

The analysis of the geographical distribution of the incidence of a disease and its relationship with risk factors are sources of relevant information in epidemiological studies and public health. They suggest hypotheses that lead to investigation and to monitoring of the possible causes of the disease. In disease mapping, Bayesian methods are widely used, especially by the possibility of adopting a hierarchical structure for modeling data. The purpose of this dissertation is to estimate the mortality for acute myocardial infarction (AMI) in the Brazilian microregions. Since infarcts are few in small microregions, lethality ends up being poorly estimated in these cases. Thus, we study together another phenomenon associated with mortality, hospitalization for AMI. This phenomenon is more common and this allows the Bayesian model to borrow strength to better estimate the lethality. We analyze jointly the data of hospitalization and mortality due to AMI from Brazilian microregions. We used the shared component model proposed by Knorr-Held e Best (2001) specified by a new formulation here proposed, assessed and validated through a simulation study. We used the INLA (*Integrated Nested Laplace Approximations*) method to estimate the model. The smoothed estimates for mortality decreased the effect of random fluctuations not associated with risk, making visible the microregions with estimates above the national average. Furthermore, we found that the relationship between mortality and hospitalization rate is nonlinear, having a variability dependency between lethality and the logarithm of the admission rate.

Keywords: Spatial epidemiology, spatial statistics, disease mapping, INLA, MCMC.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Estudos epidemiológicos visam caracterizar uma doença na população para identificar fatores de risco e prevenir o aumento da incidência da doença, além de auxiliar órgãos públicos a definir políticas de saúde de maneira mais precisa e objetiva.

Uma técnica comumente usada nesse tipo de estudo é o mapeamento de doença na qual é possível produzir mapas da distribuição geográfica da incidência de doenças. Assim, podemos identificar padrões espaciais significativos e formular hipóteses a respeito das causas da doença. Consequentemente, a localização do evento é uma informação importante a ser considerada na análise dos dados epidemiológicos.

Os dados a serem analisados nesse contexto são georreferenciados, sendo sua localização espacial identificada pelas coordenadas geográficas. Pertencem à uma área da estatística denominada estatística espacial, e podem ser classificados em três categorias: dados de superfície aleatória, dados de processos pontuais e dados de área.

No presente trabalho os dados considerados são classificados como dados de área. Nesse tipo de dado os casos da doença são localizados em áreas disjuntas que particionam uma região geográfica. As informações como, por exemplo, o número de casos da doença e o número de pessoas em risco são agregadas espacialmente nas áreas disjuntas que compõem a região de estudo. Nesta dissertação, a região de estudo refere-se ao território nacional e as áreas correspondem às microrregiões brasileiras.

Nosso estudo é do tipo observacional e envolve a população brasileira no ano de 2010. O banco de dados utilizado contém informações do número de internações hospitalares por

infarto agudo do miocárdio (IAM) atendidos pelo Sistema Único de Saúde (SUS), e do número de mortes por IAM dada a internação pela mesma doença. Naquele ano foram registradas 50.987 internações e 7.811 óbitos decorrentes do infarto.

Nosso interesse é inferir a respeito da taxa de letalidade, ou simplesmente letalidade, e da taxa de internação. A taxa bruta fornece estimativas para estas quantidades. A taxa bruta de letalidade por IAM é dada pela razão entre o número de óbitos por IAM e o número de pessoas internadas pela doença, enquanto a taxa bruta de internação por IAM é a razão entre o número de internações por IAM e a população residente. No ano considerado, a taxa bruta de letalidade para todo o Brasil foi de 15% e a de internação de $0,03\%$.

Ao calcular a taxa bruta em cada microrregião brasileira obtemos estimativas não confiáveis que podem resultar em erros de interpretação. Isto acontece porque os valores mais extremos das taxas são tipicamente encontrados em áreas cuja população em risco é pequena. A ocorrência de uma morte por IAM ou uma internação nestas áreas pode gerar uma mudança significativa na estimativa da letalidade e da taxa de internação. A alta variabilidade associada a estas áreas não está relacionada ao risco do IAM. Além disso, quando nenhuma morte é observada na microrregião a taxa bruta de letalidade estimada é zero, o que não corresponde à realidade pois o risco de uma pessoa sofrer um infarto agudo do miocárdio não é nulo. Nos casos em que nenhuma internação por IAM é observada na área essa quantidade não pode ser estimada porque o denominador no cálculo da taxa bruta é zero sendo portanto indeterminada razão.

A metodologia Bayesiana pode ser usada como um meio para superar os problemas citados acima. Sua flexibilidade na aplicação aos dados e interpretação, bem como o tratamento de suavização dado às taxas, tornaram esta técnica popular na área espacial. (Assunção *et al.* , 1998; Maiti, 1998; Pascutto *et al.*, 2000)

A escolha de um modelo inferencial que considere conjuntamente a informação existente em todas as áreas simultaneamente contribui para a diminuição da variabilidade não associada ao risco da doença, melhorando as estimativas dessas áreas. A razão para este decréscimo de variância é o fato de que microrregiões geograficamente próximas tendem a apresentar taxas mais similares do que as microrregiões mais afastadas.

No cálculo da taxa bruta a informação sobre a posição espacial dos dados é ignorada, o que

torna seu uso pouco atraente diante de modelos que possibilitam incorporar essa informação bem como outras, como a correlação entre os dados dentro de uma mesma região.

1.1 Objetivos

O objetivo é estimar a taxa de letalidade e a taxa de internação por infarto agudo do miocárdio nas microrregiões brasileiras utilizando a metodologia Bayesiana. Também temos interesse em investigar se existe alguma relação entre estas duas taxas.

O modelo componente compartilhado apresentado no trabalho de Knorr-Held e Best (2001) foi usado para modelar os dados de infarto agudo do miocárdio por se tratar de um modelo que tem por finalidade separar a superfície de risco em componentes independentes de forma a identificar tendências similares e dissimilares. Este modelo tem sido aplicado e debatido em trabalhos recentes como em Best *et al.* (2005); Dabney e Wakefield (2005); Held *et al.* (2005a,b).

A distribuição a posteriori foi obtida por aproximação determinística através do método INLA (*Integrated Nested Laplace Approximations*) proposto por Rue *et al.* (2009), utilizado devido às dificuldades analíticas encontradas. Também realizamos um estudo de simulação para constatar a nova formulação proposta para o preditor linear em relação à usada em Knorr-Held e Best (2001), a fim de possibilitar a implementação do modelo componente compartilhado e a realização da inferência Bayesiana para os dados através do método INLA.

1.2 Infarto Agudo do Miocárdio

O infarto agudo do miocárdio (IAM) é classificado como uma das doenças isquêmicas do coração, tal como angina pectoris, infarto do miocárdio recorrente, etc. (DATASUS, 2012) Caracteriza-se como um evento agudo sendo necessária a internação hospitalar.

Segundo o Ministério da Saúde (Min.Saúde, 2012a) as taxas elevadas de mortalidade por doenças isquêmicas do coração estão associadas à maior prevalência de fatores de risco relacionados diretamente ao hábito e ao estilo de vida das pessoas, tais como o hábito de fumar, hipertensão, obesidade, diabetes, sedentarismo, estresse, etc.

No Brasil, o infarto agudo do miocárdio possui impacto relevante em termos de mortalidade e no número de hospitalizações. De acordo com o Portal da Saúde do Ministério da Saúde (Min.Saúde, 2012b), as principais causas de mortes entre homens são as decorrentes de doenças isquêmicas do coração, entre elas o IAM.

1.3 Fonte dos Dados

O DATASUS - Departamento de Informática do SUS - disponibiliza informações de saúde com base em diferentes fontes como, por exemplo, o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informação Hospitalar do Sistema Único de Saúde (SIH/SUS).

Os dados fornecidos pelo SIM (2012) são geridos pelo departamento da secretaria de vigilância em saúde em conjunto com as secretarias estaduais e municipais de saúde. As secretarias de saúde coletam as declarações de óbitos dos cartórios e incluem as informações no sistema. A causa básica do óbito, codificada segundo a Classificação Internacional de Doenças (CID-10), é uma das informações registradas baseada na declaração feita pelo médico atestante.

O SIH-SUS (2012) fornece informações sobre as principais causas de internações no Brasil, a quantidade de leitos existentes para cada especialidade e o tempo médio de permanência do paciente no hospital, os procedimentos mais frequentes realizados mensalmente em cada hospital, etc. Com estas informações o pagamento dos serviços hospitalares prestados pelo SUS é realizado.

Apesar da facilidade de acesso às informações de saúde, encontramos algumas limitações com relação aos dados disponibilizados pelo DATASUS. As internações registradas referem-se exclusivamente às internações pagas pelo SUS, não contendo aquelas pagas pelos planos privados de saúde e aquelas pagas diretamente pelo usuário. Com isso não é possível utilizar a informação da população residente em 2010, pois o número de internações e o número de óbitos contemplam apenas uma parte da população e, consequentemente, o resultado estaria incorreto. Para contornar esse viés subtraímos da população residente uma estimativa da população coberta por planos privados hospitalares.

A estimativa da população coberta por planos privados hospitalares é composta pelos

beneficiários de planos privados com cobertura de assistência médica. Esta informação para o ano de 2010 está disponível na Agência Nacional de Saúde Suplementar - ANS (2012), assim como para outros anos. Segundo a ANS, a maneira como a operadora informa o endereço dos beneficiários de planos de saúde coletivos pode gerar algum erro nessa informação. Isto ocorre porque o endereço repassado pela operadora refere-se ao endereço da empresa contratante do plano coletivo ao invés do endereço residencial do beneficiário.

1.4 Organização do Trabalho

No capítulo 2 falamos brevemente sobre dados espaciais, e em seguida introduzimos a metodologia Bayesiana e a abordagem hierárquica caracterizada pela especificação sucessiva das distribuições de probabilidades. Finalizamos o capítulo apresentando os métodos de aproximação estocástica e determinística usados para a obtenção da distribuição a posteriori. O capítulo 3 descreve o estudo de simulação desenvolvido com o propósito de verificar a equivalência entre duas formulações para o preditor linear. Na simulação dois cenários foram considerados, com o propósito de avaliar a equivalência entre a formulação proposta por Knorr-Held e Best (2001) e a apresentada aqui a fim de tornar possível a implementação do modelo componente compartilhado no pacote INLA do programa R.

O capítulo 4 apresenta um estudo exploratório dos dados a fim de caracterizá-los, e os resultados obtidos com a modelagem hierárquica Bayesiana de modo a inferir a respeito da taxa de internação e, principalmente, da letalidade por IAM. Encerramos este trabalho com o capítulo 5 no qual concluímos sobre a importância da análise Bayesiana na suavização das taxas e o ganho de informação com o uso do modelo componente compartilhado, que permite investigar tendências espaciais compartilhadas entre duas doenças e específicas de cada uma.

Capítulo 2

Metodologia

2.1 Análise de Dados Espaciais

A análise espacial é um ramo da estatística que estuda quantitativamente os fenômenos aleatórios considerando sua posição no espaço. Tem por finalidade descrever a distribuição de dados estatísticos contribuindo assim na identificação e modelagem de padrões espaciais. Dentre as técnicas de análise espacial destacamos o mapeamento de doenças como a mais popular no estudo do risco de doenças (ver Held *et al.* (2005b); Knorr-Held e Best (2001); Pascutto *et al.* (2000), entre outros).

Os dados na análise espacial podem ser discretos ou contínuos, e podem ser classificados nas seguintes categorias: dados de superfície aleatória, dados de processos pontuais e dados de área. Os dados podem estar agregados ou as observações dispersas em pontos no espaço. No presente trabalho, o interesse centra-se em dados de área. Um conjunto finito de áreas compõem a região de estudo como, por exemplo, os bairros que compõem o município. As áreas são disjuntas e podem ser regulares ou irregulares.

Para formalizar a descrição acima, considere um conjunto de variáveis aleatórias indexadas pelas áreas A_i fixas, Y_i, $A_i \in \{A_1, A_2, \cdots, A_n\}$, sendo que a união destas áreas $\bigcup_{i=1}^{n} A_i = R$ forma a região de estudo.

Nesta dissertação as áreas são definidas pelas microrregiões brasileiras. No total são 558 áreas disjuntas distribuídas no território nacional formando a nossa região de estudo. Cada área possui as seguintes informações: número de internações por infarto agudo do miocárdio

(IAM), número de mortes após a internação, e número de pessoas em risco, todas referentes ao ano de 2010 e relacionadas àquelas pessoas assistidas pelo SUS. O número de sobreviventes por IAM após a internação foi calculado através da subtração do número de internações pelo número de óbitos. Este dado será necessário para a modelagem da seção 2.2.1.1. As áreas (microrregiões) são delimitadas por polígonos e a união dos polígonos forma o Brasil.

Uma característica ligada às aplicações espaciais é a existência de correlação entre as unidades espaciais, principalmente entre aquelas que fazem fronteira. Trata-se de uma correlação induzida pela proximidade geográfica com áreas próximas no espaço tendendo a apresentar valores similares. Métodos de estatística espacial têm considerado a configuração espacial para detectar e quantificar padrões nos dados e investigar o grau de associação entre fatores de risco e uma doença.

A partir da localização geográfica das áreas é possível construir uma estrutura espacial para os dados, seja através da informação de fronteira ou da distância entre as áreas. Na literatura espacial a utilização desse tipo de informação é conhecida como mecanismo de "*borrow strength*" (Gelfand *et al.* (2010); Lawson (2009); Pascutto *et al.* (2000)). As áreas emprestam informações umas as outras de forma a diminuir a variabilidade não associada ao risco da doença. Com esse mecanismo as estimativas locais da doença se tornam mais precisas e mais confiáveis, principalmente naquelas áreas afetadas pelo tamanho da população em risco. Um exemplo disso são as áreas rurais cuja densidade populacional é baixa. A observação de casos de uma doença rara nestas áreas, como o infarto agudo do miocárdio, afeta diretamente o risco. Se considerarmos, por exemplo, a taxa bruta de mortalidade (razão entre o número de total de óbitos pela população total) como estimativa para o risco de morte, a ocorrência de uma morte por infarto elevará significativamente a taxa.

2.2 Inferência Bayesiana

Muitos trabalhos que abordam o mapeamento de doenças, tais como Wakefield (2007), Held *et al.* (2005a) e Best *et al.* (2005), utilizam inferência Bayesiana na análise dos dados. Isto se deve principalmente ao fato de ser uma metodologia flexível que possibilita o uso de modelos que permitam incorporar a dependência geográfica entre as áreas, fornecendo

uma estimativa mais estável para o risco da doença do que as estimativas obtidas através da taxa bruta. Além disso, o avanço computacional significativo ocorrido no final dos anos 80 permitiu o desenvolvimento de softwares, como `WinBUGS` (Lunn *et al.*, 2000), e a aplicação de métodos de aproximação para a realização desta inferência.

Essencialmente, a metodologia Bayesiana trata o desconhecido como aleatório. Por exemplo, ao assumir um modelo para os dados precisamos definir os parâmetros que caracterizam este modelo. Entretando, tais parâmetros são desconhecidos e, através da metodologia Bayesiana, é possível atribuir a estes uma distribuição de probabilidade que descreva o nosso desconhecimento. O parâmetro desconhecido será representado por θ ou, no caso multivariado, pelo vetor paramétrico por $\boldsymbol{\theta}$. Corresponde à quantidade de interesse sobre a qual desejamos fazer inferência e obter maior conhecimento.

A informação que temos a priori sobre o parâmetro θ é expressa quantitativamente por uma distribuição de probabilidade conhecida como distribuição a priori, que é parametrizada pelos hiperparâmetros $\boldsymbol{\gamma}$. Sua construção pode ser apoiada no conhecimento prévio de pesquisadores a respeito de θ, em experimentos anteriores, ou através de métodos como de Jeffreys e Bayes-Laplace (Paulino *et al.* (2003)).

Após especificarmos a distribuição a priori, combinamos esta informação prévia com a informação trazida pelos dados e obtemos uma nova distribuição chamada de distribuição a posteriori. Esta distribuição descreverá e atualizará o conhecimento sobre o parâmetro θ após observar a amostra.

A inferência a posteriori a respeito de θ é afetada pelos dados observados $\boldsymbol{y}$ através da distribuição amostral $p(\boldsymbol{y}|\theta)$. Quando considerada uma função de θ com $\boldsymbol{y}$ fixado, esta distribuição é chamada de função de verossimilhança. Ou seja, dados os valores observados, a função depende apenas dos parâmetros desconhecidos θ.

Seja y uma variável aleatória, ou vetor aleatório $\boldsymbol{y}$, com função densidade de probabilidade $p(\boldsymbol{y}|\theta)$, em que θ é o parâmetro (ou vetor de parâmetros $\boldsymbol{\theta}$) desconhecido que caracteriza a distribuição de $\boldsymbol{y}$. Assuma $p(\theta)$ como a distribuição de probabilidade a priori para θ. A distribuição de probabilidade a posteriori $p(\theta|\boldsymbol{y})$, obtida via Teorema de Bayes, é dada por

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\theta)}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\theta)\cdot p(\theta)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{y},\theta)\, dF(\theta)} = \frac{p(\boldsymbol{y}|\theta)\cdot p(\theta)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\theta)\cdot p(\theta)\, dF(\theta)} \tag{2.1}$$

No denominador $p(\boldsymbol{y})$ corresponde à distribuição preditiva a priori, considerada constante pois não depende de θ. Sendo assim, pode-se considerar para a inferência a distribuição a posteriori a menos de uma constante normalizadora como segue:

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta) \cdot p(\theta) \tag{2.2}$$

A inferência Bayesiana baseia-se na distribuição de probabilidade a posteriori $p(\theta|\boldsymbol{y})$. A informação contida nesta distribuição a respeito de θ pode ser resumida através da estimação pontual ou por intervalo, ou mesmo em termos de esperanças de funções particulares, $g(\theta)$, do parâmetro θ como segue

$$E[g(\theta)|\boldsymbol{y}] = \int_{\boldsymbol{\Theta}} g(\theta) \cdot p(\theta|\boldsymbol{y})\, dF(\theta) \tag{2.3}$$

No caso em que θ é multivariado, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$, a distribuição marginal para o parâmetro θ_k pode ser obtida a partir da integral

$$p(\theta_k|\boldsymbol{y}) = \int_{\boldsymbol{\theta_{-k}}} p(\boldsymbol{\theta}|\boldsymbol{y})\, dF(\boldsymbol{\theta}_{-k}) \tag{2.4}$$

onde $\boldsymbol{\theta}_{-k}$ é o vetor multivariado sem o k-ésimo parâmetro, isto é, $\boldsymbol{\theta}_{-k} = (\theta_1, \cdots, \theta_{k-1}, \theta_{k+1}, \cdots, \theta_p)$.

Em algumas situações encontrar a distribuição a posteriori pode se tornar uma tarefa difícil. Um exemplo disto é o caso em que a constante normalizadora que define a distribuição a posteriori não é facilmente derivável, ou mesmo quando as integrais em (2.3) e (2.4) não são tratáveis analiticamente. Nesses casos são necessários métodos computacionais baseados em aproximações estocásticas ou determinísticas para que a distribuição possa ser obtida. Na seção 2.3 alguns desses métodos serão relatados.

A concepção Bayesiana de atribuir distribuição de probabilidade ao parâmetro de interesse pode estabelecer uma hierarquia natural em sua construção que permite fácil especificação das distribuições. Além disso, a estrutura hierárquica é bastante utilizada na modelagem espacial por possibilitar que modelos mais realísticos possam ser incorporados na análise. A maioria dos trabalhos na literatura relacionados à modelagem espacial emprega a estrutura hierárquica (Best *et al.* (2005); Dabney e Wakefield (2005); Held *et al.* (2005b)).

2.2.1 Abordagem Hierárquica

Nosso interesse é fazer inferência sobre a letalidade e a taxa de internação através da metodologia Bayesiana. Para isto, precisamos construir uma estrutura hierárquica para os dados que possibilite especificar modelos que incorporem a dependência espacial entre as observações y_i como distribuição a priori.

A modelagem hierárquica está baseada na decomposição da distribuição conjunta das variáveis aleatórias em distribuições condicionais e distribuições marginais. Visto que, às vezes, pode ser complicado especificar tal distribuição conjunta, a abordagem hierárquica se torna uma alternativa.

Sob a perspectiva Bayesiana, a estrutura hierárquica é dividida em estágios (ou níveis). Abaixo os estágios são apresentados de forma esquematizada, sendo possível verificar a facilidade na construção e no entendimento das etapas da modelagem a partir da estrutura hierárquica.

- 1º estágio - Modelo dos dados: $[\boldsymbol{Y}|\theta, \lambda] \sim p(\boldsymbol{y}|\theta, \lambda)$
- 2º estágio - Modelo dos parâmetros: $[\theta|\lambda] \sim p(\theta|\lambda)$
- 3º estágio - Modelo dos hiperparâmetros: $[\lambda] \sim p(\lambda)$

No primeiro estágio, especificamos a distribuição dos dados dado o parâmetro desconhecido θ. Neste estágio as observações y_i são condicionalmente independentes dado os valores de θ. No segundo estágio da especificação do modelo definimos a distribuição a priori para θ, e no terceiro estágio atribuímos uma distribuição hiperpriori para o(s) parâmetro(s) que definem a distribuição a priori denominado de hiperparâmetro, representado aqui por λ.

Aplicamos a modelagem hierárquica aos dados de infarto agudo do miocárdio. No primeiro estágio definimos a distribuição Poisson como sendo o modelo gerador dos dados, cuja média é dada pelo número esperado de casos vezes o risco relativo, onde o risco relativo é dado por θ. No segundo estágio atribuímos a distribuição a priori espacial introduzida por Besag *et al.* (1991), conhecida como priori de convolução BYM. E, para o terceiro estágio assumimos a distribuição Gama como distribuição hiperpriori para os parâmetros de precisão associados à distribuição a priori. Estes modelos serão descritos nas subseções a seguir.

2.2.1.1 Modelo dos Dados

No mapeamento de doenças, a contagem de casos da doença, $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$, nas diferentes áreas particionadas sob a região de estudo, é modelada como variáveis aleatórias Binomial ou Poisson. Se a doença for considerada rara, o modelo Poisson é assumido para os dados, uma vez que a distribuição Poisson é uma aproximação do modelo Binomial quando n é grande e a probabilidade de sucesso p é pequena. (Gelfand *et al.*, 2010; Lawson, 2009)

Em relação aos dados de infarto agudo do miocárdio a distribuição Poisson foi definida como o modelo dos dados. Classificamos a doença como rara por causa da baixa contagem de casos observada nas microrregiões brasileiras, como será visto no capítulo 4.

Estabelecemos que a média μ_i da distribuição Poisson para a i-ésima área é dada pelo produto entre o número esperado de casos E_i, conhecido e fixo, e o risco relativo θ_i, independentemente para $i = 1, 2, \cdots, n$. Logo, a média é constituída por dois componentes: um relacionado ao efeito da população e outro relacionado ao risco, ambos dentro da área i. O número de pessoas em risco n_i poderia ser considerado em vez de E_i.

O número esperado de casos E_i é obtido do produto entre o número de pessoas em risco n_i e o risco constante r. O risco r representa a taxa global observada da doença, seu cálculo é dado por

$$r = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} n_i. \tag{2.5}$$

Note que, o cálculo do número esperado de casos, $E_i = r \cdot n_i$, considera que o risco na área i é igual ao risco em toda a região.

Temos como objetivo estimar o risco relativo θ_i em cada área i. O estimador de máxima verossimilhança para θ_i é dado pela razão entre o número de casos da doença e o número esperado de casos na área i, y_i/E_i, conhecida em inglês como *Standardized Mortality Ratio* (SMR), e seu desvio padrão é estimado por $s_{\hat{\theta}} = \sqrt{y_i}/E_i$. As estimativas geradas por este estimador são muito afetadas por poucos casos a mais ou a menos na área i. Uma possibilidade de tratar este tipo de influência é modelar o risco relativo através da ligação logarítmica na qual é possível incorporar efeitos aleatórios que suavizem as estimativas do risco. Esta função liga o risco relativo $\boldsymbol{\theta}$ a um preditor linear $\boldsymbol{\eta}$, i.e., $\log(\theta_i) = \eta_i$.

Como dito anteriormente, a estrutura hierárquica foi empregada no estudo do mapeamento do infato agudo do miocárdio (IAM). Assumimos o modelo Poisson para os dados no primeiro estágio, assim a contribuição das observações y_i é independente condicionado ao vetor paramétrico $\boldsymbol{\theta}$, que são os riscos relativos específicos de cada área i, para $i = 1, \cdots, n$. O log risco relativo foi modelado seguindo o modelo componente compartilhado proposto por Knorr-Held e Best (2001).

O modelo componente compartilhado tem como objetivo realizar a análise conjunta de duas ou mais doenças com o propósito de descobrir similaridades, ou dissimilaridades, na distribuição geográfica do risco dessas doenças. A ideia do modelo é separar a superfície de risco em dois componentes, um componente compartilhado pelas doenças e outro componente específico de cada doença.

A suposição do modelo em relação às doenças é de que estas compartilham fatores de risco comum. O componente compartilhado capta os fatores de risco comum entre as doenças, enquanto o componente específico capta os fatores de risco exclusivos da doença. Os componentes estão relacionados com as covariáveis não-observáveis, representam os fatores de risco que agem sobre o risco porém não podem ser medidos ou identificados. Além disso, os componentes são assumidos independentes entre si (Knorr-Held e Best, 2001).

Uma característica deste modelo é que as estimativas da superfície de risco são obtidas separadamente para cada componente. Assim, mapas dessas estimativas podem ser usados para exibir diferentes padrões espaciais dos fatores de risco comum e específicos ao longo da região de estudo.

Como mencionado acima, o modelo componente compartilhado analisa conjuntamenta múltiplas medições, como por exemplo a ocorrência de doenças. Para aplicarmos este modelo aos dados de IAM, definimos o número de óbitos como sendo uma das medições e o número de sobreviventes como sendo a outra medição, ambas após a internação por IAM. As observações para cada uma das medições y_{id}, para $d = 1, 2$ e $i = 1, 2, \cdots, n$, foram assumidas variáveis aleatórias Poisson condicionalmente independentes dado θ_{id} com média igual a $\mu_{id} = E_{id} \cdot e^{\eta_{id}}$

na área i. A formulação do log risco relativo para cada medição é descrito abaixo

$$\begin{aligned} \eta_{i1} &= \alpha_1 + \phi_i \cdot \delta + \psi_{i1} \\ \eta_{i2} &= \alpha_2 + \phi_i / \delta + \psi_{i2} \end{aligned} \tag{2.6}$$

em que α_d é o efeito da média global da medição d através da região de estudo, δ é o parâmetro que pondera o componente compartilhado permitindo diferenciar o risco associado aos fatores de risco comum para cada medição. Os parâmetros ϕ_i, ψ_{i1} e ψ_{i2} correspondem, respectivamente, aos componentes compartilhado, específico da medição 1, número de mortes, e específico da medição 2, número de sobreviventes, na área i. Estes componentes são independentes entre si e correspondem aos efeitos aleatórios que representam as covariáveis não-observadas.

O interessante do modelo compartilhado é que nos permite investigar tendências espaciais compartilhadas entre o número de mortes e de sobreviventes após a internação por IAM, além das tendências que são específicas de cada um. Obteremos conhecimento adicional, pois com este modelo as estimativas da taxa de mortalidade e da taxa de sobrevivência serão fornecidas. No entanto, nosso interesse principal é inferir sobre a taxa de letalidade por IAM, e também investigar sua relação com a taxa de internação decorrente desta doença.

A média da distribuição Poisson da medição 1 e medição 2 são dadas por μ_{i1} e μ_{i2} como segue

$$\begin{aligned} \mu_{i1} &= E_{i1} \cdot \theta_{i1} = E_{i1} \cdot e^{\eta_{i1}} = E_{i1} \cdot e^{\alpha_1 + \phi_i \cdot \delta + \psi_{i1}} \\ \mu_{i2} &= E_{i2} \cdot \theta_{i2} = E_{i2} \cdot e^{\eta_{i2}} = E_{i2} \cdot e^{\alpha_2 + \phi_i / \delta + \psi_{i2}} \end{aligned} \tag{2.7}$$

A taxa de internação por IAM, ϑ_i, é dada pela soma das médias μ_{i1} e μ_{i2}. Isto é, $\vartheta_i = \mu_{i1} + \mu_{i2}$. A letalidade por IAM é a probabilidade de morrer dado que foi internado pela doença. Assim, a letalidade, φ_i, é dada pela razão entre a média μ_{i1} e ϑ_i.

$$\varphi_i = \frac{\mu_{i1}}{\mu_{i1} + \mu_{i2}} = \frac{\mu_{i1}}{\vartheta_i} \tag{2.8}$$

Essas quantidades são consequência dos resultados probabilísticos obtidos da modelagem

das variáveis aleatórias, número de mortes e número de sobreviventes por IAM, assumindo que estas variáveis são independentes. A demonstração de tais resultados encontram-se no Apêndice A.

2.2.1.2 Modelo dos Parâmetros

O primeiro estágio foi abordado na seção anterior. Nesta seção, o segundo estágio do modelo será especificado a partir da definição das distribuições a priori para os parâmetros que compõem o modelo componente compartilhado que representa o log risco relativo.

Atribuímos a cada parâmetro α_d $(d = 1, 2)$ uma distribuição a priori não informativa, representada pela distribuição normal com parâmetro de variância com um valor grande, com o propósito de não inserir nenhuma informação precisa sobre este efeito fixo. Assumimos que o logaritmo do parâmetro δ tem distribuição normal com média e variância conhecidos (valores serão apresentados na seção 4.1), como sugerido por Knorr-Held e Best (2001). Para cada um dos três componentes, compartilhado, ϕ_i, e específicos da medição 1 e 2, ψ_{i1} e ψ_{i2}, foi atribuída uma priori de convolução BYM.

A priori de convolução BYM introduzida por Besag *et al.* (1991), ou método BYM, emprega o mecanismo "*borrow strength*" visando controlar a variabilidade com base na informação não apenas global como também local através das áreas. Lawson *et al.* (2000) realizou um estudo de simulação e concluiu que o modelo BYM foi o modelo mais robusto dentre os que incorporavam estrutura espacial em sua formulação. Esta priori é constituída da soma de dois efeitos aleatórios independentes: um não estruturado espacialmente e outro estruturado espacialmente. O primeiro efeito mede a heterogeneidade local, enquanto o segundo mede a similaridade espacial.

Assumimos para o efeito aleatório não estruturado a distribuição normal com média μ e parâmetro de precisão τ_{uns} independente para cada área i. O parâmetro de precisão τ_{uns} correspondente ao inverso da variância σ^2_{uns}. Para o efeito espacial estruturado assumimos o modelo ICAR (*Intrinsic Conditional Autoregressive*) no qual uma estrutura de vizinhança é especificada de modo a incluir a informação da estrutura espacial entre as áreas. Este modelo é um caso particular do modelo CAR, apresentado a seguir, quando o parâmetro ρ de correlação espacial é igual a um: $\rho = 1$.

O modelo CAR é definido a partir das distribuições condicionais normais da variável aleatória U como segue

$$U_i|U_j = u_j, j \neq i \sim N\left(\rho\bar{u}_j, \frac{\sigma^2_{car}}{m_i}\right), \qquad i = 1, \cdots, n \tag{2.9}$$

em que $\bar{u}_j$ é a média dos vizinhos da área i, i.e., $\bar{u}_j = \frac{1}{m_i}\sum_{j\in\partial_i} u_j$, onde ∂_i denota o conjunto de índices dos vizinhos associado a área i. A média da área i dado as áreas restantes é proporcional à média de seus vizinhos vezes a constante de proporcionalidade ρ, sendo ρ tratado em seguida. A variância condicional σ^2_{car} desconhecida é diretamente proporcional ao número total de vizinhos da área i, m_i, e representa a variação local estruturada espacialmente. Assim, quanto mais vizinhos a área tiver mais informação temos sobre a mesma.

A distribuição conjunta do vetor aleatório $\boldsymbol{U}$ é dada pela distribuição Normal multivariada k-dimensional com vetor de médias $\boldsymbol{\mu}$ igual a zero e matriz de precisão $\boldsymbol{Q} = \frac{1}{\sigma^2_{car}}[\text{diag}(\boldsymbol{m}) - \rho\boldsymbol{A}]$. A matriz $\boldsymbol{A} = (a_{ij})$, $i, j = 1, \cdots, n$, é uma matriz de adjacência que representa a estrutura de vizinhança a partir da informação de fronteira entre as áreas, ou seja, $a_{ij} = 1$ se a área j é vizinha da área i, $a_{ij} = 0$ caso contrário. Além disso, nenhuma área é vizinha dela mesma: $a_{ii} = 0$ para todo i. O vetor $\boldsymbol{m}$ possui em sua i-ésima posição o número de vizinhos, m_i, da área i.

Abaixo é apresentada uma matriz de adjacência $\boldsymbol{A}$ construída a partir do mapa da Figura 2.1 cujo grafo de vizinhança é representado pelas linhas sobrepostas ao mapa ligando as áreas que fazem fronteira.

$$\boldsymbol{A} = \begin{bmatrix} & \text{Chapecó} & \text{Xanxerê} & \text{Concórdia} & \text{Joaçaba} & \text{Canoinhas} & \text{Curitibanos} \\ \text{Chapecó} & 0 & 1 & 1 & 0 & 0 & 0 \\ \text{Xanxerê} & 1 & 0 & 1 & 1 & 0 & 0 \\ \text{Concórdia} & 1 & 1 & 0 & 1 & 0 & 0 \\ \text{Joaçaba} & 0 & 1 & 1 & 0 & 1 & 1 \\ \text{Canoinhas} & 0 & 0 & 0 & 1 & 0 & 1 \\ \text{Curitibanos} & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

A matriz de precisão da distribuição conjunta $\boldsymbol{U}$ pode ser redefinida considerando a matriz de pesos espaciais padronizados $\boldsymbol{W}$, que corresponde à matriz de adjacência padronizada por linhas de forma que as linhas somem 1. Assim, a matriz de precisão é dada por

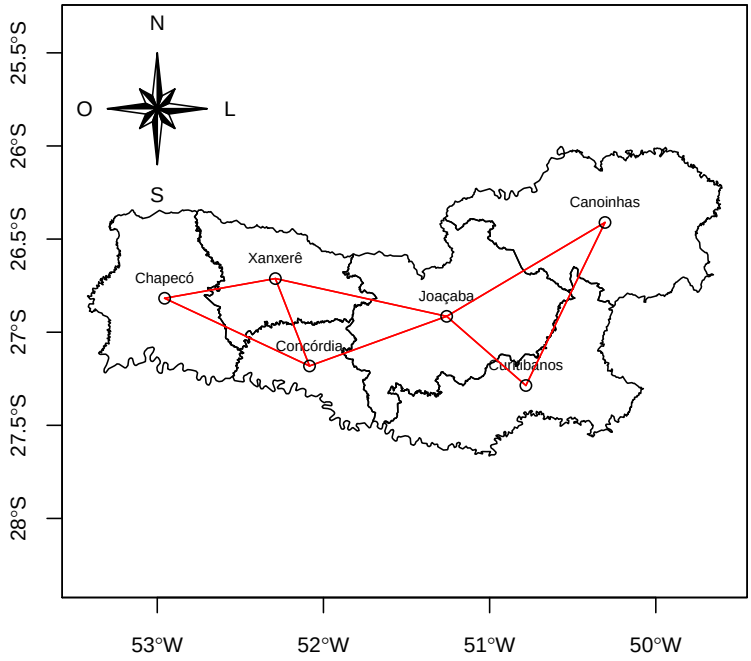

Figura 2.1: *Mapa com algumas microrregiões de Santa Catarina com o grafo de vizinhança representado pelas linhas.*

$\boldsymbol{Q} = \frac{1}{\sigma^2_{car}}\text{diag}(\boldsymbol{m})[\boldsymbol{I} - \rho\boldsymbol{W}]$, onde $\boldsymbol{W} = \text{diag}(\boldsymbol{m})^{-1}\boldsymbol{A}$.

A densidade conjunta do modelo CAR existe se a matriz de covariâncias $\boldsymbol{Q}^{-1}$ é simétrica e positiva. Esta condição é satisfeita se o parâmetro de correlação espacial ρ pertencer ao intervalo $\left(\frac{1}{\lambda}, 1\right)$, sendo $\frac{1}{\lambda} < 0$ e λ o menor autovalor da matriz $\boldsymbol{W}$. Quando ρ é igual a 1 a distribuição torna-se imprópria pois a matriz de precisão não é invertível, e assim a matriz de covariância não existe. Este é um caso particular do modelo CAR denominado ICAR. Uma distribuição é dita imprópria quando não cumpre as propriedades necessárias que definem uma distribuição de probabilidade.

No terceiro estágio da modelagem hierárquica, especificamos as distribuições hiperpriori para os parâmetros de precisão da priori de convolução BYM. Para a parte não estruturada espacialmente atribuímos para o parâmetro τ_{uns} a distribuição Gama com parâmetros $\alpha = 0.5$ e $\beta = 0.0005$. A precisão τ_{car} da parte estruturada espacialmente, em que $\tau_{car} = 1/\sigma^2_{car}$, também foi definida pela distribuição Gama com os mesmos parâmetros α e β citados acima. Esta distribuição hiperpriori foi recomendada por Kelsall e Wakefield (1999)

para representar a variação do parâmetro de precisão do efeito aleatório espacial.

Na seção 4.1 a modelagem hierárquica para os dados de infarto agudo do miocárdio será apresentada de forma esquematizada com o objetivo de mostrar as distribuições de probabilidade definidas em cada um dos estágios da modelagem com os respectivos valores que as especificam.

No mapeamento de doenças o uso de modelos espaciais que considerem a correlação espacial através da abordagem hierárquica geralmente resulta em distribuições a posteriori complexas, intratáveis matematicamente, que necessitam de métodos de aproximação para sua obtenção. Na próxima seção tais métodos serão abordados, mais especificamente os métodos de aproximação estocástica e determinística.

2.3 Métodos Computacionais

Os resumos inferenciais para a distribuição a posteriori podem ser obtidos através de esperanças de funções particulares como visto em 2.3, porém esta distribuição frequentemente não é tratável matematicamente. Apesar das vantagens da metodologia Bayesiana, alguns problemas surgem no momento de se obter a distribuição a posteriori, tais como a dificuldade em resolver as integrais ou em encontrar a constante normalizadora que define a distribuição. Os métodos de aproximação estocástica e de aproximação determinística são muito usados como meios de solucionar estes problemas.

Os métodos de aproximação estocástica são os métodos mais comumente usados na obtenção da distribuição a posteriori baseado em simulações. O avanço computacional ocorrido no final dos anos 80 possibilitou o desenvolvimento de técnicas mais extensivas baseadas em simulação que colaboraram para o crescimento do número de trabalhos estatísticos na área. Na literatura referente ao mapeamento de doenças a inferência frequentemente é realizada através da simulação de Monte Carlo via Cadeias de Markov (MCMC) com base em amostras da distribuição a posteriori, Best *et al.* (2005); Held *et al.* (2005b); Pascutto *et al.* (2000), entre outros.

Nas subseções seguintes, os métodos de aproximação estocástica e determinística serão abordados.

2.3.1 Métodos estocásticos (MCMC)

Os métodos estocásticos são técnicas baseadas em simulação estocástica para gerar amostras da distribuição a posteriori, e assim, obter informações da distribuição de forma a produzir estimativas do parâmetro de interesse, ou seja, realiza inferências a partir da amostra simulada da distribuição a posteriori. Dentre as técnicas estão o método de reamostragem e o método Monte Carlo via cadeias de Markov.

O método Monte Carlo via cadeias de Markov (MCMC) são métodos de simulação iterativa dos valores dos parâmetros via cadeia de Markov cuja distribuição estacionária, atingida após a convergência da cadeia, corresponde à distribuição a posteriori. Dessa forma, obtemos medidas resumo que caracterizam a distribuição a posteriori através das amostras simuladas desta distribuição. Neste método, amostras são geradas para qualquer distribuição de probabilidade, independente do número de parâmetros ou do conhecimento da constante de normalização. Os algoritmos mais usados na construção das cadeias de Markov são o amostrador de Gibbs e o algoritmo Metropolis-Hastings.

No algoritmo Metropolis-Hastings, cujo nome é devido aos trabalhos de Metropolis *et al.* (1953) e Hastings (1970), um valor é gerado de uma distribuição auxiliar, chamada de distribuição proposta, e aceito ou não com base em uma determinada probabilidade.

O amostrador de Gibbs, tratado inicialmente por Geman e Geman (1984), utiliza as distribuições condicionais completas como núcleo de transição da cadeia. A distribuição condicional completa $p(\theta_i|\boldsymbol{\theta}_{-i})$ é a distribuição do i-ésimo parâmetro de $\boldsymbol{\theta}$ condicionado aos parâmetros restantes $\boldsymbol{\theta}_{-i}$, onde $\boldsymbol{\theta}_{-i} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_p)$. Resumidamente, o algoritmo fornece um valor para cada parâmetro θ_i em cada iteração a partir da sua distribuição condicional completa dado os valores atualizados dos parâmetros restantes. Ao invés de especificar uma distribuição conjunta para $\boldsymbol{\theta}$ especificamos a distribuição condicional de cada parâmetro.

Por se tratar de um método bastante usado no mapeamento de doenças, apenas o método Monte Carlo via cadeias de Markov foi citado. Detalhes sobre outros métodos de simulação estocástica podem ser encontrados em Gamerman e Lopes (2006).

A dificuldade de diagnosticar a convergência da cadeia e o tempo computacional necessário para atingir tal convergência são características da técnica MCMC que a tornam custosa

computacionalmente. Atualmente o método de aproximação determinística tem sido explorado por pesquisadores como um meio alternativo de determinar a distribuição a posteriori.

2.3.2 Métodos determinísticos (INLA)

O método determinístico é outro método de aproximação usado na solução das integrais em (2.3) e (2.4) a fim de obter a distribuição a posteriori. As aproximações determinísticas, cuja aplicação era maior antes do crescente uso dos métodos estocásticos, voltaram a ser uma alternativa aos métodos de simulação estocástica a partir do trabalho de Rue *et al.* (2009) com o método de aproximação INLA - *Integrated Nested Laplace Approximations*. Este método possibilita realizar inferência Bayesiana sobre os modelos latentes Gaussianos, aproximando de forma determinística as distribuições a posteriori marginais. Além disso, apresenta vantagens em relação às simulações MCMC, tais como rapidez computacional, precisão e não necessitar analisar a convergência da cadeia.

Os modelos latentes Gaussianos são uma classe de modelos hierárquicos que assumem um campo latente Gaussiano de dimensão n indiretamente observado através dos dados. Muitos modelos estatísticos conhecidos pertencem à classe dos modelos latentes Gaussianos, como por exemplo modelos espaço-temporais, modelo de regressão semi-paramétrica, modelos espaciais, modelos geoestatísticos (ver mais em Rue *et al.* (2009)). Nestes modelos, o campo latente serve como uma ferramenta para modelar efeitos de covariáveis, heterogeneidade específica do grupo, bem como a dependência espacial e temporal entre os dados.

No INLA a variável observada y_i é assumida pertencer à uma família de distribuições onde a média μ_i é ligada a um preditor aditivo estruturado η_i através de uma função de ligação $g(\cdot)$, tal que $g(\mu_i) = \eta_i$ com

$$\eta_i = \alpha + \sum_{j}^{n_f} f^{(j)}(u_{ji}) + \sum_{k}^{n_\beta} \beta_k z_{ki} + \epsilon_i \tag{2.10}$$

Os parâmetros β_k's são os efeitos lineares das covariáveis $\boldsymbol{z}$. O termo $f^{(j)}(\cdot)$ são funções desconhecidas das covariáveis $\boldsymbol{u}$ que representam, por exemplo, efeitos aleatórios espaciais, efeitos aleatórios específicos do grupo, efeitos sazonais, tendências temporais, entre outros.

O parâmetro ϵ_i são efeitos aleatórios não estruturados.

Um modelo latente Gaussiano é obtido quando atribuímos ao vetor paramétrico $\boldsymbol{\theta} = \{\eta_i, \alpha, f^{(j)}(\cdot), \beta_k\}$ uma distribuição a priori Gaussiana. A parametrização do campo latente Gaussiano é feita de forma que inclua os preditores η_i's em vez dos ϵ_i's. Assim, alguns elementos de $\boldsymbol{\theta}$ são observados através dos dados $\boldsymbol{y}$.

O modelo latente pode ser escrito em uma estrutura hierárquica. No primeiro estágio os dados são assumidos condicionalmente independentes dado o campo latente Gaussiano $\boldsymbol{\theta}$ e os hiperparâmetros γ_1, quando estes existirem.

Primeiro estágio: $\boldsymbol{y}|\boldsymbol{\theta}, \gamma_1 \sim p(\boldsymbol{y}|\boldsymbol{\theta}, \gamma_1) = \prod_{i=1}^{n} p(y_i|\theta_i, \gamma_1)$

O segundo estágio do modelo é formado pela distribuição condicional do campo latente Gaussiano $\boldsymbol{\theta}$ dado os hiperparâmetros γ_2 de dimensão m, e tem a distribuição normal multivariada com vetor de médias $\boldsymbol{\theta}$ e matriz de precisão $\boldsymbol{Q}$ não singular.

Segundo estágio: $\boldsymbol{\theta}|\gamma_2 \sim p(\boldsymbol{\theta}|\gamma_2) = N(\boldsymbol{\mu}, \boldsymbol{Q})$

Após atribuir uma distribuição hiperpriori para os hiperparâmetros $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, i.e, $\boldsymbol{\gamma} \sim p(\boldsymbol{\gamma})$, obtemos a seguinte distribuição a posteriori

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y}) \propto p(\boldsymbol{\gamma}) p(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{i=1}^{n} p(y_i|\theta_i, \boldsymbol{\gamma}) \tag{2.11}$$

O objetivo no método INLA é obter as distribuições marginais $p(\theta_i|\boldsymbol{y})$, para $i = 1, \cdots, n$, e $p(\gamma_j|\boldsymbol{y})$, para $j = 1, \cdots, m$ $(m \leq 6)$. Temos

$$p(\theta_i|\boldsymbol{y}) = \int \underbrace{p(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y}) p(\boldsymbol{\gamma}|\boldsymbol{y})}_{p(\theta_i, \boldsymbol{\gamma}|\boldsymbol{y})} d\boldsymbol{\gamma}, \tag{2.12}$$

$$p(\gamma_j|\boldsymbol{y}) = \int p(\boldsymbol{\gamma}|\boldsymbol{y}) d\boldsymbol{\gamma_{-j}}. \tag{2.13}$$

As quantidades acima são difíceis de serem obtidas analiticamente, por isso as aproximações determinísticas $\tilde{p}(\theta_i|\boldsymbol{y})$ e $\tilde{p}(\gamma_j|\boldsymbol{y})$ são usadas:

$$\tilde{p}(\theta_i|\boldsymbol{y}) \approx \sum_k \tilde{p}(\theta_i|\boldsymbol{\gamma}_k, \boldsymbol{y})\tilde{p}(\boldsymbol{\gamma}_k|\boldsymbol{y})\Delta_k, \tag{2.14}$$

$$\tilde{p}(\gamma_j|\boldsymbol{y}) \approx \sum_k \tilde{p}(\boldsymbol{\gamma}_k|\boldsymbol{y})\Delta_{jk}. \tag{2.15}$$

Entretanto, para resolver as equações (2.14) e (2.15) é preciso obter as aproximações $\tilde{p}(\boldsymbol{\gamma}|\boldsymbol{y})$ e $\tilde{p}(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$, e avaliar $\tilde{p}(\boldsymbol{\gamma}|\boldsymbol{y})$ em uma grade construída a partir da informação de onde se concentra a maior parte da massa de probabilidade da distribuição conjunta dos parâmetros. Esta grade é importante quando se trabalha com aproximações determinísticas, e em Rue *et al.* (2009) encontra-se de forma detalhada como a grade é construída no INLA.

Sabendo que a distribuição conjunta a posteriori é

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y}) = p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})p(\boldsymbol{\gamma}|\boldsymbol{y}) \tag{2.16}$$

podemos reescrever (2.16) como:

$$p(\boldsymbol{\gamma}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})} \tag{2.17}$$

A aproximação usada para a distribuição conjunta a posteriori dos hiperparâmetros $p(\boldsymbol{\gamma}|\boldsymbol{y})$ é dada por

$$\tilde{p}(\boldsymbol{\gamma}|\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{y})}{\tilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})}\right|_{\theta=\theta^*(\gamma)} \tag{2.18}$$

em que o denominador é uma aproximação Gaussiana para a distribuição condicional completa de $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})$, e $\theta^*(\gamma)$ é a moda da distribuição condicional completa de $\boldsymbol{\theta}$ para um dado γ.

Para a distribuição marginal $p(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$ Rue *et al.* (2009) propõem três tipos de aproximação determinística que variam em termos de rapidez e precisão.

- Aproximação Gaussiana: $\tilde{p}_G(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$.
- Aproximação de Laplace: $\tilde{p}_{LA}(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$.
- Aproximação Simplificada de Laplace: $\tilde{p}_{SLA}(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$.

A aproximação Gaussiana computada no denominador de (2.18) é usada para obter a distribuição marginal. A aproximação de Laplace é encontrada de forma similar à (2.18)

$$\tilde{p}_{LA}(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{y})}{\tilde{p}_G(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\gamma}, \boldsymbol{y})}\Bigg|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\gamma})} \tag{2.19}$$

onde $\boldsymbol{\theta}^*_{-i}(\theta_i, \boldsymbol{\gamma})$ é a moda da distribuição condicional $\pi(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\gamma}, \boldsymbol{y})$, e $\tilde{p}_G(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\gamma}, \boldsymbol{y})$ é a aproximação Gaussiana de $p(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\gamma}, \boldsymbol{y})$.

A aproximação simplicada de Laplace é obtida através da expansão de Taylor no numerador e denominador em (2.19) até o termo de terceira ordem. Uma descrição mais detalhada a respeito das aproximações pode ser encontrada no trabalho de Rue *et al.* (2009).

Capítulo 3

Estudo de Simulação

Desenvolvemos um estudo de simulação com o propósito de realizar inferência Bayesiana através do método INLA para os dados de infarto agudo do miocárdio (IAM) visando usufruir as potencialidades que o método proporciona, tais como a redução de tempo computacional e o fato de não ser necessária a análise de convergência de cadeias, uma vez que o método é determinístico. O estudo objetivou validar uma nova formulação para o preditor linear de forma que o modelo componente compartilhado fosse executável no pacote INLA do programa R.

A modelagem do preditor linear por meio do modelo componente compartilhado definido em (2.6) não pode ser implementada no INLA porque o modelo especificado no pacote INLA admite a estimação de apenas um parâmetro desconhecido multiplicativo. Assim, propomos uma formulação alternativa para o preditor linear a fim de possibilitar o uso do método INLA através do pacote INLA no R. Além disso, no segundo estágio hierárquico especificamos uma distribuição a priori Gaussiana para cada um dos parâmetros desconhecidos obtendo então um modelo latente Gaussiano, tornando o método INLA aplicável como apresentado na seção 2.3.2.

No estudo de simulação dois cenários foram considerados: um com dados simulados e outro com dados reais. Em ambos cenários a modelagem hierárquica Bayesiana foi realizada, sendo a distribuição a posteriori obtida através dos métodos de aproximação estocástica e determinística. Utilizamos o software OpenBUGS para o método de aproximação estocástica, pois é um programa para análise Bayesiana que usa as técnicas de Monte Carlo via cadeias

de Markov (MCMC). Para o método de aproximação determinística usamos o pacote INLA disponível no programa R, denominaremos de R-INLA, que pode ser baixado no website www.r-inla.org. O programa R é uma linguagem e um ambiente computacional para desenvolvimento de técnicas e análises estatísticas. A seguir, a formulação original e a formulação proposta são apresentadas.

3.1 Formulação para $\boldsymbol{\eta}$

O preditor linear η_i é ligado ao risco relativo θ_i através da função logarítmica, i.e., $\eta_i = \log(\theta_i)$. Definimos para o preditor linear o modelo componente compartilhado que analisa conjuntamente duas medições. A interpretação dos parâmetros do modelo componente compartilhado foi relatada na subseção 2.2.1.1. A formulação original para o preditor linear é dada por

$$\begin{aligned} \eta_{i1} &= \alpha_1 + \phi_i \cdot \delta + \psi_{i1} \\ \eta_{i2} &= \alpha_2 + \phi_i / \delta + \psi_{i2} \end{aligned} \tag{3.1}$$

A formulação proposta neste trabalho, visando a implementação no INLA, é a seguinte

$$\begin{aligned} \eta_{i1} &= \alpha_1 + \phi_i \cdot \beta + \psi_{i1} \\ \eta_{i2} &= \alpha_2 + \phi_i + \psi_{i2} \end{aligned} \tag{3.2}$$

Avaliamos através deste estudo de simulação a equivalência entre as duas formulações, pois supomos que a formulação original para o preditor linear pode ser reescrita da forma apresentada na formulação proposta, sem perder as características e interpretação que o modelo componente compartilhado fornece. Com a nova formulação ainda é possível separar as estimativas dos componentes compartilhado e específicos para a superfície de risco em mapas individuais, de forma a expressar os padrões do risco comum entre as medições e específicos de cada uma.

As características dos componentes continuam mantidas, ou seja, o componente compartilhado captura os fatores de risco compartilhados pelas medições, e os componentes

específicos capturam os fatores de risco específicos de cada medição. Além disso, a contribuição do componente compartilhado sobre o risco relativo global é, agora, ponderado pelo parâmetro β que, assim como δ, permite que riscos gradientes diferentes estejam associados a este componente para cada medição. Os componentes, da mesma forma que na formulação original, são assumidos independentes entre si, o que possibilita usar as distribuições marginais que o INLA retorna.

Um ponto importante que devemos ressaltar é que o modelo ajustado no INLA corresponde ao da formulação proposta em (3.2). Não é possível implementar a formulação original, como dito inicialmente. Por isso, para compararmos as estimativas dos parâmetros δ e β realizamos a seguinte transformação:

$$\frac{\phi_i \cdot \delta}{\phi_i/\delta} = \frac{\phi_i \cdot \beta}{\phi_i} \quad \Rightarrow \quad \delta^2 = \beta \quad \Rightarrow \quad \delta = \sqrt{\beta} \tag{3.3}$$

3.2 Simulação

Nosso objetivo com este estudo é verificar a equivalência entre a formulação original (3.1) e a formulação alternativa (3.2) que estamos propondo. As ferramentas computacionais usadas foram o software `OpenBUGS` e o pacote INLA do programa `R` (R-INLA).

Os dados reais considerados no estudo referem-se à incidência de câncer da cavidade oral e câncer de pulmão em 126 zonas eleitorais da cidade de Yorkshire, no oeste da Inglaterra. Estes dados foram retirados de um exemplo de aplicação do modelo componente compartilhado implementado no módulo `GeoBUGS` do software `OpenBUGS`. O módulo `GeoBUGS` foi desenvolvido por uma equipe do Departamento de Epidemiologia e Saúde Pública de Londres para o ajuste de modelos espaciais e para a produção e exportação de mapas (GeoBUGS , 2012).

A modelagem hierárquica foi empregada neste estudo de simulação. As distribuições a priori definidas para os parâmetros α_1, α_2, δ, ϕ_i, ψ_{i1} e ψ_{i2} da formulação original são igualmente válidas para os parâmetros da formulação proposta, sendo a distribuição a priori para o logaritmo do δ a mesma para o logaritmo do β. A parametrização da distribuição normal é a mesma usada no software `OpenBUGS`, Normal(μ, τ), sendo τ o inverso da variância

σ^2. As distribuições a priori especificadas no segundo estágio, e as distribuições hiperpriori e os valores para os hiperparâmetros especificados no terceiro estágio são apresentadas abaixo. São válidas tanto para o cenário com dados simulados quanto para o cenário com dados reais.

- $\alpha_d \sim \text{Normal}(0, 10^{-7})$
- $\log \delta \sim \text{Normal}(0, 5.9)$
- $\phi_i \sim \text{BYM} : \text{Normal}(0, \tau_{uns}) + Normal(u_j, \tau_{sp}/m_i)$
    - $\tau_{uns} \sim \text{Gama}(0.5, 0.0005)$
    - $\tau_{sp} \sim \text{Gama}(0.5, 0.0005)$
- $\psi_{id} \sim \text{BYM} : \text{Normal}(0, \tau_{unsd}) + \text{Normal}(u_j, \tau_{spd}/m_i)$
    - $\tau_{unsd} \sim \text{Gama}(0.5, 0.0005)$
    - $\tau_{spd} \sim \text{Gama}(0.5, 0.0005)$

Ao atribuírmos aos parâmetros α_d, ϕ_i, ψ_{id} e δ, ou β, a distribuição normal, a distribuição conjunta destes parâmetros é gaussiana, consequentemente teremos um modelo latente Gaussiano que pode ser usado no INLA. Com estas distribuições é possível realizar a análise tanto no R-INLA quanto no `OpenBUGS`. Especificamos uma matriz de vizinhança binária baseada na informação de fronteira das zonas eleitorais para a parte estruturada espacialmente da priori de convolução BYM atribuída aos componentes compartilhado e específicos. A seguir descrevemos os dois cenários considerados e os dois estudos de simulação realizados em cada um, e também os resultados obtidos.

3.2.1 Cenário 1

Neste primeiro cenário realizamos dois estudos usando dados simulados. Em ambos geramos os dados aleatoriamente para duas medições, $d = 1, 2$, a partir da distribuição Poisson com média μ_{id} igual ao produto entre o valor esperado E_{id} e o risco relativo, para $i = 1, \cdots, 126$. O risco relativo é ligado a um preditor linear η_{id}, $\theta_{id} = e^{\eta_{id}}$, que é definido pelo modelo componente compartilhado, sendo a formulação original (3.1) definida para o

primeiro estudo de simulação, cujo objetivo é verificar a equivalência entre as duas formulações, e a formulação alterada (3.2) definida para o segundo estudo, cujo objetivo é comparar as estimativas obtidas pelos métodos de aproximação. O valor esperado para cada medição E_{id} nas 126 áreas foram gerados aleatoriamente da distribuição Uniforme com parâmetros de mínimo e máximo diferentes, de modo a distinguir o risco relativo global entre as medições. As etapas empregadas na geração dos dados simulados são apresentadas abaixo:

① Especificamos valores para os parâmetros: α_1, α_2 e δ (mesmo para β);

② Especificamos valores para os parâmetros de precisão: τ_{uns}, τ_{sp}, τ_{unsd} e τ_{spd}; $\{d = 1, 2\}$

③ Geramos a estrutura de vizinhança para as áreas;

④ Geramos aleatoriamente ϕ_i através do modelo BYM: Normal$(0, \tau_{uns})$+NMV$(\mathbf{0}, \tau_{sp}\boldsymbol{H})$;

⑤ Geramos aleatoriamente ψ_{id} através do modelo BYM: Normal$(0, \tau_{unsd})$+NMV$(\mathbf{0}, \tau_{spd}\boldsymbol{H})$;

⑥ Geramos o risco relativo: $e^{\eta_{id}}$, especificando a formulação original ou alterada para η_{id}, dependendo do estudo;

⑦ Geramos aleatoriamente o valor esperado para cada medição E_{id};

⑧ Geramos os dados: $y_{id} \sim$ Poisson$(E_{id} \cdot e^{\eta_{id}})$.

Sendo $\boldsymbol{H}$ igual a $[\text{diag}(\boldsymbol{m}) - \rho\boldsymbol{A}]$ que multiplicado com o parâmetro de precisão formam a matriz de precisão $\boldsymbol{Q}$ mostrada na subseção 2.2.1.2, e NMV é a distribuição normal multivariada. Os dados gerados a partir da formulação original foram usados no primeiro estudo de simulação, e os gerados a partir da formulação alterada foram usados no segundo estudo. Os valores para os parâmetros do modelo componente compartilhado e os parâmetros de precisão escolhidos são mostrados na Tabela 3.1. O valor de β é igual ao valor de δ.

Tabela 3.1: *Valores iniciais para os parâmetros do modelo componente compartilhado e parâmetros de precisão usados para simular os dados y_{i1} e y_{i2}.*

α_1	α_2	δ, β	τ_{uns}	τ_{sp}	τ_{uns1}	τ_{sp1}	τ_{uns2}	τ_{sp2}
0.20	0.20	1.20	264	462	963	567	745	353

O passo seguinte nos dois estudos é realizar a análise Bayesiana através da modelagem hierárquica a fim de obter as estimativas para os parâmetros α_1, α_2, ϕ_i, ψ_{i1}, ψ_{i2} e δ, ou β, além das estimativas para o risco relativo. No primeiro estudo de simulação definimos no primeiro estágio do modelo a distribuição Poisson para os dados, cuja média é o produto entre o valor esperado gerado e o risco relativo. O modelo componente compartilhado foi especificado para o preditor linear, e para os seus parâmetros foram determinadas as distribuições mostradas no início desta seção de modo a definir o segundo e terceiro estágios do modelo. Quando a análise foi realizada no software `OpenBUGS` definimos a formulação (3.1) para o preditor linear. Consequentemente as estimativas para os parâmetros eram baseadas no método de aproximação estocástica MCMC. Se a análise fosse realizada no R-INLA a formulação (3.2) deve ser considerada. Assim, as estimativas seriam baseadas no método INLA. Neste primeiro estudo queremos verificar a equivalência entre as formulações original e alterada.

No segundo estudo de simulação os dados gerados baseiam-se na formulação alterada. A distribuição Poisson foi novamente definida para os dados no primeiro estágio do modelo, sendo sua média o produto entre o valor esperado gerado e o risco relativo. Consideramos a formulação alterada para o preditor na análise Bayesiana realizada tanto no software `OpenBUGS` quanto no R-INLA. As distribuições a priori, as distribuições hiperpriori e os valores para os hiperparâmetros usadas no segundo e terceiro estágios da modelagem hierárquica correspondem àquelas mostradas inicialmente. Nosso objetivo é comparar o método de aproximação estocástica MCMC e o método INLA através de suas estimativas.

O software `OpenBUGS` usado nos estudos de simulação dos dois cenários é baseado em simulações iterativas MCMC. A partir das amostras simuladas da distribuição a posteriori é possível estimar quantidades que resumam a distribuição, tais como média, mediana e quantis. Para a análise dos dados simulados, duas cadeias com 25.000 realizações cada uma foram geradas, sendo que destas descartamos (*burn-in*) as primeiras 5.000 realizações. A análise de convergência das cadeias foi feita através de métodos gráficos (não apresentaremos os gráficos aqui).

O resultado da combinação, via Teorema de Bayes, dos dados simulados para cada medição e das distribuições a priori para os parâmetros é a distribuição a posteriori. As estimativas Bayesianas pontuais e os intervalos de credibilidade a posteriori para os parâmetros obtidos

pelo método de aproximação estocástica MCMC e pelo método INLA para o primeiro estudo de simulação são apresentadas na Tabela 3.2, assim como o verdadeiro valor para os interceptos de cada medição e o parâmetro de ponderação δ. Analisando os resultados da tabela constatamos que os parâmetros foram recuperados de forma satisfatória pelos dois métodos, e que o desvio padrão dos parâmetros foi bastante parecido. Além disso, em ambos os métodos os intervalos de credibilidade contém o verdadeiro valor e a magnitude dos intervalos foi praticamente a mesma.

Gráficos de dispersão foram gerados para os componentes compartilhado ϕ_i e específicos ψ_{i1} e ψ_{i2}, e para os riscos relativos θ_{i1} e θ_{i2}, para as medições 1 e 2. O objetivo era verificar a relação entre o verdadeiro valor e a estimativa Bayesiana obtida pelo método de aproximação estocástica MCMC em cada área quando a formulação original foi definida para o preditor linear, entre o verdadeiro valor e a estimativa Bayesiana obtida pelo método INLA em cada área quando a formulação alterada foi definida para o preditor, e por último a relação entre as estimativas obtidas pelos dois métodos. Os resultados gráficos para os componentes e para os riscos encontram-se no Apêndice B.

Visualmente percebemos que a dispersão dos pontos nos gráficos do verdadeiro valor versus o MCMC e do verdadeiro valor versus o INLA são similares, principalmente para os parâmetros ϕ_i, θ_{i1} e θ_{i2} em que as estimativas resultantes dos dois métodos são altamente correlacionadas. Com exceção do componente ψ_{i1}, os métodos recuperaram bem o verdadeiro valor em cada área. O gráfico gerado para verificar a relação entre as estimativas obtidas via MCMC e INLA, último gráfico em todas as figuras, mostra que elas são correlacionadas, indicando consistência na equivalência entre as formulações original (3.1) e alterada (3.2). Destacamos a alta correlação existente entre as estimativas para o risco constituído por todos os parâmetros do modelo componente compartilhado. Com base nos resultados numéricos e gráficos concluímos com este primeiro estudo de simulação que as formulações são equivalentes.

A Tabela 3.3 exibe os resultados para o segundo estudo de simulação no qual os dados simulados foram ajustados considerando a formulação alterada. Os verdadeiros valores correpondem aos valores iniciais, e as estimativas pontuais e intervalares a posteriori para os parâmetros α_1, α_2 e β, correspondem às obtidas pelo método de aproximação estocástica

Tabela 3.2: *Verdadeiro valor dos parâmetros e suas estimativas pontuais e intervalares a posteriori obtidas via MCMC e INLA, considerando a formulação original e alterada respectivamente (primeiro estudo de simulação).*

Verdadeiro		MCMC					INLA				
		Média	Desvio	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$	Média	Desvio	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$
α_1	0.2	0.025	0.039	-0.052	0.025	0.099	0.019	0.040	-0.061	0.019	0.097
α_2	0.2	0.026	0.014	-0.001	0.026	0.053	0.026	0.013	-0.0003	0.026	0.053
δ, β	1.2	1.127	0.184	0.750	1.129	1.473	1.244	-	0.935	1.240	1.506

MCMC e pelo método INLA. Ao analisarmos a tabela verificamos que α_1 foi recuperado de forma satisfatória pelos dois métodos em relação aos outros dois parâmetros, porém todos os intervalos de credibilidade contém os valores verdadeiros. Quando comparamos as estimativas constatamos que os métodos fornecem valores próximos para os parâmetros e para os desvios padrão também. Além disso, o intervalo de credibilidade de um método contém a estimativa obtido no outro método para o mesmo parâmetro.

A análise gráfica também foi realizada através dos gráficos de dispersão para os componentes compartilhado ϕ_i e específicos ψ_{i1} e ψ_{i2}, e para os riscos relativos θ_{i1} e θ_{i2}, das medições 1 e 2. Investigamos a relação entre os verdadeiros valores e as estimativas obtidas pelo método de aproximação estocástica MCMC, entre os verdadeiros valores e as estimativas obtidas pelo método INLA, e por último entre as estimativas fornecidas pelos dois métodos. Os gráficos encontram-se no Apêndice B. A dispersão dos pontos nos gráficos verdadeiro versus MCMC e verdadeiro versus INLA é mais similar para os parâmetros ϕ_i, θ_{i1} e θ_{i2}, o quer dizer que os métodos recuperaram os valores verdadeiros de forma similar. O componente específico ψ_{i1} não foi recuperado bem por nenhum dos métodos. As estimativas obtidas pelos métodos são bastante similares para todos os parâmetros, como verificado no último gráfico em cada figura, gráfico MCMC versus INLA, no qual é possível perceber a alta correlação entre elas. Com os resultados numéricos e gráficos concluímos que qualquer um dos métodos pode ser usado com a formulação alterada (3.2).

3.2.2 Cenário 2

A simulação realizada neste cenário envolve os dados reais relacionados ao número de casos de câncer da cavidade oral e câncer de pulmão nas 126 zonas eleitorais da cidade de

Tabela 3.3: *Verdadeiro valor dos parâmetros e suas estimativas pontuais e intervalares a posteriori obtidas via MCMC e INLA, considerando a formulação alterada (segundo estudo de simulação).*

Verdadeiro		MCMC					INLA				
		Média	Desvio	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$	Média	Desvio	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$
α_1	0.2	0.026	0.038	-0.049	0.026	0.100	0.021	0.039	-0.057	0.021	0.098
α_2	0.2	0.010	0.014	-0.019	0.010	0.037	0.010	0.014	-0.018	0.010	0.038
β	1.2	1.147	0.296	0.634	1.13	1.792	1.338	0.306	0.752	1.331	1.955

Yorkshire, disponíveis no exemplo de aplicação do modelo componente compartilhado no módulo `GeoBUGS`. Além dos dados, os valores esperados padronizados por sexo e idade para cada doença em cada área também estão disponibilizados. Neste cenário desenvolvemos dois estudos de simulação cujos objetivos eram verificar a equivalência entre a formulação original e a alterada (primeiro estudo de simulação) e comparar o método de aproximação estocástica MCMC e o método INLA através de suas estimativas (segundo estudo de simulação).

Inicialmente, a análise Bayesiana dos dados reais foi realizada usando a técnica MCMC por meio do software `OpenBUGS` com o propósito de obter as estimativas para os parâmetros do modelo componente compartilhado definido para o preditor linear, pois estas foram usadas posteriormente para a geração de novos dados. Seguindo a modelagem hierárquica, no primeiro estágio assumimos que o modelo gerador dos dados é o modelo Poisson com média dada pelo produto entre o valor esperado do exemplo e o risco relativo, representado por $e^{\eta_{id}}$, para $i = 1, \cdots, 126$. O modelo componente compartilhado foi usado para modelar o preditor linear η_{id}, sendo a formulação original definida para o primeiro estudo de simulação e a formulação alterada definida para o segundo. As distribuições assumidas no segundo e terceiro estágios correspondem àquelas mostradas no início desta seção. Geramos uma cadeia com 10.000 realizações, sendo descartadas as primeiras 4.000 realizações. Não avaliamos a convergência da cadeia pois o objetivo desta análise era apenas obter as estimativas para α_1, α_2, ϕ_i, ψ_{i1}, ψ_{i2} e δ ou β, dependendo da formulação, para que fossem usados como valores iniciais na geração de novos dados para duas doenças. As estimativas resultantes da análise dos dados do exemplo estão na Tabela 3.4 e na Tabela 3.5.

Usamos as estimativas pontuais da média a posteriori para os parâmetros α_1, α_2 e δ, ou β, enquanto as estimativas para os componentes compartilhado, ϕ_i, e específicos, ψ_{i1} e ψ_{i2}, correspondem aos últimos valores gerados pela cadeia para cada área. Após determinarmos

Tabela 3.4: *Valores usados para gerar os dados do primeiro estudo de simulação, obtidos da análise Bayesiana para os dados reais via MCMC considerando a formulação original.*

α_1	α_2	δ
-0.009	-0.023	0.892

Tabela 3.5: *Valores usados para gerar os dados do segundo estudo de simulação, obtidos da análise Bayesiana para os dados reais via MCMC considerando a formulação alterada.*

α_1	α_2	β
-0.012	-0.023	0.923

os valores para os parâmetros do primeiro estudo de simulação (Tabela 3.4) e do segundo estudo (Tabela 3.5), seguimos as etapas descritas abaixo para gerar novos dados para duas doenças.

① Especificamos os valores para α_1, α_2, ϕ_i, ψ_{i1}, ψ_{i2} e δ, ou β;

② Geramos o risco relativo: $e^{\eta_{id}}$, especificando a formulação original ou alterada para η_{id};

③ Usamos os valores esperados do exemplo E_{id};

④ Geramos os dados: $y_{id} \sim \text{Poisson}(E_{id} \cdot e^{\eta_{id}})$.

Os dados gerados a partir da formulação original para o preditor η_{id} foram usados no primeiro estudo de simulação. Quando gerados a partir da formulação alterada foram usados no segundo estudo. O passo seguinte nos dois estudos é realizar a análise Bayesiana e comparar as estimativas dos parâmetros do modelo componente compartilhado e do risco relativo em cada área com seus verdadeiros valores, os valores ditos iniciais. Realizamos a modelagem hierárquica cujas distribuições para o segundo e terceiro estágios foram especificadas no início desta seção.

No primeiro estudo de simulação o modelo Poisson foi definido para os dados no primeiro estágio, cuja média é o produto entre o valor esperado do exemplo e o risco relativo. O modelo componente compartilhado foi especificado para o preditor linear η_{id}. Se a análise fosse realizada no software `OpenBUGS` a formulação original era definida para o preditor, logo as estimativas obtidas para os parâmetros eram baseadas no método de aproximação estocástica MCMC. Se fosse realizada no R-INLA a formulação alterada era considerada,

assim as estimativas eram baseadas no método INLA. As distribuições usadas no segundo e terceiro estágios da modelagem são as mesmas definidas anteriormente. Nosso objetivo com este estudo foi verificar a equivalência entre as duas formulações.

No segundo estudo de simulação o modelo Poisson foi especificado aos dados novamente no primeiro estágio, sendo a média o produto entre o valor esperado do exemplo e o risco relativo. O modelo componente compartilhado com a formulação alterada foi definida para o preditor η_{id} na análise Bayesiana realizada tanto no software `OpenBUGS` quanto no R-INLA. As distribuições a priori, as distribuições hiperpriori e os valores para os hiperparâmetros usadas no segundo e terceiro estágios da modelagem são aquelas definidas inicialmente. Neste segundo estudo o objetivo é comparar os métodos através das suas estimativas.

Os resultados da análise Bayesiana para o primeiro estudo de simulação neste cenário são mostrados na Tabela 3.6. As estimativas pontuais a posteriori obtidas pelos dois métodos não se aproximaram do verdadeiro valor dos parâmetros, porém todos os intervalos de credibilidade contém o valor verdadeiro. A magnitude dos intervalos foi praticamente a mesma nos dois métodos, exceto pelo parâmetro δ no qual a amplitude foi maior para a técnica MCMC. A análise dos componentes compartilhado ϕ_i e específicos ψ_{i1} e ψ_{i2}, e dos riscos relativos θ_{i1} e θ_{i2} foi baseada nos gráficos de dispersão apresentados no Apêndice B. Verificamos a relação entre os valores verdadeiros e as estimativas obtidas pelo método de aproximação estocástica MCMC quando a formulação original foi considerada para o preditor η_{id}, entre os valores verdadeiros e as estimativas obtidas pelo método INLA quando a formulação alterada foi considerada para η_{id}, e entre as estimativas obtidas pelos dois métodos.

Visualmente o método de aproximação estocástica em comparação ao método INLA recuperou o parâmetro ϕ_i de forma mais satisfatória (Figura B.11) pois as estimativas são mais correlacionadas com os valores verdadeiros. O desempenho dos métodos na recuperação dos parâmetros restantes foi similar, tendo alcançado resultados melhores para os riscos. Analisando as estimativas obtidas pelos métodos, gráfico MCMC versus INLA para todos os parâmetros, constatamos que elas são sempre correlacionadas positivamente, indicando concordância entre os métodos e validando a equivalência entre as formulações original e alterada.

Tabela 3.6: *Valor verdadeiro para os parâmetros e suas estimativas pontuais e intervalares obtidas via MCMC e INLA, considerando a formulação original e alterada respectivamente (primeiro estudo de simulação).*

Verdadeiro		MCMC					INLA				
		Média	Desvio	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$	Média	Desvio	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$
α_1	-0.009	-0.035	0.038	-0.110	-0.034	0.039	-0.041	0.039	-0.117	-0.040	0.034
α_2	-0.023	-0.045	0.011	-0.067	-0.045	-0.023	-0.045	0.010	-0.065	-0.045	-0.025
δ, β	0.892	0.947	0.163	0.672	0.933	1.319	1.159	-	0.958	1.156	1.344

O último estudo de simulação foi avaliado a partir dos resultados numéricos da Tabela 3.7 e dos gráficos finais do Apêndice B. Analisando a tabela notamos que as estimativas pontuais não foram satisfatórias quando comparadas aos valores verdadeiros, mas novamente os valores verdadeiros estão contidos no intervalos de credibilidade. Mais uma vez os métodos apresentam estimativas pontuais próximas para α_1 e α_2, e diferentes para β, porém as estimativas pontuais de um método está contido no intervalo do outro para o mesmo parâmetro indicando conformidade entre eles. Avaliando graficamente os componentes compartilhado ϕ_i e específicos ψ_{i1} e ψ_{i2}, e também os riscos relativos, destacamos a forte correlação existente entre as estimativas obtidas pelos métodos, gráfico MCMC versus INLA. Apesar do método INLA apresentar pontos mais dispersos, o seu desempenho e do outro método na recuperação dos verdadeiros valores foi satisfatório, com exceção do componente específico ψ_{i1}. A forte associação verificada entre as estimativas do método de aproximação estocástica MCMC e as do método INLA indica que qualquer método poderá ser usado com a formulação alterada (3.2).

Tabela 3.7: *Valor verdadeiro para os parâmetros e as estimativas pontuais e intervalares obtidas via MCMC e INLA, considerando a formulação alterada (segundo estudo de simulação).*

Verdadeiro		MCMC					INLA				
		Média	Desvio	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$	Média	Desvio	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$
α_1	-0.012	-0.059	0.037	-0.134	-0.059	0.014	-0.063	0.038	-0.140	-0.063	0.012
α_2	-0.023	-0.032	0.010	-0.052	-0.032	-0.013	-0.033	0.010	-0.053	-0.033	-0.013
β	0.923	0.908	0.251	0.528	0.872	1.498	1.185	0.242	0.721	1.179	1.676

Analisando os resultados para os dois cenários algumas diferenças são identificadas. Os dois estudos de simulação que usaram dados simulados apresentaram um desempenho melhor do que aqueles que usaram dados reais. Isto acontece por se tratar de um ambiente controlado

em que o problema real é reproduzido de forma planejada. Já os dados reais não podem ser descritos de forma exata pois não conhecemos o verdadeiro processo que os gerou, por isso suposições são feitas e expressas através de uma distribuição de probabilidade.

Comparando as estimativas obtidas pelos métodos de aproximação estocástica MCMC e o método INLA verificamos que seus valores foram sempre similares, assegurando a correspondência entre os métodos. Os resultados para o primeiro estudo de simulação nos dois cenários validaram a equivalência entre as formulações original e alterada. Portanto podemos escrever o modelo componente compartilhado a partir da formulação alterada, e realizar a análise Bayesiana através do método INLA, visto que o modelo pode ser implementado no R-INLA.

Capítulo 4

Análise dos Dados

O infarto agudo do miocárdio (IAM) é um evento agudo que necessita internação hospitalar. No Brasil sua ocorrência tem significativo impacto no número de hospitalizações, além de ser a principal causa de mortes entre homens (Min.Saúde, 2012b).

Os dados de infarto agudo do miocárdio foram analisados para cada microrregião brasileira. No total são 558 microrregiões não sobrepostas, distribuídas em todo território nacional. Em cada microrregião temos a informação sobre a população total, o número de internações por IAM, o número de mortes e de sobreviventes após a internação, todas referentes às pessoas assistidas pelo SUS no período de 2010.

Dentre o total de microrregiões, 28% apresentam população inferior a 100 mil habitantes, e em mais da metade das microrregiões a população está entre 100 e 500 mil habitantes (Tabela 4.1). Microrregiões com população acima de 500 mil habitantes representam menos de 10% do número total. A distribuição da população é assimétrica como mostra a Tabela 4.3 e a Figura 4.1. A média e a mediana não são próximas, e o valor do quantil 0.9 comparado à maior população são bastante distantes.

No ano de 2010 não foram registradas internações por IAM em nove microrregiões. O máximo de internações observadas foi de 5.145 nas 549 microrregiões restantes. Destas restantes, 503 registraram a ocorrência de pelo menos uma morte após a internação, e o número de mortes foi igual ao número de internações em apenas três microrregiões. A microrregião com a menor população registrou duas internações e nenhuma morte, enquanto a com maior população registrou 2.715 internações e 389 mortes.

Tabela 4.1: *População nas microrregiões no ano de* 2010 *estratificadas segundo as classes de tamanho da população.*

Classes de tamanho da população	Número de microrregiões	Proporção de microrregiões
Até 30.000 hab.	11	2%
De 30.001 a 50.000 hab.	21	4%
De 50.001 a 100.000 hab.	123	22%
De 100.001 a 500.000 hab.	355	64%
De 500.001 a 1.000.000 hab.	34	6%
Mais de 1.000.000 hab.	14	2%
Total	558	100%

Em 55 microrregiões nenhuma morte por IAM foi registrada, embora o maior número de internações observado tenha sido 28 em uma delas. O número de internações e de mortes é maior na maioria das microrregiões que englobam as capitais quando comparadas com as outras microrregiões, o que é esperado uma vez que os municípios em torno das capitais tendem as ser grandes também. Entretanto observamos algumas exceções, como mostra a Tabela 4.2, em que as microrregiões que não contemplam as capitais apresentaram o maior número de internação (três primeiras linhas) e o maior número de mortes (linhas restantes) no Estado as quais pertencem. No Estado da Paraíba a microrregião com maior número de internações e mortes foi Campina Grande que não contempla a capital João Pessoa. Verificamos também que a microrregião do Estado que registrou muitas internações não necessariamente é a mesma que registrou o maior número de mortes. Um exemplo disto é o Estado do Maranhão apresentou duas microrregiões diferentes com valores altos para internação e morte. A Tabela C.1, no Apêndice C, apresenta as capitais e as microrregiões às quais pertencem, e também o número de mortes e de internações, bem como a população destas microrregiões.

Os dados referentes ao número de internações e de mortes, assim como à população, são assimétricos, como verificado na Tabela 4.3 e pela Figura 4.1. Os valores da mediana são inferiores aos da média, indicando que as observações são assimétricas. Além disso, a média é próxima do quantil 0.75 indicando que a presença de valores atípicos também é outra característica encontrada nos dados de IAM. Representamos graficamente os dados através do histograma exibido na Figura 4.1. O gráfico mostra que a distribuição é assimétrica para todas as variáveis, e que valores discrepantes são observados.

Tabela 4.2: *Microrregiões com maior número de internações (três primeiras linhas) e maior número de mortes (linhas restantes) que não contemplam a capital do Estado referido.*

UF	Microrregião	Núm. Mortes	Núm. Internações	População
MA	Imperatriz	11	61	525.034
PI	Picos	10	164	192.871
PB	Campina Grande	33	161	426.248
PA	Bragantina	25	66	375.202
MA	Caxias	19	52	401.041
PB	Campina Grande	33	161	426.248
AL	Arapiraca	23	76	394.971
BA	Feira de Santana	60	310	889.853
SC	Tubarão	47	242	306.902

Tabela 4.3: *Resumo dos dados para as microrregiões através da média, mediana e quantis.*

	Mínimo	25%	Mediana	Média	75%	90%	Máximo
Número de Internações	0	14	36,5	91,37	85,75	179,10	5.145
Número de Mortes	0	2	6	14	14	27	871
Número de Sobreviventes	0	11	30	77,37	73	156,2	4.274
População	1.767	95.276	152.523	260.055	260.706	458.565	6.838.242

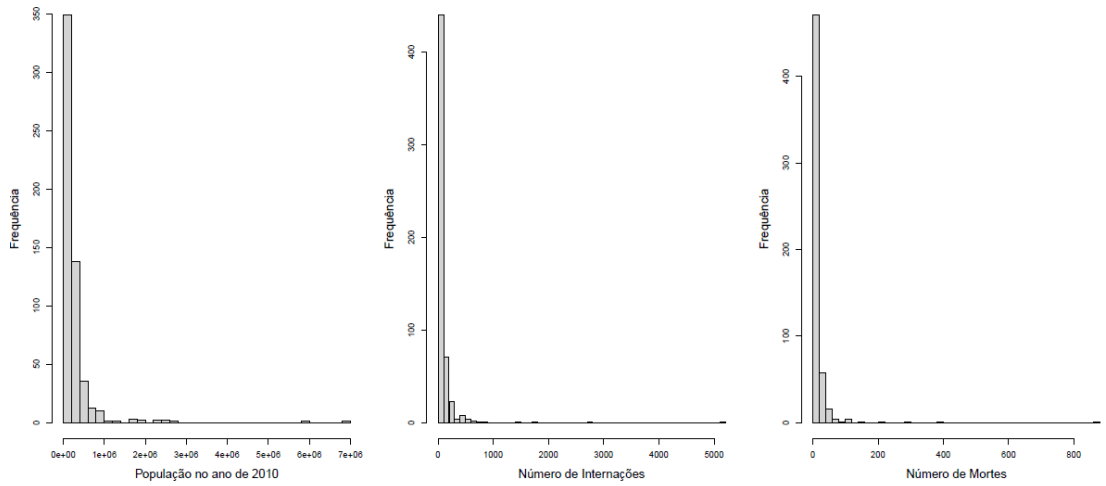

Figura 4.1: *Distribuição das variáveis: população, número de internações e número de mortes.*

4.1 Modelagem dos Dados

A metodologia Bayesiana foi empregada na análise dos dados de infarto agudo do miocárdio, e através da estrutura hierárquica especificados os modelos em cada um dos estágios. Supomos que, para cada medição, a distribuição Poisson representa o processo gerador dos dados dado o risco relativo desconhecido, θ_{id}, para $d = 1, 2$. A média do modelo Poisson é

o produto entre valor esperado e risco relativo, sendo o valor esperado E_{id} o produto entre o número de pessoas em risco e a taxa global. O risco relativo é ligado a um preditor linear, η_{id}, por meio da função logarítmica, $\log(\theta_{id}) = \eta_{id}$. Este preditor foi modelado através do modelo componente compartilhado (Knorr-Held e Best, 2001) expresso pela formulação alterada (equação 4.1). Para cada parâmetro que compõem este modelo atribuímos as distribuições a priori para o segundo estágio mostradas a seguir, e as distribuições hiperpriori para os parâmetros de precisão no terceiro estágio.

A formulação alterada para o preditor linear é dada por

$$\begin{aligned} \eta_{i1} &= \alpha_1 + \phi_i \cdot \beta + \psi_{i1} \\ \eta_{i2} &= \alpha_2 + \phi_i + \psi_{i2} \end{aligned} \tag{4.1}$$

- 1º estágio:

  $y_{id} \sim \text{Poisson}(E_{id} \cdot e^{\eta_{id}})$

- 2º estágio:

  $\alpha_d \sim \text{Normal}(0, 10^{-7})$

  $\phi_i \sim \text{BYM} : \text{Normal}(0, \tau_{uns}) + \text{Normal}(u_j, \tau_{sp}/m_i)$

  $\log \beta \sim \text{Normal}(0, 5.9)$

  $\psi_{id} \sim \text{BYM} : \text{Normal}(0, \tau_{unsd}) + \text{Normal}(u_j, \tau_{spd}/m_i)$

- 3º estágio:

  $\tau_{uns} \sim \text{Gama}(0.5, 0.0005)$

  $\tau_{sp} \sim \text{Gama}(0.5, 0.0005)$

  $\tau_{unsd} \sim \text{Gama}(0.5, 0.0005)$

  $\tau_{spd} \sim \text{Gama}(0.5, 0.0005)$

Ao atribuírmos aos parâmetros α_1, α_2, ϕ_i, ψ_{i1}, ψ_{i2} e β a distribuição normal, a distribuição conjunta destes parâmetros é gaussiana, dessa forma temos um modelo latente Gaussiano que corresponde ao modelo com o qual o INLA trabalha. A priori de convolução BYM definida

para os componentes compartilhado, ϕ_i, e específicos, ψ_{id}, permite incorporar, através do efeito aleatório estruturado espacialmente, informação da vizinhança das áreas fornecendo estimativas suavizadas para o risco relativo em cada área. Definimos a matriz de vizinhança como sendo uma matriz binária que considera a informação de fronteira entre as áreas. Ou seja, $a_{ij} = 1$ se a área j é vizinha da área i, $a_{ij} = 0$ caso contrário, e também nenhuma área é vizinha dela mesma, $a_{ii} = 0$ para todo i. A ilha de Fernando de Noronha foi ligada ao seu vizinho mais próximo com relação a distância.

4.2 Resultados

No período estudado não ocorreram internações por IAM em 9 microrregiões, logo, a taxa bruta de internação nestas microrregiões é zero. Lembrando que, a taxa bruta para qualquer doença é dada pela razão entre o número de eventos na área e o número de pessoas em risco na mesma área. A taxa bruta de letalidade, ou simplesmente letalidade bruta, para essas microrregiões é indeterminada, pois o numerador e o denominador são zero. Nos casos em que observou-se pelo menos uma internação (denominador diferente de zero) e nenhuma morte após a internação, a letalidade bruta seria zero. Com isso, tanto para a taxa bruta de internação quanto para a letalidade bruta seria errado inferir que os indivíduos nessas microrregiões tem risco de internação e de morte zero. Mesmo que a chance desses eventos ocorrerem seja pequena, ela não é exatamente igual a zero. A taxa bruta nestas situações não é um estimador adequado para a taxa de internação nem para a letalidade na microrregião, assim como o estimador SMR para o risco relativo. O ajuste das taxas através da metodologia Bayesiana considerando a configuração espacial das áreas ajudou a estimar melhor as taxas nos locais em que o valor observado foi zero ou indeterminado.

Com o propósito de usar o modelo componente compartilhado para modelar o preditor linear, definimos o número de mortes, y_{i1}, como a medição 1 e o número de sobreviventes, y_{i2}, como a medição 2. Após realizar a análise conjunta das duas medições através da modelagem hierárquica, as estimativas para o risco relativo de morte, θ_{i1}, e risco de sobreviver, θ_{i2}, assim como para as taxas de mortalidade, μ_{i1}, e de sobrevivência, μ_{i2}, foram obtidas. Além das estimativas para os componentes compartilhado e específicos que permitem investigar

tendências compartilhadas entre as medições e tendências específicas de cada medição.

A Figura 4.2 mostra os componentes específico da mortalidade ψ_{i1} e específico da sobrevivência ψ_{i2}. O componente da mortalidade tem um padrão espacial distinto em relação ao da sobrevivência com valores altos no nordeste, no Estado de São Paulo e em algumas microrregiões da região norte, e baixos em algumas microrregiões da região nordeste e centro-oeste. Isso indica a existência de fatores de risco adicional que são relevantes somente para a mortalidade. Talvez esse padrão esteja relacionado à dificuldade de acesso dos indivíduos ao local de atendimento hospitalar devido a extensão territorial no norte, ou à falta de conhecimento das pessoas em relação à doença, como por exemplo identificação dos sintomas; tratam-se de especulações de possíveis explicações para o padrão observado. Notamos que o padrão espacial no mapa da componente da mortalidade é mais claramente definido.

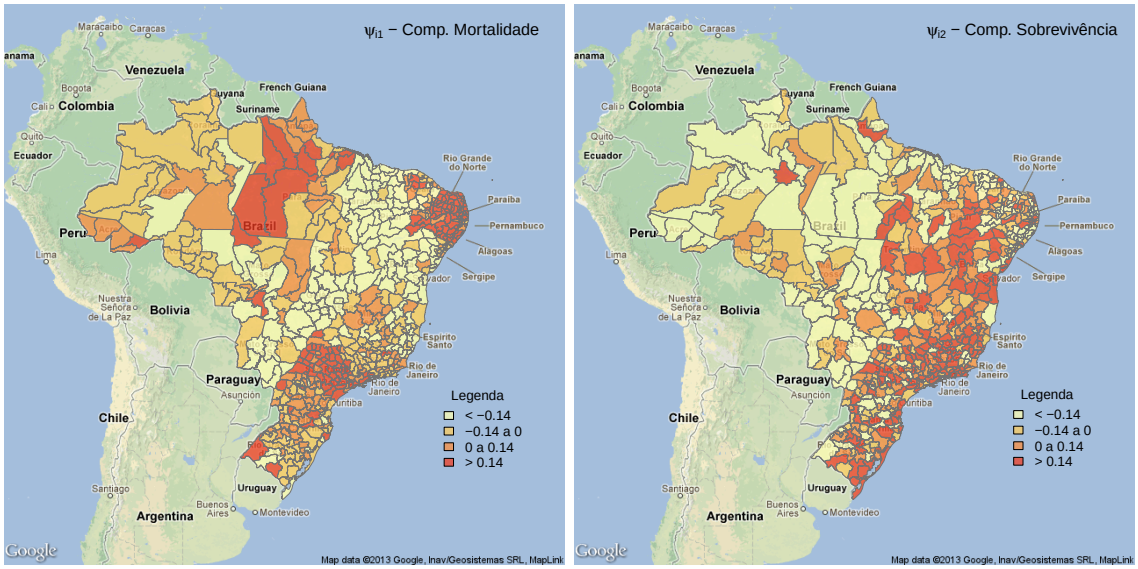

Figura 4.2: *Mapa com as estimativas para os componentes específicos relativo à mortalidade (esquerda) e à sobrevivência (direita).*

O resultado para o componente compartilhado é exibido na Figura 4.3, percebemos dois grandes clusters distintos, um na região norte do país com valores baixos e outro no sul e em grande parte da região sudeste com valores altos. Estes clusters podem estar relacionados com o estilo de vida e hábitos da população nesses locais que fornecem evidência de fatores de risco comum compartilhado entre a mortalidade e sobrevivência por IAM. A população no sul e sudeste podem estar mais expostos à fatores de risco nocivos como o sedentarismo, estresse e fumo, por exemplo.

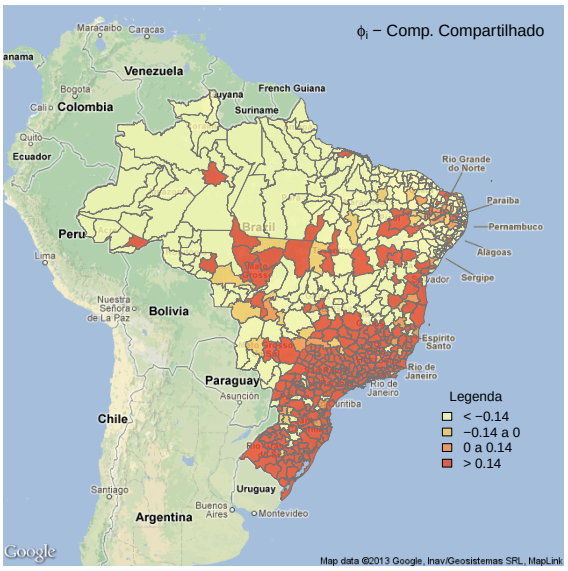

Figura 4.3: *Mapa com as estimativas para o componente compartilhado* ϕ_I.

Na Figura 4.4 as estimativas para o risco de morte e de sobrevivência em cada microrregião são exibidas. A estrutura espacial dos riscos são parecidas, com risco maior de morte no sul e sudeste, moderada no nordeste e baixa no norte, locais em que o risco de sobreviver por IAM também é alto. Além disso, notamos que microrregiões com maior risco de morte são também aquelas com maior risco do indivíduo sobreviver. A taxa de mortalidade e de sobrevivência por IAM apresentaram padrão espacial similares, com altos valores na faixa que vai da região sul até uma parte da região nordeste, e valores baixos e moderados para as microrregiões restantes.

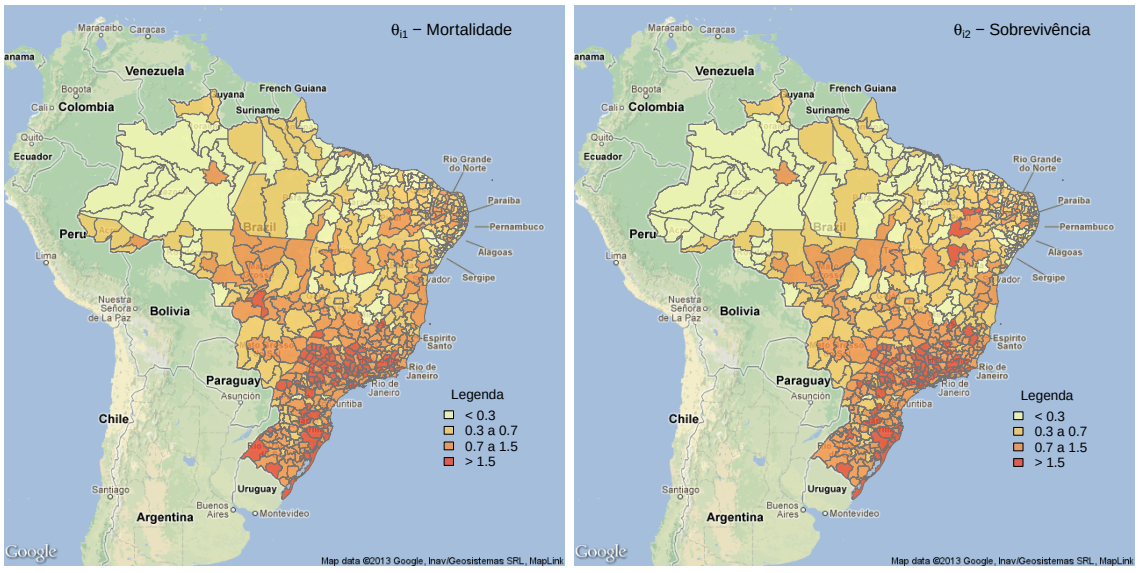

Figura 4.4: *Estimativas do risco de morte, à esquerda, e do risco de sobrevivência, à direita.*

A taxa bruta de internação, Figura 4.6 à esquerda, mostra a existência de um padrão espacial predominante da faixa que engloba a região sul até a sudeste com taxas elevadas. Na medida em que a faixa se desloca em direção ao norte observamos ainda algumas taxas elevadas mas que no extremo norte da região nordeste variam mais sem um padrão definido. No restante do mapa a taxa foi inferior a 37 na maioria das microrregiões. Algumas microrregiões no norte foram pintadas de branco porque a taxa bruta estimada foi zero. Com o ajuste Bayesiano das taxas, estas áreas puderam ser estimadas considerando a informação de seus vizinhos. Percebemos pela Figura 4.6 à direita que algumas microrregiões no noroeste e nordeste mudaram suas taxas principalmente aquelas cuja estimativa era zero. Nestas, as estimativas foram suavizadas tornando-se similares com as taxas das microrregiões que a cercam.

A letalidade bruta e a letalidade suavizada, estimada através da metodologia Bayesiana, encontram-se na Figura 4.7. Nitidamente percebemos no mapa da letalidade bruta a presença de algumas microrregiões com valores diferentes de seus vizinhos, o que dificulta a visualização das tendências espaciais. Além disso, valores extremos podem não estar relacionados diretamente ao risco da letalidade. Nas microrregiões pintadas de branco a letalidade estimada foi zero, e nas pintadas de verde ela é indeterminada (Figura 4.7 à esquerda). Ao gerar o mapa com as estimativas suavizadas para a letalidade notamos que o padrão espacial muda significativamente, a aparência do mapa torna-se mais suave. As áreas próximas agora apresentam estimativas mais similares tornando a correlacão espacial mais evidente. As estimativas nas áreas pintadas de branco e verde foram suavizadas e estimadas, respectivamente, com base na informação da letalidade de seus vizinhos. Além disso, a distribuição da letalidade suavizada nas microrregiões é simétrica como verificado pela Figura 4.5 através do histograma e do box-plot, e pelas medidas resumo apresentadas na Tabela 4.4.

Analisando a Tabela 4.4 notamos que 90% das microrregiões estão abaixo de 0.21, correspondem à 510 microrregiões sendo que destas a maioria pertence ao sudeste do Brasil. Mais da metade das microrregiões, mais especificamente 285, apresentaram letalidade acima da média nacional de 0.16. Dentre estas, 119 pertencem à região Nordeste e a região Centro-Oeste é a que registrou a menor frequência de ocorrências acima da média. A microrregião da região Nordeste que apresentou a maior letalidade foi Petrolina, 0.31, que também é a

microrregião com a maior letalidade entre as que estão acima da média nacional.

Poucas microrregiões tem letalidade abaixo de 0.1 ou acima de 0.22, mais especificamente 6 e 30 microrregiões respectivamente. A microrregião com a menor letalidade foi São Raimundo Nonato no Estado do Piauí. Não é a microrregião com a menor população do Estado, e também não apresentou os maiores valores para mortes e internações por IAM. Já a microrregião com maior letalidade, como visto acima, é Petrolina do Estado de Pernambuco, também não é a microrregião com a maior população do Estado e não possui o maior número de mortes e internações por IAM.

Tanto na parte central da região nordeste quanto em algumas microrregiões do sudeste as estimativas parecem ser mais homogêneas, e estão abaixo 0.16 da média nacional para a letalidade. Dentre as microrregiões que contemplam as capitais, duas da região nordeste e três da norte apresentaram letalidade acima da média nacional, assim como a microrregião que engloba São Paulo e Cuiabá. No Apêndice C apresentamos a letalidade suavizada para as microrregiões que contemplam as capitais brasileiras.

O gráfico da Figura 4.8 foi gerado a fim de investigar uma possível relação entre a taxa de internação e a letalidade. Percebemos pela dispersão dos pontos que as quantidades não possuem relação linear, a relação existente entre elas tem a forma de funil. Na medida em que a taxa de internação aumenta, aumenta a variabilidade da letalidade.

Tabela 4.4: *Medidas resumo da letalidade suavizada para as microrregiões brasileiras.*

Mínimo	25%	Mediana	Média	75%	90%	Máximo
0,077	0,138	0,161	0,163	0,184	0,207	0,314

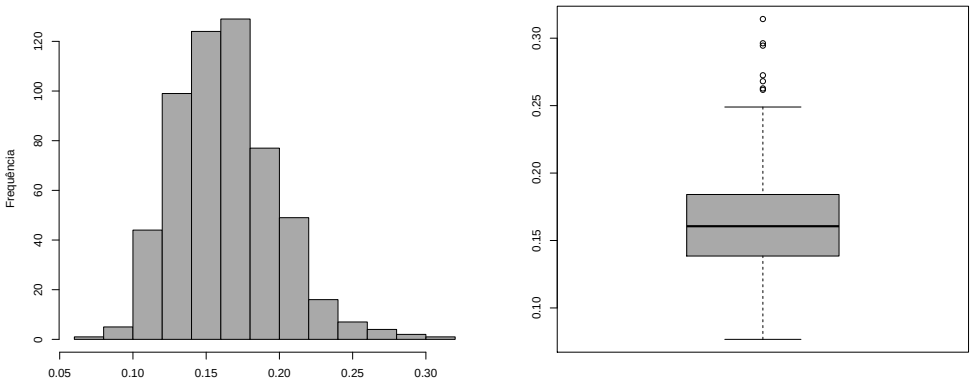

Figura 4.5: *Histograma e Box-Plot para a letalidade suavizada por IAM.*

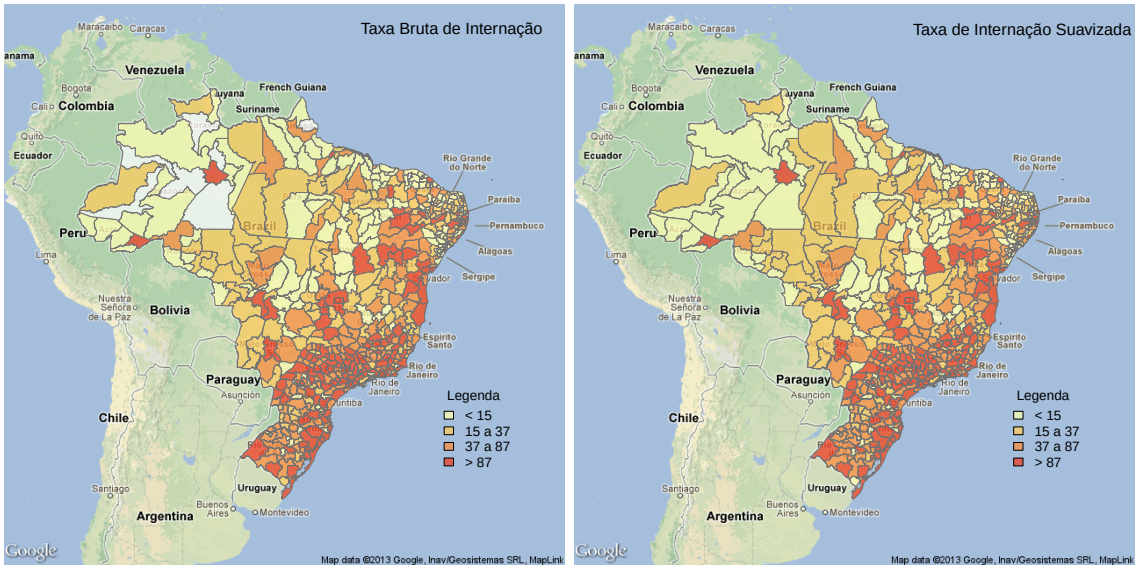

Figura 4.6: *Estimativas para a taxa bruta de internação, esquerda, e para a taxa de internação suavizada obtida através da metodologia Bayesiana.*

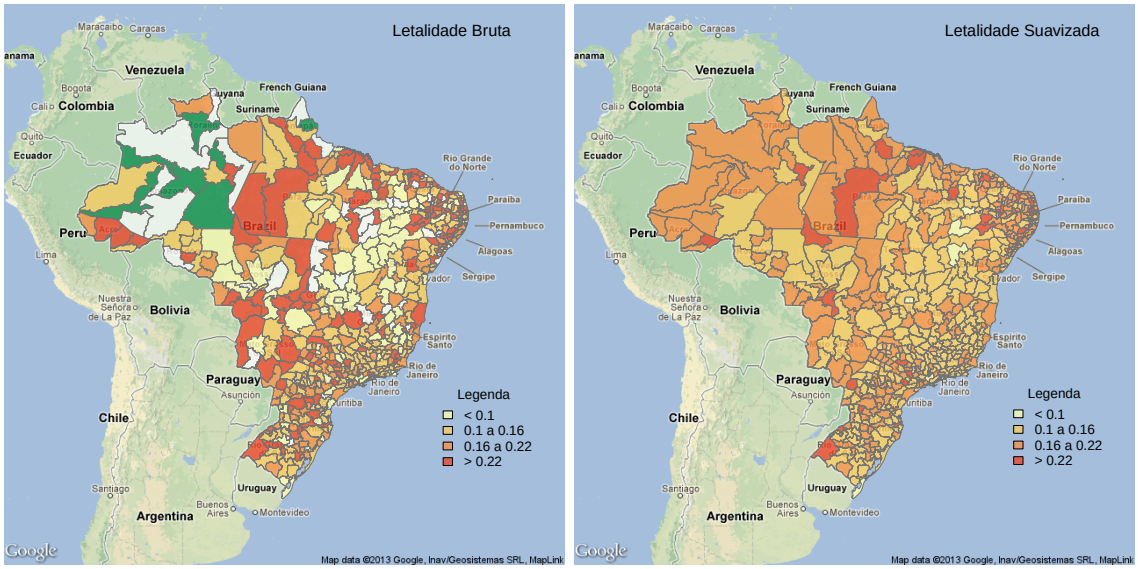

Figura 4.7: *Estimativas para a letalidade bruta, esquerda, e letalidade estimada através da metodologia Bayesiana.*

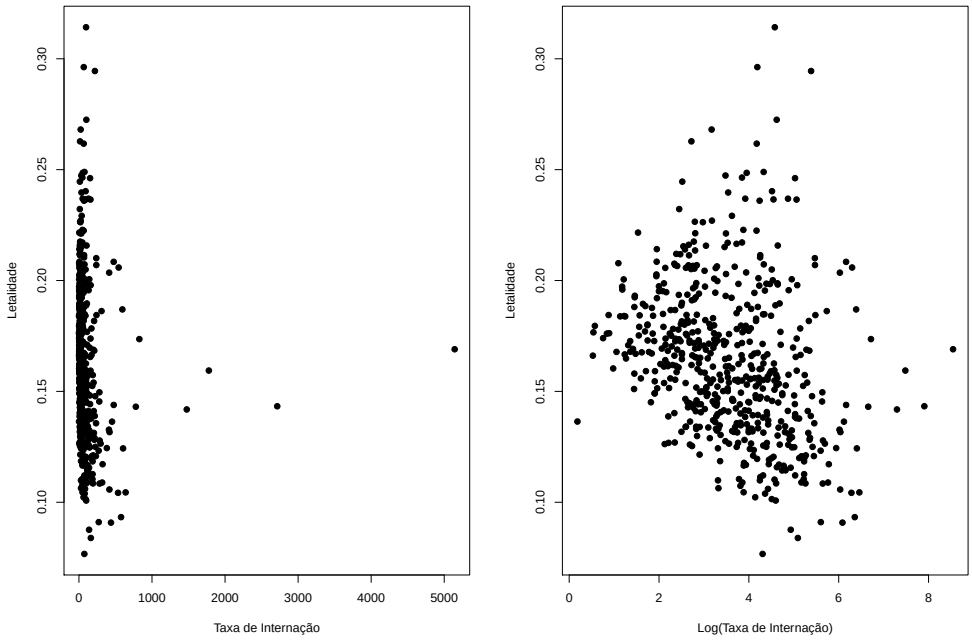

Figura 4.8: *Relacão entre a taxa de internação e letalidade por IAM ajustadas.*

Capítulo 5

Conclusões

Neste trabalho utilizamos a metodologia Bayesiana para estimar a letalidade por infarto agudo do miocárdio (IAM) nas microrregiões brasileiras. O método INLA foi o método de aproximação usado para obter as distribuições marginais a posteriori devido a sua rapidez computacional e por ser um método determinístico em que não há cadeias de Markov para serem analisadas. Desenvolvemos um estudo de simulação em que avaliamos e verificamos a equivalência entre a formulação aqui proposta e a formulação original utilizada por Knorr-Held e Best (2001) para o modelo componente compartilhado, sendo mantidas as características e interpretação dos parâmetros do modelo componente compartilhado. O resultado obtido neste estudo possibilitou a implementação do modelo no programa `R` através do pacote INLA e no software `OpenBUGS`, variando o método de aproximação utilizado na obtenção da distribuição a posteriori, determinístico e estocástico respectivamente.

Com o modelo componente compartilhado modelamos conjuntamente o número de mortes e o número de sobreviventes por IAM. Separamos a superfície de risco em um componente compartilhado pelas duas medições e componentes específicos de cada uma, com o propósito de identificar covariáveis estruturadas espacialmente não observáveis que afetam o risco de ambas medições simultaneamente e individualmente. Observamos pelo mapa que descreve a variação espacial compartilhada pela mortalidade e sobrevivência por IAM que existem dois clusters distintos, um na região norte e nordeste e outro na região sul e sudeste do Brasil, fornecendo forte evidência de fatores de risco não observados comuns entre as medições nestes locais. As altas estimativas do componente compartilhado encontram-se nas

microrregiões mais urbanizadas, ocorrendo o contrário para as baixas estimativas. O padrão espacial para os componentes específicos de mortalidade e de sobrevivência são contrários, isto é, em microrregiões em que o risco de sobreviver é alto o risco de morrer é baixo. O mapa com a variação do componente específico de mortalidade apresenta uma configuração espacial mais definida, com dois clusters com estimativas elevadas no nordeste do Brasil e no estado de São Paulo.

A distribuição das estimativas Bayesiana para a taxa de internação mostra valores moderados e elevados ao longo da faixa que engloba as microrregiões da região sul até uma parte da região nordeste. Este padrão muda para regiões restantes do Brasil. A suavização da letalidade nas microrregiões através do ajuste Bayesiano, usando a informação das microrregiões para diminuir o efeito das flutuações aleatórias não associadas ao risco, tornou mais clara a tendência espacial distribuída ao longo do Brasil. A maioria das microrregiões está na faixa intermediária que contempla a média nacional, no entanto a tendência observada na região norte se diferencia do restante com valor igual ou acima da média nacional predominantes, corroborando com o padrão encontrado na análise dos componentes específicos. Isto pode ser devido à extensão territorial que talvez dificulte o rápido acesso da população ao atendimento hospitalar, visto que o infarto agudo do miocárdio caracteriza-se como um evento agudo que necessita de socorro médico rápido. Verificamos que a relação entre a taxa de internação e de letalidade não é linear, porém há uma dependência da variabilidade da letalidade na medida em que o logaritmo da taxa de internação aumenta.

A análise dos dados de IAM pode ser melhorada em um estudo futuro com a inclusão de covariáveis. Uma possibilidade seria incorporar no número de casos esperados a informação da população estratificada por sexo e idade, já que a mortalidade por infarto é maior entre homens mais velhos.

Apêndice A

Sejam X e Y variáveis aleatórias independentes com distribuição Poisson cujas taxas são iguais a λ e φ, respectivamente. A soma dessas variáveis segue uma distribuição Poisson, e a distribuição condicional de uma das variáveis dado a soma entre elas é Binomial.

$$
\begin{aligned}
P(X+Y=n) &= \sum_{k=0}^{X+Y=n} P_X(X=k)P(Y=n-k) \\
&= \sum_{k=0}^{n} \frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{e^{-\varphi}\varphi^{(n-k)}}{(n-k)!} \\
&= \sum_{k=0}^{n} \frac{e^{-(\lambda+\varphi)}\lambda^k\varphi^{(n-k)}}{k!(n-k)!} \\
&= \sum_{k=0}^{n} \frac{e^{-(\lambda+\varphi)}\lambda^k\varphi^{(n-k)}}{k!(n-k)!} \cdot \frac{n!}{n!} \\
&= \frac{e^{-(\lambda+\varphi)}}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda^k\varphi^{(n-k)} \\
&= \frac{e^{-(\lambda+\varphi)}(\lambda+\varphi)^n}{n!} \\
&= \text{Poisson}(\lambda+\varphi)
\end{aligned}
$$

$$
\begin{aligned}
P(X=k|X+Y=n) &= \frac{P(X=k, X+Y=n)}{P(X+Y=n)} \\
&= \frac{P(X=k, Y=n-k)}{P(X+Y=n)} \\
&= \frac{\frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{e^{-\varphi}\varphi^{n-k}}{(n-k)!}}{\frac{e^{-(\lambda+\varphi)}(\lambda+\varphi)^n}{n!}} \\
&= \frac{n!\lambda^k\varphi^{n-k}}{k!(n-k!)(\lambda+\varphi)^n} \\
&= \binom{n}{k} \cdot \lambda^k \cdot \frac{\varphi^{n-k}}{(\lambda+\varphi)^n} \cdot \frac{(\lambda+\varphi)^k}{(\lambda+\varphi)^k} \\
&= \binom{n}{k} \left(\frac{\lambda}{\lambda+\varphi}\right)^k \left(\frac{\varphi}{\lambda+\varphi}\right)^{n-k} \\
&= \text{Binomial}\left(n, \frac{\lambda}{\lambda+\varphi}\right)
\end{aligned}
$$

Apêndice B

A linha entre os pontos nos gráficos representa a regressão ortogonal (minimiza-se a soma dos quadrados das distâncias ortogonais dos pontos observados à reta de regressão) para a variável dependente no eixo Y e a variável independente no eixo X. Nas figuras exibidas a seguir, as variáveis independentes correspondem às estimativas fornecidas pelo método de aproximação estocástica MCMC e o método INLA, enquanto as variáveis dependentes correspondem aos valores verdadeiros e as estimativas fornecidas pelo método INLA.

Abaixo estão os resultados gráficos para o primeiro estudo de simulação usando dados simulados. Os gráficos de dispersão apresentados foram gerados considerando as seguintes informações em cada área: verdadeiro valor, estimativa Bayesiana obtida pelo método MCMC quando a formulação original foi definida para o preditor linear, e estimativa Bayesiana obtida pelo método INLA quando a formulação alterada foi definida para o preditor. As informações são identificadas no gráfico por: “Verdadeiro”, “MCMC” e “INLA”, respectivamente.

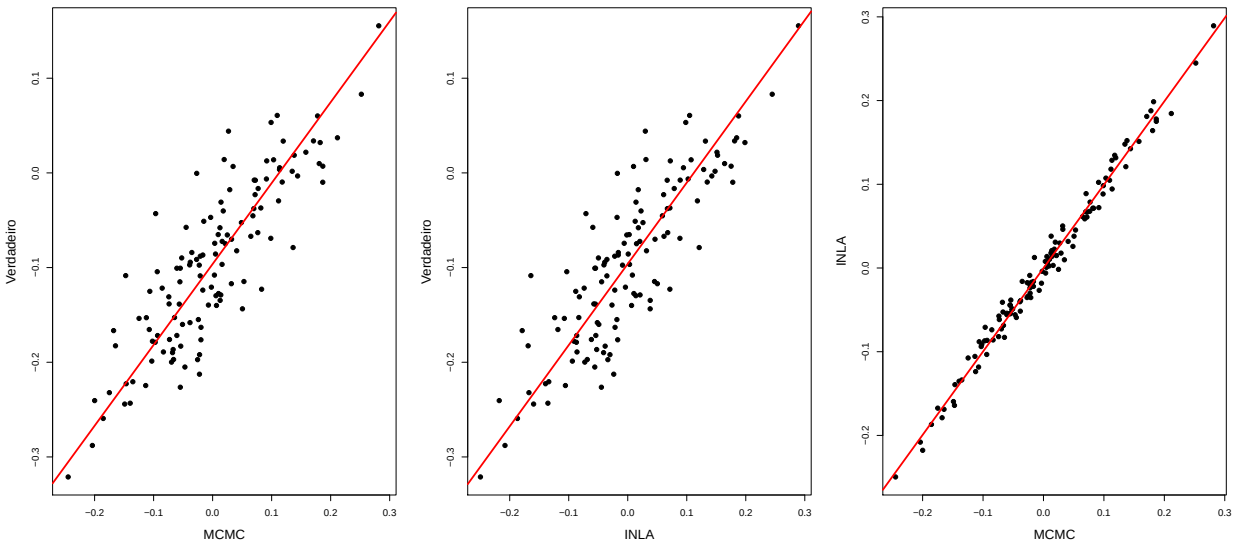

Figura B.1: *Gráfico de dispersão do componente compartilhado ϕ_i. Referentes ao primeiro estudo de simulação.*

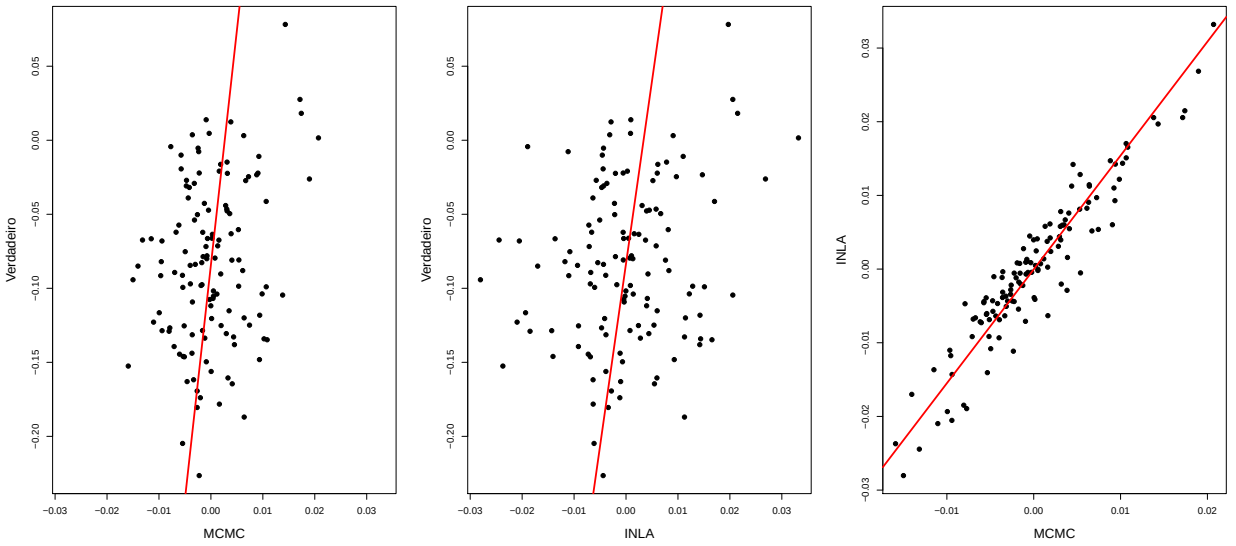

Figura B.2: *Gráfico de dispersão do componente específico da medição 1, ψ_{i1}. Referentes ao primeiro estudo de simulação.*

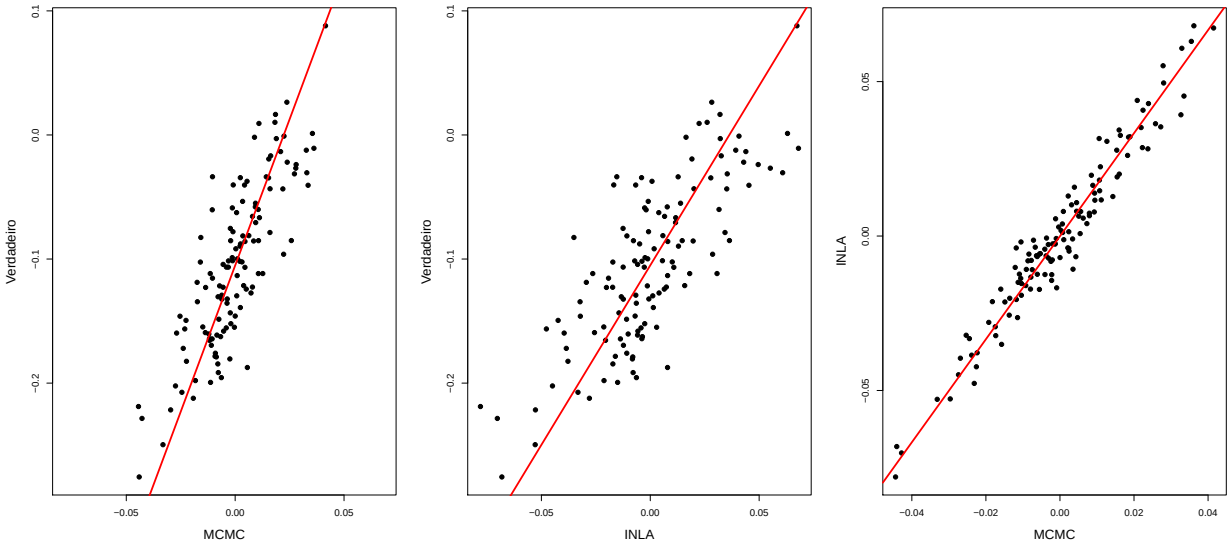

Figura B.3: *Gráfico de dispersão do componente específico da medição* 2, ψ_{i2}. *Referentes ao primeiro estudo de simulação.*

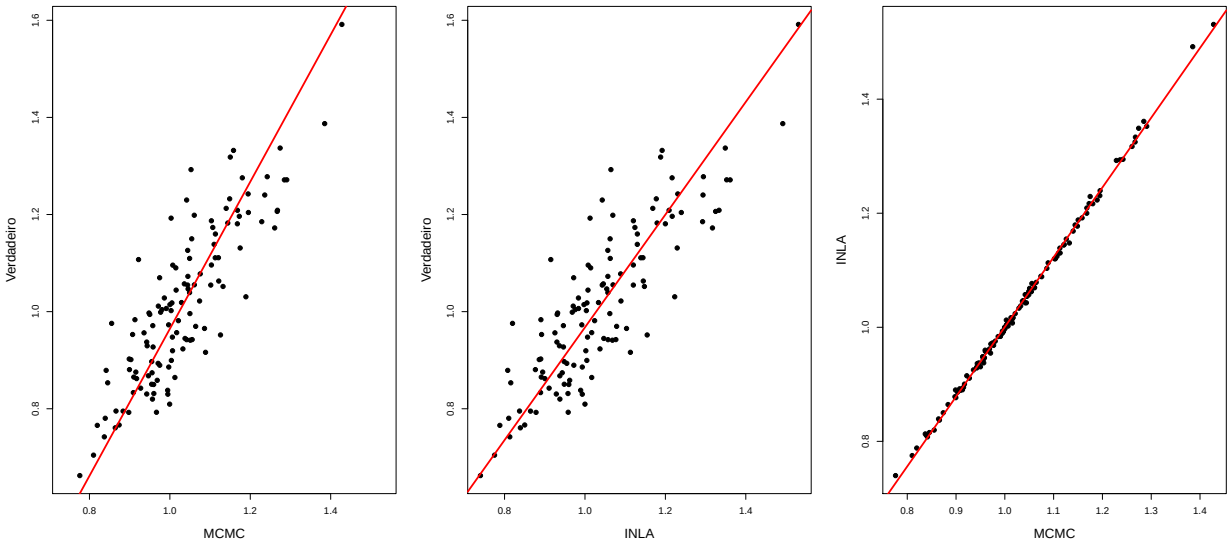

Figura B.4: *Gráfico de dispersão do risco relativo da medição* 1, θ_{i1}. *Referentes ao primeiro estudo de simulação.*

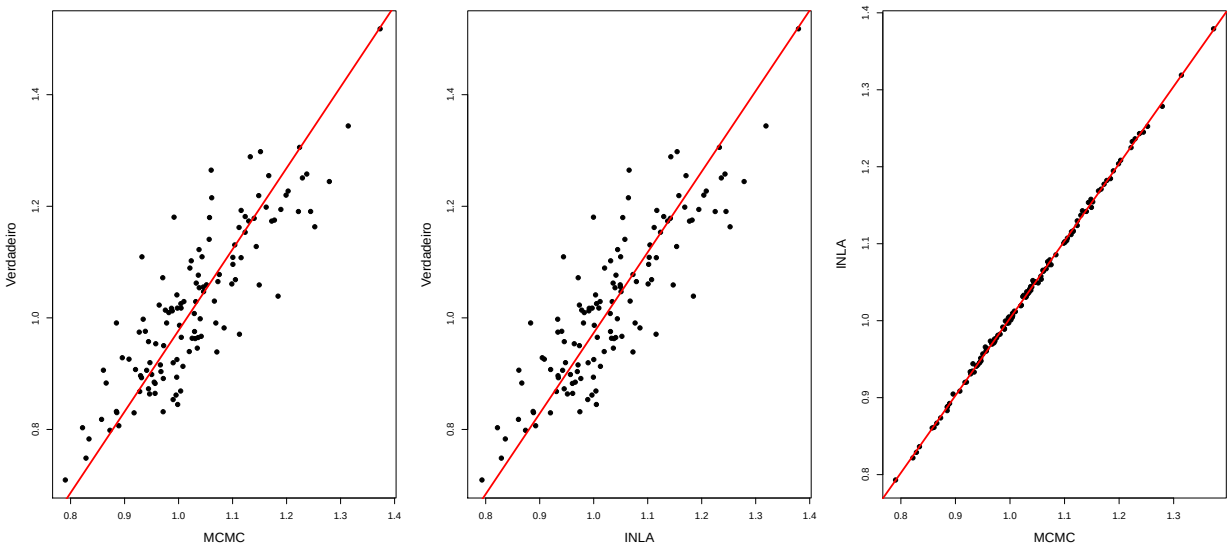

Figura B.5: *Gráfico de dispersão do risco relativo da medição 2, θ_{i2}. Referentes ao primeiro estudo de simulação.*

Resultados gráficos para o segundo estudo de simulação usando dados simulados. Os gráficos de dispersão gerados consideraram as informações em cada área referentes ao verdadeiro valor, à estimativa Bayesiana obtida pelo método MCMC e pelo método INLA considerando a formulação alterada para o preditor linear. São identificadas no gráfico por: "Verdadeiro", "MCMC" e "INLA", respectivamente.

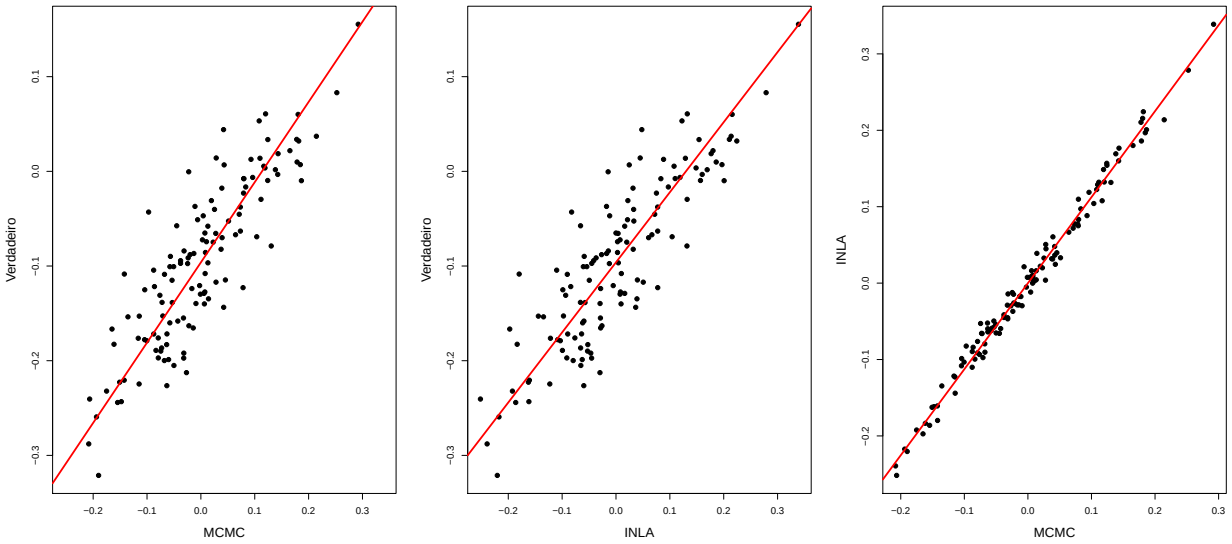

Figura B.6: *Gráfico de dispersão do componente compartilhado ϕ_i. Referentes ao segundo estudo de simulação.*

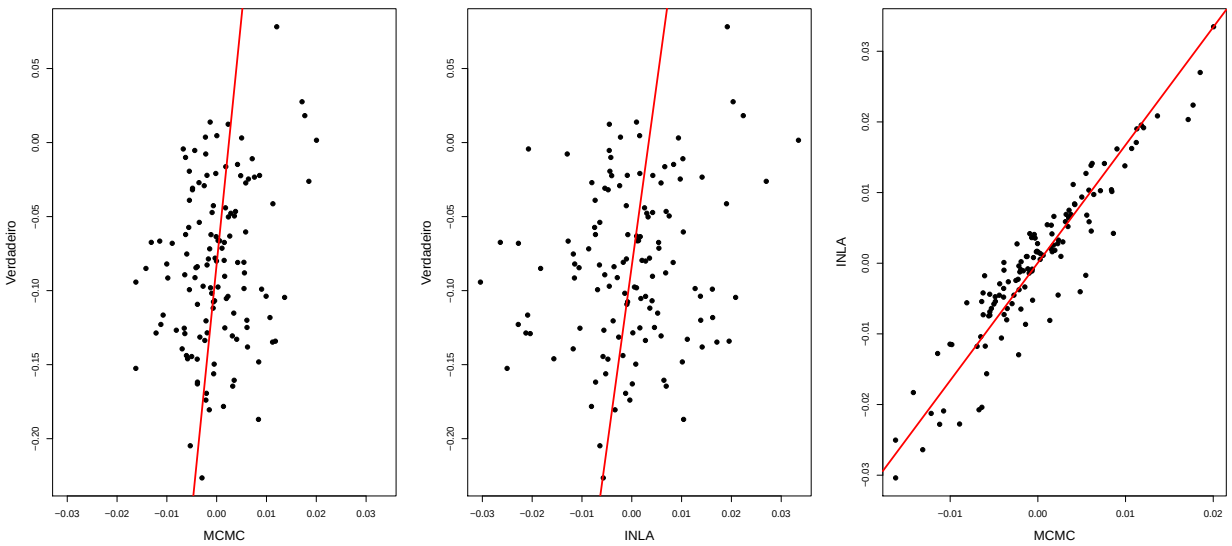

Figura B.7: *Gráfico de dispersão do componente específico da medição* 1, ψ_{i1}. *Referentes ao segundo estudo de simulação.*

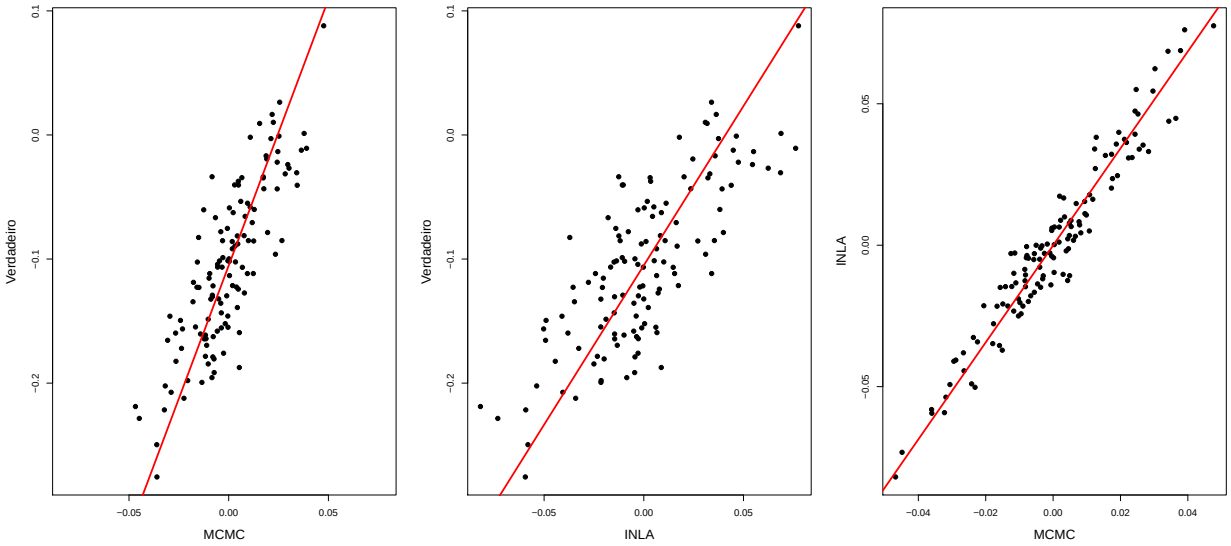

Figura B.8: *Gráfico de dispersão do componente específico da medição* 2, ψ_{i2}. *Referentes ao segundo estudo de simulação.*

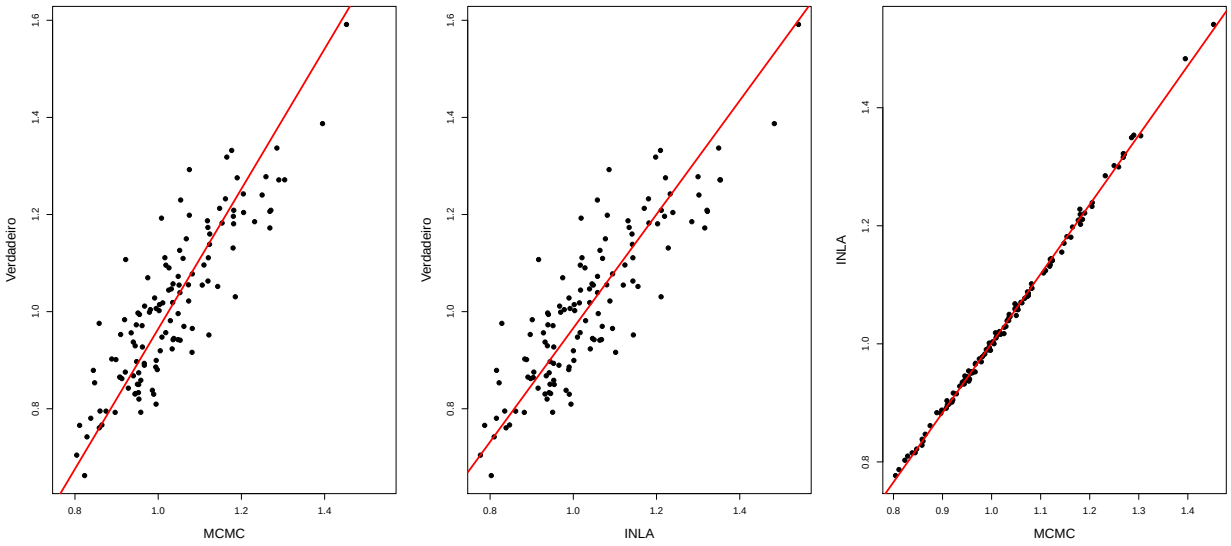

Figura B.9: *Gráfico de dispersão do risco relativo da medição* 1, θ_{i1}. *Referentes ao segundo estudo de simulação.*

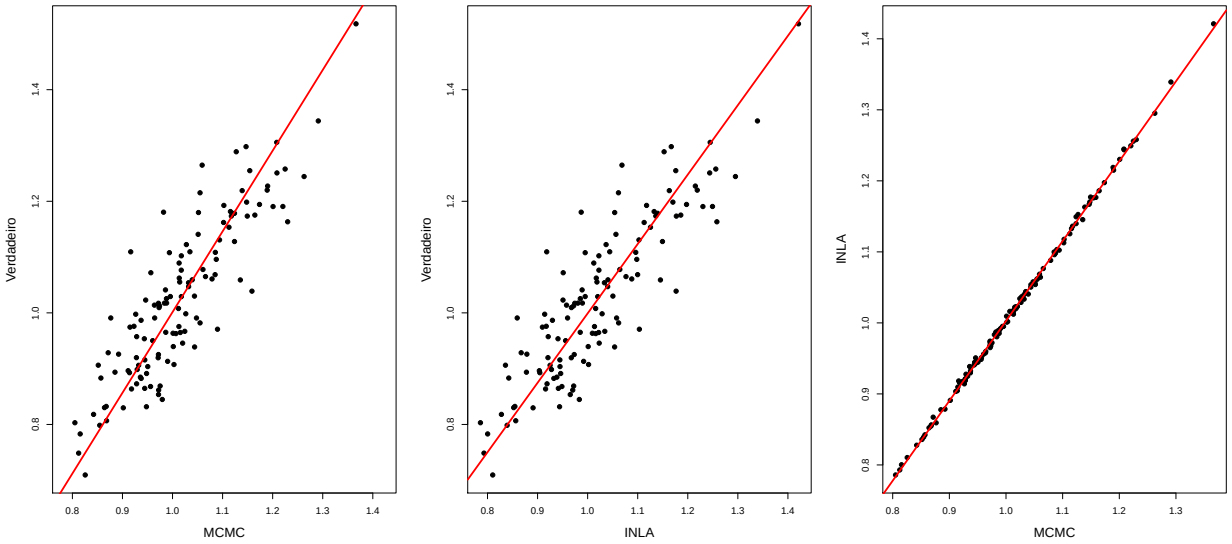

Figura B.10: *Gráfico de dispersão do risco relativo da medição 2, θ_{i2}. Referentes ao segundo estudo de simulação.*

Resultados gráficos para o primeiro estudo de simulação usando dados gerados a partir da análise dos dados reais. Os gráficos de dispersão apresentados foram gerados considerando as seguintes informações em cada área: verdadeiro valor, estimativa Bayesiana obtida pelo método MCMC quando a formulação original foi definida para o preditor linear, e estimativa Bayesiana obtida pelo método INLA quando a formulação alterada foi definida para o preditor. As informações são identificadas no gráfico por: "Verdadeiro", "MCMC" e "INLA", respectivamente.

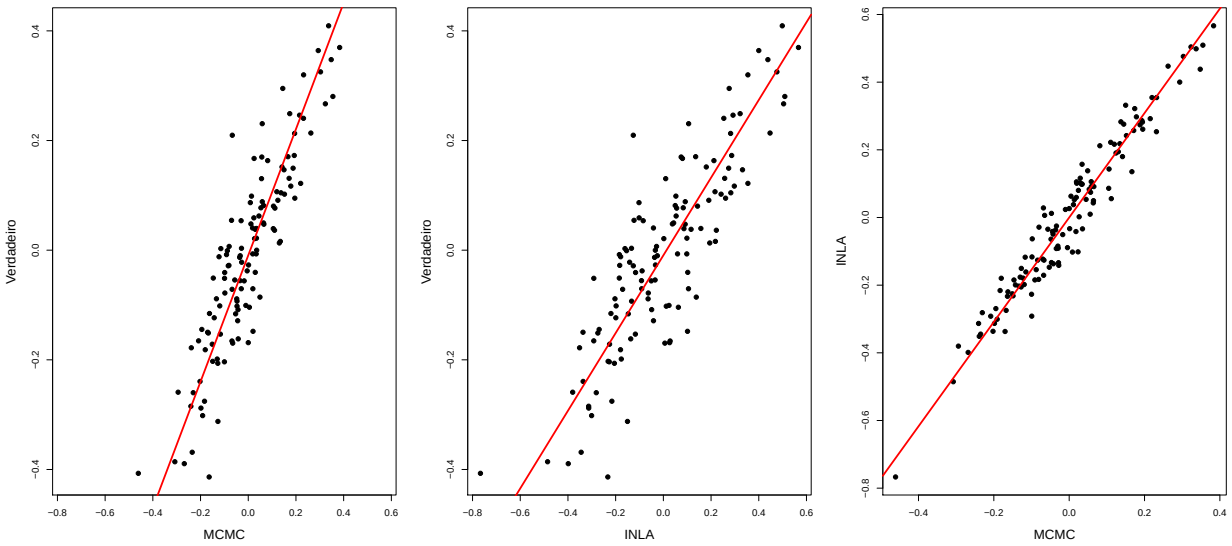

Figura B.11: *Gráfico de dispersão do componente compartilhado ϕ_i. Referentes ao primeiro estudo de simulação.*

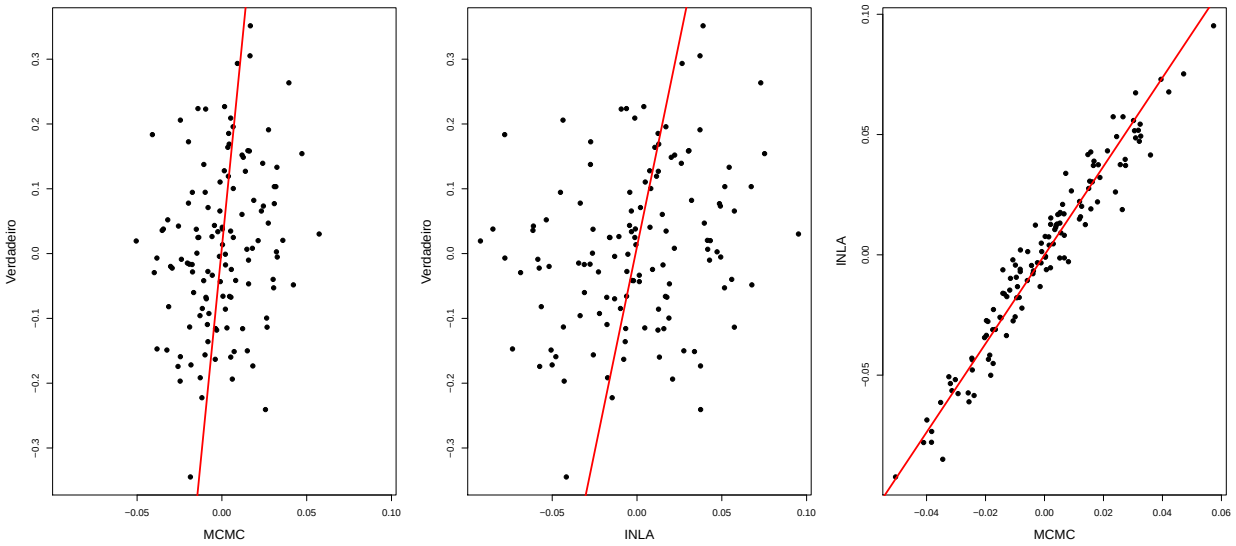

Figura B.12: *Gráfico de dispersão do componente específico da medição* 1, ψ_{i1}. *Referentes ao primeiro estudo de simulação.*

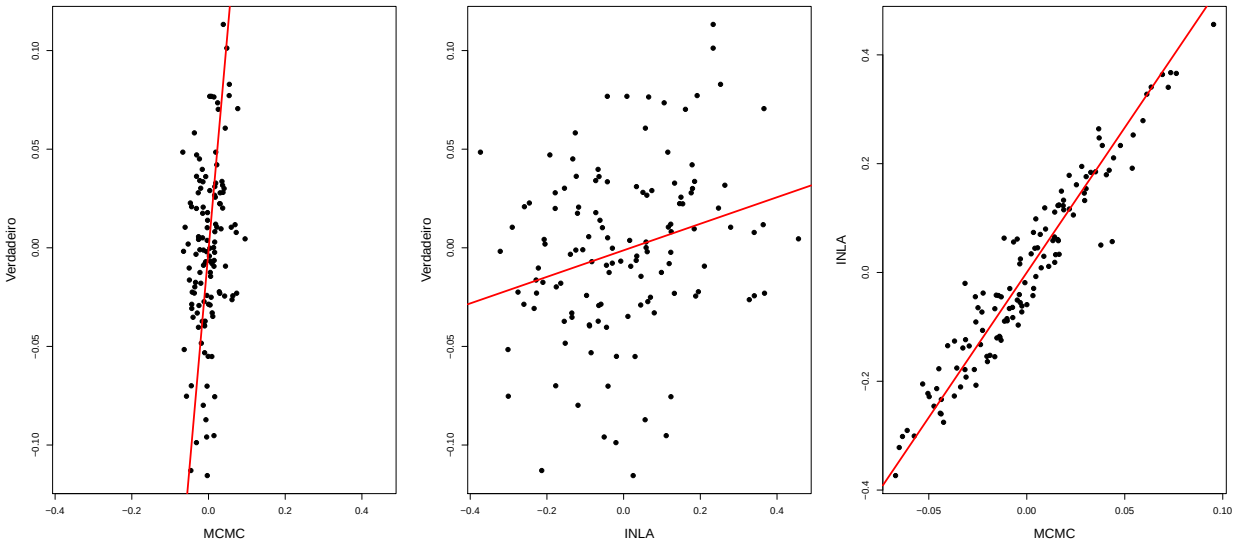

Figura B.13: *Gráfico de dispersão do componente específico da medição* 2, ψ_{i2}. *Referentes ao primeiro estudo de simulação.*

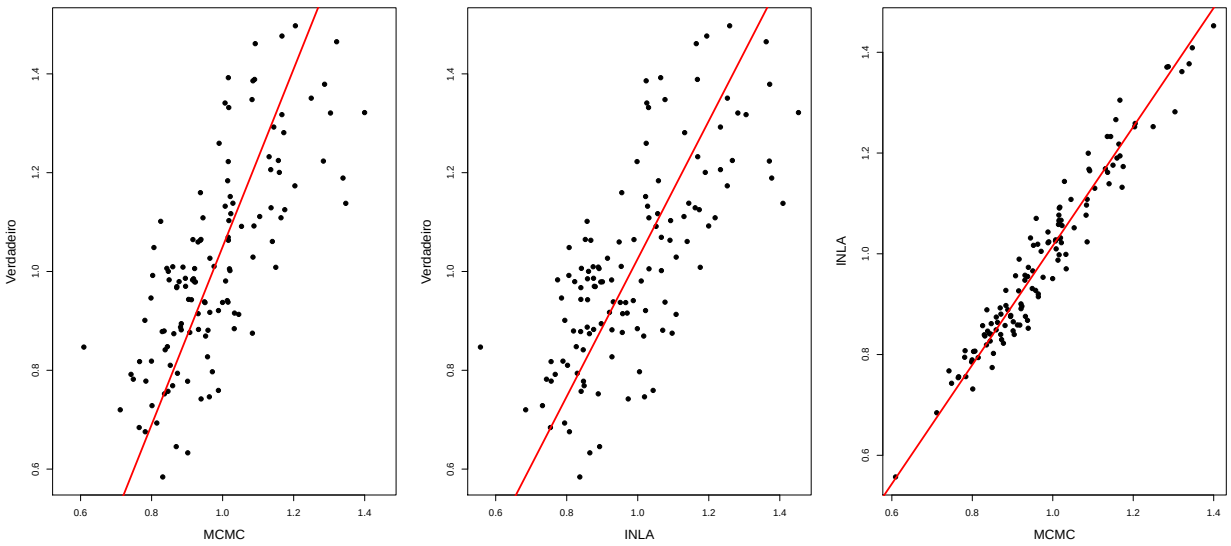

Figura B.14: *Gráfico de dispersão do risco relativo da medição* 1, θ_{i1}. *Referentes ao primeiro estudo de simulação.*

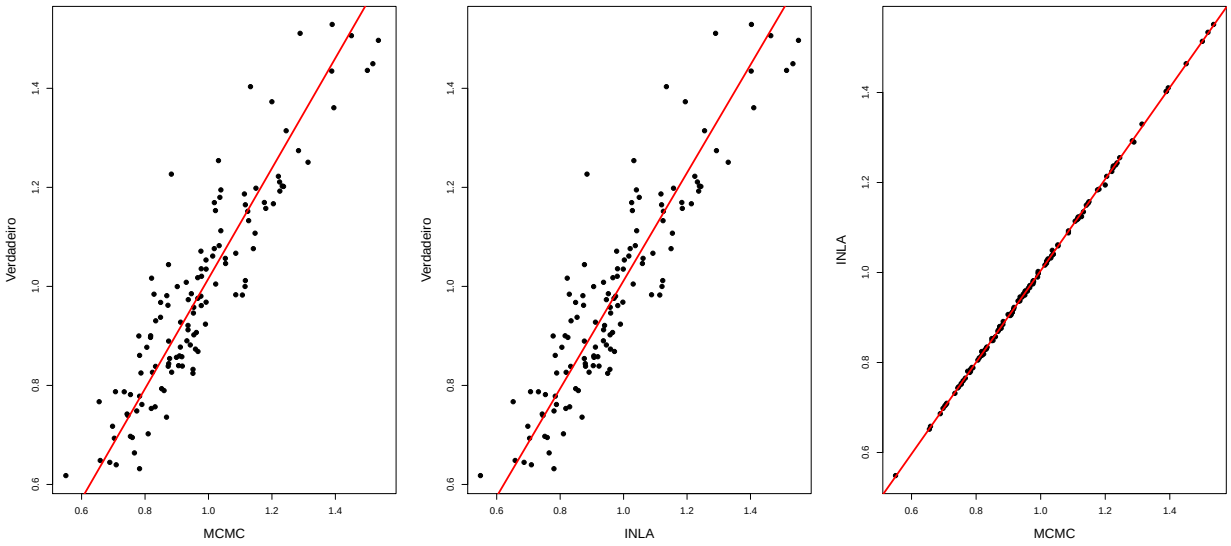

Figura B.15: *Gráfico de dispersão do risco relativo da medição* 2, θ_{i2}. *Referentes ao primeiro estudo de simulação.*

Resultados gráficos para o primeiro estudo de simulação usando dados gerados a partir da análise dos dados reais. Os gráficos de dispersão gerados consideraram as informações em cada área referentes ao verdadeiro valor, à estimativa Bayesiana obtida pelo método MCMC e pelo método INLA considerando a formulação alterada para o preditor linear. São identificadas no gráfico por: "Verdadeiro", "MCMC" e "INLA", respectivamente.

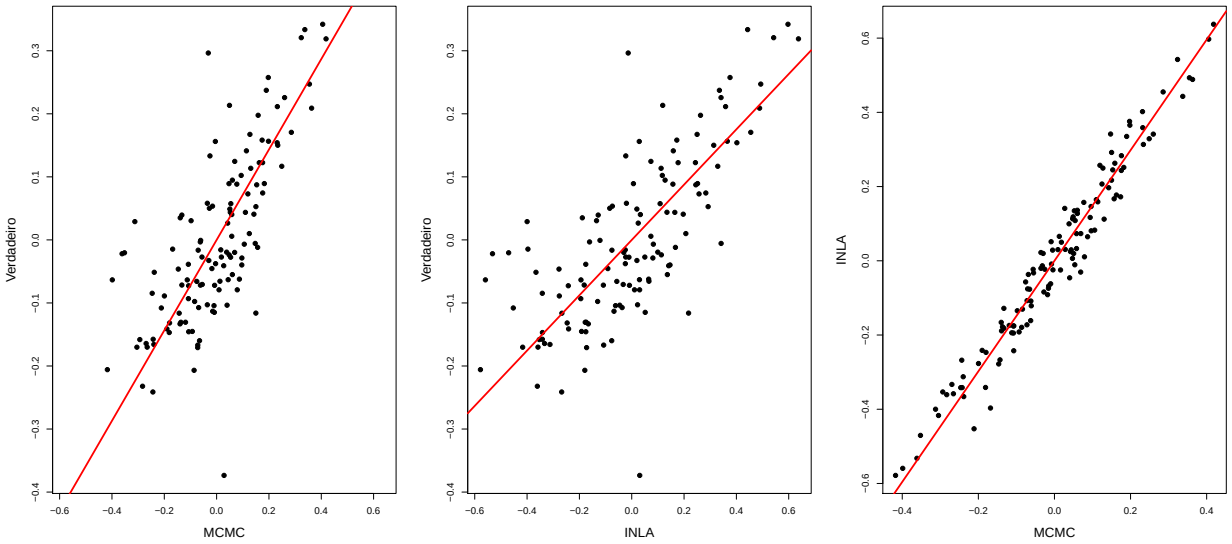

Figura B.16: *Gráfico de dispersão do componente compartilhado ϕ_i. Referentes ao segundo estudo de simulação.*

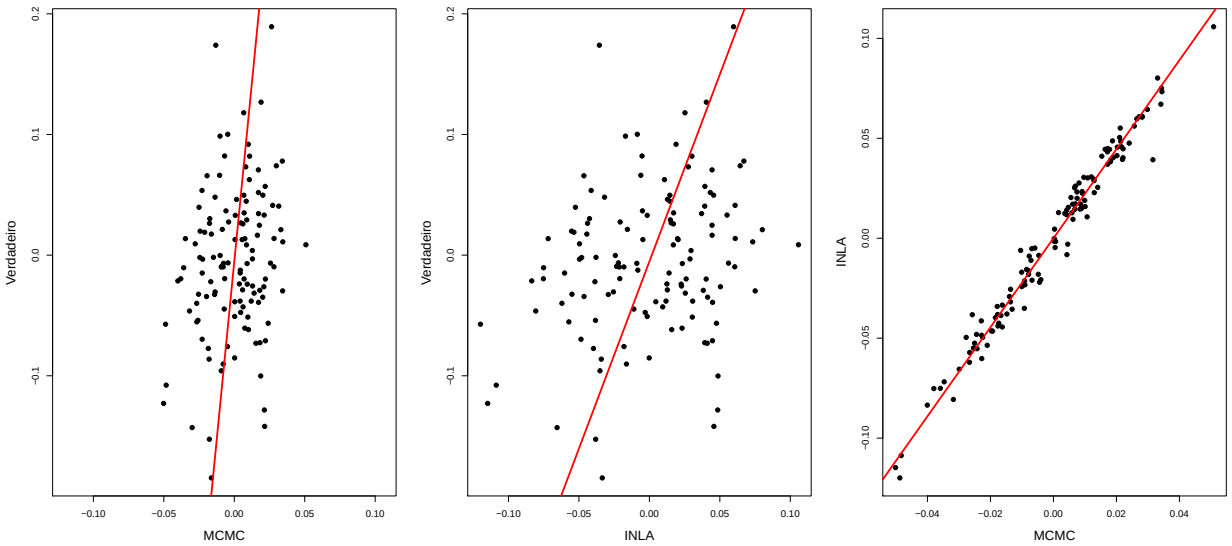

Figura B.17: *Gráfico de dispersão do componente específico da medição* 1, *ψ_{i1}. Referentes ao segundo estudo de simulação.*

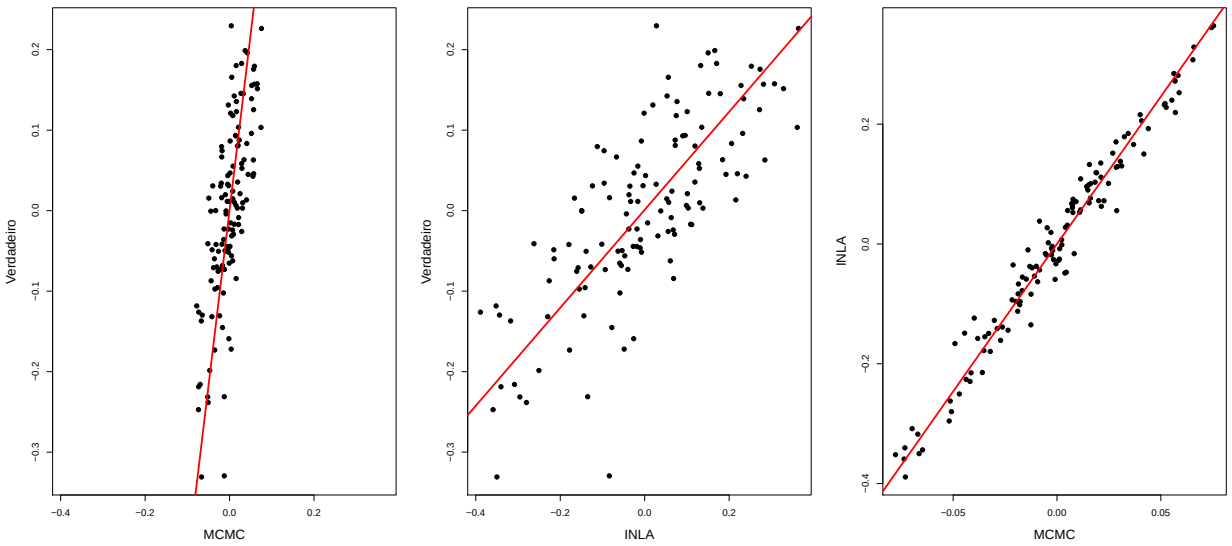

Figura B.18: *Gráfico de dispersão do componente específico da medição* 2, ψ_{i2}. *Referentes ao segundo estudo de simulação.*

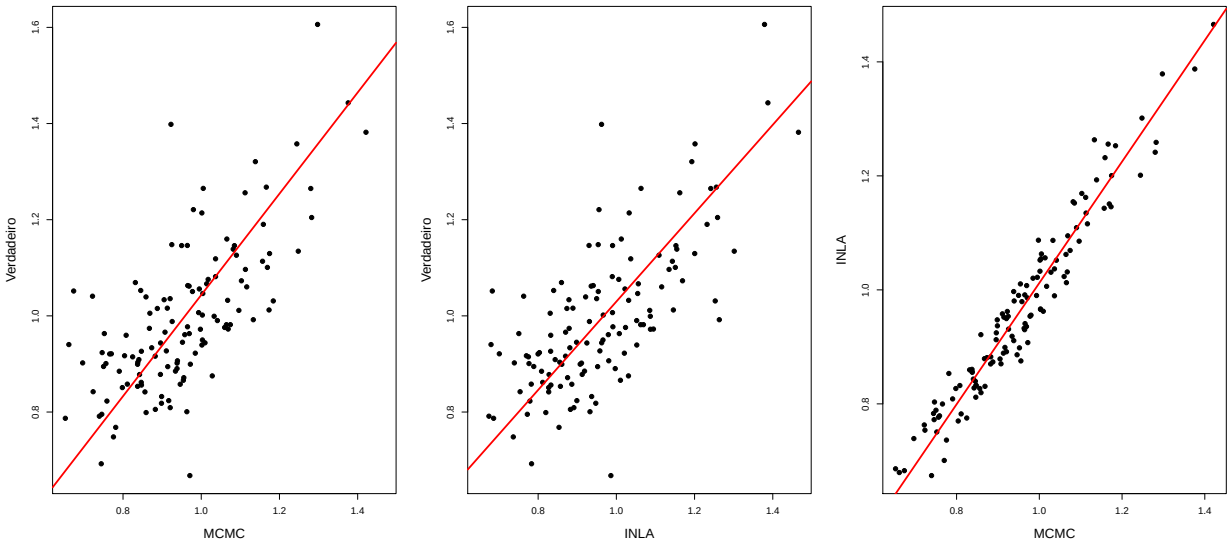

Figura B.19: *Gráfico de dispersão do risco relativo da medição* 1, θ_{i1}. *Referentes ao segundo estudo de simulação.*

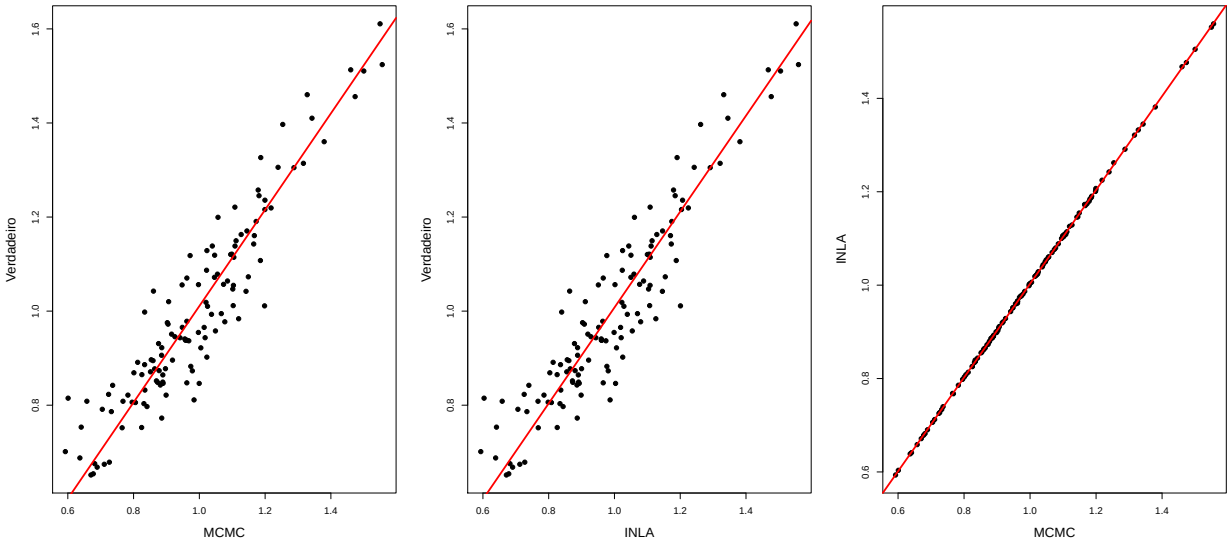

Figura B.20: *Gráfico de dispersão do risco relativo da medição 2, θ_{i2}. Referentes ao segundo estudo de simulação.*

Apêndice C

Tabela C.1: *Sigla do Estado e as respectivas capitais. A microrregião que contempla a respectiva capital, os dados observados no ano de* 2010 *e a letalidade suavizada.*

UF	Capital	Nome da Microrregião	Núm. Mortes	Núm. Internações	Núm. Sobreviventes	População	Letalidade
RO	Porto Velho	Porto Velho	8	66	58	446.174	0.14
AC	Rio Branco	Rio Branco	26	92	66	377.165	0.24
AM	Manaus	Manaus	73	605	532	1.606.731	0.12
RR	Boa Vista	Boa Vista	6	32	26	292.665	0.17
PA	Belém	Belém	19	148	129	1.635.593	0.15
AP	Macapá	Macapá	8	74	66	480.367	0.14
TO	Palmas	Porto Nacional	14	64	50	279.358	0.18
MA	São Luís	Aglomeração Urbana de São Luís	7	45	38	1.055.780	0.15
PI	Teresina	Teresina	21	153	132	811.002	0.14
CE	Fortaleza	Fortaleza	52	410	358	242.7918	0.13
RN	Natal	Natal	26	213	187	680.425	0.13
PB	João Pessoa	João Pessoa	33	130	97	823.103	0.24
PE	Recife	Recife	111	779	668	2.246.129	0.14
AL	Maceió	Maceió	21	103	82	881.512	0.20
SE	Aracaju	Aracaju	10	123	113	629.659	0.11
BA	Salvador	Salvador	54	535	481	2.540.298	0.10
MG	Belo Horizonte	Belo Horizonte	210	1477	1267	2.793.353	0.14
ES	Vitória	Vitória	42	418	376	793910	0.11
RJ	Rio de Janeiro	Rio de Janeiro	389	2715	2326	6.838.242	0.14
SP	São Paulo	São Paulo	871	5145	4274	5.818.134	0.17
PR	Curitiba	Curitiba	67	476	409	1.814.816	0.14
SC	Florianópolis	Florianópolis	32	319	287	557.892	0.11
RS	Porto Alegre	Porto Alegre	285	1780	1495	2.331.191	0.16
MS	Campo Grande	Campo Grande	36	298	262	668.181	0.13
MT	Cuiabá	Cuiabá	71	220	149	650.492	0.29
GO	Goiânia	Goiânia	64	640	576	168.5997	0.10
DF	Brasília	Brasília	50	577	527	1904093	0.09

Referências Bibliográficas

ANS(2012) ANS. *Agência Nacional de Saúde Suplementar*, 2012. URL http://www.ans.gov.br/anstabnet/anstabnet/materia_novo.htm. Citado na pág. 5

Assunção *et al.*(**1998**) R.M. Assunção, S.M. Barreto, H.L. Guerra e E. Sakurai. Mapas de taxas epidemiológicas: uma abordagem bayesiana. *Caderno de Saúde Pública*, 14(4): 713–723. Citado na pág. 2

Besag *et al.*(**1991**) J. Besag, J. York e A. Mollie. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43: 1–59. Citado na pág. 10, 14

Best *et al.*(**2005**) N. Best, S. Richardson e A. Thomson. A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59. Citado na pág. 3, 7, 9, 17

Dabney e Wakefield(2005) A.R. Dabney e J.C. Wakefield. Issues in the mapping of two diseases. *Statistical Methods in Medical Research*, 14:83–112. Citado na pág. 3, 9

DATASUS(2012) DATASUS. *Departamento de Informação do SUS*, 2012. URL http://www.datasus.gov.br. Citado na pág. 3

Gamerman e Lopes(2006) D. Gamerman e H. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman e Hall/CRC. Citado na pág. 18

Gelfand *et al.*(**2010**) A. E. Gelfand, P. J. Diggle, M. Fuentes e P. Guttorp. *Handbook of Spatial Statistics*. Chapman e Hall/CRC. Citado na pág. 7, 11

Geman e Geman(1984) S. Geman e D. Geman. Stochastic relaxation, gibbs distributions and the bayesian. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6: 721–741. Citado na pág. 18

GeoBUGS(2012) GeoBUGS. *The BUGS Project - GeoBUGS*, 2012. URL http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml. Citado na pág. 25

Hastings(1970) W.K. Hastings. Monte carlo sampling methods using markov chains and their. *Biometrika*, 57:97–109. Citado na pág. 18

Held *et al.*(**2005a**) L. Held, G. Graziano, C. Frank e H. Rue. Joint spatial analysis of gastrointestinal infectious diseases. *Statistical Methods in Medical Research*, 15:465–480. Citado na pág. 3, 7

Held *et al.*(**2005b**) L. Held, I. Natário, S.E. Fenton, H. Rue e N. Becker. Towards joint disease mapping. *Statistical Methods in Medical Research*, 14:61–82. Citado na pág. 3, 6, 9, 17

Kelsall e Wakefield(1999) J. E. Kelsall e J. C. Wakefield. *Discussion of Bayesian models for spatially correlated disease and exposure data Em: Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. Bayesian Statistics.*, volume 16. Oxford University Press. Citado na pág. 16

Knorr-Held e Best(2001) L. Knorr-Held e N.G. Best. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society*, 164:73–85. Citado na pág. ii, iii, 3, 5, 6, 12, 14, 39, 47

Lawson(2009) A.B. Lawson. *Bayesian Disease Mapping Hierarquical Modeling in Spatial Epidemiology*. Chapman e Hall/CRC. Citado na pág. 7, 11

Lawson *et al.***(2000)** A.B Lawson, A.B. Biggeri, D. Boehning, E. Lesaffre, J.F. Viel, A. Clark, P. Schlattmann e F. Divino. Disease mapping models an empirical evaluation. *Statistics in Medicine*, 19:2217–2241. Citado na pág. 14

Lunn *et al.***(2000)** D. Lunn, A. Thomas, N.G. Best e D.J. Spiegelhalter. Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337. Citado na pág. 8

Maiti(1998) T. Maiti. Hierarquical bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69:339–348. Citado na pág. 2

Metropolis *et al.***(1953)** N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller e E. Teller. Equation of state calculations by fast computing machine. *Journal of Chemical Physics*, 21:1087–1091. Citado na pág. 18

Min.Saúde(2012a) Min.Saúde. *Ministério da Saúde*, 2012a. URL http://tabnet.datasus.gov.br/cgi/idb2000/fqc11_1.htm. Citado na pág. 3

Min.Saúde(2012b) Min.Saúde. *Portal da Saúde do Ministério da Saúde*, 2012b. URL http://portal.saude.gov.br/portal/saude/visualizar_texto.cfm?idtxt=33353. Citado na pág. 4, 36

Pascutto *et al.***(2000)** C. Pascutto, J.C. Wakefield, N.G. Best, S. Richardson, L. Bernardinelli, A. Staines e P. Elliott. Statistical issues in the analysis of disease mapping data. *Statistical in Medicine*, 19:2493–2519. Citado na pág. 2, 6, 7, 17

Paulino *et al.***(2003)** C.D.M. Paulino, M.A.A. Turkman e B. Murteira. *Estatística Bayesiana*. Fundação Calouste Gulbenkian. Citado na pág. 8

Rue *et al.***(2009)** H. Rue, S. Martino e N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B*, 71:319–392. Citado na pág. 3, 19, 21, 22

SIH-SUS(2012) SIH-SUS. *Sistema de Informação Hospitalar do Sistema Único de Saúde*, 2012. URL http://www.datasus.gov.br/catalogo/sihsus.htm. Citado na pág. 4

SIM(2012) SIM. *Sistema de Informações sobre Mortalidade*, 2012. URL http://www.datasus.gov.br/catalogo/sim.htm. Citado na pág. 4

Wakefield(2007) J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183. Citado na pág. 7