UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

ANÁLISE DE CORRELAÇÃO ECOLÓGICA: UMA ABORDAGEM INTEIRAMENTE BAYESIANA PARA A MORTALIDADE INFANTIL NO RIO GRANDE DO SUL

Sérgio Kakuta Kato

Orientador: Profa. Dra. Jandyra Maria Guimarães Fachel

Porto Alegre, 2007.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

ANÁLISE DE CORRELAÇÃO ECOLÓGICA: UMA ABORDAGEM INTEIRAMENTE BAYESIANA PARA A MORTALIDADE INFANTIL NO RIO GRANDE DO SUL

Sérgio Kakuta Kato

Orientador: Profa. Dra. Jandyra Maria Guimarães Fachel

A apresentação desta dissertação é exigência do Programa de Pós-graduação em Medicina: Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Mestre.

Porto Alegre, Brasil.
2007

BANCA EXAMINADORA

Profa. Dra. Maria Inês Azambuja, Programa de Pós Graduação em Epidemiologia, UFRGS.

Prof. Dr. Sérgio Luiz Bassanesi, Programa de Pós Graduação em Epidemiologia, UFRGS.

Profa. Dra. Sidia Maria Callegari Jacques, Instituto de Matemática, Depto. de Estatística, UFRGS.

Profa. Dra. Jandyra Maria Guimarães Fachel, Programa de Pós Graduação em Epidemiologia, UFRGS (Orientadora).

AGRADECIMENTOS

À professora Dra. Jandyra Maria Guimarães Fachel pela orientação, pela amizade, pelos conselhos e ensinamentos ao longo de toda a minha vida acadêmica e profissional.

Aos professores Dra. Sidia Maria Callegari Jacques, Dra. Maria Inês Azambuja e Dr. Sérgio Luiz Bassanesi pela leitura e pelas importantes sugestões apresentadas para a qualificação do trabalho.

A minha esposa Susana e ao meu filho Rodrigo pelo amor, compreensão, apoio e presença constante neste período no qual necessitei dedicar-me a construção deste estudo.

Ao Diego, pela amizade e pelo auxílio principalmente na compreensão do funcionamento do WinBUGS e do TerraView.

Aos meus pais Mikiko e Minoru e irmãos Luísa e Hugo por fazerem parte da minha rede de apoio social.

A todos os funcionários, professores e principalmente aos colegas da Epidemio.

Aos colegas do Laboratório de Pesquisa em HIV/AIDS da UCS pela companhia, apoio e aprendizado.

Aos meus sócios Jéferson e Rafael da KLM Estatística e Pesquisa pela amizade e apoio.

A todos os amigos que contribuíram para o meu desenvolvimento.

SUMÁRIO

RESUMO

A taxa de mortalidade infantil é um dos indicadores mais usados para medir a qualidade de vida da população. Um dos indicadores sócio-econômico do Rio Grande do Sul é o Índice de Desenvolvimento Sócio-econômico (IDESE) da Fundação de Economia e Estatística (FEE) que tem como um de seus componentes a taxa de mortalidade infantil. Geralmente os estudos relacionam a taxa de mortalidade infantil com fatores de risco associados às áreas em estudo de forma descritiva, ou seja, de forma apenas visual através de mapas. O presente trabalho apresenta uma aplicação de um dos métodos de Epidemiologia Espacial: Estudos de Correlação Ecológica, através de modelos hierárquicos e métodos inteiramente Bayesianos, utilizando covariáveis. Os principais problemas presentes nas taxas de mortalidade brutas ou nas SMR (*Standardised Mortality Ratio*) como a auto-correlação espacial e a instabilidade dos estimadores para pequenas áreas são discutidos. Para superar estas dificuldades as estimativas do risco relativo obtidas pela análise de regressão espacial, utilizando modelagem inteiramente Bayesiana, são apresentados como alternativa, pois além de incorporar componente espacialmente estruturado ao modelo, permite também a inclusão de covariáveis. No artigo são analisados os riscos de mortalidade infantil nos 496 municípios do Rio Grande do Sul para dados acumulados entre os anos de 2001 a 2004. Foram comparados vários modelos com diferentes especificações de componente espacial e covariáveis provenientes do IDESE-FEE/2003. Verificou-se que os modelos que utilizam a estrutura espacial além de covariáveis apresentaram melhor performance, quando comparado pelo critério DIC (*Deviance Information Criterion*). Comparando as SMRs com os riscos relativos obtidos pela modelagem inteiramente Bayesiana foi possível observar um ganho substancial na interpretação e na detecção de padrões de variação no risco de mortalidade infantil nos municípios do Rio Grande do Sul.

ABSTRACT

The infant mortality rate is one of the indicators used to measure the population's life quality. The Rio Grande do Sul State has a social and economic indicator called *Índice de Desenvolvimento Sócio-econômico* (IDESE), maintained by the Economic and Statistics Foundation (FEE), which also uses the infant mortality rate. Usually, most studies relate the infant mortality rate with risk factors visually, aided by maps. This study presents the methodology and an application of one of the Spatial Epidemiology methods, the Ecologic Correlation, using Hierarchical Bayesian procedures. The main problems found in Ecologic correlations, such as the spatial autocorrelation and the estimator's instability for small areas, are discussed. To overcome these difficulties, the relative risk estimate obtained by spatial regression analysis using fully Bayesian estimation method is presented. Presently, the rate of infant mortality is analysed in all 496 municipalities of the Rio Grande do Sul State, between the years 2001 to 2004. Several models with different specifications of spatial components and different variables from the IDESE-FEE/2003 were compared. It was found that the model with spatial structure and the Education variable showed better performance than other models. With this methodology was possible to obtain a more interpretable pattern of infant mortality risk in the Rio Grande do Sul State.

APRESENTAÇÃO

Este trabalho consiste na dissertação de mestrado intitulada “Análise de Correlação Ecológica: Uma Abordagem Inteiramente Bayesiana para a Mortalidade Infantil no Rio Grande do Sul” apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul. O trabalho é apresentado em duas partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos
2. Artigo(s)

Documentos de apoio, incluindo o Projeto de Pesquisa, estão apresentados nos anexos.

1. INTRODUÇÃO

1.1 Mortalidade Infantil

A taxa de mortalidade infantil é um dos indicadores mais usados para medir a qualidade de vida da população (Maia e Souza, 2004) além de compor o Índice de Desenvolvimento Humano (IDH), divulgado pela ONU, e outros indicadores sócio-econômicos. O Rio Grande do Sul possui um indicador sócio-econômico, denominado Índice de Desenvolvimento Sócio-econômico (IDESE), da Fundação de Economia e Estatística (FEE), que também utiliza a taxa de mortalidade infantil para compor um índice referente à saúde.

A mortalidade infantil consiste no óbito de nascidos vivos durante o primeiro ano de vida, sendo a taxa de mortalidade infantil calculada pela razão entre o número de óbitos e o número total de nascidos vivos em uma região definida e em um espaço de tempo limitado.

Para facilidade de comparação entre os diferentes países ou regiões do globo esta taxa é normalmente expressa em número de óbitos de crianças com menos de um ano, a cada mil nascidos vivos. O índice considerado aceitável pela Organização Mundial da Saúde (OMS) é de 10 mortes / mil nascimentos.

No Brasil segundo dados do SINASC (Sistema de Informações sobre Nascidos Vivos) e SIM (Sistema de Informações sobre Mortalidade) a taxa de mortalidade infantil têm diminuído a cada ano, passando de 31,90 em 1997 para 22,58 em 2004. É importante salientar que em alguns estados a taxa de mortalidade é

estimada a partir de métodos demográficos indiretos, em função da baixa cobertura dos sistemas SIM e SINASC.

Segundo o ranking da UNICEF em 2006, o Brasil figura na 86ª posição entre os países com pior taxa de mortalidade até os 5 anos (Under 5 mortality rate ou U5MR), conceito definido pela OMS como a probabilidade de uma criança morrer até os 5 anos de idade. Mas nada se compara aos primeiros da lista como Angola e Afeganistão na qual a taxa de mortalidade é superior a 160 mortos para cada mil nascidos vivos.

No Brasil também existe uma grande variação na taxa de mortalidade: de acordo com dados disponíveis no DATASUS em 2004 o estado com maior taxa de mortalidade infantil foi o de Alagoas com 47,09 mortes até um ano para cada 1000 nascidos vivos. Por outro lado Santa Catarina foi o destaque positivo com apenas 13,62. Em geral, os municípios da região Norte e Nordeste do Brasil são os piores. No Rio Grande do Sul entre os anos de 1997 a 2004 observa-se uma estabilidade em relação à taxa de mortalidade, variando entre 15,10 a 17,30, porém entre os municípios é possível observar grande variabilidade.

Na literatura existem estudos propondo ações ou intervenções para a redução da mortalidade infantil, apontando que sua ocorrência é determinada, em ultima instância, por fatores sociais, econômicos e culturais - como renda, educação e posse de terra, que constituem os determinantes distais. Na maioria dos estudos estes fatores são relacionados descritivamente à mortalidade infantil. Estes fatores podem influenciar a ocorrência das causas imediatas de morte através de determinantes de nível intermediário, que incluem tanto a exposição a fatores de risco, como condições inadequadas de nutrição, saneamento, aglomeração, etc, quanto à falta de

acesso a fatores de proteção como, por exemplo, vacinas, manejo adequado das doenças infecciosas, atenção pré-natal, etc (Victora, 2001).

Se a doença é a manifestação no indivíduo, as ocorrências são manifestações no lugar (Barcellos, 2000). Os estudos que tem como foco a comparação de grupos ao invés de pessoas são ditos estudos ecológicos. Os estudos ecológicos são em geral rápidos, de baixo custo, ideais para variáveis mensuradas ao nível de área, além de úteis para levantar hipóteses. As informações utilizadas nos estudos ecológicos relacionados à área da saúde, como a mortalidade infantil, estão em grande parte disponíveis no DATASUS além de outras bases oficiais como o IBGE e FEE.

Um dos principais índices da FEE é o IDESE que consiste em um índice sintético composto por 12 indicadores divididos em quatro blocos temáticos: Educação, Renda, Saneamento e Domicílio e Saúde. Esses indicadores são transformados em índices, um para cada bloco, ou seja, o índice é o resultado da agregação dos indicadores desses blocos. O índice resulta da média ponderada dos indicadores que o compõe, além de utilizar reparametrização, tornando-o um índice que varia entre zero e um, quanto mais próximo de zero menor o desenvolvimento e quanto mais próximo de 1, maior. O quadro a seguir apresenta os blocos do IDESE, a descrição dos índices que o compõem além de apresentar os pesos utilizados para sua construção e a fonte dos dados brutos.

Quadro 1- Blocos do IDESE, índices componentes de cada bloco, peso do índice no bloco e fonte dos dados brutos.

Blocos	Índices	Peso no Bloco	Fontes dos Dados Brutos
Educação	Taxa de abandono no ensino fundamental	0,25	Edudata do INEP, Ministério da Educação
	Taxa de reprovação no ensino fundamental	0,20	Edudata do INEP, Ministério da Educação
	Taxa de atendimento no ensino médio	0,20	Censo Demográfico 2000 do IBGE; Edudata do INEP, Ministério da Educação; FEE
	Taxa de analfabetismo de pessoas de 15 anos e mais de idade	0,35	Censo Demográfico 2000 e PNAD do IBGE
Renda	Geração de renda - PIBpc	0,50	FEE.
	Apropriação de renda - VABpc do comércio, alojamento e alimentação	0,50	FEE
Condições de Saneamento e Domicílio	Percentual de domicílios abastecidos com água: rede geral	0,50	Censo Demográfico 2000 do IBGE
	Percentual de domicílios atendidos com esgoto sanitário: rede geral de esgoto ou pluvial	0,40	Censo Demográfico 2000 do IBGE
	Média de moradores por domicílio	0,10	Censo Demográfico 2000 e PNAD do IBGE;FEE
Saúde	Percentual de crianças com baixo peso ao nascer	0,33	DATASUS do Ministério da Saúde.
	Taxa de mortalidade de menores de 5 anos	0,33	DATASUS do Ministério da Saúde.
	Esperança de vida ao nascer	0,33	IDHM 2000 do PNUD, IPEA e Fundação João Pinheiro

Atualmente, boa parte dos estudos relaciona a doença/incidência/mortalidade como, por exemplo, a mortalidade infantil, com fatores de risco de forma descritiva, ou seja, de forma apenas visual através de mapas, ou simplesmente consideram áreas como indivíduos.

Para entender melhor a distribuição espacial da mortalidade infantil, alguns estudos têm sido realizados (Assunção *et al*., 1998; Santos e Noronha, 2001), inclusive no Rio Grande do Sul, através de estimação espaço temporal (Vieira, 2006).

Os estudos que relacionam a mortalidade com localizações geográficas na forma de mapa enquadram-se na área da Epidemiologia Espacial, tema que tem motivado inclusive trabalhos de conclusão de curso de estatística (Caumo, 2006; Stein, 2006).

1.2 Epidemiologia Espacial

A Epidemiologia Espacial compreende o estudo de modelos para avaliar a distribuição geográfica de taxas de incidência ou alguma outra medida epidemiológica de risco. Como exemplo, pode-se citar doenças contagiosas onde unidades mais próximas no espaço possuem maiores chances de contágio e, a medida que esta distância aumenta, o risco de contágio tende a diminuir.

A demanda por análises que envolvam informações geográficas é muito grande. A Organização Pan-americana de Saúde estima que cerca de 80% das necessidades de informações dos dirigentes políticos estão relacionados com a localização geográfica – e a utilização da Epidemiologia Espacial por parte dos pesquisadores está em ampla expansão, principalmente devido aos desenvolvimentos recentes do Sistema de Informação Geográfica (SIG), dos métodos estatísticos e da maior disponibilização de informações de saúde (Assunção, 2001).

Estes avanços possibilitam a confecção rápida de mapas que podem contribuir para formular hipóteses a respeito da distribuição espacial além de sua relação com indicadores sócio-econômicos (Rojas *et al*., 1999).

Entretanto, estudos na área da epidemiologia espacial não são muito freqüentes, em geral a epidemiologia ainda se baseia em abordagens individuais na busca de fatores de risco para doenças crônicas e não no ambiente sócio-cultural nas quais os indivíduos estão inseridos.

De acordo com a natureza da distribuição espacial dos dados observados e de acordo com o objetivo da análise, a Epidemiologia Espacial está dividida em três grandes áreas (Lawson, 2001):

1.2.1. Mapeamento de Doenças

O Mapeamento de Doenças tem como objetivo descrever e sintetizar a distribuição e a variação da taxas de mortalidade ou incidência de doenças no mapa. Os modelos fornecem estimativas para o parâmetro de interesse associado à doença em estudo e permitem geração de mapas informativos, gerando informações sintéticas e objetivas, para identificarem diferenciais de risco, realizar predições de epidemias, suspeitar da etiologia da doença, auxiliar em tomadas de decisões e levantar hipóteses.

1.2.2. Análise Ecológica

Também chamado de estudo de correlação ecológica (Elliot *et al.*, 2001), a análise ecológica tem características semelhantes ao Mapeamento de Doenças fornecendo a distribuição da doença no mapa, porém a sua grande diferença está no objetivo do estudo. O objetivo é estudar a relação entre potenciais fatores, como variáveis ambientais ou sócio-econômicas (variáveis explicativas), com a ocorrência

de doenças ou mortalidade, ao contrário do objetivo descritivo do mapeamento de doenças.

1.2.3. Detecção de Clusters de Doenças

Consiste em identificar um agrupamento geográfico com elevada ocorrência da doença, sem que haja hipóteses etiológicas pré-definidas; o objetivo principal é detectar um foco de doença no mapa. Segundo Elliott *et al.* (2001) também pode ser chamado "Avaliação de Fatores de Risco" e são utilizados quando, por exemplo, suspeita-se que uma fonte, como uma indústria nuclear, pode causar um acréscimo no risco de ocorrência da doença ou apresentar potencial perigo ambiental.

As duas primeiras áreas, Mapeamento de Doença e Análise Ecológica, muitas vezes se confundem na literatura; mais recentemente, métodos de Mapeamento de Doenças quando usados para levantar hipóteses relacionando fatores sócio-econômicos ou fatores de risco (covariáveis) são chamados de Análise de Regressão Ecológica ou Estudos de Correlação Ecológica (Elliot e Wartenberg, 2004).

Esta dissertação aborda a área de análise ecológica que também consta na literatura como estudos ecológicos, estudo de correlação ecológica, estudos geográficos ou estudos de regressão ecológica, dentre outros; mas diferentemente dos estudos descritivos, nela serão apresentados modelos de regressão ecológica utilizando métodos Bayesianos, que permitem a utilização das informações espaciais explicitamente na construção do modelo e obtenção de estimativas de interesse.

2. REVISÃO DE LITERATURA

2.1. Estudo de Correlação Ecológica

Estudos de Correlação Ecológica constituem uma área da Epidemiologia Espacial que por conseqüência é uma área da Estatística Espacial. O aprimoramento dos Estudos Ecológicos tem feito com que ela seja uma das principais técnicas para a análise de dados da saúde quando se trabalha com dados populacionais agregados ao nível de área geográfica.

O foco de estudos de correlação ecológica é relacionar medidas de covariáveis com a incidência de doenças, em nível geográfico (Lawson, 2001). Usualmente são examinadas hipóteses a respeito de fatores etiológicos da doença e o seu risco.

Nos estudos ecológicos o objetivo, em geral, não é apenas avaliar as causas de doenças, mas sim as causas da ocorrência da doença na população, caracterizada como moradores de uma mesma área geográfica, como região, município, bairro, etc. Em outras palavras o interesse dos estudos ecológicos focaliza-se não “na doença em populações, mas na doença de populações”; o objetivo é avaliar a “floresta e não as árvores” (Carvalho e Santos, 2005).

Para compreender melhor como o contexto afeta a saúde de grupos populacionais por meio de seleção, distribuição, interação, adaptação, entre outros, torna-se necessário medir efeitos em nível de grupo, uma vez que medidas em nível individual não podem dar conta destes processos. Recentes avanços metodológicos

no campo da estatística, particularmente os denominados modelos espaço-temporais Bayesianos trazem perspectivas inovadoras para a análise.

Estudos de correlação ecológica são muito similares ao mapeamento de doenças; a grande diferença é que no mapeamento de doenças o objetivo é descrever o fenômeno em estudo e, num estudo de correlação ecológica, pretende-se levantar hipóteses relacionando fatores sócio-econômico ou fatores de risco com a doença (Elliot e Wartenberg, 2004).

Richardson e Monfort (2001) indicam as restrições pelas quais os estudos de correlação ecológica são muitas vezes criticados e apontam considerações para a sua interpretação.

Restrições	Considerações
Dados geográficos são fáceis de serem obtidos, porém deve-se ter cuidado com a "qualidade dos dados";	Saber como se relaciona a variável resposta no nível individual e no nível agregado;
A escala da variável de exposição é potencialmente maior dentro da população do que em estudos individuais, que conduzem a um aumento de poder das análises;	A importância de considerar confundidores e incluí-los apropriadamente na modelagem dos dados agregados;
O efeito de medidas de erro de exposição é suavizado pelo cálculo de médias ou mediana das regiões;	Saber a influência que terão na variação espacial dos dados e a influência da não inclusão destas covariáveis no modelo
Correspondem a "experimentos naturais" onde cada fator de exposição pode trazer novas hipóteses para futuras explorações.	

A qualidade da interpretação de estudos de larga escala (nível de escala nacional ou internacional) é limitada em função de poder haver diversos confundidores propiciando viéses, mas por outro lado também existem viéses ao se analisar indicadores onde a escala é pequena.

Estes viéses podem ocorrer quando o objetivo do investigador é a análise dos indivíduos, mas não existem dados disponíveis a nível individual e utilizam-se dados de áreas, levando à possibilidade de conclusões impróprias. Este problema é denominado falácia ecológica ou viés ecológico (*ecological bias*). Freedman (1999) utiliza como exemplo o estudo realizado por Robinson (1950), que avaliam a relação entre a origem (ser estrangeiro ou não) com a alfabetização em 48 estados nos EUA, em 1930. A correlação de Pearson para os 48 pares de números é 0,53 entre o percentual de estrangeiros e o percentual de alfabetizados. Este resultado mostra uma associação positiva entre o percentual de estrangeiros e alfabetização, sugerindo que os estrangeiros sejam mais alfabetizados. Mas na realidade, a correlação feita a nível individual mostra uma relação negativa de -0,11. A correlação com os dados agregados proporciona uma inferência equivocada. Na verdade, o sinal da correlação foi positivo, porque os estrangeiros tendem a viver nos estados em que os nativos são mais letrados, ou seja, lugares mais desenvolvidos.

Para minimizar o viés ecológico grandes estudos estão sendo desenvolvidos com o uso de bases de dados mais acurados e sistemas de informação para áreas de pequena escala geográfica no centro intitulado “Small Area Health Statistics Unit (SAHSU)”, no Departamento de Epidemiologia e Saúde Pública da Faculdade de Medicina do Imperial College, na Inglaterra (Elliot *et al*, 2001).

Os indicadores ecológicos geralmente são taxas ou riscos. A avaliação geograficamente mais elementar é através da taxa bruta denotada por T_i, que para cada área é a razão entre a quantidade observada (doença/incidência) Y_i e a população exposta N_i.

$$T_i = Y_i / N_i$$

Essa razão torna possível a comparação de resultados entre áreas, sendo também possível pondera-la por outros fatores, como sexo e faixa etária. Além disso, em muitas ocasiões, além da comparação de diferentes áreas no mesmo mapa, é possível a comparação entre doenças/incidências diferentes para a mesma área, ou até mesmo a comparação da mesma doença/incidência em períodos diferentes - nestas situações são geralmente utilizadas medidas de risco relativo.

2.2. Modelagem para o Risco Relativo

Utilizando o método de máxima verossimilhança, estimativas para os verdadeiros riscos relativos de interesse são atribuídas às áreas do mapa, a partir da SMR (*Standardised Mortality Ratio*) que é a razão entre a taxa da região i e a taxa da região toda. A SMR da região i, é dada por:

$$SMR_i = R_i = \frac{y_i}{e_i}$$

onde

y_i = é a contagem de eventos na área i

$e_i = rN_i$, número esperado de ocorrência do evento na área i.

N_i = é a população em risco na área i.

A taxa global do risco em toda a área de estudo é $r = \sum y_i \Big/ \sum N_i$

A SMR varia de zero a infinito, sendo que R_i=1 indica que a área i teve tantos casos observados quanto seria esperado caso seu risco fosse idêntico ao de toda área avaliada.

A variância da SMR_i (y_i / e_i^2) é inversamente proporcional ao número esperado de casos na área i (e_i), e ao tamanho da população na área i (N_i), ou seja, quanto menor for o tamanho da população em risco, maior a incerteza na estimativa do risco, produzindo maior flutuação aleatórias (Olsen *et al*., 1996).

As flutuações aleatórias dos indicadores ecológicos causam uma instabilidade ao modelo, tornando difícil a interpretação dos mapas, e podem ser ainda maiores em estudos com baixa ocorrência, por exemplo, mortalidade segundo causa específica. Choynowski (1959), considerando a ocorrência de tumores cerebrais em condados poloneses, notou que os dois casos observados em Lesko com população de 17000 habitantes, geravam uma taxa bruta extrema de 11,8 por 100 mil. Se tivesse ocorrido apenas um caso, ao invés dos dois observados, essa taxa seria de 5,9 por mil - um valor consistente com a taxa de outros condados. As instabilidades das taxas brutas em áreas com pequena população, assim como as das SMRs, fazem com que os mapas apresentem taxas distintas e extremas devido à flutuação aleatória, sem associação com o risco subjacente. Em termos estatísticos, estas taxas/riscos entre as áreas não são comparáveis já que possuem variâncias muito diferentes (Assunção *et al*., 1998). Dessa forma, na questão da mortalidade infantil, por exemplo, é provável que os municípios ou as áreas com maior taxa de mortalidade tenham também poucos nascimentos. Uma taxa de mortalidade obtida simplesmente dividindo o total de óbitos infantis pelo número de nascidos vivos não é o melhor estimador. Na literatura basicamente três alternativas são propostas:

2.2.1. Agregar áreas

Agregar áreas é uma alternativa para contornar a instabilidade quando as áreas são pequenas ou com pequena população, mas neste caso muita cautela é exigida na realização das inferências (Druck *et al*., 2004). Isto contraria um dos objetivos principais de se fazer mapas que é o de fornecer resolução geográfica adequada.

2.2.2. Mapas de probabilidade

É um procedimento proposto por Choynowski (1959) para evitar efeitos drásticos quando pequenas mudanças no número de ocorrências são registradas, através da substituição das taxas por probabilidades similares ao valor p de um teste. A idéia básica também é permitir comparações de áreas através da padronização das taxas em uma escala de probabilidade, mas isto não é possível se alguma população em risco for grande, pois ocorrerão valores extremos de probabilidade, não sendo possível detectar pequenos afastamentos. Além disso, a interpretação epidemiológica não é clara (Cressie, 1991).

Apesar de tentarem controlar a instabilidade dos estimadores apresentados na modelagem clássica do risco relativo, as alternativas acima propostas não consideram a possibilidade de autocorrelação espacial.

2.2.3. Modelagem Bayesiana

Na modelagem Bayesiana no contexto de estudos ecológicos, uma informação *a priori* é atribuída para o risco relativo das áreas, e as estimativas posteriores são o resultado da combinação entre estimativa *a priori* e a

verossimilhança, informações oriundas dos dados. Esta modelagem tem tido maior destaque na literatura como uma boa alternativa para superar a instabilidade dos indicadores em áreas com população em risco pequena. A modelagem Bayesiana possibilita que as informações derivadas de áreas vizinhas sejam utilizadas na construção das estimativas, ou seja, incorpora a autocorrelação espacial ao permitir uma conexão de áreas, o que não existe na modelagem clássica. Nas situações citadas, os métodos Bayesianos podem ser empíricos ou inteiramente Bayesianos. O estimador Bayesiano empírico (Marshall, 1991; Bernardinelli e Montonolli, 1992) produz boas estimativas quando comparado aos métodos inteiramente Baysesianos (Besag *et al*, 1991), porém, apresenta desvantagens, pois parte da hipótese de que a distribuição da variável aleatória é a mesma para todas as áreas e que a média e a variância para cada uma das áreas sejam iguais, o que nem sempre é uma suposição realista, pois os dados de saúde geralmente são muito heterogêneos. Além disso, não pode ser utilizado em situações mais complexas como inclusão de covariáveis e a interação entre espaço e tempo (Assunção, 2001). Portanto dentre as alternativas apresentadas será utilizada apenas a modelagem inteiramente Bayesiana.

2.3 Análise de Regressão Espacial em Estudos Ecológicos

Os modelos de regressão espacial empregado em estudos ecológicos, além de serem ferramentas básicas na análise de eventos de saúde, podem ser utilizados para avaliar a necessidade de alocação de recursos para a saúde de acordo com a variação geográfica das estimativas. O foco é relacionar covariáveis, ou variáveis potencialmente explicativas, com a incidência da doença em nível geográfico, seja no

campo da análise exploratória ou buscando modelos explicativos a fim de apontar medidas preventivas (Lawson, 2001; Carvalho e Santos, 2005).

O modelo de regressão tradicional, que não leva em conta a estrutura espacial, em termos matriciais é dado por:

$$Y = X\beta + \varepsilon, \varepsilon \sim N\left(\underset{\sim}{0}, I\sigma^2\right),$$

ou

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k-1} \\ 1 & X_{21} & \cdots & X_{2k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

onde Y é um vetor de tamanho $(n \times 1)$ de variáveis respostas medidas em cada uma das n áreas analisadas. A matriz X de dimensão $(n \times k)$ representa os valores das $(k-1)$ variáveis aleatórias explicativas, β é o vetor $(n \times 1)$ de parâmetros de regressão a serem estimados e ε é o vetor $(n \times 1)$ de resíduos do modelo.

Quando se faz uma análise de regressão, além de determinar que variáveis explicativas contribuam de forma significativa para este relacionamento linear, deseja-se encontrar um bom ajuste entre os valores preditos pelo modelo e os valores observados da variável dependente. Para tanto, a hipótese principal é a de que as observações não são correlacionadas, e, conseqüentemente, que os resíduos (diferença entre os valores observados e preditos) também sejam independentes e não-correlacionados com a variável dependente, além de terem variância constante e distribuição *Normal* com média zero (Druck *et al*., 2004).

No caso de dados espaciais, a hipótese de independência das observações geralmente é falsa, os resíduos continuam apresentando a autocorrelação espacial

presente nos dados. A dependência entre as observações altera o poder explicativo do modelo, podendo produzir associações erradas (Bailey e Gattrel, 1995). Nos dados espaciais “a dependência está presente em todas as direções e fica mais fraca à medida que aumenta a dispersão na localização dos dados” (Cressie,1991).

Além da dependência espacial, outra questão essencial na modelagem da regressão espacial é a flutuação aleatória dos indicadores ecológicos o que também já foi discutido anteriormente.

Uma proposta adequada é considerar as amostras como realizações de processos estocásticos, ao invés de considerá-las independentes, utilizando-se assim todas as observações de forma conjunta para responder ao fenômeno estudado (levando em conta a estrutura de dependência espacial). Os modelos devem apresentar estrutura espacial que considere os efeitos da média do processo no espaço assim como o da covariância entre duas áreas adjacentes. Considerações importantes neste tipo de estudo são a estacionariedade e a isotropia. O processo é considerado estacionário se ambos os efeitos são constantes em toda a região estudada, não apresentando tendência. Um processo é isotrópico se, além de estacionário, a covariância depende somente da distância entre os pontos e não da direção entre eles. Nos modelos de regressão espacial, as informações derivadas de áreas vizinhas são incorporadas ao modelo a partir da especificação da matriz de vizinhança.

2.3.1 Análise de Regressão Utilizando Modelos Inteiramente Bayesianos

Para superar as dificuldades de instabilidade das estimativas e autocorrelação espacial na variável resposta, várias modelagens têm sido propostas na literatura, porém as que apresentam melhores resultados são os métodos de estimação inteiramente Bayesianos (Assunção, 2001). Estes modelos também são conhecidos na literatura como modelos Bayesianos hierárquicos espaciais (Bernardinelli e Montomoli, 1992) ou simplesmente como métodos de suavização Bayesiana.

Supõe-se que o número de eventos observados em cada área possui distribuição Binomial, mas como a maioria dos dados epidemiologicamente mapeados são raros ou com grande variação das taxas/riscos entre diferentes áreas, o modelo Binomial pode ser aproximado pela distribuição *Poisson* (Richardson at. al., 2004), Assim, se o número de eventos observados (Y_i) se refere ao número de óbitos no município i.

$$Y_i \sim Poisson(\mu_i)$$

onde $\mu_i = E_i\theta_i$, sendo θ_i o risco relativo de óbito infantil na i-ésima área e E_i a quantidade esperada de óbitos infantis na i-ésima área sob a hipótese de que o risco seja constante em todas as áreas e igual ao risco geral da região. Sendo assim, o primeiro nível hierárquico do modelo é dado por:

$$Y_i | E_i \sim Poisson(E_i\ \theta_i)$$

No segundo nível do modelo são especificados os componentes utilizados na estimação do logaritmo do risco θ_i.

$$\log(\theta_i) = \beta_0 + \beta_p X_{ip} + u_i + v_i$$

onde β_0 é uma constante, $\beta_p = \{\beta_1, ..., \beta_k\}$ é um vetor de constantes que, quando colocados na função exponencial representam os efeitos de cada uma das k covariáveis no $log(\theta_i)$, X_{ip} é um vetor de covariáveis e u_i e v_i são vetores de efeitos aleatórios.

O modelo é construído de modo a relacionar os componentes com o logaritmo de θ_i e não diretamente com o risco θ_i, pelas inúmeras vantagens matemáticas e computacionais que a transformação logarítmica pode proporcionar, pois a distribuição *Poisson* é uma distribuição de probabilidade enquadrada na família de distribuições *Exponenciais*.

Depois de obtidas as estimativas de $log(\theta_i)$ aplica-se a função exponencial para a obtenção das estimativas do risco θ_i.

No terceiro nível hierárquico do modelo estão as distribuições *a priori* para cada um dos parâmetros do modelo β_0, β_p, u_i e v_i.

Em geral à β_0 se atribui uma distribuição do tipo priori $Uniforme(-\infty ; +\infty)$, e a cada um dos elementos do vetor β_p se atribui independentemente uma distribuição *Normal* com média igual a zero e parâmetro de dispersão $\boldsymbol{\tau}_\beta$ o menor possível, de modo que estas distribuições *a priori* sejam pouco informativas.

Quanto aos componentes aleatórios, a definição da distribuição *a priori* apresenta diferenças em relação aos componentes já expostos.

O componente u_i representa um efeito aleatório não espacial, que pode ser visto como particularidades de cada área, ou seja, efeitos de pequena escala que não ultrapassam as fronteiras das áreas.

A cada um dos componentes u_i se atribui independentemente uma distribuição *a priori* do tipo *Normal* com média zero e parâmetro de precisão τ_u. Sendo assim o vetor $(u_1,...,u_n)$ segue uma distribuição *Normal Multivariada*, composta de termos independentes e depende de um único parâmetro τ_u, em geral desconhecido e denominado Hiperparâmetro, já que é um parâmetro de uma distribuição *a priori*.

A este Hiperparâmetro τ_u também se deve atribuir uma distribuição *a priori*, denominada Hiperpriori. Geralmente para parâmetros de dispersão ou precisão se atribui uma distribuição *Gamma* com parâmetros *a* e *b* atribuídos de forma que a Hiperpriori seja pouco informativa, ou seja, uma distribuição com grande variabilidade.

Já o componente v_i incorpora a estrutura espacial, captando a influência das áreas vizinhas. Reflete os efeitos de larga escala através de uma distribuição *a priori* espacialmente estruturada através da definição de uma matriz de vizinhança entre as áreas, denotada por $W_{(n\times x)}^{(1)}$ que pode ser definida de diversas formas:

Para um conjunto de *n* áreas $(A_1,\ldots,A_n)$, construímos uma matriz $W_{(n\times x)}^{(1)}$, onde cada um dos elementos W_{ij} representa uma medida de proximidade entre A_i e A_j.

Esta medida de proximidade pode ser calculada a partir de diversos critérios, sendo os mais utilizados:

1. Uma matriz binária, mais simples e comumente utilizada na literatura, que também será utilizada nesta dissertação: considera apenas o fato das áreas fazerem fronteira ou não, assumindo W_{ij} =1 caso i e j sejam regiões vizinhas ou W_{ij} =0 , caso contrário;

2. Uma matriz que considera maior peso (maior influência) para tamanhos de fronteira maiores, ou seja, baseia-se nos tamanhos das fronteiras. Por exemplo, W_{ij} pode ser igual ao tamanho, em km, da fronteira entre as áreas i e j. Esta opção não é muito utilizada pela dificuldade de se medir a fronteira;

3. Uma matriz W_{ij} que é uma extensão da segunda alternativa e considera não só o tamanho das fronteiras como a presença de barreiras naturais. Por exemplo, dependendo da variável de estudo, pode ser razoável considerar que duas áreas não sejam tão correlacionadas quando entre elas existe uma montanha, um rio ou qualquer outra característica geográfica que possa interferir no deslocamento dos indivíduos (Mollié, 1996).

É necessário destacar que diferentes especificações na matriz de vizinhança gerarão diferentes estimativas, sendo uma importante competência do pesquisador definir a matriz de vizinhanças mais adequada ao contexto do problema. Em muitos casos diferentes matrizes de vizinhança são testadas para verificar qual delas melhor se ajusta à realidade dos dados.

Voltando à especificação da distribuição *a priori*, ao componente v_i se atribui uma distribuição denominada CAR (Condicional Auto-Regressiva) Normal:

$$v_i \mid v_j \sim Normal\left(\frac{\sum_j W_{ij} v_j}{\sum_j W_{ij}}, \frac{\left(\frac{1}{\tau_v}\right)}{\sum_j W_{ij}}\right)$$

Assim sendo o efeito espacial médio da i-ésima área (v_i) é dado pela média ponderada (W_{ij} é a matriz de pesos) dos efeitos dos seus vizinhos e a variância é inversamente proporcional à soma dos pesos das áreas consideradas vizinhas. Pode-se notar que as estimativas de v_i possuem maior precisão à medida que a soma dos pesos aumenta.

Em particular, quando se utiliza a estrutura de vizinhança binária, a qual é a estrutura mais comumente utilizada, o efeito espacial médio da i-ésima área v_i é dado pela média aritmética dos efeitos dos seus vizinhos e a variância é inversamente proporcional à quantidade de áreas vizinhas. Sendo assim, quanto maior o número de vizinhos, maior é a precisão da estimativa de v_i.

Além disso, a especificação completa da distribuição CAR Normal depende de um único parâmetro τ_v (o inverso da variância de v_i) ao qual também deve ser atribuída uma distribuição *a priori Gamma*, analogamente a distribuição *a priori* para τ_u.

A distribuição conjunta condicional *a priori* do efeito espacial v é dada por:

$$(v \mid \tau_v) \propto \frac{1}{\left(\frac{1}{\tau_v}\right)^n} \left\{ -\frac{1}{2\left(\frac{1}{\tau_v}\right)^2} \sum_i \sum_j W_{ij} \left(v_i - v_j\right)^2 \right\},$$

Esta é uma distribuição *a priori* imprópria já que é baseada nas diferenças pareadas entre os v_i (Schmidt *et al*., 2002) e, como distribuições *a priori* impróprias podem levar a distribuições *a posteriori* impróprias, na prática se impõe uma restrição para que esses efeitos v_i somem zero. Contudo a estrutura espacial pode ser atribuída à distribuição *a priori* de diversas formas.

A inferência Bayesiana será baseada na distribuição *a posteriori* de θ:

$$\pi(\theta \mid y) \propto l(y_1,...,y_n \mid \theta)\pi(\log(\theta_i)),$$

onde $l(y_1,...,y_n \mid \theta)$ é a função de verossimilhança e a distribuição *a priori* é $\pi(\theta)$ para o $\log(\theta_i)$.

A distribuição *a posteriori* para θ será proporcional a:

$$\left(\prod_{i=1}^{n} \frac{(\theta_i N_i)^{y_i}}{y_i!} \exp\{-\theta_i N_i\}\right) \pi(u_1,\ldots,u_n \mid \tau_u)\pi(v_1,\ldots,v_n \mid \tau_v)\pi(\beta_{11},\ldots,\beta_{1n} \mid \tau_\beta)\pi(\tau_u)\pi(\tau_v)\pi(\tau_\beta)$$

Assim temos uma distribuição *a posteriori* que não pode ser encontrada analiticamente. Desta forma é necessário utilizar métodos de simulação estocástica chamados MCMC (Markov Chain Monte Carlo).

Simulando uma quantidade suficiente de valores para os hiperparâmetros, podemos conhecer toda a distribuição *a posteriori* para o $\log(\theta_i)$, assim como qualquer característica da distribuição conjunta *a posteriori* dos parâmetros e

hiperparâmetros. Para as simulações podem ser usados diversos métodos MCMC que são encontrados em Gamerman (1997).

A inferência *a posteriori* baseada nas amostras simuladas pelo procedimento MCMC segue métodos simples como o método de momentos, justificando sua validade através de argumentos frequentistas tais como a Lei dos Grandes Números (Assunção, 2001).

Um software que atualmente está sendo utilizado para modelagem de dados com enfoque Bayesiano, utilizando métodos MCMC, é o **WinBUGS** (Win Bayesian inference Using Gibbs Sampling). É um Software livre implementado por Thomas *et al*. (1992) que possui uma biblioteca para análise de dados espaciais e que será utilizado na aplicação dos dados neste trabalho.

Utilizando-se a metodologia Bayesiana em estudos de correlação ecológica se deve fazer a análise dos parâmetros do modelo, para encontrar fatores realmente explicativos para o desfecho em estudo. Neste contexto avaliaremos as distribuições *a posteriori* para os parâmetros de regressão do segundo nível do modelo, o uso de critérios para escolha do modelo, e a análise das taxas suavizadas. Um critério utilizado é o *Deviance Information Criterion* (DIC), uma forma semelhante ao critério *Akaike's Information Criterion* (AIC) utilizado em modelos de regressão "clássica". O DIC é usado quando se modelam os dados através de MCMC, além de ser amplamente utilizado para comparar modelos com diferentes níveis de complexidade (Spiegelhalter *et al*. 2002).

O DIC assume que a média *a posteriori* é uma boa estimativa dos parâmetros do modelo. Nos casos em que estas distribuições são multimodais e nos casos em que existe acentuada assimetria, ele não é recomendado.

O modelo que apresentar o menor DIC é considerado como aquele que melhor pode predizer um novo conjunto de dados com a mesma estrutura dos dados observados.

3. REFERÊNCIAS BIBLIOGRÁFICAS

Assunção, R.M.; Barreto, S.M; Guerra, H.L.; Sakura, E. (1998). Mapas de taxas epidemiológicas: uma abordagem Bayesiana. *Cadernos de Saúde Pública*, Rio de Janeiro, 14(4):713-723.

Assunção, R.M. (2001). *Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia*. São Carlos, ABE. Disponível em http://www.est.ufmg.br/leste/publicacoes.htm. Data de acesso: 08/05/2006.

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

Barcellos, C. (2000). Organização espacial, saúde e qualidade de vida. In: Seminário Nacional Saúde e Ambiente no Processo de Desenvolvimento (Fundação Oswaldo Cruz, org.), Série *Fiocruz: Eventos Científicos* 2, pp. 27-34, Rio de Janeiro: Fundação Oswaldo Cruz.

Bernardinelli, L; Montomoli, M. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 11, 983-1007.

Besag, J.; York, J.; Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Carvalho, M.S.; Santos, R.S. (2005). Análise de dados espaciais em Saúde Pública: métodos, problemas, perspectivas. *Cadernos de Saúde Pública*, Rio de Janeiro, 21(2): 361-378.

Caumo, R.B.; (2006). *Estatística Espacial em Dados de Área:Uma Modelagem Inteiramente Bayesiana para o Mapeamento de Doenças Aplicada à Dados Relacionados com a Natalidade em Mulheres Jovens de Porto Alegre*. Monografia de Conclusão do Bacharelado em Estatística. Intituto de Matemática, UFRGS.

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

Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

Druck, S.; Carvalho, M.S. ; Câmara, G.; Monteiro, A.M.V. (2004). *Análise Espacial de Dados Geográficos*. Brasília: Embrapa.

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

Elliott, P.; Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environ Health Perspect*; 112:998-1006.

Freedman, D.A. (1999). Ecological Inference and the Ecological Fallacy. *International Encyclopedia of the Social & Behavior Sciences Technical Report*, California, 549.

Gamerman, D. (1997). *Markov Chain Monte Carlo:. Stochastic Simulation for Bayesian Inference*. London: Chapman&Hall.

Lawson A. (2001). *Statistical Methods in Spatial Epidemiology*. Sussex: John Wiley & Sons.

Maia, S.F.; Sousa, T.R.V. (2004). Uma investigação dos determinantes da redução da taxa de mortalidade infantil nos estados da Região Nordeste do Brasil. In: I Congresso da Associação Latino Americana de População - ALAP, 2004, Caxambu.

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

Mollié, A. (1996). Bayesian mapping of disease. In: W.R. Gilks, S. Richardson and D.J. Spiegelhalter (editors). *Monte Carlo Markov Chain in Practice*. London: Chapman and Hall.

Olsen, S.; Martuzzi, M.; Elliott, P. (1996). Cluster analysis and disease mapping-Why, when, how? A step by step guide. *British Medical Journal*, 313, 863-865.

Richardson, S.; Monfort, C. (2001). Ecological correlation studies. In: Elliot, P.; Wakefield, J.; Best, N.; Briggs, D. *Spatial Epidemiology: Methods and Applications*. New York; Wiley, p. 205-220.

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

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, Los Angeles, 15: 351-57.

Rojas, L. I.; Barcellos, C. & Peiter, P., (1999). Utilização de mapas no campo da epidemiologia no Brasil. *Informe Epidemiológico do SUS*, 8:25-35.

Santos, S.M.; Noronha, C.P. (2001). Padrões espaciais de mortalidade e diferenciais sócio-econômicos na cidade do Rio de Janeiro. *Cadernos de Saúde Pública*, Rio de Janeiro, 17(5):1099-1110.

Schmidt, A.M.; Nobre, A.A.; Ferreira, G.S. (2002). Alguns aspectos da modelagem de dados espacialmente referenciados, *Revista Brasileira de Estatística*, vol. 63, n.220, pp. 59-88.

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

Stein, M.C.; (2006). *Estudos de correlação ecológica:definição, métodos, problemas e aplicação*. Monografia de Conclusão do Bacharelado em Estatística. Intituto de Matemática, UFRGS.

Thomas, A.; Spiegelhalter, D.J.; Gilks, W.R. (1992). BUGS: A program to perform Bayesian inference using Gibbs Sampling. *Bayesian Statistics 4*, 837-842.

Victora C.G. (2001). Intervenções para reduzir a mortalidade infantil pré-escolar e materna no Brasil. *Revista Brasileira de Epidemiologia*, São Paulo, v. 4, p. 3-69.

Vieira, D.M.; (2006). *Análise Espaço-Temporal na Estimação de Taxas de Incidência/ Mortalidade*. Dissertação de Mestrado em Epidemiologia, Faculdade de medicina, Universidade Federal do Rio Grande do Sul.

4. OBJETIVOS

4.1 Objetivo Geral

O objetivo geral deste trabalho é apresentar Modelos de Regressão Espacial Bayesianos para Estudos de Correlação Ecológica para estimação do risco relativo com covariáveis, além de mostrar a utilização de mapas para melhor visualização desta distribuição.

4.2 Objetivos Específicos

(1) Ilustrar os métodos através da utilização dos dados espaciais de mortalidade infantil nos municípios do Rio Grande do Sul entre os anos de 2001 a 2004 e relacioná-los com o Índice de Desenvolvimento Sócio-econômico de 2003 da Fundação de Economia e Estatística Siegfried Emanuel Heuser (IDESE-FEE).

(2) Comparar a modelagem clássica do risco relativo (SMR) com o melhor modelo obtido pelo método de Regressão Espacial Bayesiano.

(3) Apresentar as estimativas de risco através de mapas.

5. ARTIGO

Utilização da Modelagem Inteiramente Bayesiana na Detecção de Padrões de Variação de Risco Relativo de Mortalidade Infantil no Rio Grande do Sul

Sérgio Kakuta Kato, Programa de Pós-Graduação em Epidemiologia, UFRGS;

Jandyra Maria Guimarães Fachel, Pós-Graduação em Epidemiologia, UFRGS e Departamento de Estatística, Instituto de Matemática, UFRGS

A ser enviado para: ***Cadernos de Saúde Pública*****.**

Utilização da Modelagem Inteiramente Bayesiana na Detecção de Padrões de Variação de Risco Relativo de Mortalidade Infantil no Rio Grande do Sul

Sérgio Kakuta Kato[1]

Jandyra Maria Guimarães Fachel[1,2]

1 – Programa de Pós-Graduação em Epidemiologia da Faculdade de Medicina da Universidade Federal do Rio Grande do Sul.

Rua Ramiro Barcelos, 2600, sala 419, Porto Alegre, RS, Brasil, CEP:90035-003

klmsergio@terra.com.br

2 – Departamento de Estatística, Instituto de Matemática, UFRGS

Resumo

A taxa de mortalidade infantil é um dos indicadores mais usados para medir a qualidade de vida da população. Em grande parte dos estudos ecológicos a análise da associação entre as covariáveis e a taxa de mortalidade infantil é realizada apenas descritivamente, comparando mapas que apresentam a taxa de mortalidade infantil com mapas que apresentam a distribuição espacial das covariáveis. Os Estudos de Correlação Ecológica tem como foco relacionar covariáveis ou variáveis potencialmente explicativas, com a incidência da doença ou com a variabilidade de taxas de mortalidade a nível geográfico.

Neste artigo são analisados os fatores possivelmente associados à mortalidade infantil nos 496 municípios do Rio Grande do Sul, através de dados acumuladas entre os anos de 2001 a 2004, obtidos pela análise de regressão utilizando modelagem inteiramente Bayesiana como alternativa para superar a auto correlação espacial e a instabilidade dos estimadores clássicos como a Taxa Bruta e a SMR (*Standardised Mortality Ratio*). Foram comparados 9 modelos com diferentes especificações de componente espacial e covariáveis, provenientes dos blocos do Índice de Desenvolvimento Sócio-econômico da Fundação de Economia e Estatística (IDESE-FEE/2003). Verificou-se que o modelo que utiliza a estrutura espacial além da covariável Educação apresenta melhor desempenho, quando comparado pelo critério DIC (*Deviance Information Criterion*). Comparando as estimativas das SMRs com os riscos relativos obtidos pela modelagem inteiramente Bayesianas, foi possível observar um ganho substancial na interpretação e na detecção de padrões de variação do risco de mortalidade infantil nos municípios do Rio Grande do Sul, ao utilizar esta modelagem. A região da Serra Gaúcha destacou-se com baixo risco relativo e estimativas muito homogêneas.

Palavras chave: Epidemiologia Espacial; Modelos Inteiramente Bayesianos; Risco de Mortalidade Infantil; Correlação Ecológica; MCMC; WinBUGS.

ABSTRACT

The infant mortality rate is one of the indicators used to measure the population's life quality. The Rio Grande do Sul State has a social and economic indicator called *Índice de Desenvolvimento Sócio-econômico* (IDESE), maintained by the Economic and Statistics Foundation (FEE), which also uses the infant mortality rate. Usually, most studies relate the infant mortality rate with risk factors visually, aided by maps. This study presents the methodology and an application of one of the Spatial Epidemiology methods, the Ecologic Correlation, using Hierarchical Bayesian procedures. The main problems found in Ecologic correlations, such as the spatial autocorrelation and the estimator's instability for small areas, are discussed. To overcome these difficulties, the relative risk estimate obtained by spatial regression analysis using fully Bayesian estimation method is presented. Presently, the rate of infant mortality is analysed in all 496 municipalities of the Rio Grande do Sul State, between the years 2001 to 2004. Several models with different specifications of spatial components and different variables from the IDESE-FEE/2003 were compared. It was found that the model with spatial structure and the Education variable showed better performance than other models. With this methodology was possible to obtain a more interpretable pattern of infant mortality risk in the Rio Grande do Sul State.

Key words: Spatial Epidemiology; Fully Bayesian Method; Infant Mortality Rate; Ecologic Correlation; MCMC.

Introdução

A taxa de mortalidade infantil é um dos indicadores mais usados para medir a qualidade de vida da população[1], além de compor o Índice de Desenvolvimento Humano (IDH), divulgado pela ONU, e outros indicadores sócio-econômicos. O Rio Grande do Sul possui um indicador sócio-econômico próprio, denominado Índice de Desenvolvimento Sócio-econômico (IDESE) da Fundação de Economia e Estatística (FEE) que também utiliza a taxa de mortalidade infantil para compor um índice referente à saúde. Nos estados do Brasil a taxa de mortalidade é obtida a partir de dados do SINASC (Sistema de Informações sobre Nascidos Vivos) e SIM (Sistema de Informações sobre Mortalidade), porém, é importante salientar que em alguns estados, a taxa de mortalidade é estimada a partir de métodos demográficos indiretos, em função da baixa cobertura dos sistemas SIM e SINASC.

No Brasil existe uma grande variação na taxa de mortalidade: de acordo com dados disponíveis no DATASUS, em 2004 o estado com maior taxa de mortalidade foi Alagoas, com 47,09 mortes até um ano de idade para cada 1000 nascidos vivos, por outro lado, Santa Catarina foi o destaque positivo com apenas 13,62 mortes para cada 1000 nascidos vivos. Em geral os municípios da região Norte e Nordeste do Brasil são os que apresentam as piores taxas. No Rio Grande do Sul entre os anos de 1997 a 2004 observou uma estabilidade em relação à taxa de mortalidade, que variou entre 15,10 e 17,30, porém é possível observar grande variabilidade entre municípios.

Existem estudos propondo ações ou intervenções para a redução da mortalidade infantil, apontando que sua ocorrência é determinada em ultima instância por fatores sociais, econômicos e culturais - como renda, educação e posse de terra[2].

Os estudos que têm como foco a comparação de grupos ao invés de pessoas são denominados ecológicos. Os estudos ecológicos são ideais para dados de área, como taxas por área geográfica, além de serem úteis para levantar hipóteses. As informações utilizadas nos estudos ecológicos relacionados à área da saúde, como a mortalidade infantil, estão em grande parte disponíveis no DATASUS, além de outras bases oficiais como as do IBGE e da FEE. Os estudos que utilizam modelos para avaliar a distribuição geográfica de taxas de mortalidade ou de incidência ou alguma outra medida epidemiológica de risco se enquadram na área da Epidemiologia Espacial.

A demanda por análises que envolvam informações geográficas é muito grande. A Organização Pan-americana de Saúde estima que cerca de 80% das necessidades de informações dos dirigentes políticos estão relacionados com a localização geográfica. Neste mesmo sentido a utilização da Epidemiologia Espacial por parte dos pesquisadores está em ampla expansão, principalmente devido aos desenvolvimentos recentes do Sistema de Informação Geográfica (SIG), da maior disponibilidade de informações de saúde e dos avanços dos métodos estatísticos, destacando-se a inferência Bayesiana, a qual ainda é pouco empregada em aplicações na área da saúde[3]. Estes avanços possibilitam a confecção de mapas apresentando estimativas mais precisas dos indicadores de saúde, bem como a obtenção de modelos capazes de mensurar o efeito de cováriaveis, que podem contribuir para formular hipóteses a respeito da distribuição espacial destes indicadores, além de avaliar sua relação com indicadores sócio-econômicos[4].

Em grande parte dos estudos ecológicos a análise da associação entre as covariáveis e a taxa de mortalidade infantil é realizada apenas descritivamente, ou utilizando técnicas de modelagem muitas vezes inadequadas.

A avaliação descritiva, apenas visual, é bastante fácil de realizar e consiste em comparar mapas que apresentam a taxa de mortalidade infantil por município ou por bairro com mapas que apresentam a distribuição espacial das covariáveis. A grande dificuldade neste tipo de análise é interpretar o mapa que apresenta a taxa de mortalidade infantil, pois este indicador geralmente tem alta variabilidade (pouca precisão) em regiões com população reduzida, apresentando em muitos casos valores extremamente superiores ou inferiores à taxa ou risco médio da área como um todo, sem que estas regiões realmente se caracterizem como regiões de alto ou baixo risco. Sendo assim, estes mapas podem não representar o verdadeiro processo subjacente a estes indicadores[5,6].

Outras vezes são utilizados modelos de regressão linear para relacionar as covariáveis à taxa de mortalidade, ou seja, considera-se as áreas como indivíduos. O grande problema neste tipo de análise, que também é bastante acessível e por isso muito utilizada, reside no fato de que as unidades amostrais (municípios ou bairros) geralmente não podem ser considerados independentes e, sendo assim, os resíduos do modelo continuam apresentando a autocorrelação espacial presente nos dados. Nos dados espaciais, áreas vizinhas são mais similares do que áreas mais distantes. A interdependência entre as unidades amostrais altera o poder explicativo do modelo, podendo produzir associações errôneas[7].

Para superar as dificuldades de instabilidade das estimativas e autocorrelação espacial na variável resposta, várias modelagens têm sido propostas, porém as que

apresentam melhores resultados são os métodos de estimação inteiramente Bayesianos[9]. Nessa linha, para entender melhor a distribuição espacial da mortalidade infantil, alguns estudos foram já realizados em Minas Gerais[10], e no Rio de Janeiro[11]. No Rio Grande do Sul, estes métodos foram utilizados em uma estimação espaço-temporal[12].

O objetivo deste artigo é demonstrar a utilização de modelos de regressão espacial inteiramente Bayesianos na detecção de padrões de variação, através do mapeamento do risco de mortalidade infantil nos municípios do Rio Grande do Sul, acumulado para os anos de 2001 a 2004, utilizando como covariáveis sub-escalas do Índice de Desenvolvimento Sócio-econômico de 2003 da Fundação de Economia e Estatística Siegfried Emanuel Heuser (IDESE-FEE).

Metodologia

Dados Utilizados

As informações sobre o número de nascimentos e óbitos em recém-nascidos nos 496 municípios do Rio Grande do Sul entre os anos de 2001 e 2004 foram obtidas no site do DATASUS.

O IDESE-FEE produz um índice sintético composto por 12 indicadores divididos em quatro blocos temáticos: Educação, Renda, Saneamento e Domicílio e Saúde. Esses indicadores são transformados em índices, um para cada bloco, ou seja, o índice é o resultado da agregação dos indicadores desses blocos. O índice resulta da média ponderada dos indicadores que o compõe. Devido a reparametrização, é um índice que varia entre zero e um: quanto mais próximo de zero menor o desenvolvimento e quanto mais próximo de 1 maior o desenvolvimento. Utilizaremos como covariáveis os índices relativos aos blocos temáticos Educação, Renda e Saneamento e Domicilio. Não foi utilizado o bloco Saúde visto que a taxa de mortalidade infantil compõe este bloco. O quadro a seguir apresenta os blocos do IDESE, a descrição dos índices que o compõem além de apresentar os pesos utilizados para sua formulação e a fonte dos dados brutos.

Quadro 1- Blocos do IDESE, índices componentes de cada bloco, peso do índice no bloco e fonte dos dados brutos.

Blocos	Índices	Peso no Bloco	Fontes dos Dados Brutos
Educação	Taxa de abandono no ensino fundamental	0,25	Edudata do INEP, Ministério da Educação
	Taxa de reprovação no ensino fundamental	0,20	Edudata do INEP, Ministério da Educação
	Taxa de atendimento no ensino médio	0,20	Censo Demográfico 2000 do IBGE; Edudata do INEP, Ministério da Educação; FEE
	Taxa de analfabetismo de pessoas de 15 anos e mais de idade	0,35	Censo Demográfico 2000 e PNAD do IBGE
Renda	Geração de renda - PIBpc	0,50	FEE.
	Apropriação de renda - VABpc do comércio, alojamento e alimentação	0,50	FEE
Condições de Saneamento e Domicílio	Percentual de domicílios abastecidos com água: rede geral	0,50	Censo Demográfico 2000 do IBGE
	Percentual de domicílios atendidos com esgoto sanitário: rede geral de esgoto ou pluvial	0,40	Censo Demográfico 2000 do IBGE
	Média de moradores por domicílio	0,10	Censo Demográfico 2000 e PNAD do IBGE;FEE
Saúde	Percentual de crianças com baixo peso ao nascer	0,33	DATASUS do Ministério da Saúde.
	Taxa de mortalidade de menores de 5 anos	0,33	DATASUS do Ministério da Saúde.
	Esperança de vida ao nascer	0,33	IDHM 2000 do PNUD, IPEA e Fundação João Pinheiro

Considerando o risco de mortalidade infantil como variável resposta no modelo de regressão, neste artigo serão avaliados os 9 modelos a seguir: Sem efeito espacial e controlando por educação, renda e condições de saneamento e domicílio, com efeito espacial e sem covariáveis, com efeito espacial controlando por educação, com efeito espacial controlando por saneamento e domicílio, com efeito espacial controlando por renda, com efeito espacial controlando por educação e renda, com efeito espacial controlando por educação e saneamento, com efeito espacial controlando por renda e saneamento, e com efeito espacial controlando por educação, renda e saneamento.

Análise Estatística

A análise de regressão utilizando modelos inteiramente Bayesianos é própria para a modelagem dos diferentes tipos de efeitos espaciais, temporais e espaço-temporais. Nessa modelagem supõe-se que o número de eventos observados em cada área possui distribuição Binomial. Como a maioria dos dados epidemiológicos mapeados são raros ou com grande variação das taxas/riscos entre diferentes áreas, o modelo Binomial pode ser aproximado pela distribuição *Poisson*[13], com número de eventos observados (Y_i) e com valor esperado (μ_i):

$$Y_i \sim Poisson(\mu_i)$$

onde, $\mu_i = E_i\theta_i$ sendo θ_i o risco relativo de óbito infantil na i-ésima área e E_i a quantidade esperada de óbitos infantis na i-ésima área sob a hipótese de que o risco seja constante em todas as áreas e igual ao risco geral da região. O risco relativo θ_i é uma quantidade que pode assumir valores reais entre 0 e $+\infty$, sendo igual a 1 nas áreas onde o risco possui exatamente o mesmo risco médio para toda a região.

O primeiro nível hierárquico do modelo é dado por:

$$Y_i | E_i \sim Poisson(E_i\, \theta_i)$$

No segundo nível do modelo são especificados os componentes utilizados na estimação do logaritmo do risco θ_i.

$$\log(\theta_i) = \beta_0 + \beta_p X_{ip} + u_i + v_i$$

onde β_0 é uma constante, $\beta_p = \{\beta_1, ..., \beta_k\}$ é um vetor de constantes que, quando colocados na função exponencial, representam os efeitos de cada uma das k covariáveis no $log(\theta_i)$, X_{ip} é um vetor de covariáveis e u_i e v_i são vetores de efeitos

aleatórios. O modelo é construído de modo a relacionar os componentes com o logaritmo de θ_i e não diretamente com o risco θ_i, pelas inúmeras vantagens matemáticas e computacionais que a transformação logarítmica pode proporcionar no presente caso, pois a distribuição *Poisson* é uma distribuição de probabilidade enquadrada na família de distribuições *Exponenciais.* Depois de obtidas as estimativas de $log(\theta_i)$ aplica-se a função exponencial para a obtenção das estimativas do risco θ_i.

No terceiro nível hierárquico do modelo estão as distribuições *a priori* para cada um dos parâmetros do modelo β_0, β_p, u_i e v_i. Neste artigo serão testados nove diferentes modelos, desde modelos sem covariáveis até modelos com os três blocos do IDESE. Em todos os modelos para β_0 será atribuída uma distribuição *a priori Uniforme(*$-\infty$ *;* $+\infty$*)*, e a cada um dos elementos do vetor β_p, uma distribuição *Normal* com média igual a zero e parâmetro de dispersão $\boldsymbol{\tau}_\beta$ o menor possível, de modo que estas prioris sejam pouco informativas.

Quanto aos componentes aleatórios, o componente u_i representa um efeito aleatório não espacial, que pode ser visto como decorrente de particularidades de cada área, ou seja, efeitos de pequena escala que não ultrapassam as fronteiras das áreas. A cada um dos componentes u_i será atribuída independentemente uma distribuição *a priori* com distribuição *Normal* com média zero e parâmetro de precisão τ_u. Sendo assim o vetor $(u_1,...,u_n)$ segue uma distribuição *Normal Multivariada*, composta de termos independentes e depende de um único parâmetro τ_u, em geral desconhecido e denominado Hiperparâmetro, já que é um parâmetro de uma *a priori*. Para τ_u será atribuído *a priori*, uma distribuição *Gamma* com parâmetros 0,5 e 0,0005 atribuídos

de forma que a Hiperpriori seja pouco informativa, ou seja, uma distribuição com grande variabilidade.

Já o componente v_i incorpora a estrutura espacial, captando a influência das áreas vizinhas referentes aos efeitos de larga escala através de uma *a priori* espacialmente estruturada através da definição de uma matriz de vizinhança entre as áreas. Neste estudo será utilizada apenas a matriz de vizinhança binária, na qual cada elemento assumirá valor 1 quando as áreas são vizinhas e 0, caso contrário. Ao componente v_i se atribui uma distribuição *a priori* denominada CAR (*Condicional Auto-Regressiva*) Normal. Como a matriz de vizinhança utilizada é binária, o efeito espacial médio da i-ésima área v_i é dado pela média aritmética dos efeitos dos seus vizinhos e a variância é inversamente proporcional à quantidade de áreas vizinhas; sendo assim, quanto maior o número de vizinhos, maior é a precisão da estimativa de v_i.

Além disso, a especificação completa da distribuição CAR Normal depende de um único parâmetro τ_v (o inverso da variância de v_i) ao qual também será atribuída uma distribuição *a priori Gamma*, analogamente à distribuição *a priori* para τ_u.

A distribuição conjunta condicional *a priori* do efeito espacial v é uma distribuição *a priori* imprópria, já que é baseada nas diferenças pareadas entre os v_i [14] e como distribuições *a priori* impróprias podem levar a distribuições *a posteriori* impróprias, na prática se impõe uma restrição para que esses efeitos v_i somem zero. Contudo a estrutura espacial pode ser atribuída à distribuição *a priori* de diversas formas.

A inferência Bayesiana será baseada na distribuição *a posteriori* de θ:

$$\pi(\theta \mid y) \propto l(y_1,...,y_n \mid \theta)\pi(\log(\theta_i)),$$

onde $l(y_1,...,y_n \mid \theta)$ é a função de verossimilhança e a distribuição *a priori* é $\pi(\theta)$ para o $\log(\theta_i)$.

A distribuição *a posteriori* para θ será proporcional a:

$$\left(\prod_{i=1}^{n} \frac{(\theta_i N_i)^{y_i}}{y_i!} \exp\{-\theta_i N_i\}\right) \pi(u_1,\ldots,u_n \mid \tau_u)\pi(v_1,\ldots,v_n \mid \tau_v)\pi(\beta_{11},\ldots,\beta_{1n} \mid \tau_\beta)\pi(\tau_u)\pi(\tau_v)\pi(\tau_\beta)$$

Assim temos uma distribuição *a posteriori* que não pode ser encontrada analiticamente. Desta forma é necessário utilizar métodos de simulação estocástica chamados MCMC (Markov Chain Monte Carlo). Um software que atualmente está sendo utilizado para modelagem de dados com enfoque Bayesiano, utilizando métodos MCMC é o **WinBUGS** (*Win Bayesian inference Using Gibbs Sampling*)[15]. É um Software livre que possui uma biblioteca para análise de dados espaciais e que será utilizado na aplicação dos dados neste trabalho, além do software TerraView, também um software livre, para a elaboração dos mapas.

Esta distribuição conjunta *a posteriori* é uma distribuição de probabilidade de todos os parâmetros do modelo conjuntamente, e de onde são obtidas as estimativas de cada um dos parâmetros. Serão consideradas como estimativas dos parâmetros (que serão apresentadas nos mapas) as médias obtidas desta distribuição, sendo que para cada parâmetro pode-se obter informações adicionais como intervalos de

credibilidade, que não costumam ser apresentados em mapas, mas que são úteis na avaliação da associação entre as covariáveis e a taxa de mortalidade infantil.

Utilizando-se a metodologia Bayesiana em estudos de correlação ecológica deve-se fazer a análise dos parâmetros do modelo para encontrar fatores realmente explicativos para a variável em estudo. Neste contexto avaliaremos as distribuições *a posteriori* para os parâmetros de regressão do segundo nível do modelo, e o uso de critérios para escolha do modelo, e a análise dos riscos. Um critério utilizado é o *Deviance Information Criterion* (DIC), uma forma semelhante ao *Akaike's Information Criterion* (AIC) utilizado em modelos de regressão "clássica". O DIC é usado quando se modelam os dados através de MCMC, além de ser amplamente utilizado para comparar modelos com diferentes níveis de complexidade[16]. O DIC assume que a média *a posteriori* é uma boa estimativa dos parâmetros do modelo nos casos em que estas distribuições são multimodais e nos casos em que existe acentuada assimetria, ele não é recomendado. O modelo que apresentar o menor DIC é considerado como aquele que melhor pode predizer um novo conjunto de dados com a mesma estrutura dos dados observados.

Resultados

O Rio Grande do Sul, no período estudado de 2001 a 2004 tinha em sua composição 496 municípios com características muito diferentes em relação ao risco de mortalidade infantil e em relação á indicadores sócio-econômicas. Com o objetivo de buscar estimativas mais robustas, foi utilizada a soma de nascimentos e óbitos no período entre 2001 e 2004 para o cálculo dos riscos. O número de nascidos vivos por município no Rio Grande do Sul, totalizando os anos de 2001 a 2004, variou de 35 a

79672. Além do risco relativo estimado pelo modelo de regressão utilizando modelagem inteiramente Bayesiana foi também considerada a estimação de máxima verossimilhança para o risco, SMR (*Standardised Mortality Ratio*), também denominada estimativa clássica do risco ou risco relativo bruto, que é a razão entre a taxa da região i e a taxa geral para o conjunto de regiões estudadas. No caso do Rio Grande do Sul entre os anos de 2001 a 2004, a taxa de mortalidade geral é de 15,63 óbitos para cada mil nascidos vivos. Na modelagem Bayesiana as covariáveis utilizadas, como descrito anteriormente, foram os blocos do IDESE de 2003: Educação, Renda, e Condições de Saneamento e Domicílio. Os mapas da Figura 1 foram construídos com base nos indicadores do IDESE utilizados.

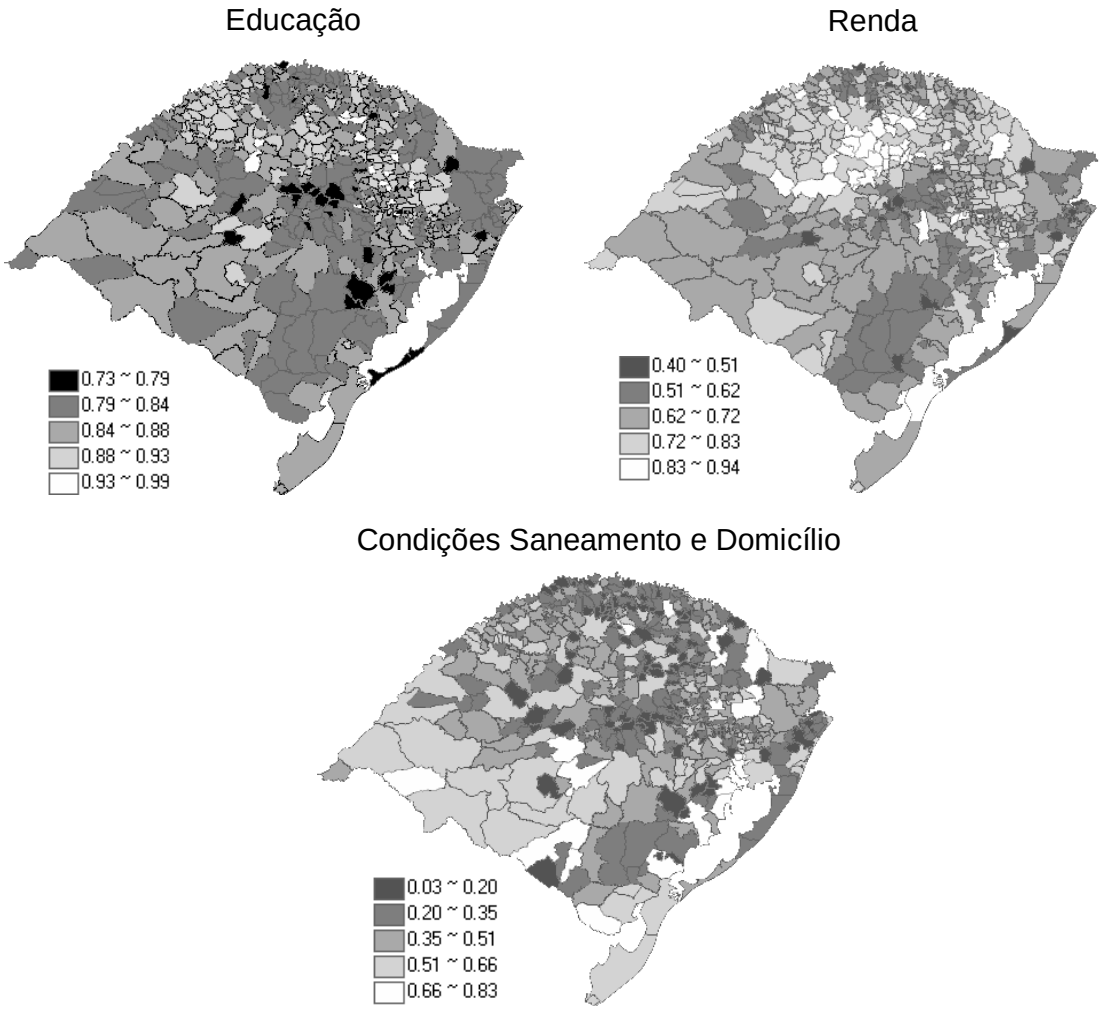

Figura 1- Mapas dos índices Educação, Renda e Condições de Saneamento e Domicílio do IDESE por municípios do Rio Grande do Sul em 2003.

Na Figura 2 é apresentado o mapa com as SMRs de cada município, e na Figura 3 a dispersão entre a SMR e o número de nascidos vivos acumulado (2001 a 2004). O município de Porto Alegre não foi considerado neste diagrama em função da grande quantidade de nascidos vivos, o que poderia distorcer os resultados. O que se pode perceber é que a SMR apresenta flutuações aleatórias muito grandes, influenciadas pelo tamanho da população em risco, particularmente quando o número de nascimentos é inferior a 200. Além disso, não considera as informações dos municípios vizinhos na estimativa do risco.

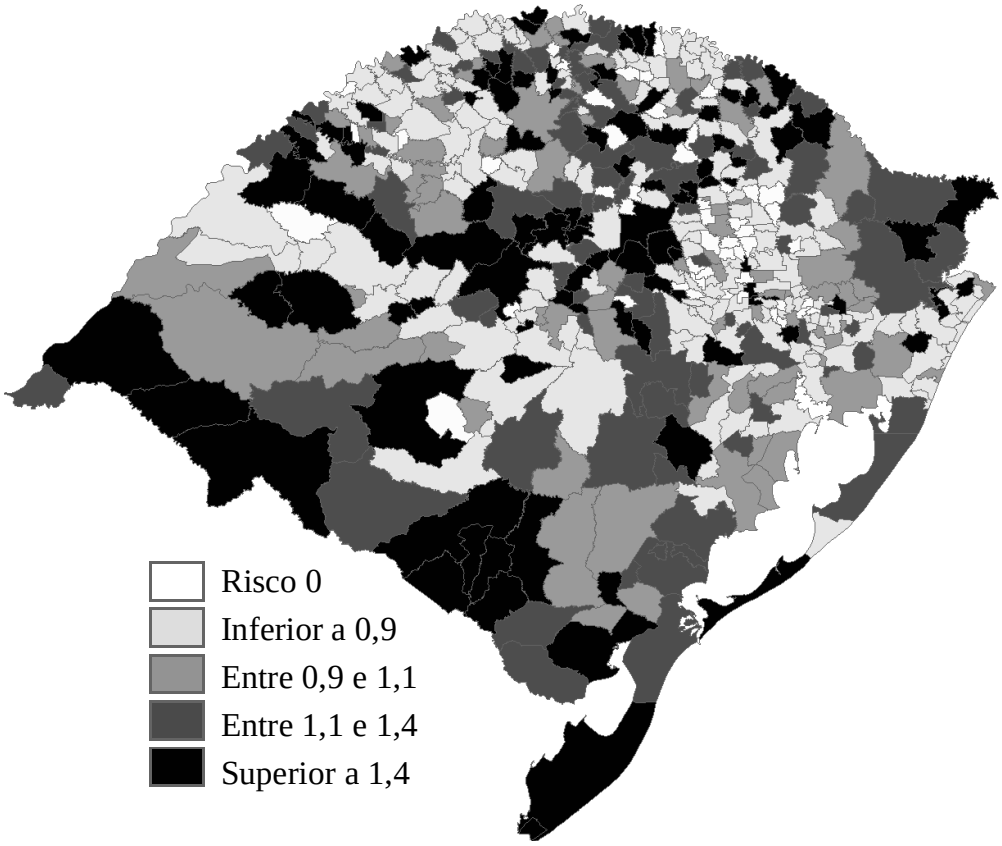

Figura 2- Mapa do risco relativo da mortalidade infantil no Rio Grande do Sul (2001 a 2004), obtido pelo método de máxima verossimilhança (SMR).

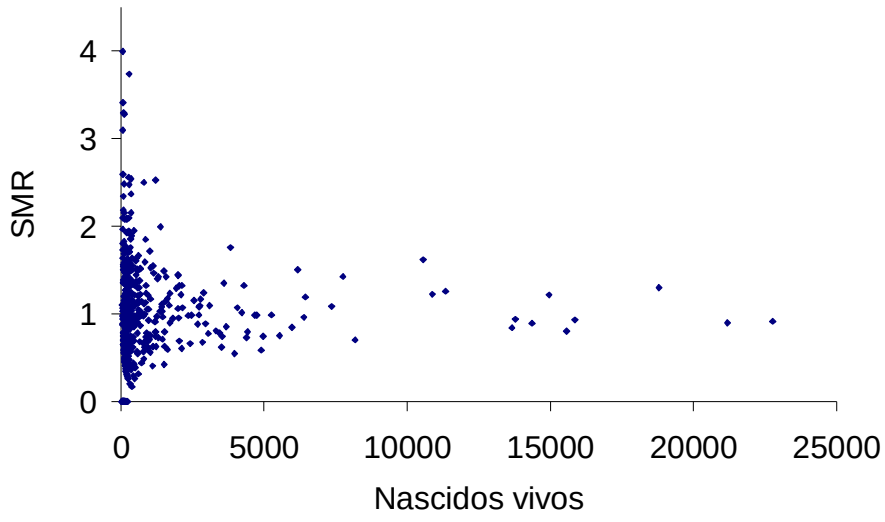

Figura 3 - Relação entre SMR e número de nascidos vivos entre 2001 e 2004.

Para superar as dificuldades de instabilidade das estimativas e incluir no modelo a autocorrelação espacial da variável resposta, todas as regressões foram realizadas utilizando métodos de estimação inteiramente Bayesianos e o critério DIC foi utilizado para a escolha do modelo que melhor estima o risco relativo. Em todos os modelos utilizaram-se 150000 simulações, descartando as 5000 primeiras iterações (*burn-in*) e com espaçamento de 50 entre os pontos amostrados (*thin*), a fim de retirar possíveis autocorrelações na simulação, restando uma amostra de tamanho 2900. A Tabela 1 apresenta os modelos avaliados com seus respectivos valores de DIC. Pode-se perceber que o modelo sem efeito espacial, controlando para Educação, Renda e Condições de Saneamento e Domicílio é o que apresenta o maior valor de DIC, ou seja, é o que apresenta as piores estimativas de risco relativo. Por outro lado, os demais modelos, que incorporam o efeito espacial apresentam valores muito próximos com destaque para o modelo controlando com Educação do IDESE.

Tabela 1- Modelos avaliados e respectivos valores de DIC obtidos pelo WinBUGS baseado em 150000 simulações.

Modelo	DIC
Sem efeito espacial e controlando por educação, renda e saneamento	2622,96
Com efeito espacial e sem covariável	2332,99
Com efeito espacial controlando por educação	2306,17
Com efeito espacial controlando por renda	2334,04
Com efeito espacial controlando por saneamento	2334,34
Com efeito espacial controlando por educação e renda	2307,25
Com efeito espacial controlando por educação e saneamento	2310,89
Com efeito espacial controlando por renda e saneamento	2335,71
Com efeito espacial controlando por educação, renda e saneamento	2307,28

A partir de agora, sempre que nos referirmos ao risco relativo, obtido através da análise de regressão utilizando modelagem inteiramente Bayesiana, estaremos utilizando apenas o melhor modelo, segundo o critério DIC, o com efeito espacial que utiliza como covariável o bloco Educação do IDESE. A interpretação é similar à das SMRs.

A Figura 4 compara as estimativas do risco de mortalidade infantil na modelagem clássica (SMR) e na Bayesiana para os municípios gaúchos. As estimativas inteiramente Bayesianas do risco relativo apresentam bem menos variabilidade, ou seja, são mais concentradas do que as fornecidas pela SMR. Riscos altos obtidos na modelagem clássica foram reduzidos, enquanto que valores próximos de zero ficaram próximos da média da distribuição. Na modelagem Bayesiana aqueles municípios com grande quantidade de nascimentos apresentam pequena variação entre as novas estimativas e as SMR. Por outro lado, as estimativas dos municípios com pequena população em risco são mais afetadas pelas informações dos

municípios vizinhos além de ser influenciada pela covariável Educação, sendo assim observadas diferenças maiores entre o SMR e o risco relativo estimado.

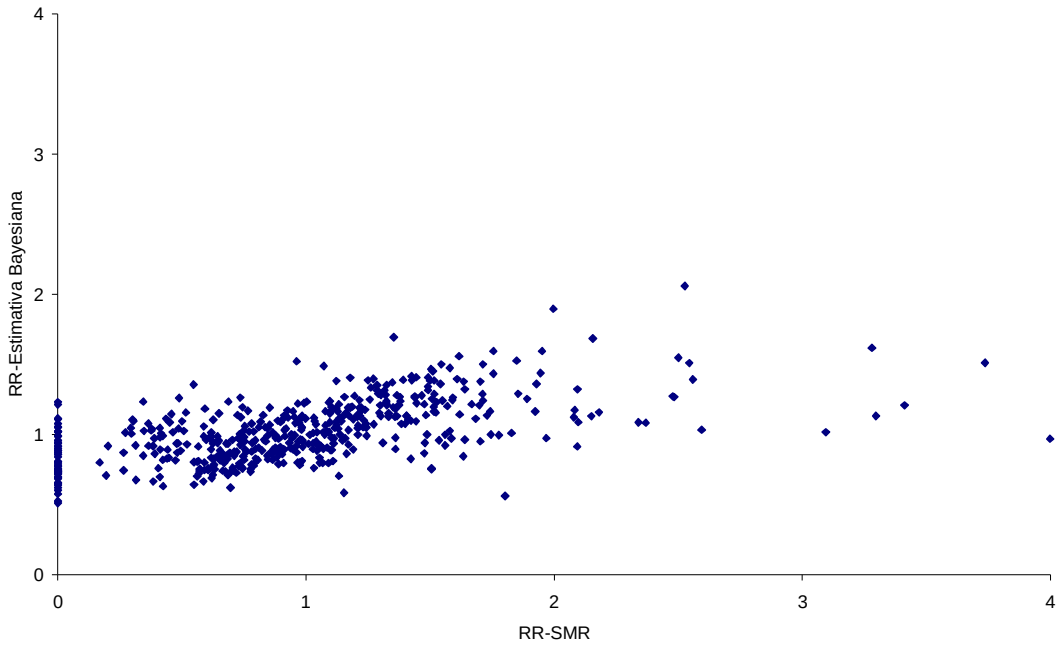

Figura 4 – Relação entre SMR e risco relativo inteiramente Bayesiano.

O efeito visual das estimativas obtidas pela modelagem inteiramente Bayesiana pode ser verificado na Figura 5. O mapa apresenta padrões de risco com estimativas suavizadas e melhor interpretáveis, além de menores flutuações aleatórias. De acordo com a Figura 6 não existem mais municípios com risco de mortalidades nula ou extremamente superiores ao risco médio estadual.

O efeito do tamanho da população sob os riscos relativos de SMR e de estimativa inteiramente Bayesiana pode ser exemplificado com o município Alto Alegre. Neste município, ocorreram 64 nascimentos entre os anos de 2001 a 2004, sendo a SMR aproximadamente 4 vezes (3,998) o risco do estado; usando a estimativa Bayesiana, o risco neste município apresentou praticamente o mesmo risco do estado (0,968).

Além do mapa utilizando modelagem inteiramente Bayesiana apresentar padrões de risco mais bem definidos, as estimativas municipais podem ser analisadas com maior segurança, pois as estimativas apresentam maior estabilidade. Assim os valores extremos podem ser analisados desconsiderando a hipótese de que sejam flutuações aleatórias causadas por pequena população em risco.

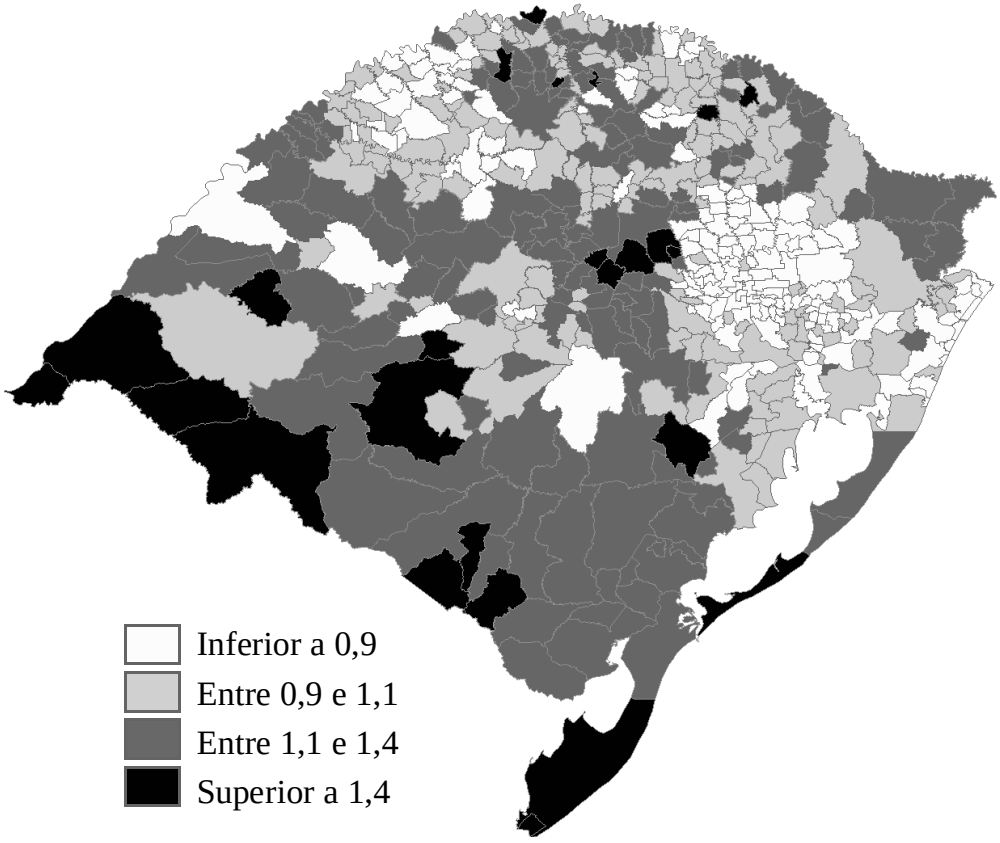

Figura 5- Mapa do risco relativo da mortalidade infantil no Rio Grande do Sul (2001 a 2004), obtido pela modelagem inteiramente Bayesiana.

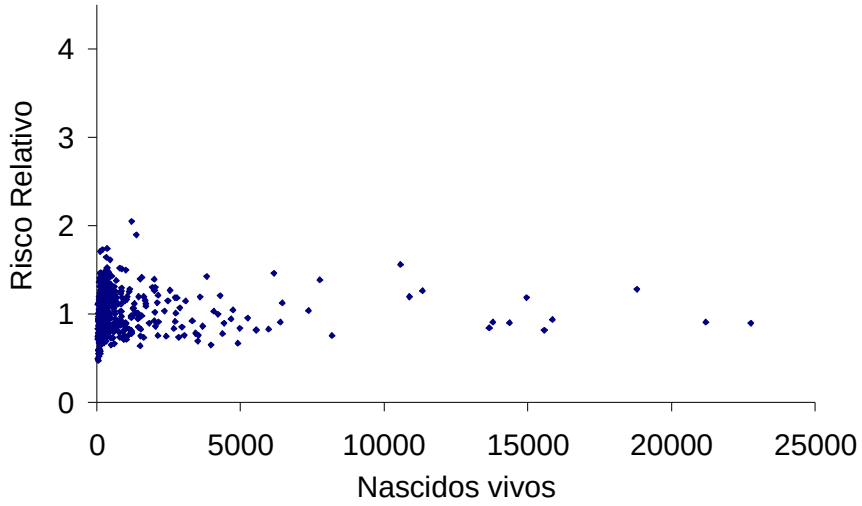

Figura 6- Relação entre o risco relativo obtido pela modelagem inteiramente Bayesiana e número de nascidos vivos entre os anos de 2001 a 2004.

Na modelagem clássica do risco (SMR), 100 municípios apresentaram risco 40% superior ao risco médio do Estado, enquanto que no melhor modelo inteiramente Bayesiano estes foram apenas 25, eles estão identificados nos mapas na tonalidade mais escura.

Os municípios sem óbito infantil, entre 2001 e 2004 e geralmente com poucos nascimentos no período são em número de 50. A estimativa clássica (SMR) para o risco relativo destes municípios seria zero. Esta é uma estimativa teoricamente incorreta, pois sendo assim, nestes municípios não haverá risco de ocorrência de óbito infantil, quando na realidade este risco existe, apenas não foi manifestado na prática. Nestes casos, através da modelagem inteiramente Bayesiana, os municípios passaram a ter pelo menos a metade do risco de mortalidade infantil do Estado, chegando a atingir 1,231, ou seja, um risco 23% maior do que a média estadual, no município Herveiras, cujo índice Educação do IDESE foi o menor entre os municípios que não apresentaram mortalidade infantil no período.

A Tabela 2 apresenta a relação dos 25 municípios com os maiores risco de mortalidade infantil, além da classificação da região segundo o COREDE. No mapa estes municípios estão na tonalidade mais escura.

Tabela 2 – Municípios com os maiores risco relativo de mortalidade infantil estimado pela modelagem inteiramente Bayesiana no Rio Grande do Sul (valores acumuladas de 2001 a 2004)

Município	Região	2001-2004		RR
		Nascimentos	Óbitos	
Redentora	Noroeste Colonial	1215	48	2,058
São José do Norte	Sul	1378	43	1,895
Barra do Quaraí	Fronteira Oeste	189	4	1,693
Charrua	Norte	356	12	1,685
Engenho Velho	Rio da Várzea	117	6	1,617
Chuí	Sul	328	9	1,594
Hulha Negra	Campanha	459	14	1,593
Uruguaiana	Fronteira Oeste	10566	267	1,557
Fontoura Xavier	Alto da Serra do Botucaraí	793	31	1,546
Barros Cassal	Alto da Serra do Botucaraí	865	25	1,526
Dilermando de Aguiar	Central	133	2	1,522
Cacique Doble	Nordeste	291	17	1,512
Manoel Viana	Fronteira Oeste	352	14	1,509
Vicente Dutra	Médio Alto Uruguai	414	10	1,500
Dom Feliciano	Centro-Sul	1008	27	1,500
Lagoão	Alto da Serra do Botucaraí	418	7	1,488
Pedras Altas	Sul	162	4	1,474
Santana do Livramento	Fronteira Oeste	6168	145	1,466
São José do Herval	Alto da Serra do Botucaraí	127	3	1,452
Tunas	Vale do Rio Pardo	263	8	1,437
São Gabriel	Fronteira Oeste	3826	105	1,433
Quaraí	Fronteira Oeste	1572	35	1,415
Santa Vitória do Palmar	Sul	1994	45	1,408
Aceguá	Campanha	343	8	1,407
Lajeado do Bugre	Rio da Várzea	163	3	1,405

Ainda avaliando os padrões de risco é possível destacar positivamente a região da Serra gaúcha com baixas taxas de risco e muito homogênea. Outras regiões com agrupamentos de baixo risco de mortalidade infantil são: região metropolitana de

Porto Alegre, Litoral, Norte, Missões e Fronteira Nordeste. Por outro lado, algumas regiões apresentam alto risco de mortalidade infantil, principalmente nos municípios localizados nas regiões da Fronteira Oeste, Sul e Alto da Serra do Botucaraí.

Conclusões e Considerações Finais

Neste artigo, utilizamos modelos de regressão espacial inteiramente Bayesiano para estimar o risco de mortalidade infantil nos municípios do Rio Grande do Sul acumulado para os anos de 2001 a 2004, utilizando como co-variáveis sub-escalas do Índice de Desenvolvimento Sócio-econômico de 2003 da Fundação de Economia e Estatística Siegfried Emanuel Heuser (IDESE-FEE). O método é baseado em simulações Monte Carlo Via Cadeias de Markov (MCMC) para estimar as distribuições posteriores dos riscos relativos, incorporando efeitos de origem aleatória na ocorrência do evento em cada município, testando diferentes funções de regressão para o risco, além de possibilitar a estimação dos parâmetros. Como resultado temos mapas com padrões de risco mais precisos, com estimativas suavizadas e com menos flutuações aleatórias.

Na modelagem inteiramente Bayesiana, foram avaliadas as possíveis influências das covariáveis, além do efeito espacial. Observou-se que os modelos que incorporam a estrutura espacial além de covariáveis apresentaram melhor desempenho, padrão também detectado em estudo que utilizou a abordagem inteiramente Bayesiana para avaliar a distribuição de homicídios na cidade de Curitiba[17].

Comparando as SMRs com o risco relativo obtido pela modelagem inteiramente Bayesiana foi possível observar um ganho substancial na interpretação e na detecção de padrões de variação no risco de mortalidade infantil nos municípios do Rio

Grande do Sul. Na modelagem inteiramente Bayesiana as estimativas municipais puderam ser analisadas com maior segurança, pois as estimativas apresentaram maior estabilidade. Assim os valores extremos puderam ser analisados desconsiderando a hipótese de flutuações aleatórias causadas por pequena população em risco, ou seja, os municípios com altos riscos relativos de mortalidade infantil devem sofrer ações prioritárias.

Além das vantagens já apontadas da modelagem inteiramente Bayesiana, o método permite a incorporação da modelagem espaço-temporal, de modo a aumentar a precisão das estimativas do risco de mortalidade infantil. Comparando os achados com uma análise espaço-temporal[12] avaliando também a mortalidade infantil no Rio Grande do Sul, porém sem o uso de covariáveis, é possível observar semelhança nos padrões de risco.

Na epidemiologia espacial existe uma grande flexibilidade para explorar melhor os métodos de mapeamento de doenças e de correlação espacial, a partir de diferentes estruturas de vizinhança. Existem várias formas de definir a matriz de vizinhança. Neste artigo utilizamos apenas a matriz binária, mas poderia ser considerado o tamanho da fronteira, o tamanho das fronteiras com ou sem a presença de barreiras naturais ou outra característica geográfica que possa interferir a ligação entre áreas[18]. Diferentes especificações na matriz de vizinhança gerarão diferentes estimativas. Na literatura, a comparação de estimativas utilizando diferentes matrizes de vizinhança ainda é pouco explorada.

Atualmente também estão sendo conduzidos estudos de simulação[13] com o objetivo de comparar os modelos Bayesianos com modelos semi-paramétricos.

A partir dos resultados, acredita-se que os governantes e as Secretarias de Saúde municipais e estaduais podem ter uma melhor visão do risco de Mortalidade Infantil subjacente para os municípios, sem as flutuações aleatórias inerente aos dados brutos. Com isso possibilitando gerar novas políticas públicas para a mortalidade infantil no Rio Grande do Sul.

Várias áreas do conhecimento podem se beneficiar dos métodos aqui propostos, porém a obtenção das estimativas a partir da modelagem inteiramente Bayesiana ainda é um processo complexo e exige tempo na execução das iterações, além de ainda não existir um software que combine a modelagem com a criação de mapas.

Referências

1. Maia, S.F.; Sousa, T.R.V. (2004). Uma investigação dos determinantes da redução da taxa de mortalidade infantil nos estados da Região Nordeste do Brasil. In: I Congresso da Associação Latino Americana de População – ALAP. .

2. Victora C.G. (2001). Intervenções para reduzir a mortalidade infantil pré-escolar e materna no Brasil. *Revista Brasileira de Epidemiologia*, São Paulo, 4: 3-69.

3. Carvalho, M.S.; Santos, R.S. (2005). Análise de dados espaciais em Saúde Pública: métodos, problemas, perspectivas. *Cadernos de Saúde Pública*, Rio de Janeiro, 21(2): 361-378.

4. Rojas, L. I.; Barcellos, C. & Peiter, P. (1999). Utilização de mapas no campo da epidemiologia no Brasil. *Informe Epidemiológico do SUS*, 8:25-35.

5. Lawson A. (2001). *Statistical Methods in Spatial Epidemiology*. Sussex: John Wiley & Sons.

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

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

8. Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

9. Assunção, R.M. (2001). *Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia*. São Carlos, ABE. Disponível em http://www.est.ufmg.br/leste/publicacoes.htm. Data de acesso: 08/05/2006.

10. Assunção, R.M.; Barreto, S.M; Guerra, H.L.; Sakura, E. (1998). Mapas de taxas epidemiológicas: uma abordagem Bayesiana. *Cadernos de Saúde Pública*, Rio de Janeiro, 14(4):713-723.

11. Santos, S.M.; Noronha, C.P. (2001). Padrões espaciais de mortalidade e diferenciais sócio-econômicos na cidade do Rio de Janeiro. *Cadernos de Saúde Pública*, Rio de Janeiro, 17(5):1099-1110.

12. Vieira, D.M. (2006). *Análise Espaço-Temporal na Estimação de Taxas de Incidência/ Mortalidade*. Dissertação Mestrado em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul.

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

14. Schmidt, A.M.; Nobre, A.A.; Ferreira, G.S. (2002). Alguns aspectos da modelagem de dados espacialmente referenciados, *Revista Brasileira de Estatística*, 63(220): 59-88.

15. Thomas, A.; Spiegelhalter, D.J.; Gilks, W.R. (1992). BUGS: A program to perform Bayesian inference using Gibbs Sampling. *Bayesian Statistics 4*: 837-842.

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

17. Ehlers, R.S.; Silva, S.A.; Melo, L.L.M. (2006). Fully Bayesian spatial analysis of homicide rates. *Journal Estadistica*, Santiago de Chile, 58.

18. Mollié, A. (1996). Bayesian mapping of disease. In: W.R. Gilks, S. Richardson and D.J. Spiegelhalter (editors). *Monte Carlo Markov Chain in Practice*. London: Chapman and Hall.

ANEXO 1 – PROJETO DE PESQUISA

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

FACULDADE DE MEDICINA

PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

PROJETO DE PESQUISA

ANÁLISE DE CORRELAÇÃO ECOLÓGICA: UMA ABORDAGEM INTEIRAMENTE BAYESIANA PARA A MORTALIDADE INFANTIL NO RIO GRANDE DO SUL

AUTOR: SERGIO KAKUTA KATO

ORIENTADORA: JANDYRA MARIA GUIMARÃES FACHEL

Porto Alegre, 2007

1. INTRODUÇÃO

1.1 Mortalidade Infantil

A taxa de mortalidade infantil é um dos indicadores mais usados para medir a qualidade de vida da população (Maia e Souza, 2004), além de compor o Índice de Desenvolvimento Humano (IDH) e outros indicadores sócio-econômicos. O Rio Grande do Sul possui um indicador sócio-econômico, denominado Índice de Desenvolvimento Sócio-econômico (IDESE) da Fundação de Economia e Estatística (FEE) que também utiliza a taxa de mortalidade infantil para compor um índice referente à saúde.

Na literatura existem estudos propondo ações ou intervenções para a redução da mortalidade infantil, porém, na maioria deles fatores são relacionados descritivamente à mortalidade infantil.

Os estudos que tem como foco a comparação de grupos ao invés de pessoas são ditos estudos ecológicos. Os estudos ecológicos são em geral rápidos, de baixo custo, ideais para amostras integrais, como poluição, altitude e clima, além de úteis para levantar hipótese. As informações utilizadas nos estudos ecológicos, relacionados à área da saúde como, por exemplo, a mortalidade infantil, estão, em grande parte, disponíveis no DATASUS além de outras bases oficiais como o IBGE e FEE.

Um dos principais índices da FEE é o IDESE que consiste em um índice sintético composto por 12 indicadores divididos em quatro blocos temáticos: Educação, Renda, Saneamento e Domicílio e Saúde.

Para entender melhor a distribuição espacial da mortalidade infantil, alguns estudos tem sido realizado (Assunção *et al.*, 1998; Santos e Noronha, 2001),

inclusive no Rio Grande do Sul, através de estimação espaço temporal (Vieira e Fachel, 2006).

Os estudos que relacionam a mortalidade com localizações geográficas na forma de mapa se enquadram na área da Epidemiologia Espacial.

1.2 Epidemiologia Espacial

Os indicadores ecológicos geralmente são taxas ou riscos. A avaliação geograficamente mais elementar é através da taxa bruta denotada por T_i, que para cada área é a razão entre a quantidade observada (doença/incidência) Y_i e a população exposta N_i, tornando possível a comparação de resultados entre áreas, porém, em muitas ocasiões, além da comparação de diferentes áreas no mesmo mapa, é necessário a comparação entre doenças/incidências diferentes para a mesma área, ou até mesmo a comparação da mesma doença/incidência em períodos diferentes, e nestas situações é geralmente utilizada medidas de risco relativo.

1.3. Modelagem Clássica para o Risco Relativo

Utilizando o método de máxima verossimilhança, a estimativa para o verdadeiro risco relativo de interesse são atribuídas ás áreas do mapa, a partir da razão entre a taxa da região i e a taxa da região toda, que é denominada SMR (*Standardised Mortality Ratio*) da região i, que é dada por:

$$SMR_i = R_i = \frac{y_i}{e_i}$$

Onde y_i = é a contagem de eventos na área i; N_i = é a população em risco na área i; $e_i = rN_i$, número esperado de ocorrência do evento na área i; $r = \sum y_i / \sum N_i$, taxa global do risco total em toda a área de estudo.

A SMR varia de zero a infinito, sendo que quando R_i=1 indica que a área i teve tantos casos observados quanto seria esperado caso seu risco fosse idêntico ao de toda área avaliada.

A variância da SMR_i (y_i / e_i^2) é inversamente proporcional ao número esperado de casos na área i (e_i), proporcional ao tamanho da população na área i (N_i), ou seja, quanto menor for o tamanho da população em risco, maior a incerteza na estimativa do risco, produzindo maior flutuações aleatórias (Olsen *et al*., 1996). Estas flutuações aleatórias dos indicadores ecológicos causam uma instabilidade ao modelo, tornando difícil a interpretação dos mapas, e pode ser ainda maior em estudos com baixa incidência, por exemplo, mortalidade segundo causa específica. Na literatura basicamente três alternativas são propostas: agregar áreas, mapas de probabilidade e modelagem Bayesiana. Atualmente as duas primeiras não estão sendo mais utilizadas em função das vantagens demonstradas pela modelagem Bayesiana (Assunção, 2001).

1.4 Análise de Regressão Espacial em Estudos Ecológicos

Os modelos de regressão espacial empregado em estudos ecológicos além de uma ferramenta básica na análise de eventos de saúde podem ser utilizados para avaliar a necessidade de alocação de recursos para a saúde de acordo com a variação geográfica. O foco é relacionar covariáveis, ou variáveis potencialmente explicativas, com a incidência da doença a nível geográfico, seja no campo da análise exploratória

ou buscando modelos explicativos a fim de apontando medidas preventivas (Lawson, 2001; Carvalho e Santos, 2005).

Geralmente, quando se faz uma análise de regressão, procura-se encontrar um bom ajuste entre os valores preditos pelo modelo e os valores observados da variável dependente, além de determinar quais as variáveis explicativas contribuem de forma significativa para este relacionamento linear. Para tanto, a hipótese principal é que as observações não são correlacionadas, e, conseqüentemente, que os resíduos (diferença entre os valores observados e preditos) também sejam independentes e não-correlacionados com a variável dependente, além de ter variância constante, e apresentar distribuição *Normal* com média zero (Druck *et al*., 2004).

No caso de dados espaciais, a hipótese de independência das observações geralmente é falsa, os resíduos continuam apresentando a autocorrelação espacial presente nos dados. A dependência entre as observações altera o poder explicativo do modelo, podendo produzir associações erradas (Bailey e Gattrel, 1995). Nos dados espaciais "a dependência está presente em todas as direções e fica mais fraca à medida que aumenta a dispersão na localização dos dados" (Cressie,1991).

1.5 Análise de Regressão Utilizando Modelos Inteiramente Bayesiano

Para superar as dificuldades de instabilidade das estimativas e autocorrelação espacial na variável resposta várias modelagens têm sido propostas na literatura, porém as que apresentam melhores resultados são os métodos de estimação inteiramente Bayesianos (Assunção, 2001).

Assim supõe-se que o número de eventos observados em cada área possui distribuição Binomial, mas como a maioria dos dados epidemiologicamente

mapeados são raros ou com grande variação das taxas/riscos entre diferentes áreas, o modelo Binomial pode ser aproximado pela distribuição *Poisson* (Richardson at. al., 2004), com número de eventos observados (Y_i) e com valor esperado (μ_i):

$$Y_i \sim Poisson(\mu_i)$$

onde, $\mu_i = E_i\theta_i$ sendo θ_i o risco relativo de óbito infantil na i-ésima área e E_i a quantidade esperada de óbitos infantis na i-ésima área, sob a hipótese de que o risco seja constante em todas as áreas e igual ao risco geral da região, e sendo assim, o primeiro nível hierárquico do modelo é dado por:

$$Y_i/E_i \sim Poisson(E_i\ \theta_i)$$

No segundo nível do modelo são especificados os componentes utilizados na estimação do logaritmo do risco θ_i.

$$\log(\theta_i) = \beta_0 + \beta_p X_{ip} + u_i + v_i$$

Onde β_0 é uma constante, $\beta_p = \{\beta_1, ..., \beta_k\}$ é um vetor de constantes que, quando colocados na função exponencial representam os efeitos de cada uma das k covariáveis no $log(\theta_i)$, X_{ip} é um vetor de covariáveis e u_i e v_i são vetores de efeitos aleatórios.

O modelo é construído de modo a relacionar os componentes com o logaritmo de θ_i e não diretamente com o risco θ_i, pelas inúmeras vantagens matemáticas e computacionais que a transformação logarítmica pode proporcionar no

presente caso, pois a distribuição *Poisson* é uma distribuição de probabilidade enquadrada na família de distribuições *Exponenciais.*

Depois de obtidas as estimativas de $log(\theta_i)$ aplica-se a função exponencial para a obtenção das estimativas do risco θ_i.

No terceiro nível hierárquico do modelo estão as distribuições *a priori* para cada um dos parâmetros do modelo β_0, β_p, u_i e v_i.

Em geral a β_0 se atribui uma *a priori Uniforme*$(-\infty\ ;\ +\infty)$, e a cada um dos elementos do vetor β_p se atribui independentemente uma distribuição *Normal* com média igual a zero e parâmetro de dispersão $\boldsymbol{\tau}_\beta$ o menor possível, de modo que estas prioris sejam pouco informativas.

Quanto aos componentes aleatórios a definição da distribuição *a priori* apresenta diferenças em relação aos componentes já expostos.

O componente u_i representa um efeito aleatório não espacial, que podem ser visto como as particularidades de cada área, ou seja, efeitos de pequena escala que não ultrapassam as fronteiras das áreas.

A cada um dos componentes u_i se atribui independentemente uma *a priori* com distribuição *Normal* com média zero e parâmetro de precisão τ_u. Sendo assim o vetor $(u_1,...,u_n)$ segue uma distribuição *Normal Multivariada*, composta de termos independentes e depende de um único parâmetro τ_u, em geral desconhecido e denominado Hiperparâmetro, já que é um parâmetro de uma distribuição *a priori*.

A este Hiperparâmetro τ_u também se deve atribuir uma distribuição *a priori*, denominada Hiperpriori. Geralmente para parâmetros de dispersão ou precisão se atribui uma distribuição *Gamma* com parâmetros *a* e *b* atribuídos de forma que a Hiperpriori seja pouco informativa, ou seja, uma distribuição com grande variabilidade.

Já o componente v_i incorpora a estrutura espacial, captando a influência das áreas vizinhas referentes aos efeitos de larga escala através de uma priori espacialmente estruturada através da definição de uma matriz de vizinhança entre as áreas, denotada por $W_{(n\times x)}^{(1)}$.

Seja um conjunto de *n* áreas $(A_1,\ldots,A_n)$, construímos uma matriz $W_{(n\times x)}^{(1)}$, onde cada um dos elementos W_{ij} representa uma medida de proximidade entre A_i e A_j.

Esta medida de proximidade pode ser calculada a partir de diversos critérios, sendo os mais utilizados porém neste estudo será utilizada apenas a matriz binária, na qual W_{ij} =1 ij w = caso faça fronteira, ou W_{ij} =0 , caso contrário.

Voltando à especificação da priori, ao componente v_i se atribui uma priori denominada CAR (Condicional Auto-Regressiva) Normal:

$$v_i \mid v_j \sim Normal\left(\frac{\sum_j W_{ij} v_j}{\sum_j W_{ij}}, \frac{\left(\frac{1}{\tau_v}\right)}{\sum_j W_{ij}}\right)$$

Em particular, quando se utiliza a estrutura de vizinhança binária,a qual é a estrutura mais comumente utilizada, o efeito espacial médio da i-ésima área v_i é dado pela média aritmética dos efeitos dos seus vizinhos e a variância inversamente proporcional a quantidade de áreas vizinhas, sendo assim, quanto maior o número de vizinhos, maior é a precisão da estimativa de v_i.

Além disso, a especificação completa da distribuição CAR Normal depende de um único parâmetro τ_v (o inverso da variância de v_i) ao qual também deve ser atribuída uma distribuição *a priori Gamma*, analogamente a distribuição *a priori* para τ_u.

Logo a distribuição conjunta condicional *a priori* do efeito espacial v é dada por:

$$(v \mid \tau_v) \propto \frac{1}{\left(\frac{1}{\tau_v}\right)^n} \left\{ -\frac{1}{2\left(\frac{1}{\tau_v}\right)^2} \sum_i \sum_j W_{ij} (v_i - v_j)^2 \right\},$$

Esta é uma distribuição *a priori* imprópria já que é baseada nas diferenças pareadas entre os v_i (Schmidt *et al.*, 2002) e, como prioris impróprias podem levar *a posteriori* impróprias, na prática se impõe uma restrição para que esses efeitos v_i somem zero. Contudo a estrutura espacial pode ser atribuída à distribuição *a priori* de diversas formas.

A inferência bayesiana será baseada na distribuição *a posteriori* de θ:

$$\pi(\theta \mid y) \propto l(y_1,...,y_n \mid \theta)\pi(\log(\theta_i)),$$

onde $l(y_1,...,y_n \mid \theta)$ é a função de verossimilhança e a distribuição *a priori* é $\pi(\theta)$ para o $\log(\theta_i)$.

A distribuição *a posteriori* para θ será proporcional a:

$$\left(\prod_{i=1}^{n} \frac{(\theta_i N_i)^{y_i}}{y_i!} \exp\{-\theta_i N_i\}\right) \pi(u_1,\ldots,u_n \mid \tau_u)\pi(v_1,\ldots,v_n \mid \tau_v)\pi(\beta_{11},\ldots,\beta_{1n} \mid \tau_\beta)\pi(\tau_u)\pi(\tau_v)\pi(\tau_\beta)$$

Assim temos a distribuição *a posteriori* que não pode ser encontrada analiticamente. Desta forma é necessária utilizar métodos de simulação estocástica chamados MCMC (Markov Chain Monte Carlo).

Um software que atualmente está sendo utilizado para modelagem de dados com enfoque bayesiano, utilizando métodos MCMC é o **WinBUGS** (Win Bayesian inference Using Gibbs Sampling). É um Software livre implementado por Thomas *et al* (1992) que possui uma biblioteca para análise de dados espaciais e que será utilizado na aplicação dos dados neste trabalho.

Este modelo será utilizado na construção dos mapas, de modo que estes apresentem informações de maneira mais clara do que os mapas realizados com as taxas brutas ou SMRs.

2. OBJETIVOS

2.1 Objetivo Geral:

O objetivo geral deste trabalho é apresentar Modelos de Regressão Espacial Bayesiano para Estudos de Correlação Ecológica, para estimação de taxas de incidência ou risco relativo onde serão incluídas covariáveis no modelo, além da utilização de mapas para melhor visualização desta distribuição.

2.1 Objetivo Específico:

Tem como objetivo específico ilustrar os métodos, através da utilização da mortalidade infantil nos municípios do Rio Grande do Sul entre os anos de 2001 a 2004 relacionados com Índice de Desenvolvimento Sócio-econômico de 2003 da Fundação de Economia e Estatística Siegfried Emanuel Heuser (IDESE-FEE), além da comparação entre a modelagem clássica do risco relativo (SMR) e o melhor modelo obtido pelo método de Regressão Espacial Bayesiano.

3. METODOLOGIA

Serão utilizados dados publicados no site do DATASUS e da Fundação de Economia e Estatística (FEE), e para a realização das estimativas do risco de mortalidade infantil serão utilizados os softwares estatísticos WinBugs e R, além do Terra View para ilustrar os achados através de mapas.

Serão construídos diferentes modelos para a explicação do fenômeno, sendo escolhido o mais adequado, de acordo com critérios DIC, todos os modelos propostos serão métodos inteiramente bayesianos.

4. CRONOGRAMA

Atividade	Ano / Semestre								
	2006/1	2006/2	2007/1	2007/2					
				Jul	Ago	Set	Out	Nov	Dez
Revisão da literatura									
Familiarização softwares									
Análise dos dados									
Redação									
Defesa preliminar									
Correções									
Defesa final									

5. REFERÊNCIAS

Assunção, R.M. (2001). *Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia*. São Carlos. ABE. Disponível em http://www.est.ufmg.br/leste/publicacoes.htm. Data de acesso: 08/05/2006.

Assunção, R.M.; Barreto, S.M; Guerra, H.L.; Sakura, E. (1998). Mapas de taxas epidemiológicas: uma abordagem Bayesiana. *Cadernos de Saúde Pública*, Rio de Janeiro, 14(4):713-723.

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

Carvalho, M.S.; Santos, R.S. (2005). Análise de dados espaciais em Saúde Pública: métodos, problemas, perspectivas. *Cadernos de Saúde Pública*, Rio de Janeiro, 21(2): 361-378.

Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

Druck, S.; Carvalho, M.S. ; Câmara, G.; Monteiro, A.M.V. (2004). Análise Espacial de Dados Geográficos. Brasília: Embrapa.

Lawson A. (2001). *Statistical methods in spatial epidemiology*. Sussex: John Wiley & Sons.

Maia, S.F.; Sousa, T.R.V. (2004). Uma investigação dos determinantes da redução da taxa de mortalidade infantil nos Estados da Região Nordeste do Brasil. In: I Congresso da Associação Latino Americana de População - ALAP, 2004, Caxambu. I Congresso da Associação Latino Americana de População - ALAP. Caxambu : ALAP.

Olsen, S.; Martuzzi, M.; Elliott, P. (1996). Cluster analysis and disease mapping-Why, when, how? A step by step guide, *British Medical Journal*, 313, 863-865.

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

Santos, S.M.; Noronha, C.P. (2001). Padrões espaciais de mortalidade e diferenciais sócio-econômicos na cidade do Rio de Janeiro. *Cadernos de Saúde Pública*, Rio de Janeiro, 17(5):1099-1110.

Schmidt, A.M.; Nobre, A.A.; Ferreira, G.S. (2002). Alguns Aspectos da Modelagem de Dados Espacialmente Referenciados, *Revista Brasileira de Estatística*, vol. 63, n.220, pp. 59-88.

Thomas, A.; Spiegelhalter, D.J.; Gilks, W.R. (1992). BUGS: A program to perform Bayesian inference using Gibbs Sampling. *Bayesian Statistics 4*, 837-842.

Vieira, D.M. (2006). *Análise Espaço-Temporal na Estimação de Taxas de Incidência/ Mortalidade*. Dissertação Mestrado em Epidemiologia - Universidade Federal do Rio Grande do Sul.

ANEXO – LINHAS DE COMANDO

WinBugs - Modelo 1– Sem efeito espacial

```
model
{

for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-log(e[i])+alpha+beta1*edu[i]+beta2*renda[i]+beta3*sanea[i]
        # Risco Relativo
        theta[i]<-exp(alpha+beta1*edu[i]+beta2*renda[i]+beta3*sanea[i])
    # Taxa de incidência
    tx[i]<- 15.632*theta[i]

        }



alpha~dflat()
beta1~dnorm(0.0, 1.0E-5)
beta2~dnorm(0.0, 1.0E-5)
beta3~dnorm(0.0, 1.0E-5)

}
```

WinBugs - Modelo 2– Com efeito espacial e sem covariáveis

```
model
{
for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-log(e[i])+alpha+v[i]+u[i]
        # Risco Relativo
        theta[i]<-exp(v[i]+u[i])
      # Taxa de incidência
      tx[i]<- 15.632*theta[i]
        }


# Priori Car Normal
u[1:m]~car.normal(adj[],weights[],num[],tau.u)


# Priori
for(i in 1:m)
{
v[i]~dnorm(0,tau.v)


}

#Weights
for(j in 1:sumNumNeig)
{
        weights[j]<-1
}

alpha~dflat()

# Hiperprioris
tau.v~dgamma(0.5, 0.0005)
tau.u~dgamma(0.5, 0.0005)

}
```

WinBugs - Modelo 3– Com efeito espacial e covariável Educação

```
model
{

for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-log(e[i])+alpha+v[i]+u[i]+beta1*edu[i]
        # Risco Relativo
        theta[i]<-exp(alpha+v[i]+u[i]+beta1*edu[i])
    # Taxa de incidência
    tx[i]<- 15.632*theta[i]

        }


# Priori Car Normal
u[1:m]~car.normal(adj[],weights[],num[],tau.u)



# Priori
for(i in 1:m)
{
v[i]~dnorm(0,tau.v)


}

#Weights
for(j in 1:sumNumNeig)
{
        weights[j]<-1
}

alpha~dflat()
beta1~dnorm(0.0, 1.0E-5)

# Hiperprioris
tau.v~dgamma(0.5, 0.0005)
tau.u~dgamma(0.5, 0.0005)

}
```

WinBugs - Modelo 4– Com efeito espacial e covariável Renda

```
model
{

for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-log(e[i])+alpha+v[i]+u[i]+beta1*renda[i]
        # Risco Relativo
        theta[i]<-exp(alpha+v[i]+u[i]+beta1*renda[i])
    # Taxa de incidência
    tx[i]<- 15.632*theta[i]

        }

# Priori Car Normal
u[1:m]~car.normal(adj[],weights[],num[],tau.u)



# Priori
for(i in 1:m)
{
v[i]~dnorm(0,tau.v)


}

#Weights
for(j in 1:sumNumNeig)
{
        weights[j]<-1
}

alpha~dflat()
beta1~dnorm(0.0, 1.0E-5)

# Hiperprioris
tau.v~dgamma(0.5, 0.0005)
tau.u~dgamma(0.5, 0.0005)

}
```

WinBugs - Modelo 5– Com efeito espacial e covariável Saneamento

```
model
{

for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-log(e[i])+alpha+v[i]+u[i]+beta1*sanea[i]
        # Risco Relativo
        theta[i]<-exp(alpha+v[i]+u[i]+beta1*sanea[i])
    # Taxa de incidência
    tx[i]<- 15.632*theta[i]

        }


# Priori Car Normal
u[1:m]~car.normal(adj[],weights[],num[],tau.u)



# Priori
for(i in 1:m)
{
v[i]~dnorm(0,tau.v)


}

#Weights
for(j in 1:sumNumNeig)
{
        weights[j]<-1
}

alpha~dflat()
beta1~dnorm(0.0, 1.0E-5)

# Hiperprioris
tau.v~dgamma(0.5, 0.0005)
tau.u~dgamma(0.5, 0.0005)

}
```

WinBugs - Modelo 6– Com efeito espacial e covariáveis Educação e Renda

```
model
{

for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-log(e[i])+alpha+v[i]+u[i]+beta1*edu[i]+beta2*renda[i]
        # Risco Relativo
        theta[i]<-exp(alpha+v[i]+u[i]+beta1*edu[i]+beta2*renda[i])
    # Taxa de incidência
    tx[i]<- 15.632*theta[i]

        }


# Priori Car Normal
u[1:m]~car.normal(adj[],weights[],num[],tau.u)



# Priori
for(i in 1:m)
{
v[i]~dnorm(0,tau.v)

}

#Weights
for(j in 1:sumNumNeig)
{
        weights[j]<-1
}

alpha~dflat()
beta1~dnorm(0.0, 1.0E-5)
beta2~dnorm(0.0, 1.0E-5)


# Hiperprioris
tau.v~dgamma(0.5, 0.0005)
tau.u~dgamma(0.5, 0.0005)

}
```

WinBugs - Modelo 7– Com efeito espacial e covariáveis Educação e Saneamento

```
model
{

for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-log(e[i])+alpha+v[i]+u[i]+beta1*edu[i]+beta2*sanea[i]
        # Risco Relativo
        theta[i]<-exp(alpha+v[i]+u[i]+beta1*edu[i]+beta2*sanea[i])
    # Taxa de incidência
    tx[i]<- 15.632*theta[i]

        }


# Priori Car Normal
u[1:m]~car.normal(adj[],weights[],num[],tau.u)


# Priori
for(i in 1:m)
{
v[i]~dnorm(0,tau.v)

}

#Weights
for(j in 1:sumNumNeig)
{
        weights[j]<-1
}

alpha~dflat()
beta1~dnorm(0.0, 1.0E-5)
beta2~dnorm(0.0, 1.0E-5)


# Hiperprioris
tau.v~dgamma(0.5, 0.0005)
tau.u~dgamma(0.5, 0.0005)

}
```

WinBugs - Modelo 8– Com efeito espacial e covariáveis Renda e Saneamento

```
model
{

for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-log(e[i])+alpha+v[i]+u[i]+beta1*renda[i]+beta2*sanea[i]
        # Risco Relativo
        theta[i]<-exp(alpha+v[i]+u[i]+beta1*renda[i]+beta2*sanea[i])
    # Taxa de incidência
    tx[i]<- 15.632*theta[i]

        }


# Priori Car Normal
u[1:m]~car.normal(adj[],weights[],num[],tau.u)


# Priori
for(i in 1:m)
{
v[i]~dnorm(0,tau.v)


}

#Weights
for(j in 1:sumNumNeig)
{
        weights[j]<-1
}

alpha~dflat()
beta1~dnorm(0.0, 1.0E-5)
beta2~dnorm(0.0, 1.0E-5)



# Hiperprioris
tau.v~dgamma(0.5, 0.0005)
tau.u~dgamma(0.5, 0.0005)

}
```

WinBugs - Modelo 9– Com efeito espacial e covariáveis Educação, Renda e Saneamento

```
model
{
for (i in 1:m)
{
        # Verossimilhança
        y[i]~dpois(mu[i])
        log(mu[i])<-
log(e[i])+alpha+v[i]+u[i]+beta1*edu[i]+beta2*renda[i]+beta3*sanea[i]
        # Risco Relativo
        theta[i]<-exp(alpha+v[i]+u[i]+beta1*edu[i]+beta2*renda[i]+beta3*sanea[i])
    # Taxa de incidência
    tx[i]<- 15.632*theta[i]

        }

# Priori Car Normal
u[1:m]~car.normal(adj[],weights[],num[],tau.u)


# Priori
for(i in 1:m)
{
v[i]~dnorm(0,tau.v)


}

#Weights
for(j in 1:sumNumNeig)
{
        weights[j]<-1
}

alpha~dflat()
beta1~dnorm(0.0, 1.0E-5)
beta2~dnorm(0.0, 1.0E-5)
beta3~dnorm(0.0, 1.0E-5)


# Hiperprioris
tau.v~dgamma(0.5, 0.0005)
tau.u~dgamma(0.5, 0.0005)

}
```