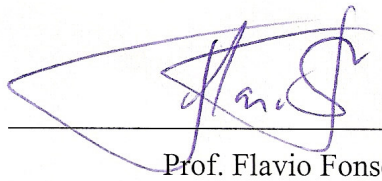
APLICAÇÃO DE MODELOS DE MISTURAS FINITAS NA CLASSIFICAÇÃO DE ÁREAS DE RISCO PARA TUBERCULOSE

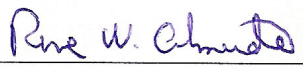
Liana Wernersbach Pinto

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

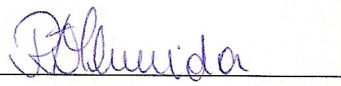
Aprovada por:



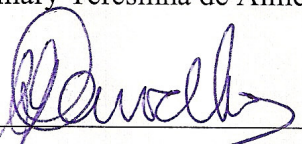
Prof. Flavio Fonseca Nobre, Ph.D.



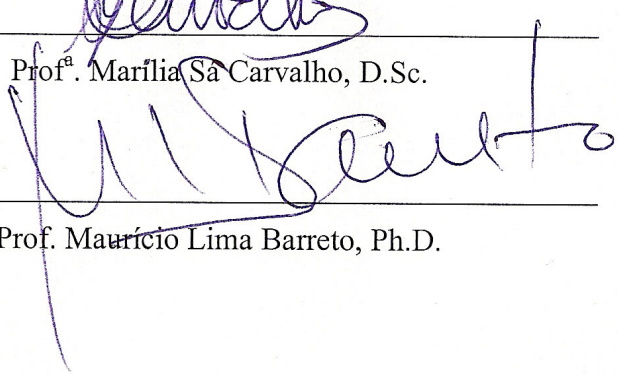
Prof. Renan Moritz Varnier Almeida, Ph.D.



Prof[a]. Rosimary Teresinha de Almeida, Ph.D.



Prof[a]. Marília Sá Carvalho, D.Sc.



Prof. Maurício Lima Barreto, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

DEZEMBRO DE 2006

PINTO, LIANA WERNERSBACH

Aplicação de modelos de misturas finitas na classificação de áreas de risco para tuberculose [Rio de Janeiro] 2006

X, 89 p. 29,7 cm (COPPE/UFRJ, D.Sc., Engenharia Biomédica, 2006)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. Vigilância à Saúde
2. Análise espacial
3. Modelos de misturas finitas
4. Mapeamento de doenças
5. Tuberculose

I. COPPE/UFRJ II. Título (série)

Agradecimentos

Ao Professor Flávio Fonseca Nobre pela orientação do trabalho e atenção dispensada ao longo de todos esses anos de convivência.

A Cosme Marcelo Passos da Silva e Julio Cesar Stacchini de Souza pela troca de idéias durante a realização deste trabalho e, acima de tudo, pela amizade e apoio em todos os momentos.

A Luciana da Silva Costa e Saint Clair dos Santos Gomes Júnior pela amizade e apoio.

Aos professores do Programa de Engenharia Biomédica por todos os ensinamentos.

Aos funcionários do Programa de Engenharia Biomédica por sua presteza em resolver todos os problemas.

Aos colegas do Programa de Engenharia Biomédica por compartilharem todos os momentos de angústias e alegrias durante a nossa convivência.

Ao CNPq pelo apoio financeiro por meio de bolsa de estudos.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

APLICAÇÃO DE MODELOS DE MISTURAS FINITAS NA CLASSIFICAÇÃO DE ÁREAS DE RISCO PARA TUBERCULOSE

Liana Wernersbach Pinto

Dezembro/2006

Orientador: Flavio Fonseca Nobre

Programa: Engenharia Biomédica

O presente trabalho teve como objetivo investigar a aplicação de modelos de misturas finitas, com e sem covariáveis, na identificação de áreas de risco para doenças, sendo composto pelas seguintes etapas: (1) análise exploratória espacial; (2) aplicação do modelo de mistura finita; e (3) aplicação da regressão mista. Empregaram-se dados relativos à mortalidade e morbidade por tuberculose nos municípios do estado do Rio de Janeiro, no período de 1997 a 2002. A técnica de mistura finita modela os dados como uma soma ponderada de *g* densidades de probabilidade, cada qual representando uma subpopulação, cujos pesos representam a força daquele grupo. Os resultados da análise espacial dos dados mostraram que a distribuição espacial dos eventos não segue uma distribuição aleatória, existindo áreas de maior concentração (Região Metropolitana do estado). Por meio dos modelos de mistura finita verificou-se que, para os dados em questão, os modelos de quatro e seis componentes levaram a um melhor ajuste. Quando se utilizou o modelo de regressão mista, foi possível identificar as covariáveis que têm efeito estatisticamente significativo, positivo ou negativo. Os resultados obtidos mostram que os modelos de mistura finita e a regressão mista consistem em ferramentas que podem ser bastante úteis para a classificação de regiões e, portanto, o ponto de partida para a avaliação do risco em mapas de doenças.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

MIXTURE MODELS APPLICATION IN TUBERCULOSIS RISK AREAS CLASSIFICATION

Liana Wernersbach Pinto

December/2006

Advisor: Flavio Fonseca Nobre

Department: Biomedical Engineering

The analysis and identification of disease clustering in space and its representation on a map is an important issue within disease surveillance. Identification of high risk groups provides valuable hints for possible exposure and targets for subsequent analytical studies and allows public health professionals to make appropriate resources allocation. The present work aimed to investigate methods for classification of disease, particularly the mixture models and mixed regression. A data set containing mortality and morbidity of tuberculosis from 1997 to 2002 in the municipalities of Rio de Janeiro State was used. In accordance with the analysis of the maximum likelihood estimates and Bayesian Information Criteria (BIC), it has been verified that mixture models with four and six components led to the best fit. When applying mixed regression, models with four components presented the best fit. The results show that the finite mixtures models and mixed regression are valuable tools for the classification of regions and can be, therefore, the starting point for the evaluation of the risk in maps of illnesses.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1: Introdução

A vigilância em Saúde Pública consiste da coleta contínua, da análise e da disseminação de dados de saúde, com o objetivo de prevenir e controlar doenças, injúrias e outros problemas de saúde (THACKER, 2000). Recentemente, a vigilância em Saúde Pública vem sofrendo transformações devido ao surgimento de novos problemas, como o bioterrorismo, e novas doenças. Além disso, persistem alguns antigos problemas de saúde, por vezes em níveis alarmantes, como a tuberculose. Adicionalmente, recentes inovações tecnológicas como a monitoração em tempo real, grandes bancos de dados em formatos acessíveis, avanços nos sistemas de informação geográfica, métodos estatísticos sofisticados e novas ferramentas computacionais, possibilitaram avanços nesta área (STROUP *et al.*, 2004).

Os dados de vigilância devem ser disseminados rapidamente para os profissionais da Saúde Pública, incluindo aqueles que os coletaram, assim como para os gestores de saúde e, dependendo da situação, a outras agências e à comunidade. A disseminação dos dados freqüentemente estimula a busca por informações adicionais vindas de outras fontes.

Os sistemas de vigilância em Saúde Pública podem ser empregados com diferentes propósitos como, por exemplo, detectar surtos e epidemias, fornecer informações sobre a história natural de uma doença, projetar a extensão de uma epidemia e avaliar o impacto de intervenções na área da saúde. No que diz respeito à detecção de surtos e epidemias, é de fundamental importância para os profissionais da Saúde Pública que modificações nos quadros usuais de uma determinada doença sejam rapidamente identificadas, visto que isto propiciará uma intervenção oportuna de forma a atenuar a morbidade e a mortalidade.

A vigilância de doenças infecto-contagiosas é realizada em muitos países e tem como objetivo a detecção de epidemias em um estágio inicial. As epidemias geralmente se iniciam em uma pequena área e subseqüentemente se espalham. No sentido de detectar epidemias em um estágio inicial, a observação de dados de pequenas áreas, como municípios, condados, bairros, distritos sanitários e escolares e setores censitários,

é fundamental. Para a análise ser realizada num nível local também é necessária a existência de informações demográficas e socioeconômicas das áreas envolvidas.

Os dados de vigilância incluem registros da incidência e da prevalência de doenças transmissíveis. O objetivo da análise destas informações é a identificação e a quantificação de padrões (tendências, padrões cíclicos, anomalias) no tempo e no espaço. Os dados podem incluir endereços de residências ou datas de diagnósticos, mas, habitualmente, devido à necessidade de sigilo, as informações são disponibilizadas para períodos de tempo específicos (semanas, meses ou anos), distritos geográficos (setores censitários, bairros, condados, estados), ou ambos.

A epidemiologia espacial consiste da descrição e análise das variações geográficas de uma doença, relacionadas a fatores de risco demográficos, ambientais, comportamentais, socioeconômicos, genéticos e infecciosos. Avanços nos sistemas de informação geográfica, surgimento de novas metodologias estatísticas e a disponibilidade de dados de saúde georeferenciados, propiciaram novas oportunidades para a investigação da variação geográfica de doenças. Esta investigação é parte de uma longa tradição de análises geográficas, datando do século XIX, quando os mapas de doenças em diferentes países começaram a emergir para a caracterização do espalhamento e descoberta de possíveis causas das ocorrências de doenças infecto-contagiosas (ELLIOTT e WARTENBERG, 2004).

A investigação espacial epidemiológica se divide em quatro áreas principais (BAILEY, 2001): (1) mapeamento de doenças, cujo foco é a construção de mapas da distribuição de uma doença ou evento de saúde; (2) estudos ecológicos, que têm por objetivo estudar as relações entre fatores etiológicos e a ocorrência de doenças; (3) estudo de aglomerados de doenças, cujo foco é a identificação de áreas de risco elevado de ocorrência de doenças ou a avaliação de aumento do risco em torno de uma fonte suspeita de risco ambiental; e (4) avaliação e monitoramento ambiental, que tem por objetivo verificar a distribuição espacial de fatores ambientais relevantes para a saúde. Na prática, contudo, essas quatro áreas principais se confundem e se mesclam. Além dos objetivos mencionados, a epidemiologia geográfica também pode ser utilizada para o planejamento da localização de unidades de saúde, para a verificação do acesso da população aos serviços de saúde, para a obtenção de estimativas da cobertura dos programas de imunização e, ainda, no auxílio à busca de novas formas de prevenção de

doenças e programas de educação voltados para a saúde da população (ELLIOTT e WARTENBERG, 2004).

Uma das ferramentas mais importantes da análise espacial são os mapas de doenças. Estes constituem uma forma rápida e resumida de visualização de informações geográficas complexas, permitindo a identificação de padrões nos dados, os quais nem sempre são facilmente identificados na apresentação tabular.

A apresentação das informações no mapa pode variar desde a simples representação de pontos (casos) no mapa e de representações pictóricas relativas ao número de casos agregados dentro de uma área, até o mapeamento de modelos complexos que descrevem a estrutura dos eventos.

A análise da variação geográfica de uma doença é feita normalmente por meio de mapas temáticos ou coropléticos, nos quais as áreas são sombreadas de acordo com os valores da variável de interesse, usualmente uma contagem (LAWSON *et al.*, 2000). Freqüentemente, assume-se uma distribuição de Poisson para os dados, o que implica em um risco λ igual para todas as áreas, impossibilitando, desta forma, a identificação de áreas de risco. Além disso, com freqüência os dados violam a suposição de igualdade entre média e variância da distribuição de Poisson. Este fenômeno é chamado de superdispersão e está comumente presente em dados relativos a pequenas unidades geográficas (como municípios, condados, bairros) ou a doenças raras na população. A superdispersão pode ser indicativa da heterogeneidade dos dados (RICHARDSON *et al.*, 2004).

Os principais objetivos do mapeamento de doenças são (LAWSON *et al.*, 2000):

(i) descrever a variação geográfica da incidência de uma doença para a formulação de hipóteses etiológicas;

(ii) identificar áreas de risco aumentado de forma que uma ação possa ser tomada;

(iii) fornecer um mapa do risco de uma doença em uma região de forma a permitir uma melhor alocação de recursos e avaliação do risco.

A identificação da heterogeneidade espacial dos riscos de ocorrência de uma doença fornece informações sobre grupos expostos e grupos-alvo para a realização de

estudos analíticos. Além disso, os mapas de doenças constituem um ponto de partida para investigações de aglomerados de casos.

Na construção dos mapas, o primeiro passo consiste na escolha da medida epidemiológica que será apresentada. Uma medida de uso freqüente é a razão de mortalidade/morbidade padronizada (*Standardized mortality/morbidity ratio* - SMR) que é dada pela razão entre o número de casos observados y_i e o número de casos esperados E_i para cada área i, $SMR_i = y_i / E_i$. Um valor de SMR igual a 1 indica que, para a área em questão, o número de casos observados foi igual ao número de casos esperados, considerando um risco igual para toda a região em estudo. Abordagens tradicionais para categorização das áreas são baseadas nos percentis da distribuição do SMR ou na construção de intervalos de igual amplitude (SCHLATTMANN & BÖHNING, 1993; REZAEIAN *et al.*, 2004).

A SMR é um estimador não tendencioso. Contudo, ela é baseada em uma única observação, consistindo então, em um modelo saturado. Seu uso apresenta três desvantagens principais (LAWSON *et al.*, 2000):

(i) A variância da SMR é grande em áreas com pequena população (e conseqüentemente E_i pequeno) e pequena em áreas com população grande. Essa instabilidade das estimativas dificulta as decisões baseadas na SMR.

(ii) Não é possível diferenciar regiões quando casos não são observados. Assim, os valores de SMR iguais a zero não permitem a distinção na variação dos valores esperados entre as áreas.

(iii) A SMR não revela a estrutura subjacente aos dados e não é parcimoniosa (por se tratar de um modelo saturado).

Como abordagens alternativas sugere-se (LAWSON *et al.,* 2000):

(i) modelos de alisamento, os quais empregam dados de áreas vizinhas com o objetivo de reduzir a variabilidade.

(ii) modelos bayesianos lineares, os quais são baseados em uma função linear da SMR. Estes foram inicialmente empregados no mapeamento de doenças por MARSHALL (1991).

(iii) modelos bayesianos, os quais assumem que os riscos relativos são realizações de alguma distribuição de probabilidade.

(iv) modelos bayesianos empíricos, que são modelos bayesianos onde a estimativa da distribuição a priori para os riscos relativos é obtida a partir dos dados observados.

Outra abordagem encontrada na literatura para o mapeamento de doenças são os modelos de mistura finita (BÖHNING & SEIDEL, 2003). Eles consistem de uma abordagem para a modelagem de dados de populações heterogêneas, isto é, constituídas por várias subpopulações (ou grupos de risco). Nos modelos de mistura finita, a distribuição de probabilidade é vista como uma soma ponderada de *g* densidades de probabilidade, cada densidade representando uma subpopulação ou grupo de risco e seus respectivos pesos representando a força daquele grupo no modelo geral. Esta modelagem permite a definição de áreas para uma ação prioritária. Os modelos de mistura podem ser estendidos por meio da inclusão de covariáveis, sendo então chamados de regressão mista (regressão de Poisson mista, regressão logística mista).

Apesar dos mapas de doenças constituírem uma forma bastante atrativa de representação da variação geográfica de uma doença, deve-se ter bastante cautela em sua interpretação, uma vez que, dependendo da forma, escala e resolução escolhidas para a apresentação da variável de interesse, os resultados podem diferir bastante. Mapas de um mesmo conjunto de dados apresentados em escalas diferentes podem resultar em padrões visuais bem diferentes. Além disso, variações nas taxas observadas no mapa podem refletir diferenças na qualidade dos dados com relação ao diagnóstico, a classificação ou ao registro dos casos.

Segundo REZAEIAN *et al.* (2004), apesar dos mapas ilustrarem relações espaciais que podem não ter sido identificadas em tabelas, não se deve apresentar somente um gráfico, pois este é um dos muitos mapas possíveis para os dados em questão. BAILEY & GATRELL (1995) enfatizam que o ponto final da análise espacial não é necessariamente um único mapa "correto". É essencial assegurar que os padrões aparentes não são simplesmente artefatos do processo de construção dos mapas. Idealmente, um investigador deve produzir uma variedade de mapas de forma a ilustrar diferentes aspectos dos dados.

Os objetivos do presente trabalho são: (1) verificar a distribuição espaço temporal da mortalidade e morbidade por tuberculose nos municípios do estado do Rio de Janeiro; (2) investigar a aplicação de modelos de misturas finitas na identificação de áreas de risco para tuberculose; e (3) verificar, por meio da regressão mista, quais as covariáveis são importantes na determinação de áreas de risco para tuberculose.

A tuberculose foi escolhida como aplicação por se tratar de um importante problema de Saúde Pública no Brasil e, em especial, no estado do Rio de Janeiro, onde a sua incidência supera a média nacional (SES/RJ, 2003). As técnicas propostas, análise espacial de dados e modelos de mistura finita serão empregadas no intuito de classificar áreas de risco para a ocorrência de casos e óbitos por tuberculose. A identificação de áreas de risco possibilitaria uma atuação mais rápida e eficiente dos serviços de saúde, além de auxiliar numa melhor alocação de recursos e de unidades de saúde e o desenvolvimento de programas direcionados para as áreas de maior risco.

Uma vez classificadas as áreas quanto ao risco, será aplicado o modelo de regressão de Poisson mista que terá por objetivo identificar nas áreas em questão as variáveis explicativas da situação de risco encontrada. Apesar dos fatores de risco para a tuberculose serem amplamente conhecidos, o objetivo de sua utilização será verificar o desempenho dos modelos de mistura finita, a sua capacidade de classificar áreas e a elas associar fatores de risco. A classificação de áreas e a caracterização destas pode ser uma ferramenta auxiliar na definição de áreas prioritárias para o desenvolvimento e implantação de programas de saúde.

A seguir é apresentada a estrutura do trabalho. No Capítulo 2 é apresentada uma revisão da literatura sobre tuberculose e mapeamento de doenças. O Capítulo 3 consiste de uma revisão teórica sobre as metodologias empregadas no trabalho, as técnicas de análise espacial de dados e os modelos de mistura finita e a regressão mista. No Capítulo 4, Materiais e Métodos, são descritos os dados utilizados, os modelos ajustados e as análises realizadas. O Capítulo 5 apresenta os resultados obtidos na análise exploratória espacial e na etapa de modelagem. No Capítulo 6 é apresentada a discussão dos resultados obtidos e no Capítulo 7, a conclusão do trabalho.

Capítulo 2: Revisão da Literatura

2.1 Tuberculose

Segundo a Organização Mundial da Saúde (OMS) a tuberculose (TB) é a maior causa de morte por doença infecto-contagiosa em adultos, infectando cerca de 30% da população mundial (MS, 1998). No ano de 2004, as regiões que registraram o maior número de casos novos foram o sudoeste da Ásia (33%) e a África (29%). Estima-se que naquele ano ocorreram 1,7 milhão de mortes em decorrência da tuberculose em todo o mundo (OMS, 2006).

O Brasil ocupa o 16º. lugar entre os 22 países responsáveis por 80% do total de casos de tuberculose no mundo. Estima-se que a prevalência na população brasileira seja de aproximadamente 50 milhões de indivíduos. De acordo com o Sistema de Informações de Agravos de Notificação (SINASC/MS) anualmente são notificados aproximadamente 85.000 casos novos e ocorrem 6.000 óbitos em decorrência da doença (MS, 2006).

Em 1993, a Organização Mundial da Saúde (OMS) declarou a situação da tuberculose como estado de urgência, tendo sido criado em decorrência disto o programa *STOP TB*, que reunia instituições como o Banco Mundial, o *Centers for Disease Control*, a *International Union Against Tuberculosis and Lung Disease*, entre outros, na luta contra a doença. A OMS aponta como principais causas para a gravidade da situação atual da tuberculose no mundo os seguintes fatos: desigualdade social, advento da AIDS, envelhecimento da população e grandes movimentos migratórios. Adicionalmente, em muitos locais o problema da tuberculose foi deixado de lado pelas autoridades, pois estas consideravam-no, equivocadamente, como um problema resolvido (RUFFINO-NETTO, 2002).

Em 1998, em virtude da grave situação em que o país se encontrava foi estabelecido o Plano Nacional de Controle da Tuberculose (PNCT). Tal plano está integrado com a rede de serviços de saúde e baseia-se em ações unificadas, executadas em conjunto pelas esferas federal, estadual e municipal. Suas metas eram: diagnosticar,

em três anos (até 2001), pelo menos 92% dos casos; tratar, com sucesso, pelos menos 85% dos casos diagnosticados; e reduzir, em nove anos (até 2007), a incidência em pelo menos 50% e a mortalidade em dois terços (RUFFINO-NETTO, 2001; MS, 1999). No entanto, pouca atenção foi dada à verificação da adequação dessas metas (SOUZA, 2003).

No PNCT, o Ministério da Saúde recomendava a implantação da Estratégia do Tratamento Supervisionado (DOTS). Esta é uma estratégia de controle da tuberculose recomendada internacionalmente e que aumenta a probabilidade de cura dos doentes em função da garantia do tratamento assistido, contribuindo para a interrupção da transmissão da doença. O DOTS consiste de cinco elementos básicos: compromisso político, diagnóstico de pacientes por baciloscopia, observação direta dos pacientes tratados, acesso a quantidades adequadas de medicamentos e acompanhamento de cada paciente diagnosticado (MS, 2006).

No ano de 2003, a região Sudeste totalizou aproximadamente 50% dos casos ocorridos no Brasil, tendo sido o Rio de Janeiro o estado brasileiro com maior taxa de incidência e mortalidade por tuberculose. O coeficiente de incidência no estado é quase duas vezes a média nacional. Observa-se, contudo, em todas as Unidades Federativas a existência de municípios com taxas altíssimas, especialmente em áreas metropolitanas e periféricas das grandes cidades (HIJAR, 2005).

Em virtude da grave situação do estado do Rio de Janeiro frente ao problema da tuberculose, o Programa de Controle de Tuberculose do estado definiu como áreas prioritárias todos os municípios com mais de 100 mil habitantes. Os 22 municípios prioritários totalizam 86% da população do estado, sendo responsáveis por 96% dos casos de tuberculose. Doze municípios, entre os 22, pertencem à região metropolitana do estado e concentram 73% da população, sendo responsáveis por 85% dos casos (SES/RJ, 2003).

Em países desenvolvidos, a doença acomete principalmente idosos, minorias étnicas e imigrantes estrangeiros. Nos países em desenvolvimento, a população economicamente ativa é o principal grupo atingido (MS, 1998). No Brasil, a doença acomete principalmente homens em idade produtiva. Algumas populações indígenas, carcerárias e sem-teto apresentam incidência muito maior que a população geral.

A tuberculose é um grave problema de Saúde Pública, intimamente ligado à pobreza e à má distribuição de renda. A tuberculose é uma doença que não apresenta variações cíclicas ou sazonais de importância prática. Sua prevalência é maior em áreas de elevada concentração populacional e precárias condições socioeconômicas e sanitárias. Em 1981, com o surgimento da Síndrome da Imunodeficiência Adquirida (AIDS) ocorreu aumento do número de casos de tuberculose no mundo. A associação AIDS/tuberculose constitui atualmente um sério problema de Saúde Pública.

Estimativas preliminares do Ministério da Saúde indicam a existência de cerca de 400.000 pessoas infectadas pelo vírus da AIDS no Brasil e que 30% destas apresentam também a infecção pelo bacilo da tuberculose (KERR-PONTES *et al.*, 1997). Observa-se que a infecção conjunta pelo HIV e pelo *Mycobacterium tuberculosis* pode aumentar em 25 vezes o risco de desenvolvimento da tuberculose-doença (KERR-PONTES *et al*, 1997), sendo o risco de morte destes pacientes duas vezes maior que dos pacientes HIV(+) sem tuberculose (LAGRANGE *et al.*, 2000). A OMS recomenda a monitoração epidemiológica e a intensificação da detecção rápida da AIDS como estratégia de controle da tuberculose, por meio da busca ativa de pessoas com tosse ativa e febre persistentes. Essa estratégia depende na integração entre os programas de saúde e entre os sistemas de informação (RODRIGUES-JR *et al.*, 2006). Estudos mostram a falta de identificação por parte dos sistemas de vigilância da co-infecção tuberculose/AIDS (LIMA *et al.*, 1997; RODRIGUES-JR *et al.*, 2006).

Outra questão que também contribui para o agravamento do problema da tuberculose é o aparecimento de focos de tuberculose multiresistente (MR). A MR consiste na resistência do bacilo à medicação do tratamento e constitui um grande problema em diversas regiões do mundo. No Brasil, entretanto, a resistência aos fármacos utilizados para tratamento não tem sido grande problema (HIJAR, 2005). A principal maneira de evitar o surgimento da MR é aumentar a aderência do paciente ao tratamento, ou seja, assegurar a cura e diminuir o abandono do tratamento. Segundo RUFFINO-NETTO (1999) a taxa de abandono do tratamento permaneceu constante em 14% no período de 1982 a 1999.

A tuberculose é causada pelo *Mycobacterium tuberculosis*, que tem o homem como seu reservatório principal. A doença é transmitida pela tosse (principalmente), fala e espirros de indivíduos infectados com o bacilo. Com relação ao quadro clínico, a

tuberculose não possui sinais e sintomas característicos, havendo comprometimento do estado geral (febre baixa vespertina com sudorese, inapetência e emagrecimento). Quando a doença atinge os pulmões, o indivíduo pode apresentar dor torácica e tosse produtiva, acompanhada ou não de escarros hemoptóicos.

A maior parte dos casos, 85%, são da forma pulmonar da doença, a mais contagiosa. No entanto, ela também pode atingir outros órgãos do organismo como rins, ossos e meninges. Apesar de grave, a tuberculose é curável em praticamente 100% dos casos. O tratamento é feito em regime ambulatorial e consiste em esquema terapêutico empregando uma combinação de drogas (Isoniazida, Rifampicina, Pirazinamida, Estreptomicina, Etambutol e Etionamida).

Segundo o Guia de Vigilância Epidemiológica (1998) na análise da mortalidade por tuberculose deve-se considerar a distribuição geográfica, os grupos etários e a associação com a AIDS. Desta forma, o presente trabalho se propôs a analisar a distribuição espaço-temporal dos dados de morbidade e mortalidade por tuberculose ocorridos no estado do Rio de Janeiro, a aplicar a técnica de mistura finita para a classificação de áreas de risco para a tuberculose, e por meio da regressão mista, identificar quais covariáveis estão associadas com cada um dos grupos de risco.

MOTA *et al.* (2003) conduziram um estudo ecológico no intuito de verificar a distribuição espacial da mortalidade atribuída à tuberculose na cidade de Salvador, Bahia, nos anos de 1991, 1994 e 1997, sendo empregada como unidade de análise o distrito sanitário (DS). Verificaram que a desigualdade da mortalidade por tuberculose entre os DS foi muito acentuada (3 a 4 vezes em algumas situações) e, que modo geral, as taxas mais elevadas correspondiam a áreas situadas na periferia da cidade, onde grande parte das famílias vive em condições precárias. Observam ainda que esses elevados riscos de morte podem ainda estar revelando desigualdades e problemas relacionados com o acesso aos serviços de saúde, a aderência ao tratamento ou na qualidade da atenção médica. Concluem que a identificação de áreas de maior risco de óbito para a tuberculose pode constituir-se em subsídio para a formulação de políticas capazes de melhorar o acesso e a qualidade dos serviços de saúde disponíveis a populações residentes nessas áreas.

RODRIGUES-JR *et al.* (2006) apresentaram um estudo que visava analisar a incidência da tuberculose entre os casos de AIDS. Uma abordagem geoestatística foi empregada com a finalidade de verificar o padrão de distribuição espacial. Realizaram estudo retrospectivo, empregando dados secundários. Foram incluídos no estudo todos os casos de AIDS notificados ao Ministério da Saúde no período de 1991 a 2001. Os casos foram estratificados segundo município, gênero, regiões administrativas de saúde, categorias de transmissão da AIDS e número de anos desde o diagnóstico. Empregou-se um modelo geoestatístico Gaussiano para a construção do mapa temático do risco, utilizando como variável resposta a incidência da tuberculose dentro dos casos de AIDS. Observaram que as regiões mais populosas apresentaram maior risco de transmissão da tuberculose e concluíram que as coordenadas geográficas estão associadas com o risco de tuberculose, mas não com o risco de AIDS. Nesse estudo puderam ser identificadas áreas de maior risco para transmissão da tuberculose.

SOUZA *et al.* (2005) buscaram analisar a ocorrência da tuberculose, identificando variáveis definidoras de situações coletivas de risco que determinam a distribuição espacial como subsídio a implantação de um sistema de vigilância de base territorial. Realizaram estudo ecológico no período de 1996-2000, em Olinda, Recife, Pernambuco. Empregaram um modelo multivariado de regressão logística na detecção de variáveis relacionadas com a transmissão da doença. Os resultados mostraram que as variáveis média de moradores por domicílio, presença de casos de retratamento e existência de famílias com mais de um caso no período foram estatisticamente significativas. Os autores argumentam que o simples mapeamento de casos de retratamento e de domicílios com ocorrências de repetidos casos permitiria refinar o foco de atenção em micro-áreas prioritárias para intervenções intensivas. Concluem que é importante e necessária a estruturação de um sistema de vigilância da tuberculose que contemple intervenções de base territorial como forma de identificar seus determinantes. Além disso, argumentam que o pressuposto para a construção de tal sistema é o conhecimento de que a distribuição das endemias é também determinada por processos sociais, os quais estão relacionados às características do espaço onde ocorrem.

VICENTIN *et al.* (2002) estudaram a mortalidade por tuberculose no município do Rio de Janeiro no ano de 1991 visando estabelecer correlações com indicadores sociais. Nesse estudo foram consideradas as regiões administrativas da cidade. A análise

dos dados mostrou correlação entre a maioria dos indicadores sócio-econômicos estudados e o coeficiente de mortalidade por tuberculose por 100 habitantes /ano. Mostraram correlação significativa direta os indicadores: índice de Robin Hood, o qual indica a proporção de renda que deveria ser retirada dos ricos e transferida para os pobres para que fosse alcançada uma distribuição eqüitativa; razão de renda entre os 10% mais ricos e os 40% mais pobres; e proporção de chefes de família com renda entre um e dois salários mínimos. A correlação foi inversa e significativa com os indicadores: proporção de residentes com mais de 10 anos com curso superior, área média por domicílio, número de cômodos por domicílio, renda média em salários mínimos e proporção de chefes de família com rendas entre 10-15, 16-20 e acima de 20 salários mínimos.

SOUZA *et al.* (2000) propõem a construção de um sistema de vigilância integrado como uma forma de atender de maneira mais adequada aos indivíduos que vivem em áreas carentes. Propõem uma mudança no sistema de vigilância de forma que o foco tradicional em indivíduos de risco seja substituído pelo foco em áreas e populações de risco. Construíram um índice de carência social, com base em variáveis sócio-econômicas e demográficas. Utilizaram ainda a análise de componentes principais seguida de uma análise de aglomerados para a identificação de fatores associados à incidência da tuberculose. Verificaram associação entre o índice de carência social e a ocorrência de TB. Os resultados mostraram grupos e áreas de maior prioridade que necessitavam de intervenção.

VENDRAMINI *et al.* (2006) realizaram um estudo cujos objetivos foram (1) explorar o risco para a tuberculose em relação com potenciais determinantes na cidade de São José do Rio Preto, São Paulo, Brasil; (2) analisar os indicadores de morbidade e mortalidade e determinar a relação entre o risco de tuberculose e o nível sócio-econômico, empregando para tal um sistema de informação geográfica (GIS) e os dados do censo 2000. O estudo foi realizado entre 1984 e 2005, tendo sido empregados os setores censitários. A seleção das variáveis socioeconômicas foi realizada empregando a técnica de componentes principais. O nível sócio-econômico foi baseado nas seguintes características: renda média, anos de educação do chefe da família, média de rendimento e anos de estudo de mulheres chefes de família, percentual de pessoas e mulheres analfabetos, percentual de famílias com mais de cinco residentes. Verificou-se que o risco para a tuberculose é duas vezes mais alto em áreas com condição

socioeconômica baixa. Concluem que a identificação de áreas com diferentes níveis de risco permitiriam a Departamento Municipal de Saúde propor intervenções que minimizassem o risco da doença no nível individual e populacional. Verificou-se que a população idosa é mais vulnerável a TB. A pobreza também retarda a procura ao sistema de saúde, levando a um aumento no período de infecção e a um aumento do risco da transmissão.

2.2 Mapeamento de doenças

A construção dos primeiros mapas de doenças data do século XIX quando, John Snow se utilizou desta ferramenta com o intuito de visualizar o endereço das vitimas de cólera juntamente com a localização das bombas de fornecimento de água. Ao representar no mapa os endereços das vítimas, Snow observou que os distritos que apresentavam maiores taxas de mortalidade eram abastecidos por duas companhias de água, as quais captavam água do rio Tâmisa, num ponto abaixo da cidade. Ao estudar os dados de outra epidemia, verificou ter ocorrido mudança de local de captação de água de uma das companhias para um ponto do rio livre dos efluentes dos esgotos da cidade. Snow conseguiu com isso estabelecer que a disseminação da doença se dava por meio de água contaminada.

A partir desta experiência inicial, registrando os casos juntamente com sua localização no espaço, seguiram-se outros trabalhos. Verificou-se que a localização do evento no espaço em muito poderia contribuir para a localização de áreas de maior ocorrência e a descoberta de fatores etiológicos e populações suscetíveis, além de permitir a atuação mais efetiva, direcionada para áreas de maior necessidade.

A construção de mapas, no entanto, não é tarefa simples. Como foi relatado em vários trabalhos na literatura (ELLIOTT e WARTENBERG, 2004; RICHARDSON *et al.*, 2004), existem muitos problemas associados à construção de mapas como a escolha da medida a ser apresentada, da escala e do padrão de cores a ser empregado. Assim, para um mesmo fenômeno podem ser construídos diferentes mapas, mostrando cenários bem diferentes.

Inicialmente deve-se escolher uma escala para a construção do mapa. Qualquer que seja a escolha, ela sempre resultará em alguma forma de alisamento da informação original. As escolhas dos símbolos a serem empregados ou a forma de representação dos eventos também podem ter efeito na percepção visual do usuário do mapa. Adicionalmente, os processamentos como a interpolação e o alisamento, por vezes necessários, também podem levar a interpretações equivocadas (LAWSON, 2005).

Agregar os dados em áreas maiores também tem efeitos sobre a interpretação dos mapas. A agregação implica em uma mudança de escala, uma vez que, acumular as observações em unidades espaciais maiores muda a escala de análise, agindo como um alisamento. Segundo LAWSON (2005), o mapa é uma ferramenta de visualização e juntamente com ele devem ser apresentadas informações estatísticas relevantes. Devem ser apresentadas, pelo menos, as taxas e as estimativas de suas variabilidades para cada região.

Algumas questões envolvendo a interpretação de mapas foram estudadas no âmbito da ciência cognitiva. Nesses estudos, certos aspectos da apresentação dos mapas foram investigados em relação à habilidade dos observadores em detectar aglomerados de casos. Esses estudos, em sua maioria, focaram na construção de mapas temáticos, isto é, mapas onde esquemas de cores são usados para representar o risco em uma dada região, tendo estabelecido que (LAWSON, 2005):

- mapas monocromáticos levam a menor variabilidade por parte dos observadores na detecção de áreas de risco; recomenda-se a construção de mapas em escalas de cinza;

- mapas construídos com o emprego de duas cores tem maior variabilidade, mas são preferidos pelos observadores;

- mapas de pontos de densidade tendem a enfatizar pequenos aglomerados e induzem à maior variabilidade na identificação de áreas de risco.

No intuito de contornar os problemas associados com a construção dos mapas, foram desenvolvidas várias técnicas de análise espacial, cujo objetivo era mostrar a distribuição das doenças. Dentre essas técnicas pode-se citar as técnicas de alisamento

(como média móvel e alisamento bayesiano empírico e hierárquico), modelos de regressão espacial e modelos de mistura finita, entre outros.

Na abordagem bayesiana empírica, assume-se que os parâmetros do modelo são conhecidos e obtém-se inicialmente a distribuição *a posteriori* dos parâmetros de interesse. A seguir, os parâmetros do modelo são estimados por meio da distribuição marginal dos dados e as inferências são baseadas na distribuição *a posteriori* estimada. No modelo Bayesiano Hierárquico é necessário especificar inicialmente uma distribuição *a priori* para os parâmetros e, a seguir, a distribuição *a posteriori* dos parâmetros de interesse é obtida. As inferências são baseadas nas distribuições *a posteriori*, sendo o parâmetro de interesse estimado por meio de sua média *a posteriori* e a sua precisão por meio da variância *a posteriori* (GHOSH e RAO, 1994). A compreensão e implementação desses modelos, no entanto, não é trivial, o que dificulta sua utilização. Os modelos de misturas finitas serão abordados com maior detalhamento na seção 3.4.

O modelo de mistura finita assume que a população é composta por g subpopulações homogêneas, cada qual com um risco diferente, que pode ser dependente de diferentes covariáveis. Eles consistem em uma das abordagens usadas para a construção de mapas de doenças. Sua primeira utilização remonta ao fim do século XIX, quando Pearson modelou medidas de caranguejos da Baía de Nápoles e conseguiu, por meio desta modelagem, verificar a existência de duas subespécies. Na estimação dos parâmetros do modelo foi empregado o método dos momentos. No entanto, a dificuldade de implementação do modelo e uma série de questões teóricas a ele relacionadas fizeram com que este tipo de modelagem permanecesse pouco utilizado por um longo período. Na década de 60 do século XX surgiram os primeiros registros do uso de modelos de mistura com estimação de parâmetros feita por máxima verossimilhança. No entanto, foi a partir da publicação do artigo de DEMPSTER, LAIR e RUBIN (1977) sobre o algoritmo Maximização da Esperança (EM) que houve maior difusão do uso de modelos de mistura finita. Isso ocorreu devido ao fato do ajuste de modelos de mistura finita com parâmetros estimados por máxima verossimilhança ser um exemplo clássico de um problema que é consideravelmente simplificado por este algoritmo. O algoritmo EM foi desenvolvido para situações onde há dados incompletos. No caso dos modelos de mistura, se desconhece a classificação de cada uma das áreas.

Os modelos de mistura finita representam uma ferramenta para a modelagem estatística de uma grande variedade de fenômenos aleatórios com distribuições de probabilidade complexas. A importância deste tipo de modelagem, seu desenvolvimento e sua freqüente aplicação ao longo dos anos recentes, se devem ao fato destes modelos proporcionarem uma forma natural de modelar dados heterogêneos, isto é, compostos por subpopulações (MILITINO *et al.*, 2001; RATTANASIRI *et al.*, 2004; BEKLAS *et al.*, 2005; LIU *et al.*, 2006, VILLATE *et al.*, 2006). Ao longo das duas últimas décadas, os modelos de mistura finita vêm recebendo contínua atenção, tanto do ponto de vista prático quanto teórico (BÖHNING *et al.*, 2000; B BÖHNING e SEIDEL, 2003; BÖHNING, 2003).

Os modelos de mistura finita foram aplicados com sucesso nas áreas de biologia, genética, geografia, medicina, economia, engenharia, entre outras. Nessas aplicações, os modelos de misturas finitas são utilizados em uma grande variedade de áreas da estatística, incluindo análise de aglomerados e de classes latentes, análise de discriminantes, análise de imagem, análise de sobrevida, entre outras. Vale também destacar seu papel direto na análise e inferência de dados, fornecendo modelos descritivos para as distribuições de probabilidade (MCLACHLAN e PEEL, 2001).

Em anos recentes, foram publicados vários trabalhos empregando esta técnica para a construção de mapas de doenças (MILITINO *et al.*, 2001; BÖHNING, 2003; RATTANASIRI *et al.*, 2004; UGARTE *et al.*, 2004). Acredita-se que esta modelagem seja uma alternativa para a construção de mapas registrando a situação real de risco das áreas, sendo capaz de diferenciar áreas e identificar diferentes níveis de risco. Nesta área, os modelos de mistura são empregados no intuito de fornecer uma melhor descrição da função densidade de probabilidade dos dados e também pode ser visto como uma técnica para detecção de aglomerados, sendo classificado como tal por alguns autores. A seguir são apresentados alguns trabalhos encontrados na literatura e que empregam a técnica de misturas finitas na etapa de modelagem e na construção de mapas de risco. Também são apresentados trabalhos onde houve inclusão de covariáveis ao modelo de mistura finita (modelos de regressão mista).

AYUTHYA e BÖHNING (1995) apresentaram estudo piloto sobre acidentes de trânsito na cidade de Bangkok, sendo os dados obtidos a partir de registros policiais de quatro delegacias. Os objetivos principais do trabalho foram o desenvolvimento de uma

medida da densidade dos acidentes e a identificação de grupos de risco. Para tanto, ajustou-se um modelo de mistura finita com componentes pertencendo a uma distribuição normal. A estimação dos parâmetros foi realizada por máxima verossimilhança com utilização do algoritmo EM. Após a estimação dos parâmetros, cada área foi classificada no grupo de risco para o qual a probabilidade *a posteriori*, obtida pela utilização do teorema de Bayes, fosse maior. Os resultados permitiram avaliar a estrutura espacial do risco e a identificação de três grupos com riscos diferenciados.

SCHLATTMANN e BÖHNING (1993) apresentaram um estudo cujo objetivo era a investigação de métodos a serem utilizados na construção de mapas de riscos de doenças. Os dados utilizados foram casos de hepatite B ocorridos em Berlim no ano de 1989 e a medida epidemiológica escolhida para ser apresentada no mapa foi a SMR. Apresentam a abordagem habitualmente utilizada na construção dos mapas, a qual consiste da utilização de percentis da distribuição da SMR. Comentam, contudo, que esta escolha é arbitrária, não sendo garantia de classificação correta das áreas segundo o risco. A seguir, apresentam os modelos de mistura finita como método alternativo para a construção de mapas. Desta forma, ajustam um modelo de quatro componentes para os quais assume-se uma distribuição de Poisson. Os resultados da aplicação deste tipo de modelagem permitiram verificar que ela representa uma forma mais adequada de construção de mapas de risco para doenças.

Ainda com o intuito de aumentar o conhecimento sobre modelos a serem empregados no mapeamento de doenças, BÖHNING *et al.* (2000) apresentaram a modelagem de casos de câncer em mulheres da Alemanha Oriental, registrados no período de 1960 a 1989. Foram empregadas as seguintes modelagens: modelo de mistura finita considerando a série como um único período de tempo; modelo de mistura considerando diferentes janelas de tempo; e uma regressão de Poisson mista, onde o tempo foi incluído como covariável. Os modelos apresentados foram comparados em termos de seus benefícios, dificuldades e facilidade de interpretação dos resultados. Os parâmetros foram estimados por máxima verossimilhança via EM e os modelos foram comparados usando o Critério de Informação Bayesiano (BIC – *Bayesian Information Criterion*). Concluiu-se que o modelo que incorporava tempo e espaço simultaneamente (regressão de Poisson mista) teve desempenho muito bom para os dados utilizados.

RATTANASIRI *et al*. (2004) aplicaram um modelo de mistura finita com o intuito de investigar a distribuição geográfica dos casos de malária ocorridos no período de 1995 a 1997 na Tailândia. Também ajustaram um modelo de mistura com incorporação do tempo. Os mapas construídos empregando a abordagem de mistura finita foram comparados com técnicas tradicionais de mapeamento. Concluíram que o modelo de mistura ajustado aos dados mostrou melhor resultado que as técnicas tradicionais, uma vez que foi capaz de reduzir a variabilidade aleatória e produzir um mapa que representava de forma mais fiel a situação de risco.

DALRYMPLE *et al.* (2003) examinaram a incidência da Síndrome de Morte Súbita na Infância (SIDS) em Catemburry, no período de 1973 a 1989, em relação ao clima. Três tipos de modelos de misturas foram empregados no intuito de evidenciar os efeitos de covariáveis climáticas como temperatura e umidade entre os meses com e sem registros de SIDS. Segundo os autores, esses métodos foram selecionados para a modelagem dos dados devido a sua capacidade de acomodar os zeros extras da série (fenômeno da superdispersão), os dados heterogêneos e a autocorrelação presente em séries de dados de SIDS. Relataram ainda que a análise realizada levou a uma melhor compreensão da associação entre o clima e as mortes por SIDS. Os resultados mostraram que os meses onde ocorreram flutuações de temperatura estavam associados com maior risco de ocorrência de SIDS, quando comparados com aqueles em que a temperatura permaneceu praticamente constante.

SCHLATTMANN *et al.* (1996) empregam a abordagem dos modelos de misturas finitas para a modelagem de casos de tuberculose ocorridos na cidade de Berlim, Alemanha, em 1991. Uma vez identificada a heterogeneidade dos dados, aplicaram um modelo de regressão mista com o intuito de incluir na análise variáveis explanatórias. Como indicador de pobreza utilizam o percentual de pessoas dependentes do seguro social. Argumentam que, como observado em outros estudos realizados em países desenvolvidos, que a tuberculose é mais prevalente em pessoas nascidas em outros países, sendo esta covariável também incluída no modelo. Ajustam um modelo de dois componentes e verificam que as duas covariáveis inseridas no modelo são significativas. Um dos principais objetivos foi obter estimativas válidas para os SMR. Apresentam um modelo para a inclusão de covariáveis no modelo de mistura e concluem que a inclusão das mesmas no modelo supera a dificuldade do modelo de mistura univariado de não ser capaz de incluir a autocorrelação.

CHANDRASEKARAN e ARIVARIGNAN (2006) relatam que estudos anteriores apresentaram dados de setores censitários relativos à doenças infecciosas como a tuberculose, categorizando-os em quartis, como forma de ajudar as autoridades a encontrar áreas de maior e menor risco. Empregaram os modelos de mistura finita, especificamente uma mistura de distribuições de Poisson, para classificar grupos de risco para tuberculose e como forma de representação da distribuição espacial da incidência desta doença na cidade de Madurai, Índia. O mapa coroplético apresentando a incidência de tuberculose segundo quantis se mostrou inadequado para representar a variação geográfica da doença pois os dados exibiam superdispersão.

Capítulo 3: Bases Teóricas

3.1 Análise Espacial de Dados Epidemiológicos

A análise de dados espaciais compreende um conjunto de técnicas para coleta, descrição, visualização e análise de fenômenos que possuem coordenadas geográficas, as quais devem estar presentes na análise exploratória e na modelagem de dados.

Nas últimas décadas, o avanço na área de Computação possibilitou o processamento e a armazenagem de grandes volumes de dados, motivando o desenvolvimento de técnicas para a sua análise. Grande parte das técnicas desenvolvidas para a análise espacial de dados se apóia na premissa de que regiões mais próximas no espaço possuirão maior correlação entre si do que com outras regiões mais afastadas. Ao contrário do que ocorre na análise de séries temporais (onde as observações possuem ordenação natural), a correlação espacial ocorre em todas as direções no espaço.

Os dados espaciais podem ser classificados em quatro categorias (BAILEY e GATRELL, 1995): dados geoestatísticos, padrões pontuais, dados de áreas e dados de interação espacial.

Os dados geoestatísticos constituem amostras de medidas contínuas, geralmente de natureza ambiental como, por exemplo, temperatura e pluviosidade. Habitualmente, existirão várias estações de coleta nas quais a superfície será observada, gerando as variáveis aleatórias de interesse. A técnica usualmente utilizada para a análise deste tipo de dado é a *krigagem*, a qual consiste de uma regressão cujos erros são correlacionados pelas distâncias entre os pontos de coleta dos dados. O variograma, uma função que mede o grau de independência entre os erros segundo a distância, é a ferramenta utilizada para a definição da correlação dos erros. Os objetivos deste tipo de análise são: predizer a variável de interesse em localizações não monitoradas, escolher localizações para novos pontos de coleta e interpolar os pontos de coleta para a obtenção da superfície de interesse.

Os padrões pontuais constituem um tipo de dado espacial onde o interesse principal está no conjunto de coordenadas geográficas representando as localizações exatas dos eventos. A análise deste tipo de dado tem como objetivo identificar padrões aleatórios e aglomerados espaciais (pontos quentes). A metodologia aplicada é denominada análise de padrões pontuais. Nela, diante da identificação de aglomerados são realizados testes de hipóteses para verificar a significância estatística dos mesmos, além da aplicação de um conjunto de técnicas que incluem, por exemplo, as funções k e métodos de suavização baseados em *kernel* (BAILEY e GATRELL, 1995). Normalmente, o objetivo da análise deste tipo de dado é verificar a existência de fontes poluentes ou outros fatores de natureza variada que estejam espacialmente próximos aos pontos quentes e que possam estar influenciando ou não a ocorrência dos mesmos.

Os dados de área constituem uma agregação das observações do fenômeno ou variável de interesse para uma área definida arbitrariamente. Esta área pode ser uma grade regular, como no caso do sensoriamento remoto, ou uma grade irregular, definida arbitrariamente (como municípios ou condados de um estado ou bairros de um município). Os objetivos da análise são: verificar se o padrão das ocorrências é aleatório, obter mapas suavizados para o fenômeno em questão e verificar a existência de áreas com valores mais elevados que o esperado. A cada área i estará associado um valor da variável de interesse Y. Exemplos deste tipo de dado são muitos: número de casos de uma doença em um município, número de homicídios em uma dada região. As metodologias para a sua análise são habitualmente denominadas de análise de dados de área. Apesar do valor y_i estar associado a toda a área de interesse, é comum, em algumas técnicas, a definição de um ponto de referência dentro da área, sendo usualmente utilizado o seu centróide.

Finalmente, o quarto tipo de dado analisado pela estatística espacial é o de interação espacial. Neste tipo, existem estações ou posições fixas, da mesma forma como nas superfícies aleatórias. As estações são consideradas como pares ordenados, sendo um dos elementos do par a origem (i) e o outro o destino (j). Desta forma, o dado a ser modelado é da forma y_{ij} e representa o fluxo da estação i para a estação j. Exemplos deste tipo de dado são o fluxo de pacientes entre centros de atendimento, o fluxo de passageiros entre cidades e o fluxo entre residências e determinadas facilidades ou serviços (hospitais, supermercados). Nestes casos, a análise tem por objetivo entender e construir modelos para o fluxo e usar esta informação para a predição sobre

como o fluxo pode se modificar mediante diferentes cenários (BAILEY e GATRELL, 1995).

Para cada um dos tipos de dados espaciais apresentados existem diferentes técnicas para a descrição do fenômeno, para o ajuste de um modelo e para a realização de inferências.

A análise espacial pode ser dividida em duas etapas: análise exploratória e modelagem. Na primeira são realizados procedimentos que visam descrever e explorar o conjunto de dados, sendo, para tanto, construídos gráficos como histogramas, diagramas de caixas, diagramas de dispersão e mapas. Estes últimos são obtidos pela representação direta da variável de interesse, ou de alguma transformação desta, no mapa da região em questão. É comum a construção de mapas suavizados, ou alisados, como uma forma de obter melhores subsídios para a exploração do fenômeno. A etapa da análise exploratória é fundamental para que se possam desenvolver hipóteses sobre o processo subjacente ao fenômeno em estudo e com isso permitir a construção de um modelo para os dados.

Na análise espacial são investigados dois tipos de efeitos: os efeitos de primeira ordem, também chamados de variações em larga escala, que se referem a alterações na média do processo no espaço; e os efeitos de segunda ordem ou de pequena escala e que refletem as variações dos valores da variável de estudo, levando em consideração os valores de unidades vizinhas. Modelos para dados espaciais devem considerar a combinação dos efeitos de larga e pequena escala.

Na análise de dados espaciais, dois conceitos importantes surgem e devem ser verificados: a estacionariedade, que ocorre quando as propriedades estatísticas independem da localização do evento, e a isotropia, que ocorre quando a covariância depende somente da distância entre os pontos e não da direção entre eles. Com relação a estacionariedade, os processo podem ainda ser classificados como: estritamente estacionários, quando todas as propriedades estatísticas são insensíveis à rotação e à translação; fracamente estacionários, quando a média é constante e a covariância é independente das localizações, dependendo somente da separação e direção entre os eventos; e estacionário em incrementos, quando a variância das diferenças de primeira ordem entre os pontos não depende de sua localização.

A aquisição e o armazenamento de dados espaciais é um tópico de desenvolvimento recente e está contido na área de Sistemas de Informação Geográfica (SIG). Estes consistem de sistemas que fornecem ferramentas para a coleta, o armazenamento, a manipulação, a análise e a visualização de dados georeferenciados. A construção das bases de dados constituem o principal problema deste tipo de sistema.

Os dados analisados no presente trabalho são dados de área (irregulares). A seguir, serão descritas as técnicas estatísticas comumente empregadas na análise deste tipo de informação. Na seqüência, será abordado o mapeamento de doenças, tema deste trabalho.

Análise de dados de área

Os dados de área têm como principal característica o fato de representarem uma agregação de valores da variável de interesse, os quais se encontram dispersos dentro de cada uma das áreas estabelecidas. Tais dados possuem grande importância para a Saúde Pública, uma vez que as informações analisadas normalmente se encontram agregadas por bairros, municípios ou estados. Isto ocorre porque a dificuldade e os custos para a coleta dos dados, além de questões de natureza ética, impõem dificuldades para a obtenção da localização exata dos eventos (padrões pontuais). A construção de mapas de atributos epidemiológicos é a maneira usual de mostrar a distribuição espacial de uma doença.

Um dos principais problemas associados à análise de dados de área irregulares, que constituem o tipo de dado mais utilizado em Saúde Pública, é a divisão arbitrária das unidades de área, sendo freqüentemente utilizadas divisões políticas (bairros, condados, municípios, estados), distritos sanitários, setores censitários, distritos escolares, entre outros. Este problema é referido na literatura como unidade de área modificável (BAILEY e GATRELL, 1995), e faz com que os resultados das análises dependam da definição de área empregada no estudo.

Com relação à análise de dados de área, inicialmente, é necessário descrever a estrutura espacial das áreas a serem trabalhadas. Usualmente, isto é feito por meio da construção de uma matriz quadrada $n \times n$ de distâncias entre as áreas e cuja notação habitual é $\boldsymbol{W}$, onde $w_{ii} = 0$, sendo chamada de matriz de proximidade espacial. Os

elementos w_{ij} desta matriz medem o grau de proximidade entre as áreas i e j. Existem vários critérios que podem ser utilizados para a construção desta matriz, a qual pode ser simétrica ou não. Uma das escolhas mais freqüentes consiste na utilização de uma matriz binária, com $w_{ij} = 1$ caso as áreas i e j compartilhem uma fronteira, e $w_{ij} = 0$, em caso contrário, obtendo-se desta forma uma matriz simétrica. Uma matriz assimétrica será construída quando, por exemplo, w_{ij} for proporcional à distância cartesiana entre os centróides ou centros populacionais das unidades geográficas i e j. Outras escolhas comuns para a construção da matriz de pesos espaciais são:

- $w_{ij} = 1$ se o centróide da área j é um dos k centróides mais próximos da área i e $w_{ij} = 0$ em caso contrário;

- $w_{ij} = 1$ se o centróide da área j está a uma distância específica de i e $w_{ij} = 0$ em caso contrário.

- $w_{ij} = 1$ se a área i tem fronteira com a área j e $w_{ij} = 0$ em caso contrário;

- $w_{ij} = I_{ij}/I_i$, onde: I_{ij} é o comprimento da fronteira comum entre i e j, I_i é o perímetro da área i.

Comumente realiza-se a padronização da matriz de proximidade espacial. Neste caso, os pesos w_{ij} associados à área i totalizarão 1, isto é, a soma de cada uma das linhas da matriz totalizará 1 (BAILEY e GATRELL, 1995). Alguma vezes, com o intuito de melhor estudar as dependências espaciais, é necessário definir matrizes de proximidade de ordens superiores, ou seja, definem-se matrizes de proximidade de primeira ordem, de segunda ordem e de ordens superiores.

Uma vez definida a matriz de proximidade espacial, o passo seguinte consiste na utilização de métodos que permitam verificar como a média do atributo de interesse varia em função da região de estudo, ou seja, analisar os efeitos de primeira ordem ou variações de larga escala. Estas podem ser facilmente identificadas por meio do ajuste de uma média móvel, a qual pode ser obtida por:

$$\hat{\mu}_j = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}} \qquad (3.1)$$

onde:

w_{ij} é a distância entre as áreas *i* e *j* segundo o critério de vizinhança adotado;

y_j é a resposta da variável de interesse.

Quando a matriz de proximidade é normalizada, a média móvel se reduz a seguinte expressão:

$$\hat{\mu}_j = \sum_{j=1}^{n} w_{ij} y_j \tag{3.2}$$

A média móvel permite estimar variações globais e tendências espaciais. No entanto, freqüentemente é necessário explorar também a dependência espacial, isto é, como os valores se correlacionam no espaço. Várias técnicas foram sugeridas no intuito de analisar as variações de pequena escala, sendo as medidas de correlação espacial ou autocorrelação as mais utilizadas. Estas medem o quanto o valor observado de um atributo em uma dada área independe dos valores desta mesma variável em áreas vizinhas. O I de Moran é a medida de uso mais freqüente, sendo obtida por meio da seguinte fórmula (BAILEY e GATRELL, 1995):

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2 \right) \left((\sum \sum_{i \neq j} w_{ij} \right)} \tag{3.3}$$

onde

n corresponde ao número de áreas;

w_{ij} são os pesos obtidos a partir da matriz de proximidade $\boldsymbol{W}$;

$\bar{y}$ é a média do atributo na região;

y_i é o valor do atributo na área *i* considerada.

O resultado do índice I de Moran não está limitado ao intervalo [-1,1], como o coeficiente de correlação de Pearson. No entanto, seus valores com freqüência se localizam dentro deste intervalo e costuma-se interpretá-los da seguinte forma: valores positivos indicam uma correlação positiva entre as áreas e valores negativos, uma correlação negativa.

Após calcular o índice I de Moran, é necessário estabelecer a sua significância estatística e, desta forma, testar a hipótese nula de independência entre as áreas. Duas são as abordagens habitualmente encontradas para a realização do teste. Uma delas considera a distribuição empírica simulada da estatística e é construída por meio da geração de permutações nos valores da variável de interesse. Se o valor obtido para o índice se localizar na área de rejeição da hipótese nula da distribuição simulada, então se considera que se trata de um evento com significância estatística. Outra forma de verificar a significância estatística do índice I de Moran consiste em supor que a sua distribuição amostral é aproximadamente normal, com média e variância dadas por (CLIFF e ORD, 1981):

$$E(I) = -\frac{1}{(n-1)} \tag{3.4}$$

$$\sigma^2 = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2} \tag{3.5}$$

onde

$$S_0 = \sum_{i=1}^{n} \sum_{j,j\neq i}^{n} w_{ij}$$

$$S_1 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j,j\neq i}^{n} (w_{ij} + w_{ji})^2$$

$$S_2 = \sum_{k} \left(\sum_{j} w_{kj} + \sum_{i} w_{ik} \right)^2$$

A generalização do I de Moran pode ser de maior interesse. Desta forma, são calculadas as correlações para diferentes faixas de distância, produzindo um gráfico chamado de correlograma. Para o índice de Moran, pode-se calcular a correlação espacial para a distância *k* como:

$$I^{(k)} \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^{(k)} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2 \right) \left(\sum \sum_{i \neq j} w_{ij}^{(k)} \right)} \tag{3.6}$$

ANSELIN (1996) propôs uma forma adicional para apresentação dos resultados do índice I de Moran. Nesta se procura visualizar espacialmente o relacionamento entre os valores observados *Z* e as médias locais *WZ*. A associação linear entre elas pode ser explorada, identificando diferentes regimes espaciais presentes nos dados. Esta abordagem apresenta o índice de Moran da seguinte forma:

$$I = \frac{Z^t WZ}{Z^t Z} \tag{3.7}$$

onde *Z* é o vetor de desvios e *WZ* é o vetor de médias.

O resultado pode ser visualizado por meio de um diagrama de espalhamento, denominado gráfico de espalhamento de Moran, ou por meio de um mapa cloroplético, chamado mapa de espalhamento de Moran. O gráfico permite a análise da variabilidade espacial, a qual é dividida nos seguintes quadrantes: Q1 (valores positivos e médias positivas) e Q2 (valores negativos e médias negativas) indicam correlação espacial positiva, isto é, vizinhos com atributos semelhantes; Q3 (valores positivos e médias negativas) e Q4 (valores negativos e médias positivas) indicam correlação espacial negativa, isto é, vizinhos com atributos diferentes. No mapa de espalhamento de Moran os rótulos “alto-alto”, “baixo-baixo”, “alto-baixo” e “baixo-alto” indicam, respectivamente, os quadrantes 1, 2, 3 e 4.

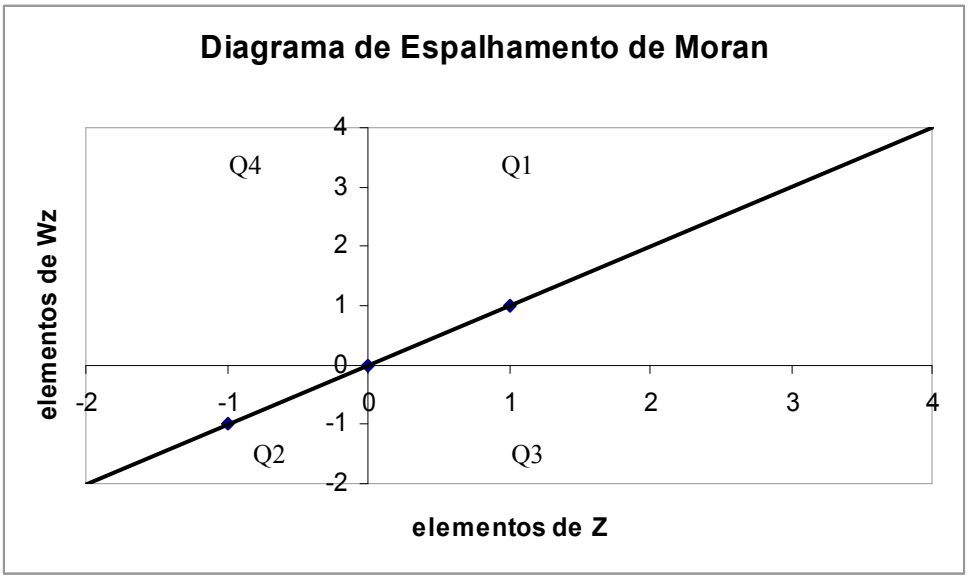


Figura 3.1: Diagrama de espalhamento de Moran

Os indicadores globais de autocorrelação, como o índice I de Moran, fornecem uma medida única da autocorrelação. Algumas vezes é interessante examinar a dependência espacial de forma mais detalhada, podendo-se calcular, para tanto, indicadores locais. Estes indicadores locais fornecem um valor específico para cada área permitindo assim a identificação de "bolsões" de dependência espacial, além da identificação de valores extremos. O índice local de Moran (ANSELIN, 1995) e os índices G_i e G_i^* (GETIS e ORD, 1992) são exemplos de indicadores locais.

Caso sejam utilizados os desvios, z_i, em relação à média, o índice local de Moran pode ser calculado por:

$$I = \frac{z_i \sum_j w_{ij} z_j}{\sum_{i=1}^{n} z_i^2} \qquad (3.8)$$

Assim como no índice I de Moran global, também é necessário verificar a significância estatística dos resultados encontrados para os indicadores locais. Os resultados podem ser apresentados em um mapa denominado "LISA MAP", onde indicadores locais são classificados em três grupos: não significantes, com significância de 95% e com significância de 99%.

3.2 Modelo linear generalizado

O modelo linear generalizado consiste de uma generalização do modelo de regressão linear, que permite grande flexibilidade na modelagem de todos os tipos de dados em situações variadas, fornecendo uma estrutura unificada para a realização das análises.

Antes de definir os pressupostos do modelo e como ajustá-lo, é fundamental comentar sobre a família exponencial de distribuições de probabilidade. Pode-se dizer que uma variável aleatória Y tem distribuição de probabilidade pertencente à família exponencial se sua função de massa (ou densidade) de probabilidade puder ser reescrita da seguinte forma (KRZANOWSKI, 1998):

$$f(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\} \tag{3.9}$$

onde *a(.)*, *b(.)* e *c(.)* são funções específicas e θ e ϕ são os parâmetros. Se ϕ é conhecido, então a família é dita exponencial e θ será o seu parâmetro natural ou canônico.

O modelo linear generalizado tem como pressupostos (KRZANOWSKI, 1998):

(i) As respostas Y_i consistem de variáveis aleatórias com distribuição pertencente à família exponencial;

(ii) As variáveis explanatórias fornecem uma série de preditores lineares dados por: $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$, para $i = 1,..., p$;

(iii) A ligação entre (i) e (ii) é dada por $g(\mu_i) = \eta_i$, onde μ_i é a média da variável Y_i para $i = 1,..., n$ e $g(.)$ é a função de ligação do modelo.

Existem várias funções de ligação que podem ser utilizadas. Contudo, o modelo pode ser simplificado se for escolhida como função de ligação a mesma função que define o parâmetro canônico. Neste caso, a função de ligação será dita canônica. A seguir são listadas as funções de ligação canônicas para as distribuições mais comumente empregadas:

- Distribuição normal: função identidade, $g(\mu) = \mu$;
- Distribuição binomial: função logit, $g(\mu) = \log\{\mu/(n-\mu)\}$
- Distribuição de Poisson: função log, $g(\mu) = \log \mu$
- Distribuição gama: função recíproca, $g(\mu) = 1/\mu$

A equação (3.9) é uma função massa (densidade) de probabilidade e, desta forma:

$$\int_R \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\}dy = 1 \tag{3.10}$$

Se um conjunto de variáveis Y_i satisfaz aos pressupostos do modelo linear generalizado, sua função de verossimilhança será dada por (KRZANOWSKI, 1998):

$$L(\theta_1,...,\theta_n;\phi) = \prod_{i=1}^{n} \exp\{[y_i\theta_i - b(\theta_i)]/a_i(\phi) + c(y_i,\phi)\} \tag{3.11}$$

e a log-verossimilhança corresponderá a:

$$l(\theta_1,...,\theta_n;\phi) = \sum_{i=1}^{n} \frac{[y_i\theta_i - b(\theta_i)]}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i,\phi) \tag{3.12}$$

Os parâmetros $\beta_1, ..., \beta_p$ do preditor linear constituem as quantidades de interesse, as quais serão obtidas por meio da solução de um sistema de equações. A solução será obtida de forma iterativa sendo utilizados métodos como o de Newton-Raphson e o método dos escores. Resumidamente, estes funcionam da seguinte forma: atribui-se um valor inicial para o parâmetro de interesse e então, a cada iteração, este valor será modificado, gerando uma nova seqüência de estimativas $\beta_1, ..., \beta_p$. Isso ocorrerá até o ponto em que a diferença entre os sucessivos vetores gerados for inferior a uma tolerância previamente estabelecida. Quando isto ocorrer diz-se que o processo convergiu (KRZANOWSKI, 1998).

O passo seguinte do processo de modelagem consiste em verificar o ajuste do modelo. Desta forma, é necessário utilizar-se de estatísticas que quantificarão a diferença entre os valores ajustados e os valores observados. Com este intuito emprega-se usualmente o teste da razão de verossimilhanças, que consiste de uma razão de

verossimilhanças entre o modelo ajustado e o modelo saturado, isto é, aquele onde o número de parâmetros é igual ao número de observações. A razão é dada por (KRZANOWSKI, 1998):

$$\Lambda = \frac{\exp \sum_{i=1}^{n} \{[y_i \hat{\theta}_i - b(\hat{\theta}_i)] / a_i(\phi) + c(y_i, \phi)\}}{\exp \sum_{i=1}^{n} \{[y_i \tilde{\theta}_i - b(\tilde{\theta}_i)] / a_i(\phi) + c(y_i, \phi)\}}$$

$$= \exp \sum_{i=1}^{n} \{[y_i (\hat{\theta}_i - \tilde{\theta}_i) - b(\hat{\theta}_i) + b(\tilde{\theta}_i)] / a_i(\phi)\} \tag{3.13}$$

Logo,

$$-2 \log \Lambda = 2 \sum_{i=1}^{n} \{[y_i (\hat{\theta}_i - \tilde{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)] / a_i(\phi)\} \tag{3.14}$$

No caso especial em que $a_i(\phi) = \phi$ tem-se que $-2 \log \Lambda = D/\phi$ onde D é dado por:

$$D = 2 \sum_{i=1}^{n} [y_i (\hat{\theta}_i - \tilde{\theta}_i) - b(\hat{\theta}_i) + b(\tilde{\theta}_i)] \tag{3.15}$$

A equação (3.15) é conhecida como *deviance* do modelo e é uma medida da discrepância existente entre o modelo ajustado e um modelo saturado. No intuito de decidir se o valor encontrado para a razão de verossimilhanças é estatisticamente significativo ou não, realiza-se um teste de hipótese, onde a hipótese nula, de que o modelo ajustado é correto, assume uma distribuição χ^2 com *(n-p)* graus de liberdade (KRZANOWSKI, 1998).

Ainda visando verificar a adequação do modelo ajustado aos dados pode ser realizada uma análise de seus resíduos, os quais consistem da diferença entre o valor observado e o valor ajustado através do modelo. Dentre os resíduos mais empregados na análise do ajuste de modelos lineares generalizados pode-se citar: o resíduo de Pearson padronizado, o resíduo *deviance* padronizado e o resíduo de verossimilhança, os quais são apresentados nas equações 3.16, 3.17 e 3.18, respectivamente.

$$r_{i(P)} = \frac{y_i - \hat{\mu}_i}{\sqrt{[(1-h_i)\,\hat{\text{var}}(y_i)]}} \tag{3.16}$$

$$r_{i(D)} = \frac{\text{sgn}(y_i - \hat{\mu}_i)/\sqrt{D_i}}{\sqrt{(1-h_i)}} \tag{3.17}$$

$$r_{i(L)} = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{h_i r_{i(P)}^2 + (1-h_i) r_{i(D)}^2} \tag{3.18}$$

onde h_i é o i-ésimo elemento da diagonal da matriz de pesos obtida ao final do procedimento iterativo para obtenção das estimativas de máxima verossimilhança e *sgn* é o sinal de $(y_i - \hat{\mu}_i)$.

A regressão de Poisson, também chamada de modelo log-linear, é um caso especial de modelo linear generalizado, sendo utilizada na modelagem do número de eventos que ocorrem em certo intervalo. Estes modelos também podem ser empregados na modelagem de dados que consistem na contagem do número de indivíduos (ou elementos) classificados de acordo com uma combinação de fatores. Esse tipo de informação é normalmente apresentado em uma tabela de contingência. A variável resposta Y_i pode ser o número de casos de uma doença em um município, o número de filhos em uma família, o número de acidentes de carro em um bairro, entre outras. O objetivo da regressão de Poisson é a modelagem da variável resposta em função de um conjunto de variáveis explanatórias x_i, chamadas de covariáveis.

Por definição, uma variável tem distribuição de Poisson se e somente se sua função de probabilidade é dada por:

$$P(Y = k) = \frac{\exp(-\lambda)\lambda^k}{k!} \qquad k = 1,2,3,... \tag{3.19}$$

Da definição acima segue que:

$$E[Y] = \lambda \quad \text{e} \quad Var[Y] = \lambda \tag{3.20}$$

A distribuição de Poisson é uma distribuição de probabilidade discreta e sua forma varia de acordo com o parâmetro λ, sendo mais assimétrica quanto menor for o

valor de λ. Para valores grandes de λ, a forma da distribuição se parecerá com a de uma distribuição Normal.

O modelo de regressão de Poisson é dado por:

$$E[Y_i|x_i] = \exp(\beta^t x_i) \tag{3.21}$$

onde x_i representa os valores das variáveis explanatórias e $\beta = (\beta_1,...,\beta_p)^T$ corresponde ao vetor de parâmetros desconhecidos que se deseja estimar. O conhecimento sobre os parâmetros β, fornece informações sobre a importância de cada uma das variáveis explanatórias para a resposta obtida. A função de máxima verossimilhança para β é dada por:

$$LogL(y_1,...,y_n \mid \beta, x_1,...,x_n) = \sum_{i=1}^{n} \log P(Y_i = y_i \mid \beta, x_i) \tag{3.22}$$

ou ainda

$$LogL(\beta) = \sum_{i=1}^{n} \left\{ -\exp(\beta^t x_i) + y_i(\beta^t x_i) - \log(y_i!) \right\} \tag{3.23}$$

Após derivar a função dada em (3.23) obtém-se:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) x_i = 0 \tag{3.24}$$

onde $\hat{y}_i$ é o valor estimado através do modelo.

Se o modelo de Poisson for adequado aos dados com que se está trabalhando, então:

$$E[Y_i \mid x_i] = Var[Y_i \mid x_i] \tag{3.25}$$

No entanto, se $E[Y_i \mid x_i] > Var[Y_i \mid x_i]$ ou $E[Y_i \mid x_i] < Var[Y_i \mid x_i]$, pode-se dizer que os dados são subdispersos e superdispersos, respectivamente.

O fenômeno da superdispersão ocorre quando a variância da variável resposta é maior que a média. Nessas situações, a suposição de que os dados se ajustam a uma

distribuição é inadequada, devendo ser utilizados modelos alternativos. O modelo de quasi-verossimilhança é uma das possíveis abordagens para dados superdispersos. Este modelo leva às mesmas estimativas obtidas por meio da regressão de Poisson, porém corrige a variabilidade das mesmas. Uma outra abordagem possível para dados com esta característica é a regressão binomial negativa, a qual permite uma análise mais completa dos dados, sendo dada por:

$$P(Y_i = y_i \mid \beta, x_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} \left(\frac{\lambda_i}{\lambda_i + \theta} \right)^{y_i} \left(1 - \frac{\lambda_i}{\lambda_i + \theta} \right)^{\theta} \tag{3.26}$$

onde λ_i corresponde a média condicional e $\lambda_i(1+\eta^2\lambda_i)$ com $\eta^2 = 1/\theta$, corresponde a variância condicional. O parâmetro η^2 não varia com as observações. Assim como para o modelo de Poisson, a média será modelada como:

$$E[Y_i \mid x_i] = \lambda_i = \exp(\beta^t x_i) \tag{3.27}$$

e a variância condicional será dada por:

$$Var[Y_i \mid x_i] = \exp(\beta^t x_i)(1 + \eta^2 \exp(\beta^t x_i)) \tag{3.28}$$

A máxima verossimilhança será utilizada na estimação dos parâmetros β e do parâmetro extra η, o qual mede o grau de superdipersão ou subdispersão dos dados. Quando $\eta = 0$, tem-se o modelo de Poisson.

3.3 Modelo de Mistura Finita

O modelo de mistura finita assume que a função densidade de probabilidade $f(x)$ pode ser modelada como a soma ponderada de *g* densidades de probabilidade. Ele é utilizado na modelagem de populações heterogêneas, isto é, constituídas por várias subpopulações. Seja $\boldsymbol{X_1, ..., X_n}$ uma amostra aleatória de tamanho *n*, onde $\boldsymbol{X_i}$ é um vetor aleatório *p*-dimensional com função de probabilidade $f(\mathbf{x}_i; \Theta)$ com domínio em R^p. A densidade de $\boldsymbol{X_i}$ pode ser escrita da seguinte forma (MCLACHLAN e PEEL, 2001):

$$f(\mathbf{x}_i;\Theta) = \sum_{j=1}^{g} \pi_j f_j(\mathbf{x}_i;\theta_j) \tag{3.29}$$

onde $f(\mathbf{x}_i;\Theta)$ são as densidades componentes da mistura, Θ corresponde aos parâmetros das distribuições de probabilidade que compõem a mistura e π_i são seus pesos ou proporções, os quais apresentam as seguintes restrições: $0 \leq \pi_j \leq 1 \ \ (i = 1,...,g)$ e $\sum_{j=1}^{g} \pi_j = 1$, isto é, são quantidades não negativas e que totalizam 1.

Para valores do número de componentes *g* entre 1 e o tamanho amostral *n*, os modelos de misturas finitas podem ser vistos como um compromisso semi-paramétrico entre: (a) um modelo paramétrico completo representado por uma única família paramétrica (g = 1) e (b) um modelo não paramétrico como representado no caso em que *g* = *n* pelo método de *kernel* para estimação da densidade. Assim, os modelos de misturas finitas têm muito da flexibilidade da abordagem não paramétrica enquanto retêm algumas das vantagens da abordagem paramétrica (MCLACHLAN e PEEL, 2001). Os modelos de mistura podem ser vistos como paramétricos uma vez que uma forma paramétrica é atribuída a cada um dos componentes da mistura. No entanto, eles também podem ser vistos como modelos não paramétricos uma vez que permitem que o número de componentes aumente até *g*.

Estimação de Parâmetros em um modelo de mistura finita

Ao longo dos anos, uma grande variedade de abordagens foi utilizada para estimar distribuições mistas. Estas incluem métodos gráficos, método dos momentos, métodos de distância mínima, método da máxima verossimilhança e abordagens bayesianas.

A estimação dos parâmetros em um modelo de mistura finita é normalmente realizada pelo algoritmo Maximização da Esperança (*Expectation Maximization* - EM; DEMPSTER *et al.*, 1977). Este consiste de um método para otimização de funções de verossimilhança, útil em situações onde há dados faltosos ou incompletos ou em casos

onde os métodos de otimização mais simples falham. O algoritmo possui duas etapas: estimação (Passo E) e maximização (Passo M). A idéia geral do algoritmo consiste em:

- Substituir os dados ausentes por valores estimados;
- Estimar os parâmetros do modelo;
- Repetir os passos anteriores até a convergência.

O ajuste de modelos de mistura por máxima verossimilhança é um exemplo de problema que é consideravelmente simplificado pelo EM, o qual propõe um processo de estimação para dados vistos como incompletos. As distribuições mistas podem ser reformuladas como um problema de dados incompletos, onde o grupo ou subpopulação ao qual cada observação pertence é desconhecido (MCLACHLAN e PEEL, 2001).

Seja o vetor Z_j uma variável aleatória independente e identicamente distribuída e que assume valores de 1 a g, com probabilidade igual a π_1, ..., π_g, e suponha que a densidade condicional de Y_j dado $Z_j = i$ seja $f_i(y_j)$ ($i = 1,..., g$). Dentro deste contexto a variável Z_j pode ser vista como a variável indicadora do grupo ao qual pertence cada uma das observações. O vetor indicador Z_j possui distribuição multinomial consistindo em g categorias com probabilidade π_1, ..., π_g.

$$Z_j \sim Mult_g(1,\pi) \quad \text{onde} \quad \pi = (\pi_1,...,\pi_g) \tag{3.30}$$

Na abordagem proposta pelo algoritmo EM, os dados são vistos como incompletos, sendo o vetor de dados completos dado por:

$$\mathbf{y}_C = (\mathbf{y}^T, \mathbf{z}^T)^T \tag{3.31}$$

onde $\mathbf{y} = (\mathbf{y}_1^T,...,\mathbf{y}_n^T)^T$ corresponde ao conjunto de dados observados e $z = (z_1^T,...,z_n^T)^T$ corresponde às variáveis indicadoras, onde $z_{ij} = 1$ se y_j pertence ao componente j e $z_{ij} = 0$ em caso contrário (MCLACHLAN e PEEL, 2001).

A i-ésima proporção da mistura π_i pode ser vista como a distribuição *a priori* de que a observação pertença ao i-ésimo componente da mistura, enquanto que a distribuição *a posteriori* de que a observação pertença ao i-ésimo componente é dada por:

$$\begin{aligned}\tau_i(y_i) &= P\{Z_{ij}=1 \mid y_j\} \\ &= \pi_i f_i(y_j)/f(y_j) \qquad \text{onde } i=1,...,g \quad \text{e} \quad j=1,...,n\end{aligned} \tag{3.32}$$

O algoritmo EM explora a simplicidade de se trabalhar com a distribuição conjunta de Y_j e Z_j para encontrar a estimativa de máxima verossimilhança com base nos dados observados. Ele constrói a função de verossimilhança utilizando o vetor de dados completo Y_c e contorna o fato de que o componente z_j não está disponível trabalhando iterativamente com a esperança condicional da log-verossimilhança para os dados completos, o que é realizado utilizando o valor corrente dos parâmetros. A função de verossimilhança para os dados completos será dada por:

$$\log L_c(\Psi) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}(\log f_i(y_j;\theta_i) + \log \pi_i) \tag{3.33}$$

onde ψ corresponde ao vetor completo de parâmetros do modelo.

A estimativa de máxima verossimilhança para os dados completos pode ser obtida iterativamente em dois passos. No passo E é calculada a esperança condicional da verossimilhança para os dados completos.

$$\text{Passo E:} \quad Q(\Psi,\Psi^{(0)}) = E_{\Psi^{(0)}}\{\log L_c(\Psi) \mid \mathbf{y}\} \tag{3.34}$$

Uma vez que a verossimilhança para os dados completos é linear com relação a z_{ij}, o passo E se reduz ao cálculo da esperança condicional de z_{ij} dados os valores observados $\boldsymbol{y}$. Como resultado obtém-se:

$$Q(\Psi,\Psi^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})(\log f_i(\mathbf{y}_j;\boldsymbol{\theta}_i) + \log \pi_i) \tag{3.35}$$

O passo M consiste da maximização de $Q(\Psi;\Psi^{(k)})$ com relação à Ψ para fornecer a estimativa atualizada $\Psi^{(k+1)}$.

O termo z_{ij} se refere à probabilidade *a posteriori* de que a observação y_j pertença ao i-ésimo componente. Como no modelo de mistura a informação sobre os grupos aos quais as observações pertencem não está disponível, a probabilidade *a posteriori* é utilizada para estimar a informação ausente.

Alguns problemas estão associados ao modelo de mistura. Tais problemas se referem à estimação dos parâmetros pelo algoritmo EM, sendo algo comum também a outros métodos. Dificuldades comumente relacionadas são: a escolha do número de componentes do modelo; a escolha das estimativas iniciais para os parâmetros do modelo e a convergência para máximos locais durante o processo de obtenção da estimativa de máxima verossimilhança.

A escolha dos valores iniciais tem grande influência na velocidade de convergência do algoritmo e na sua capacidade de encontrar o máximo global. Dentre as sugestões descritas na literatura (KARLIS & XEKALAKI, 2003) encontram-se: a definição dos valores iniciais por meio de uma busca em uma grade de valores gerados aleatoriamente; o uso de informações suplementares com o intuito de formar aglomerados cujas médias seriam então utilizadas como valores iniciais; o uso de componentes principais; o uso de estimativas iniciais obtidas a partir de outros métodos, como o dos momentos, métodos gráficos, entre outros.

KARLIS E XEKALAKI (2003), em um trabalho sobre a escolha das estimativas iniciais em modelos de misturas, concluem que se deve começar o processo com vários conjuntos de valores iniciais e realizar um pequeno número de iterações, sem necessariamente observar sua convergência. A seguir, deve-se executar o algoritmo até a convergência, utilizando-se o conjunto que obteve a melhor estimativa de máxima verossimilhança. Segundo os autores, esta proposta reduz o tempo de execução do algoritmo e evita que ele fique por longo tempo parado em um máximo local, muitas vezes bem distante da solução global do problema. Outra proposta interessante consiste da utilização de técnicas de detecção de aglomerados, como o algoritmo k-médias, cujos centróides poderiam ser usados como estimativas iniciais para os parâmetros do modelo (HAYKIN, 2001; MILITINO, 2001).

O método k-médias (*k*-means) é uma técnica para partição de dados por meio da formação de agrupamentos na base da similaridade, buscando minimizar a distância média intra-classe. Possui os seguintes passos: (1) especificar o número de agrupamentos; (2) determinar o valor inicial para os centróides (seleção aleatória); (3) calcular a distância entre cada observação e cada centróide; (4) atribuir cada observação ao aglomerado mais próximo; (5) calcular o centróide utilizando as observações que lhe foram atribuídas; repetir os passos 3 a 5 até que não ocorram mais alterações.

A avaliação do número de componentes em um modelo de mistura finita é um problema difícil e ainda não totalmente resolvido. Na maior parte dos trabalhos sobre o assunto, a abordagem tem sido separar o problema da estimação do problema de identificação do número de componentes do modelo. Assim, primeiro se verifica o número de componentes e, a seguir, se faz a estimação para um número fixo de componentes *g* (MCLACHLAN e PEEL, 2001).

A estimação da ordem de um modelo de mistura é feita de duas formas, ambas atuando sobre a verossimilhança. O acréscimo de termos a um modelo leva a um aumento da verossimilhança. Desta forma, várias abordagens largamente empregadas, como o Critério de Informação de Akaike (AIC – *Akaike´s Information Criterion*) e o Critério de Informação Bayesiano (BIC – *Bayesian Information Criterion*), atuam penalizando a verossimilhança de acordo com o número de componentes existentes no modelo. Assim, quanto maior o número de componentes, maior será o termo a ser subtraído da verossimilhança.

Uma outra abordagem também freqüente na identificação do número de componentes do modelo é a realização de um teste de hipóteses utilizando a Razão de Verossimilhanças (LRT – *Likelihood Ratio Test*), dada por $-2\log\lambda$. Assim, verifica-se a hipótese H_0: $g = g_0$ contra H_1: $g = g_1$ onde normalmente $g_1 = g_0 + 1$. Na prática, termos vão sendo adicionados ao modelo até que o aumento da log verossimilhança não represente mudança significativa que justifique a presença de mais um componente. Verificou-se, contudo que, nos modelos de misturas, a hipótese de que a LRT assume uma distribuição χ^2, com número de graus de liberdade dada pela diferença entre o número de componentes dos dois modelos testados, não é válida. Assim, MCLACHLAN e PEEL (2001) sugerem a realização de *bootstrap* (MARTINEZ E MARTINEZ, 2002), isto é, a realização de um processo de reamostragem até que sejam obtidas amostras suficientes para a obtenção da distribuição de probabilidade de H_0. Critérios como o AIC e o BIC são mais simples de serem implementados e não enfrentam esses problemas, sendo utilizados com bastante freqüência. O AIC seleciona o modelo que minimiza o seguinte termo:

$$AIC = -2\log L(\hat{\Psi}) + 2d \tag{3.36}$$

onde *d* é o número de componentes do modelo.

O método do AIC é freqüentemente utilizado na seleção do modelo. Vários autores observaram contudo, que ele tende a superestimar o número correto de componentes do modelo (SOROMENHO, 1993; CELEUX e SOROMENHO, 1996).

O Critério de Informação Bayesiano, BIC, de SCHWARZ (1978) é dado por:

$$BIC = -2\log L(\hat{\Psi}) + d\log n \tag{3.37}$$

que corresponde a duas vezes a log verossimilhança negativa, a qual será minimizada na seleção do modelo. Como log $n > 2$ para $n > 8$, pode-se observar que o termo de penalização do BIC penaliza modelos mais complexos mais que o AIC cujo termo de penalidade $2d$ não depende do tamanho amostral. Como conseqüência, reduz-se a tendência do AIC de ajustar componentes em excesso. Por outro lado, observou-se que ele tende a ajustar poucos componentes quando o modelo para as densidades é válido e o tamanho amostral não é muito grande. Verifica-se ainda que, se o modelo para as densidades componentes não é válido, o BIC tende a ajustar muitos componentes (MCLACHLAN e PEEL, 2001).

Em diversas situações a análise de covariáveis é desejável. Os modelos de misturas finitas que incorporam covariáveis são chamados de modelos lineares generalizados mistos ou simplesmente modelos de regressão mista. Nesta abordagem, as proporções ou pesos da mistura, assim como os parâmetros das distribuições que a compõem podem depender de covariáveis. Um exemplo de situação em que as misturas de GLM aparecem na prática é em problemas onde a superdispersão está presente.

Para o i-ésimo componente da mistura, seja μ_{ij} a média dos Y_j, $h_i(\mu_{ij})$ a função de ligação e $\eta_i = h_i(\mu_{ij}) = \beta_i^T x_j$ o preditor linear ($i = 1, ..., g$). A log verossimilhança para o i-ésimo componente é dada por:

$$\log f(y_j;\theta_{ij},\kappa_i) = \kappa_i^{-1}\{\theta i_j y_i - b_i(\theta_{ij})\} + c_i(y_j;\kappa_i) \tag{3.38}$$

onde κ_i é o parâmetro de dispersão do modelo. A log verossimilhança para Ψ é dada por:

$$\log L(\Psi) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_{ij} f(y_i;\theta_{ij};\kappa_i) \tag{3.39}$$

$$\text{onde} \quad \pi_{ij} = \pi_i(\mathbf{x}_j;\alpha) \quad \text{e} \quad (i=1,...,g; j=1,...,n)$$

A log verossimilhança para o vetor completo de parâmetros Ψ no modelo misto com covariáveis é dada por:

$$\log L_c = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{\log \pi_{ij} + \log f_i(y_j;\theta_{ij},\kappa_i)\} \tag{3.40}$$

O passo E é o mesmo mostrado anteriormente para o modelo de mistura sem covariáveis. O passo M envolverá a resolução de um sistema de equações, dado por:

$$\begin{aligned}
&\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij}(y_j;\boldsymbol{\Psi}^{(k)})\partial \log \pi_{ij}/\partial\alpha = 0 \\
&\text{e} \\
&\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij}(y_j;\boldsymbol{\Psi}^{(k)})\partial \log f_i(y_j;\theta_{ij},\kappa_i)/\partial\beta = 0
\end{aligned} \tag{3.41}$$

onde

$$\tau_{ij}(y_j;\boldsymbol{\Psi}^{(k)}) = \frac{\pi_{ij} f_i(y_j;\theta_{ij}^{(k)},\kappa_i)}{\sum_{h=1}^{g}\pi_{hj}^{(k)} f_h(y_j;\theta_{hj}^{(k)},\kappa_h)}$$

onde $\Psi(\alpha^T,\beta^T)^T$ corresponde ao vetor completo de parâmetros do modelo misto e τ_{ij} corresponde a probabilidade *a posteriori* de pertinência a cada um dos componentes do modelo.

O modelo de regressão de Poisson misto, um caso especial do modelo linear generalizado misto, é empregado com freqüência no mapeamento de doenças. Ele é dado por:

$$f(y_j) = \sum_{1=1}^{g} \pi_i f(y_i;\mu_{ij}) \tag{3.41}$$

$$\text{onde} \qquad \mu_{ij} = \lambda_i e_j \exp(\beta_i^T \mathbf{x}_j) \tag{3.42}$$

Na equação dada por (3.42), e_j representa o valor esperado para a área *j* e λ_i representa a taxa do componente *i* da mistura.

Capítulo 4: Materiais e Métodos

4.1 Área de Estudo

A área estudada compreende o estado do Rio de Janeiro, um dos menores estados brasileiros e que totaliza uma área de 43.696,054 km2, equivalente a apenas 0,51% da área total do país (IBGE, 2006). Este limita-se ao norte e a noroeste com o estado de Minas Gerais, a oeste com o estado de São Paulo e a nordeste com o estado do Espírito Santo. Com relação a sua divisão político-administrativa, ele é composto por 92 municípios (CIDE, 2006) divididos em 8 regiões (Tabela 4.1).

No presente trabalho foi empregada a malha de 1997, obtida junto à Fundação Instituto Brasileiro de Geografia e Estatística (IBGE). Nesta, o estado é composto por 91 municípios. O município de Mesquita foi criado em 1999, sendo anteriormente parte do município de Nova Iguaçu. Desta forma, neste trabalho as informações de mortalidade para o ano de 2002 referentes ao município de Mesquita foram adicionadas aos valores do município de Nova Iguaçu.

Tabela 4.1: Municípios pertencentes a cada uma das regiões do estado do Rio de Janeiro

Regiões			
Metropolitana	Noroeste	Norte	Serrana
Belford Roxo	Aperibé	Campos dos Goytacazes	Bom Jardim
Duque de Caxias	Bom Jesus do Itabapoana	Carapebus	Cantagalo
Guapimirim	Cambuci	Cardoso Moreira	Carmo
Itaboraí	Italva	Conceição de Macabu	Cordeiro
Japeri	Itaocara	Macaé	Duas Barras
Magé	Itaperuna	Quissamã	Macuco
Nilópolis	Laje do Muriaé	São Fidélis	Nova Friburgo
Niterói	Miracema	São Francisco de Itabapoana	Petrópolis
Nova Iguaçu	Natividade	São João da Barra	Santa Maria Madalena
Paracambi	Porciúncula		São José do Vale do Rio Preto
Queimados	Santo Antônio de Pádua		São Sebastião do Alto
Rio de Janeiro	São José de Ubá		Sumidouro
São Gonçalo	Varre-Sai		Teresópolis
São João de Meriti			Trajano de Morais
Seropédica			
Tanguá			

Tabela 4.1 (cont.): Municípios pertencentes a cada uma das regiões do estado do Rio de Janeiro

Regiões			
Baixadas Litorâneas	Médio Paraíba	Centro-Sul	Costa Verde
Araruama	Barra do Piraí	Areal	Angra dos Reis
Armação dos Búzios	Barra Mansa	Comendador Levy Gasparian	Itaguaí
Arraial do Cabo	Itatiaia	Engenheiro Paulo de Frontin	Mangaratiba
Cabo Frio	Pinheiral	Mendes	Parati
Cachoeiras de Macacu	Piraí	Miguel Pereira	
Casimiro de Abreu	Porto Real	Paraíba do Sul	
Iguaba Grande	Quatis	Paty do Alferes	
Maricá	Resende	Sapucaia	
Rio Bonito	Rio Claro	Três Rios	
Rio das Ostras	Rio das Flores	Vassouras	
São Pedro da Aldeia	Valença		
Saquarema	Volta Redonda		
Silva Jardim			

4.2 Fontes de Dados

Os dados empregados no presente trabalho consistiram do número de óbitos por tuberculose (CID-10 A15-A19) ocorridos no período de 1997 a 2002 e do número de casos notificados de tuberculose no período de 1997 a 2000 ocorridos nos municípios do estado do Rio de Janeiro. Os dados de mortalidade foram obtidos na Internet, no sítio do DATASUS (DATASUS, 2006), e os dados de morbidade foram obtidos no sítio da Fundação CIDE (CIDE, 2006). As estimativas populacionais e covariáveis empregadas no estudo foram obtidas junto ao IBGE em seu sítio na Internet (IBGE, 2006) e no Atlas do Desenvolvimento Humano no Brasil - versão 1.0.0 (ESM Consultoria, 2003). Com relação às estimativas populacionais, foram calculadas médias para cada um dos períodos mencionados.

Optou-se por trabalhar na etapa da modelagem com a razão de mortalidade/morbidade, sendo esta obtida por meio da razão entre o número de casos observados e o número de casos estimados para o período. O número de casos esperados foi obtido da seguinte forma:

(1) calculou-se o risco (PEREIRA, 2005) absoluto para todo o estado a partir da seguinte expressão:

$$R = \frac{\sum_{i=1}^{91} O_i}{\sum_{i=1}^{91} pop_i}$$

na qual Oi corresponde ao número de casos observados em cada um dos 91 municípios que compõem o estado e popi corresponde à população de cada um dos municípios. Este cálculo resultou na medida de risco para o estado do Rio de Janeiro.

(2) Calculou-se o valor esperado para cada um dos municípios por meio de:

$$E_i = pop_i \cdot R$$

isto é, o valor esperado para cada área foi obtido pela multiplicação da população da área pelo risco para o estado.

4.3 Métodos

A seguir serão descritos os procedimentos realizados na análise exploratória de dados e na modelagem usando os modelos de mistura finita, a regressão binomial negativa e o modelo de regressão de Poisson mista.

4.3.1 Análise exploratória

Na análise exploratória foram calculadas as medidas descritivas (média, mínimo, máximo, quartis e percentual de zeros nas séries) e construídos diagramas de caixa para os dados de mortalidade e morbidade, sendo empregados para este fim os programas MATLAB® e R (THE R CORE DEVELOPMENT TEAM, 2006).

Para a realização da análise exploratória espacial foi utilizado o pacote SPDEP (BIVAND, 2006) do programa R (THE R CORE DEVELOPMENT TEAM, 2006). O primeiro passo da análise espacial consistiu da construção da matriz de proximidade espacial, a qual empregou como critério de adjacência a existência de fronteira comum. Foram construídos mapas coropléticos para as razões de mortalidade e morbidade para

uma primeira visualização da distribuição espacial do evento de interesse. Na construção dos mapas coropléticos foram utilizados dois critérios para a construção dos intervalos: (1) quintis da distribuição; e (2) k-médias.

A seguir, foi ajustada uma média móvel das razões de mortalidade e de morbidade para a verificação de alterações na média do processo e a observação de tendências espaciais.

Para a verificação da autocorrelação espacial foi empregado o I de Moran, sendo sua significância estatística verificada por meio de duas possibilidades para a hipótese nula: a normalidade e a randomização. Em ambas as situações, foram considerados como estatisticamente significativos os valores-p inferiores a 0,05.

4.3.2 Modelo de Mistura Finita

Conjuntos de valores aleatórios foram utilizados como estimativas iniciais para os parâmetros dos modelos ajustados aos dados de mortalidade e morbidade. Foram construídos modelos de misturas com número de componentes variando de dois a sete. Estes componentes consistiam de distribuições Normais e de Poisson. Empregou-se a mistura Normal-Poisson em virtude da grande dispersão observada nos dados. Por se tratar de um modelo nulo, sem covariáveis, foram inseridos apenas o número de casos/óbitos observados no período em cada um dos municípios e o valor esperado de casos, o qual foi incluído no modelo na forma de um offset. A seleção dos modelos foi realizada empregando-se o Critério de Informação Bayesiano, BIC (SCHWARTZ, 1978), sendo selecionados os mais parcimoniosos, isto é, com melhor relação entre o valor do BIC e o número de componentes.

No ajuste dos modelos de misturas finitas aos dados de mortalidade e morbidade por tuberculose foi empregado o programa R (THE R CORE DEVELOPMENT TEAM, 2006) usando a biblioteca FLEXMIX (LEISCH, 2004). Para a construção dos mapas para apresentação dos resultados, empregou-se a biblioteca SPDEP (BIVAND, 2006).

4.3.3 Modelo de Regressão de Binomial Negativa

A regressão binomial negativa foi implementada no programa R. Foram empregadas as seguintes covariáveis: *proporção de famílias com renda maior que dez salários mínimos (renda > 10), número médio de moradores por domicílio (nppd), Índice de Gini (Gini), densidade demográfica (densdemog), Índice de Desenvolvimento Humano Municipal (IDH), taxa mortalidade de AIDS (tx9702) e taxa de morbidade de AIDS (tx9700)*. Para realização da modelagem normalizou-se as covariáveis. A seleção de modelos foi baseada na *deviance* do modelo, no Critério de Akaike (AIC) e em gráficos de resíduos.

O Índice de Gini mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar *per capita*. Seu valor varia de 0 quando não há desigualdade a 1 quando a desigualdade é máxima. O IDH é obtido pela média aritmética simples de três índices referentes às dimensões Longevidade (expectativa de vida ao nascer), Educação (taxa de alfabetização) e Renda (renda *per capita*) (ESM Consultoria, 2003).

4.3.4 Modelo de Regressão Mista

Uma vez detectada a heterogeneidade dos dados por meio dos modelos de misturas finitas, o passo seguinte da análise consistiu em selecionar as covariáveis que seriam empregadas na regressão de misturas finitas. Foram empregadas as mesmas covariáveis utilizadas no modelo de regressão binomial negativa: *proporção de famílias com renda maior que dez salários mínimos (renda > 10), número médio de moradores por domicílio (nppd), Índice de Gini (Gini), densidade demográfica (densdemog), Índice de Desenvolvimento Humano (IDH), taxa mortalidade de AIDS (tx9702) e taxa de morbidade de AIDS (tx9700)*.

Os modelos foram implementados no programa R (THE R CORE DEVELOPMENT TEAM, 2006) usando a biblioteca FLEXMIX (LEISCH, 2004). A seleção de modelos foi realizada empregando o BIC (SCHWARTZ, 1978). Empregou-se novamente uma mistura Normal-Poisson. A variável dependente foi o número de casos/óbitos observados no município no período, sendo empregado o valor esperado para cada uma das áreas como um *offset* do modelo.

Capítulo 5: Resultados

5.1 Análise exploratória

Na análise descritiva dos dados foram calculadas medidas de resumo e construídos diagramas de caixa. Estes foram construídos para que a distribuição dos dados pudesse ser observada e os dados de mortalidade e morbidade pudessem ser visualmente comparados.

A Tabela 5.1 mostra as medidas descritivas: mínimo, quartis, média, desvio padrão e máximo das razões de mortalidade e morbidade. Os valores máximos para a mortalidade e a morbidade são 2,04 e 1,50, respectivamente. As medidas de máximo e quartis sugerem a presença de valores extremos nos dados.

Tabela 5.1: Medidas descritivas das razões de mortalidade (período: 1997-2002) e morbidade (período: 1997-2000) por tuberculose nos municípios do estado do Rio de Janeiro

Dados	Mínimo	Q_1	Mediana	Média	Q_3	Máximo	Desvio padrão	% zeros
Mortalidade	0,00	0,41	0,62	0,69	0,87	2,04	0,42	4,40
Morbidade	0,03	0,26	0,52	0,55	0,74	1,50	0,36	0,0

Na Figura 5.1 observa-se o diagrama de caixa para os dados de mortalidade. Pode-se verificar a existência de cinco valores extremos (Engenheiro Paulo de Frontin – Região Centro-sul Fluminense, Nilópolis, Queimados, Japeri e Nova Iguaçu – Região Metropolitana do estado).

Na Figura 5.2 pode-se observar o diagrama de caixa para os dados de morbidade. Foi identificado de apenas um valor extremo, o município de Duque de Caxias, o qual pertence à Região Metropolitana do estado.

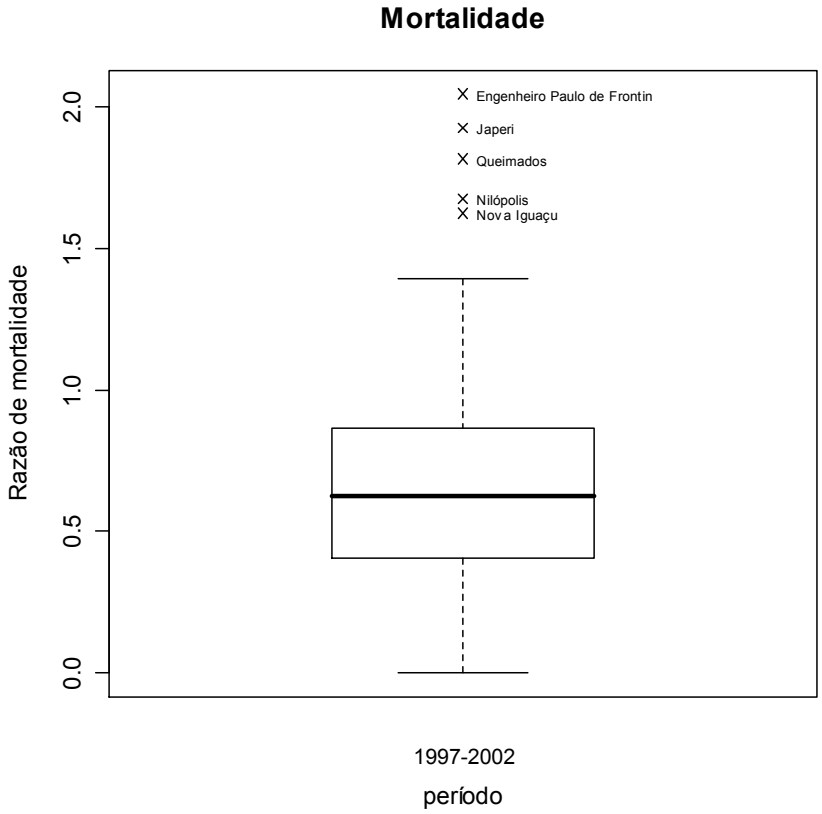


Figura 5.1: Diagrama de caixa da mortalidade por tuberculose no estado do Rio de Janeiro para o período 1997-2002

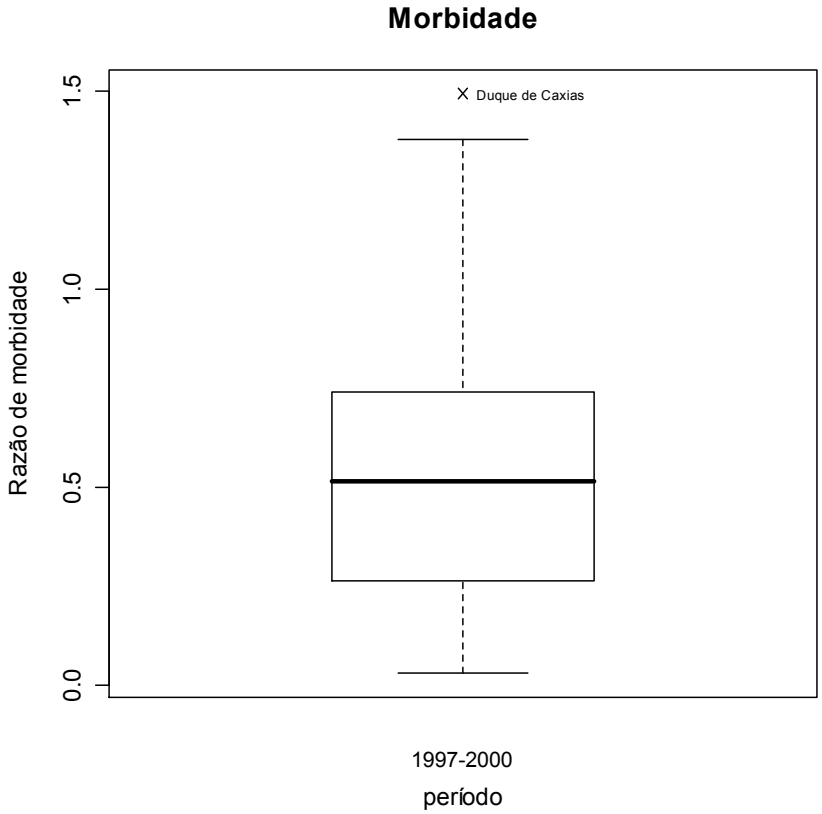


Figura 5.2: Diagrama de caixa da morbidade por tuberculose no estado do Rio de Janeiro para o período 1997-2000

A análise exploratória teve como objetivo verificar as principais características dos dados de mortalidade e de morbidade por tuberculose nos municípios do estado do Rio de Janeiro. A partir desta análise inicial pode-se verificar: (1) que a maioria dos municípios registrou poucos casos nos períodos em questão; (2) que alguns poucos municípios têm número de casos excessivamente alto e estes se localizam, em sua grande maioria, na Região Metropolitana do estado.

Como os dados se referem a ocorrências de uma doença em uma área, o passo seguinte da análise consistiu em considerar com maior detalhamento o espaço.

5.2 Análise exploratória espacial

Nesta seção é apresentada a análise exploratória espacial dos dados de mortalidade e morbidade por tuberculose no estado do Rio de Janeiro. Nas subseções a seguir são apresentados os resultados das técnicas empregadas com os objetivos de visualização dos dados no espaço, verificação de tendência e investigação de autocorrelação espacial.

5.2.1 Visualização dos dados

A seguir são apresentados os resultados das técnicas escolhidas para a visualização dos dados: quintis e k-médias. Busca-se com tais técnicas obter uma compreensão adequada do problema em questão. Com relação à técnica de k-médias, optou-se por agrupar os dados em cinco aglomerados em função dos resultados verificados na análise exploratória inicial e também de forma de permitir a comparação dos resultados das duas técnicas de visualização dos dados.

Na Figura 5.3 pode ser visualizado o mapa coroplético da razão de mortalidade segundo quintis, sendo empregadas as cores branco, para os municípios sem registro de casos, preto, para os municípios com mais casos e três tonalidades de cinza indicando municípios com número intermediário de casos.

Verifica-se na Figura 5.3 que 18 (19,9%) municípios foram classificados no primeiro quintil e que 71 (18,0%) apresentaram SMR menor que 1, isto é, o valor observado foi menor que o valor esperado para a área.

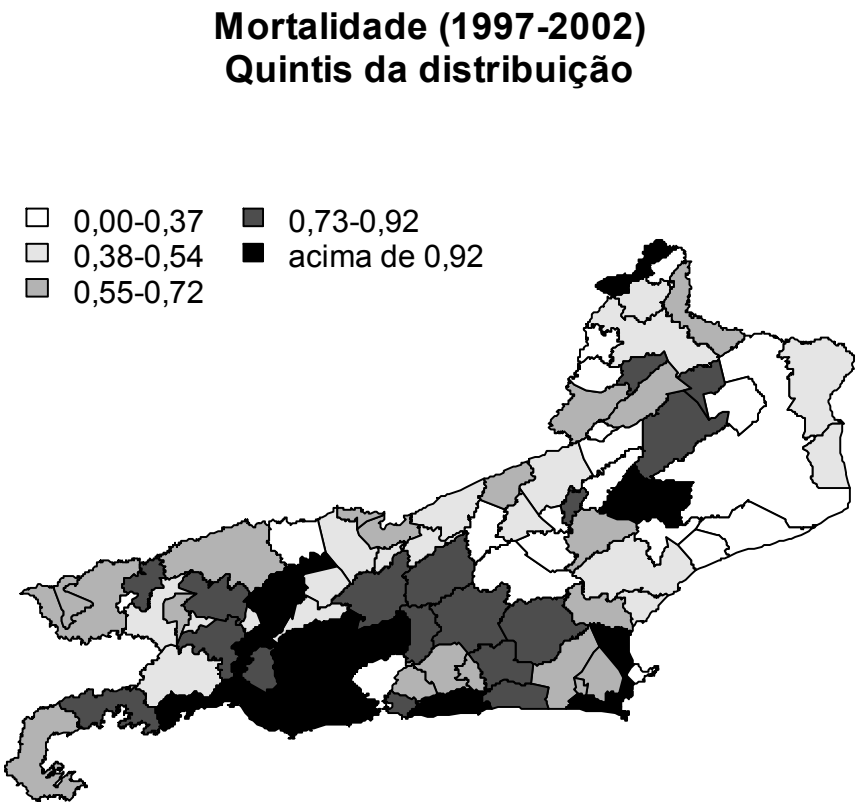


Figura 5.3: Mapas coropléticos da razão de mortalidade por tuberculose no estado do Rio de Janeiro no período 1997-2002, segundo quintis

Na Figura 5.4 podem ser observados os resultados obtidos com a técnica de k-médias. Formaram-se cinco aglomerados com centróides iguais a 0,1, 0,4, 0,7, 1,2 e 1,8 e número de componentes iguais a 8, 31, 35, 12 e 5, respectivamente. Foram identificados como pertencentes ao último intervalo os municípios de Engenheiro Paulo de Frontin (Região Centro-Sul do estado), Japeri, Nova Iguaçu, Nilópolis e Queimados (Região Metropolitana do estado).

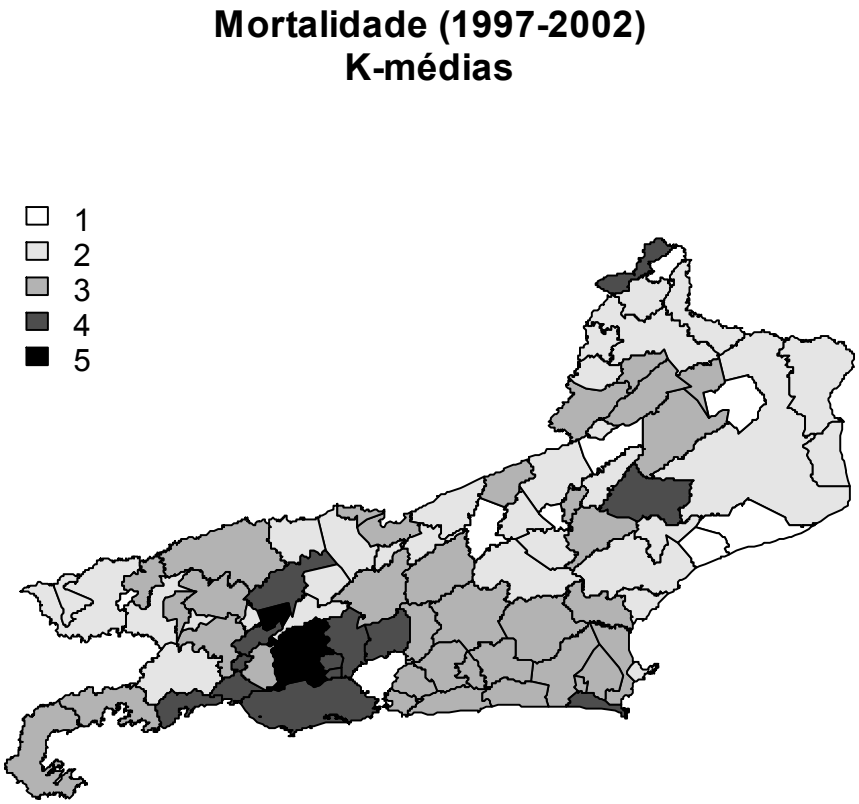


Figura 5.4: Mapas coropléticos da razão de mortalidade por tuberculose no estado do Rio de Janeiro no período 1997-2002, segundo k-médias

A seguir serão mostrados os resultados da análise espacial dos dados de morbidade por tuberculose no estado do Rio de Janeiro para o período de 1997 a 2000. Na elaboração dos mapas foram utilizados dois critérios para a construção dos intervalos de classe: quintis e k-médias.

A construção de intervalos segundo quintis (Figura 5.5) mostrou que 18 (19,8%) municípios foram classificados no primeiro intervalo e 19 (20,9%) no último intervalo da distribuição. Dentre os municípios hachurados de preto verifica-se que Duque de Caxias, Paracambi, Belford Roxo, Japeri, Queimados, Rio de Janeiro, São João de Meriti e Niterói (Região Metropolitana do estado), Volta Redonda (Região do Médio Paraíba), Vassouras e Mendes (Região Centro-Sul Fluminense), Mangaratiba, Itaguaí e Angra dos Reis (Região da Costa Verde) e Laje do Muriaé (Região Noroeste do estado) possuem valor para a razão de morbidade maior que 1. Verificou-se ainda que 77 (84,6%) municípios tiveram valores para a razão de morbidade inferiores a 1.

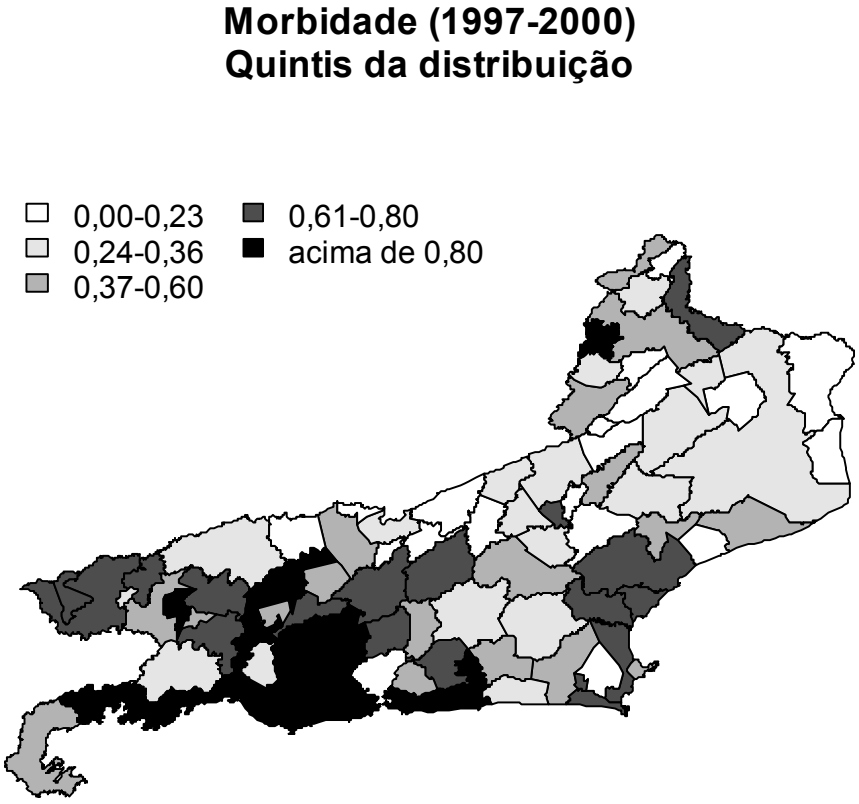


Figura 5.5: Mapas coropléticos da razão de morbidade por tuberculose no estado do Rio de Janeiro no período 1997-2000, segundo quintis

A aplicação da técnica de k-médias (Figura 5.6) levou à classificação de 16 municípios no primeiro aglomerado (centróide = 0,13) e 14 municípios no último aglomerado (centróide = 1,22). Os municípios classificados no último aglomerado foram: Laje do Muriaé, Vassouras, Mendes, Duque de Caxias, Paracambi, Japeri, Queimados, Belford Roxo, Itaguaí, São João de Meriti, Rio de Janeiro, Angra dos Reis, Niterói e Mangaratiba. Destes, apenas o município de Engenheiro Paulo de Frontin não se localiza na Região Metropolitana, sendo pertencente à Região Centro-Sul. Quanto aos demais aglomerados formados, verifica-se que foram classificados do segundo ao quarto aglomerado, 26 (centróide = 0,3), 17 (centróide = 0,6) e 18 (centróide = 0,8) municípios, respectivamente.

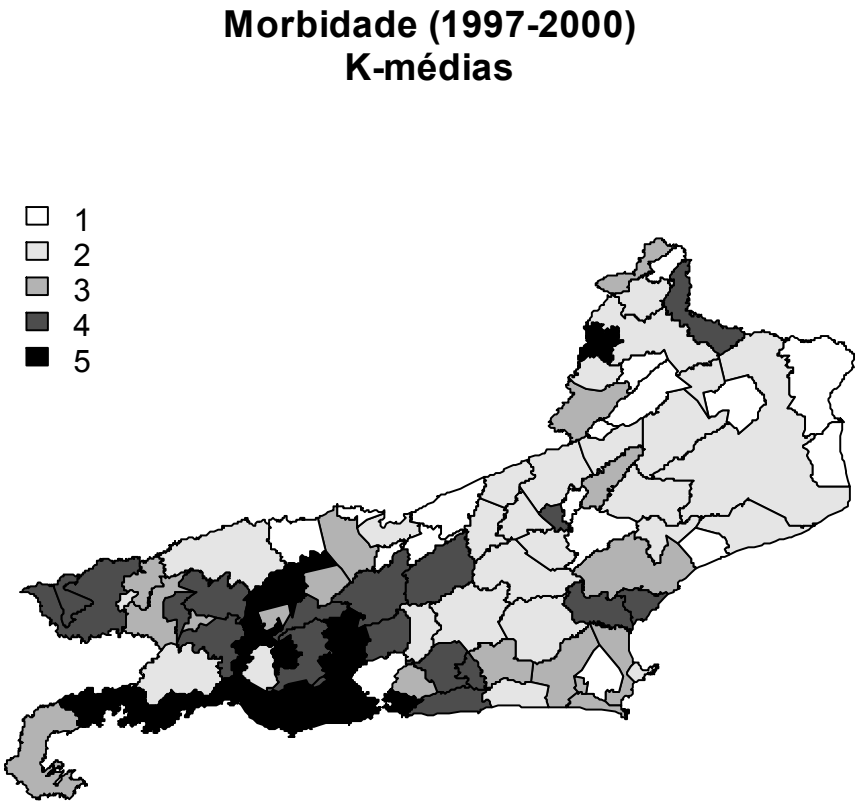


Figura 5.6: Mapas coropléticos da razão de morbidade por tuberculose no estado do Rio de Janeiro no período 1997-2000, segundo k-médias

5.2.2 Verificação da tendência

Com o intuito de verificar variações globais e tendências nos dados de mortalidade foi aplicada a média móvel aos valores das razoes de mortalidade e morbidade, cujos resultados podem ser observados nas Figura 5.7 e 5.8. Empregou-se o critério dos quintis para a construção dos intervalos de classe.

Na Figura 5.7 observa-se que 79 (86,8%) municípios registraram razão de mortalidade menor ou igual a um. Foram classificados no último quintil da distribuição 19 municípios. Na Figura 5.7 verifica-se que a área de situação mais grave corresponde a Região Metropolitana do estado, com alguns focos na Região das Baixadas Litorâneas.

Na Figura 5.8 é apresentado o mapa das médias móveis para a razão de morbidade por tuberculose nos municípios do estado do Rio de Janeiro para o período 1997-2000. Com a aplicação desta técnica, verificou-se que 84 (92,3%) municípios apresentaram valor da razão de morbidade inferior a 1, indicando um número de casos

observado inferior ao número de casos esperado. Pode-se observar ainda que 20 (22,0%) municípios foram classificados no último intervalo.

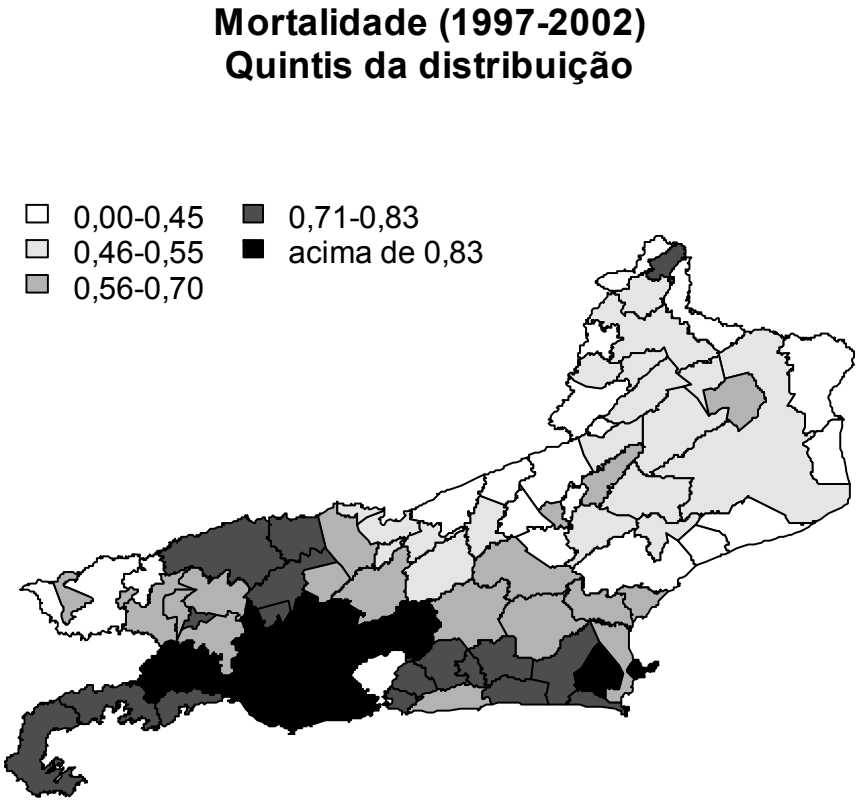


Figura 5.7: Médias móveis da razão de morbidade por tuberculose nos períodos 1997-2000

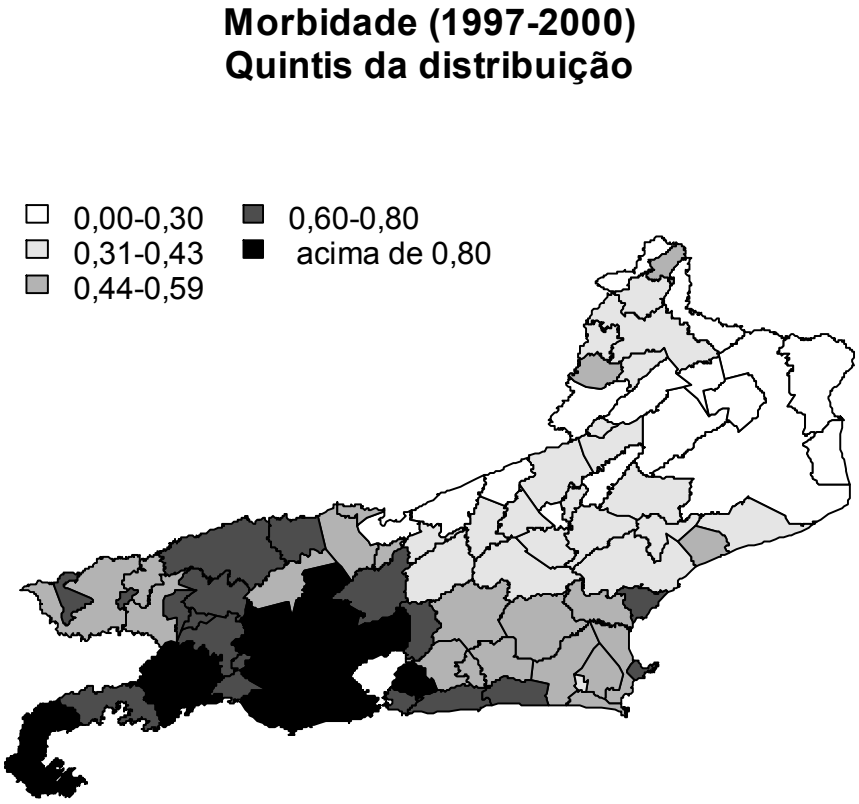


Figura 5.8: Médias móveis da razão de morbidade por tuberculose nos períodos 1997-2000

5.2.3 Verificação da autocorrelação espacial

Com o intuito de verificar a presença de autocorrelação espacial nos dados de mortalidade calculou-se o Índice I de Moran, cujos resultados são mostrados na Tabela 5.2. Empregou-se o critério da fronteira comum na definição da vizinhança. Pode-se observar, por meio dos testes de independência, supondo a normalidade ou a aleatoriedade, que a hipótese nula, isto é, de aleatoriedade espacial, foi rejeitada.

Tabela 5.2: Índice I de Moran para as razoes de mortalidade e morbidade por tuberculose nos períodos de 1997-1998, 1999-2000 e 2001-2002.

Dados	Moran I	Signif, N	Signif, R
Mortalidade	0,38	$9{,}18 \times 10^{-9}$	$7{,}42 \times 10^{-9}$
Morbidade	0,42	$3{,}51 \times 10^{-10}$	$3{,}68 \times 10^{-10}$

Nesta seção foram apresentados os resultados da análise exploratória espacial. Por meio desta pôde-se verificar a distribuição espacial dos eventos, identificando regiões de maior ocorrência e tendências. Observou-se que a aplicação da técnica de alisamento média móvel não proporcionou melhor visualização dos dados. A hipótese de aleatoriedade espacial foi rejeitada para os dados de mortalidade e morbidade.

5.3 Modelagem

5.3.1 Aplicação dos modelos de misturas finitas

Nesta seção são apresentados os resultados da aplicação dos modelos de misturas finitas aos dados de mortalidade e morbidade por tuberculose nos municípios do estado do Rio de Janeiro. Foi empregada na modelagem dos dados as razões de mortalidade e morbidade. Foram ajustados modelos de mistura com número de componentes variando de dois a sete.

Na Figura 5.9 é apresentado o Critério de Informação Bayesiano (BIC) para os modelos ajustados aos dados de mortalidade. Pode-se observar da figura que ocorre uma queda acentuada no valor do BIC a medida em que aumenta o número de componentes do modelo. A redução no valor do BIC se torna mais suave nos modelos com cinco ou mais componentes.

Observa-se que o valor do BIC sofreu grande redução do modelo com dois componentes em relação ao modelo com três componentes. Pouca diferença, em termos do BIC, pode ser observada nos modelos de quatro a componentes e, desta forma, o modelo com quatro componentes foi escolhido. Verificou-se a existência de componentes vazios, isto é, sem municípios, nos modelos mais complexos (seis ou sete componentes).

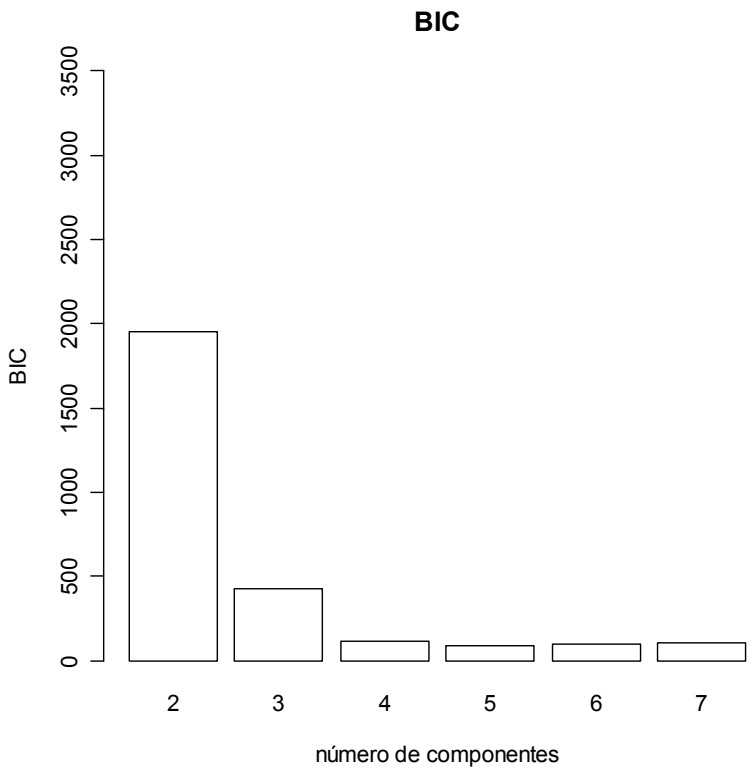


Figura 5.9: BIC para os modelos ajustados aos dados de mortalidade por tuberculose nos municípios do estado do Rio de Janeiro, período 1997-2002

Podem ser observados na Tabela 5.3 os parâmetros do modelo de mistura selecionado para os dados de mortalidade. A estimativa de máxima verossimilhança para o modelo selecionado foi de -22,09. Os pesos (π_i) do modelo variaram de 0,54 a 0,03 e o parâmetro λ_i de 0,00021 a 0,00045.

Tabela 5.3: Resultados do modelo de mistura selecionado para os dados de mortalidade nos municípios do estado do Rio de Janeiro, período 1997-2002

Nº. de componentes	Parâmetros		Máxima verossimilhança	BIC
	π_i	λ_i		
4	0,54	0,00021	-22,09	111,85
	0,28	0,00030		
	0,15	0,00041		
	0,03	0,00046		

Na Figura 5.10 é apresentado o modelo de mistura selecionado para o período de 1997-2002. Este agrupou os municípios em quatro grupos de risco. Verifica-se que do primeiro ao quarto grupo foram classificados respectivamente 49 (53,8%), 26 (28,6%), 14 (15,4%) e 2 (2,2%) municípios. Foram classificados no penúltimo intervalo municípios pertencentes à Região Metropolitana do estado e os municípios de Campos dos Goytacazes (Região Norte do estado), Petrópolis (Região Serrana), Cabo Frio (Região das Baixadas Litorâneas) e Volta Redonda (Região do Médio Paraíba). Foram classificados no último componente da mistura apenas os municípios de Nova Iguaçu e do Rio de Janeiro.

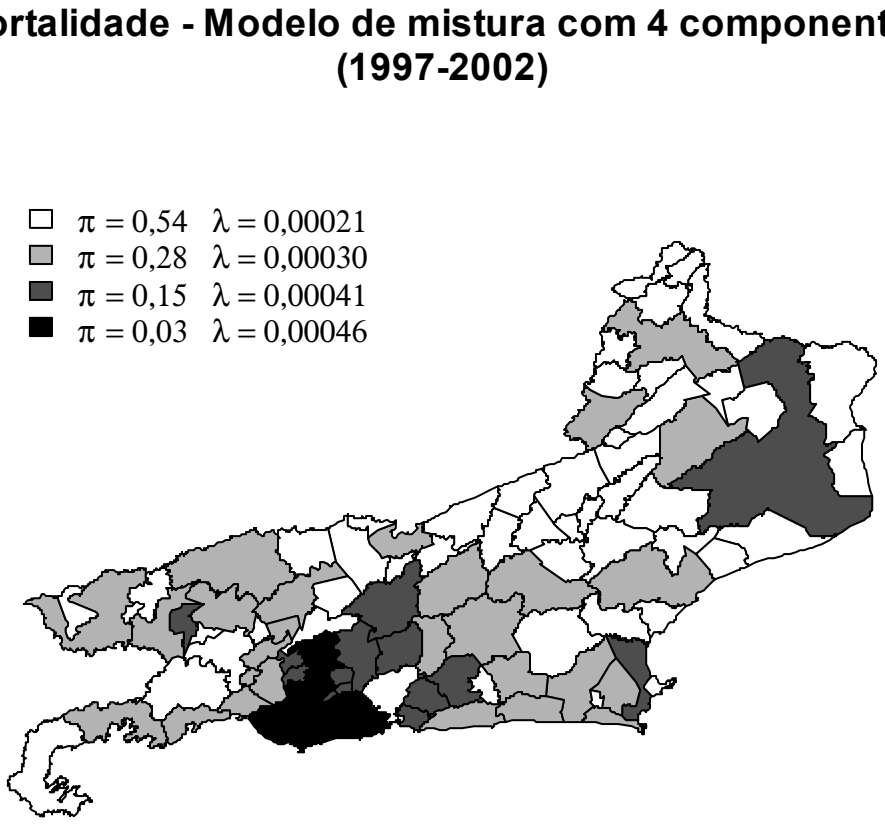


Figura 5.10: Modelo de mistura com quatro componentes selecionado para os dados de mortalidade por tuberculose no estado do Rio de Janeiro, período de 1997-2002

Na Figura 5.11 é apresentado o resultado do BIC para os modelos de misturas ajustados aos dados de morbidade por tuberculose para o período 1997-2000. De forma semelhante ao ocorrido para o período anterior, verifica-se por meio do gráfico, acentuada redução no valor do BIC do modelo com dois componentes em relação aos demais modelos ajustados. Os valores continuam se reduzindo à medida que aumenta o

número de componentes do modelo. Como o valor do BIC foi semelhante para os modelos com seis e sete componentes, escolheu-se o mais parcimonioso.

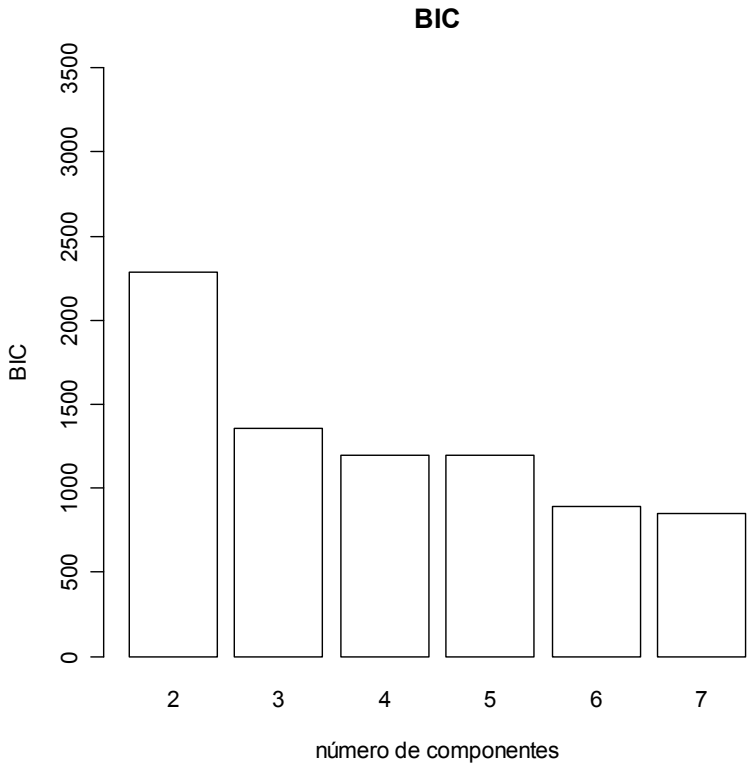


Figura 5.11: BIC para os modelos ajustados aos dados de morbidade por tuberculose nos municípios do estado do Rio de Janeiro, período 1997-2000

Observam-se na Tabela 5.4 os parâmetros do modelo de mistura selecionado para os dados de morbidade. Para este modelo verificou-se uma estimativa de máxima verossimilhança igual a -394,50. Os pesos dos componentes da mistura variaram de 0,45 a 0,01 e o parâmetro das distribuições de Poisson (λ_i) de 0,00048 a 0,00521.

A Figura 5.12 mostra os resultados da aplicação dos modelos de misturas finitas aos dados de morbidade por tuberculose. O modelo selecionado é constituído por seis componentes. Os municípios foram assim classificados: 11 (12,1%) municípios no primeiro intervalo, 11 (12,1%) no segundo, 18 (19,8%) no terceiro, 42 (46,1%) no quarto, 8 (8,8%) no quinto, e 1 (1,1%) no sexto intervalo da distribuição. Foram classificados no penúltimo intervalo os municípios de Belford Roxo, São Gonçalo e São João de Meriti, Niterói, Duque de Caxias, Nova Iguaçu, Petrópolis e Volta Redonda e no último o município do Rio de Janeiro.

Tabela 5.4: Resultados do modelo de mistura selecionado para os dados de morbidade nos municípios do estado do Rio de Janeiro, período 1997-2000

N^{o,} de componentes	Parâmetros		Máxima verossimilhança	BIC
	π_i	λ_i		
6	0,11	0,00048	-394,50	892,76
	0,13	0,00078		
	0,20	0,00143		
	0,45	0,00264		
	0,10	0,00390		
	0,01	0,00521		

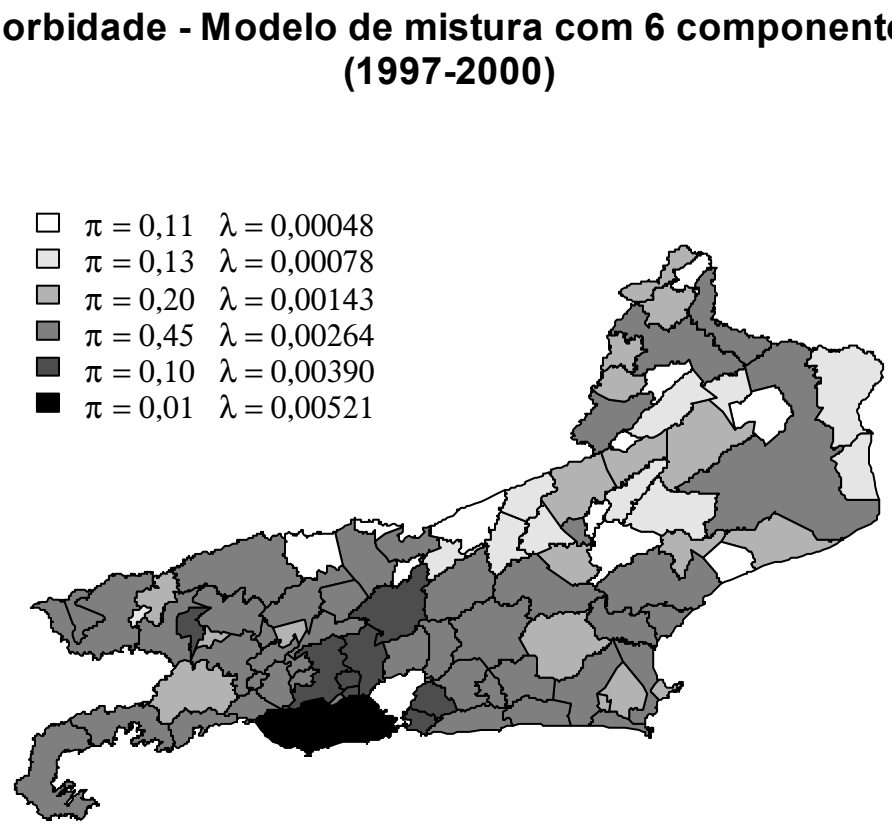


Figura 5.12: Modelo de mistura com seis componentes selecionado para os dados de morbidade por tuberculose no estado do Rio de Janeiro, período de 1997-2000

5.3.2 Aplicação do modelo linear generalizado

Foram ajustados modelos de regressão binomial negativa com o intuito de detectar as covariáveis estatisticamente significativas para a mortalidade e morbidade ocorridas no estado no Rio de Janeiro, em cada um dos períodos de tempo analisados. Selecionaram-se sete covariáveis: *proporção de famílias com renda maior que dez salários mínimos (renda > 10), número médio de moradores por domicílio (nppd), Índice de Desenvolvimento Humano (IDH), Índice de Gini (Gini), densidade demográfica (densdemog) e a taxa de AIDS(tx9702 e tx9700).* O objetivo foi confrontar os resultados do modelo de regressão binomial negativa com o modelo de regressão de Poisson mista.

Na Figura 5.13 são apresentados os diagramas de caixa das covariáveis inseridas no modelo de regressão binomial negativa e no modelo de regressão de Poisson mista. Por meio do gráfico pode-se observar a distribuição das covariáveis selecionadas.

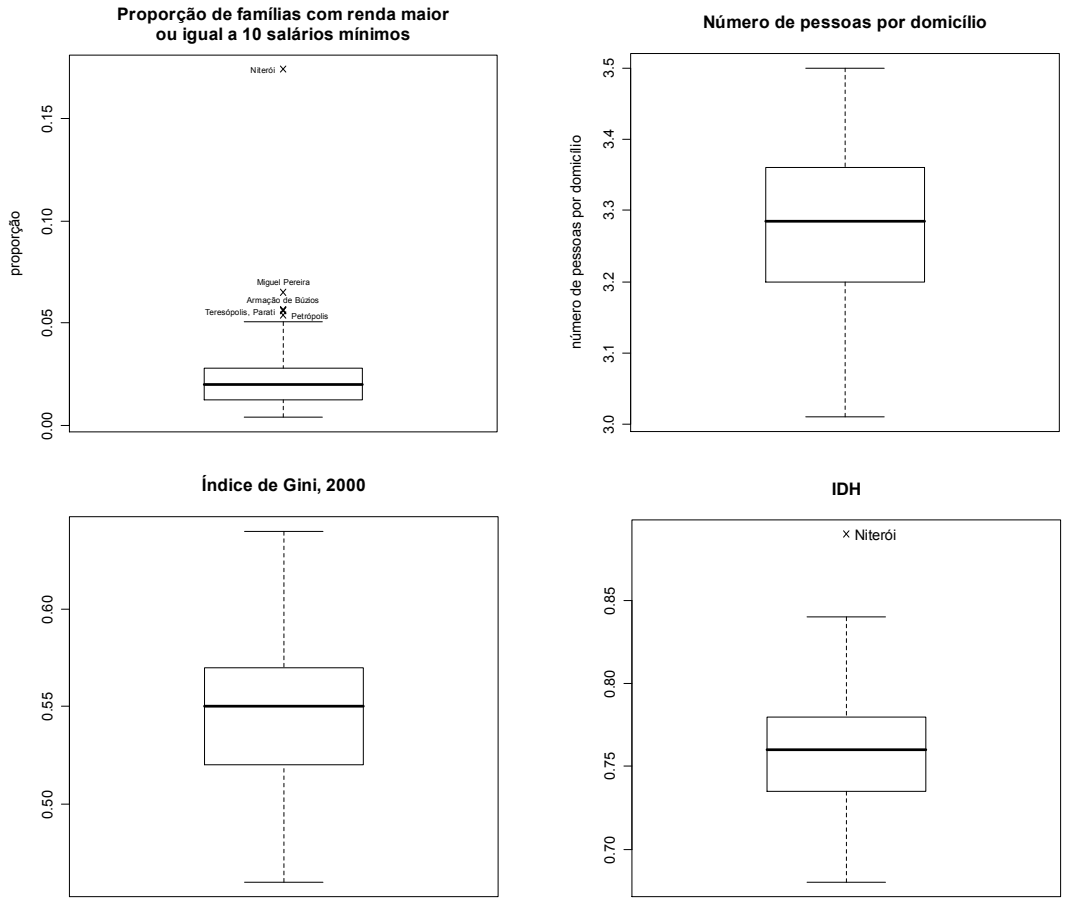


Figura 5.13: Diagrama de caixa das covariáveis empregadas na análise de regressão

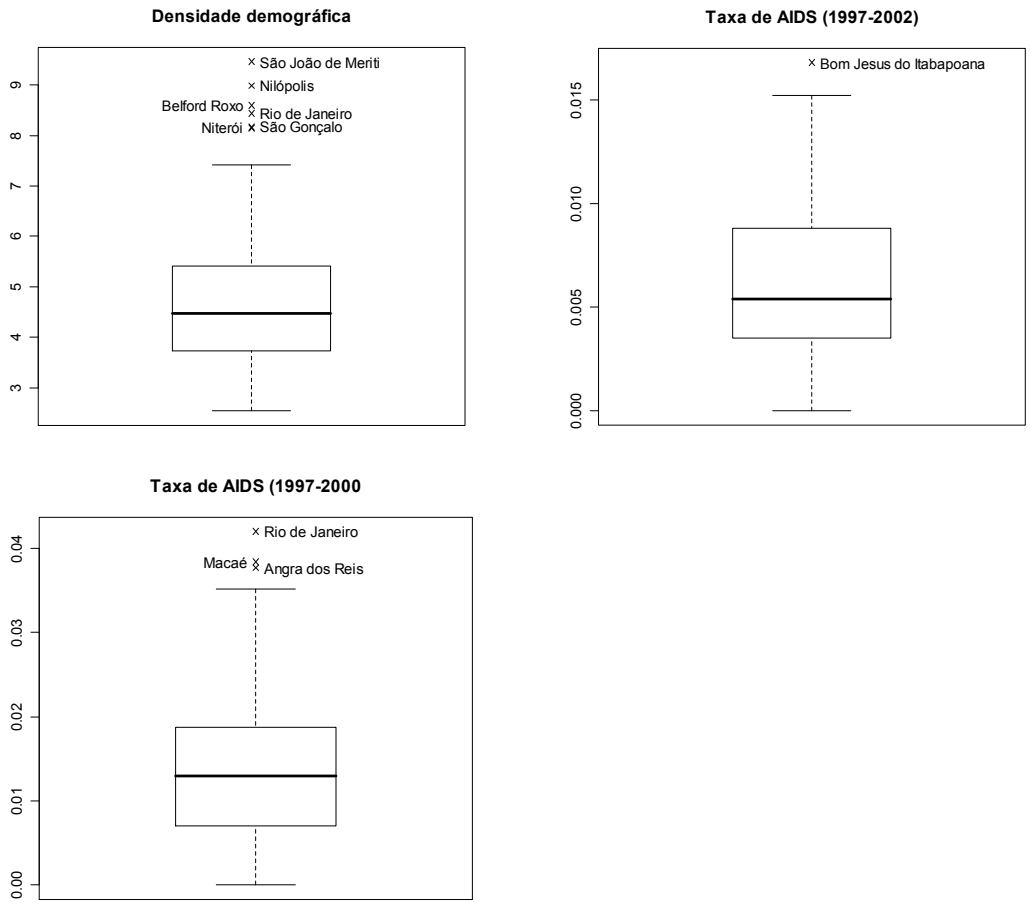


Figura 5.13 (cont.): Diagrama de caixa das covariáveis empregadas na análise de regressão

Com relação à variável *renda* > *10* observou-se uma média de 0,02, sendo detectada a presença de valores extremos (Niterói, Miguel Pereira, Armação de Búzios, Teresópolis, Parati e Petrópolis). A variável *nppd* é um valor médio para cada município e variou entre 3,0 e 3,5. O exame da variável *Gini* mostrou média de 0,5. A variável *IDH* variou entre 0,6 e 0,9, sendo detectado um valor extremo (Niterói). Verifica-se da Figura 5.13 que a variável *densdemog* variou entre 12,6 e 12.880,0, com média de 590,0. Foi calculado o logaritmo da variável *densdemog* para a construção do diagrama de caixa. Para esta variável foram identificados cinco municípios com valores extremos (São João de Meriti, Nilópolis, Belford Roxo, Rio de Janeiro, Niterói e São Gonçalo). Para a variável *tx9702*, a qual se refere à razão de mortalidade por AIDS para o período 1997-2002, verifica-se média de 0,006 e presença de um valor extremo (Bom Jesus do Itabapoana). Quanto à morbidade por AIDS no período 1997-2000, representada pela

variável *tx9700*, observa-se média de 0,01, com a presença de três municípios com valores extremos (Macaé, Rio de Janeiro e Angra dos Reis).

Na Tabela 5.3 podem ser visualizados os resultados do modelo de regressão binomial negativa aplicado aos dados de mortalidade e morbidade por tuberculose. Inicialmente, foram ajustados modelos considerando todas as variáveis. As covariáveis identificadas como não significativas em cada período foram, então, excluídas, sendo ajustado um novo modelo. As estimativas dos coeficientes, erros padrão e valor-p para as variáveis podem ser visualizadas na Tabela 5.5. Verifica-se que, para os modelos selecionados, foram consideradas estatisticamente significativas as variáveis *gini e* a taxa de AIDS (*tx9702* e *tx9700*). Adicionalmente, foi considerada estatisticamente significativa, no modelo de morbidade, a variável *IDH.*

Tabela 5.5: Modelo linear generalizado para os dados de mortalidade e morbidade por tuberculose nos municípios do estado do Rio de Janeiro

Dados	Estimativa	Erro padrão	Valor-p
Mortalidade (1997-2002)			
Intercepto	-0,41	0,05	$3{,}00 \times 10^{-12}$
Gini	-0,14	0,05	$8{,}14 \times 10^{-3}$
tx9702	0,28	0,05	$4{,}76 \times 10^{-7}$
Morbidade (1997-2000)			
Intercepto	-0,70	0,06	$< 2{,}00 \times 10^{-16}$
Gini	-0,17	0,06	$6{,}20 \times 10^{-3}$
IDH	0,20	0,07	$7{,}27 \times 10^{-3}$
tx9700	0,36	0,07	$1{,}10 \times 10^{-6}$

Pode-se observar na Figura 5.14 os resultados da análise de resíduos dos modelos selecionados para cada um dos períodos. Verifica-se por meio do gráfico *Normal Q-Q* um ajuste adequado do modelo aos dados.

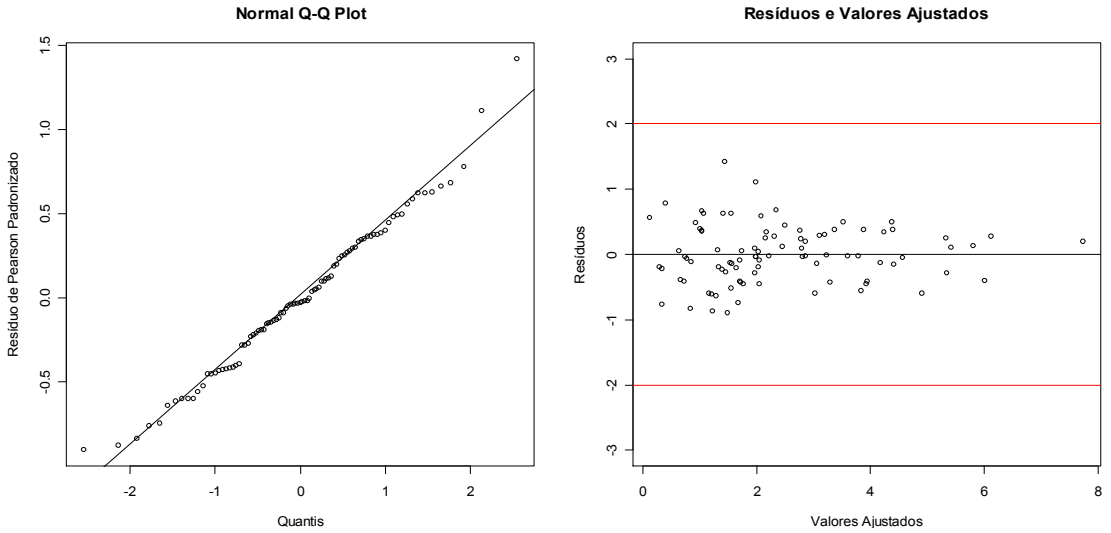


Figura 5.14: Gráficos de resíduos dos modelos de regressão binomial negativa ajustados aos dados de mortalidade por tuberculose no estado do Rio de Janeiro, período 1997-

Na Figura 5.15 são apresentados gráficos dos resíduos do modelo ajustado aos dados de morbidade. Verifica-se novamente um ajuste adequado do modelo. Pode-se observar a existência de um valor extremo no gráfico dos Resíduos e Valores Ajustados.

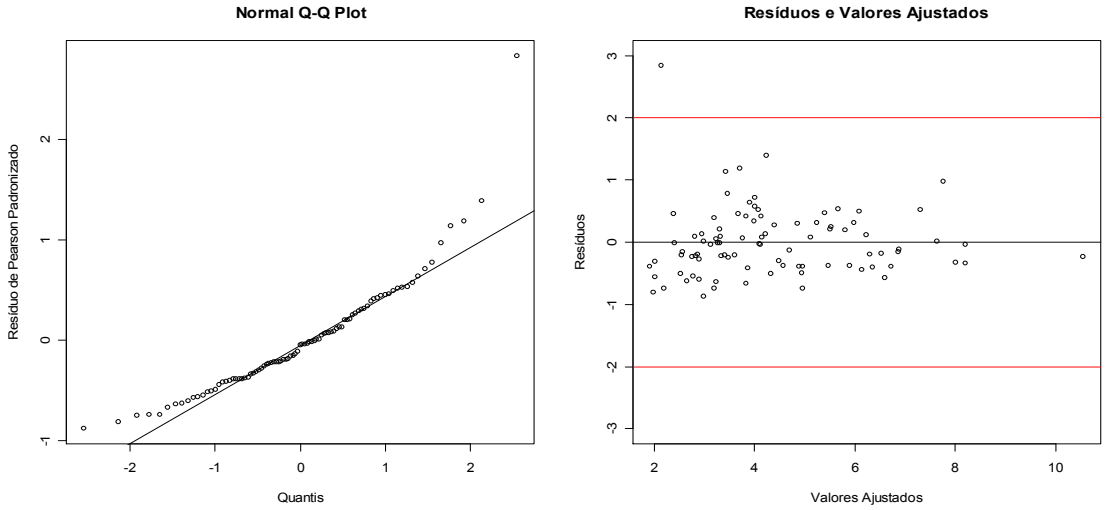


Figura 5.15: Gráficos de resíduos dos modelos de regressão binomial negativa ajustados aos dados de morbidade por tuberculose no estado do Rio de Janeiro, período 1997-2000

5.3.3 Aplicação dos Modelos Lineares Generalizados Mistos

A seguir são apresentados os resultados da aplicação do modelo linear generalizado misto. Foram ajustados modelos de regressão mista para as razões de mortalidade e morbidade. As Tabelas 5.6 a 5.7 apresentam as estimativas dos coeficientes dos modelos selecionados e, entre parênteses, os respectivos erros padrão. A seleção dos modelos foi baseada no BIC. As covariáveis que não foram consideradas estatisticamente significativas para algum dos componentes dos modelos ajustados foram identificadas com asteriscos.

Para os dados de mortalidade, foi selecionado um modelo de regressão misto composto por cinco componentes. Este classificou do primeiro ao quinto componente, respectivamente, 41, 34, 5, 5 e 6 municípios. Observa-se da Tabela 5.6 que quase todas as variáveis inseridas no modelo foram consideradas estatisticamente significativas para os cinco componentes. Verifica-se que somente a variável *Gini* não foi significativa para os componentes 1 e 2 do modelo.

Na Figura 5.16 é apresentado o resultado do modelo de regressão mista ajustado aos dados de mortalidade. Verifica-se que foram coloridos de preto os municípios de Petrópolis (Região Serrana do estado), Volta Redonda (Região do Médio Paraíba), Itaboraí, Belford Roxo, São João de Meriti e Nilópolis (Região Metropolitana do estado). Foram classificados no quarto componente da mistura e, portanto, coloridos com a tonalidade mais escura de cinza, os municípios de Duque de Caxias, Nova Iguaçu, São Gonçalo, Rio de Janeiro (Região Metropolitana do estado) e Cabo Frio (Região das Baixadas Litorâneas). Observa-se da figura que os parâmetros λ e π variaram de 0,000208 a 0,000485 e de 0,45 a 0,06, respectivamente.

Tabela 5.6: Regressão mista para os dados de mortalidade por tuberculose no estado do Rio de Janeiro no período 1997-2002

Covariável	Componente				
	1	2	3	4	5
Intercepto	-12,0	-4,7	-13,9	$-1{,}1 \times 10^{4}$	$-2{,}6 \times 10^{2}$
	(3,4)	(1,5)	$(1{,}0 \times 10^{-4})$	(7,4)	$(3{,}6 \times 10^{-4})$
IDH	16,0	8,0	$-3{,}7 \times 10^{2}$	$4{,}4 \times 10^{2}$	$1{,}8 \times 10^{3}$
	(4,2)	(1,9)	$(5{,}2 \times 10^{-4})$	$(1{,}6 \times 10^{1})$	$(9{,}4 \times 10^{-4})$
Gini	0,8*	1,8*	$6{,}5 \times 10^{2}$	$1{,}8 \times 10^{4}$	$-1{,}8 \times 10^{3}$
	(2,4)	(1,8)	$(6{,}1 \times 10^{-4})$	$(1{,}3 \times 10^{1})$	$(1{,}0 \times 10^{-5})$
Densdemog	-0,2	0,5	$1{,}5 \times 10^{2}$	$6{,}2 \times 10^{2}$	$1{,}5 \times 10^{1}$
	(0,2	(0,1)	$(2{,}1 \times 10^{-5})$	$(7{,}3 \times 10^{-1})$	$(3{,}5 \times 10^{-5})$
Tx9702	93,4	79,4	$-1{,}5 \times 10^{4}$	$6{,}7 \times 10^{4}$	$-4{,}2 \times 10^{3}$
	(20,9)	(19,4)	$(5{,}1 \times 10^{-3})$	$(1{,}9 \times 10^{2})$	$(8{,}4 \times 10^{3})$

*variável não significativa para o componente

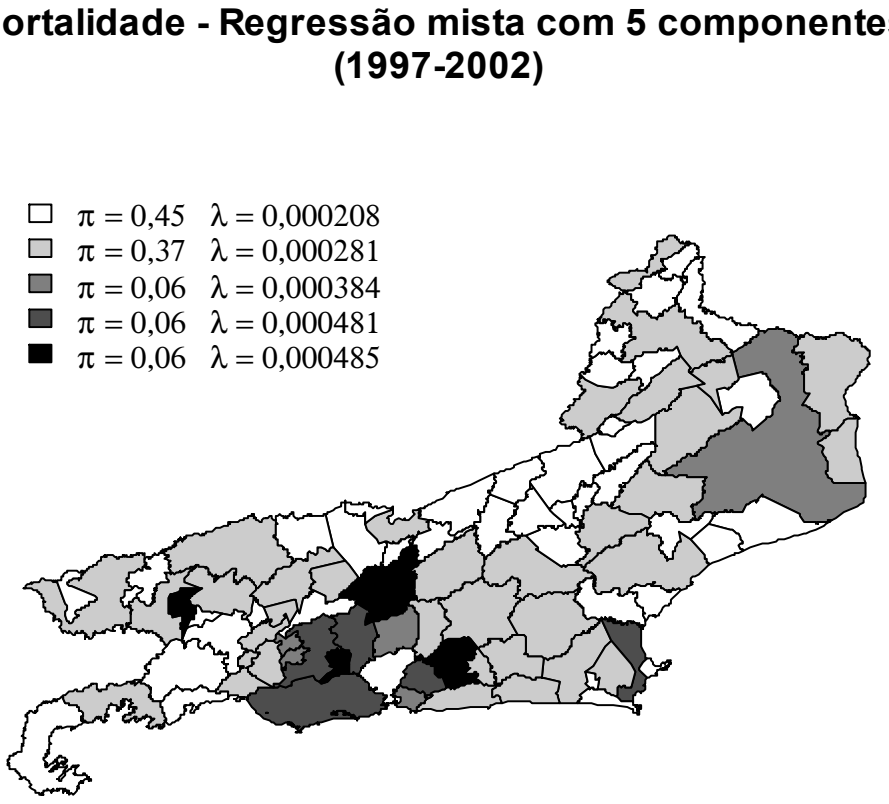


Figura 5.16: Regressão mista para os dados de mortalidade por tuberculose no estado do Rio de Janeiro no período 1999-2002

Na Tabela 5.7 são apresentados os resultados da regressão mista para os dados de morbidade por tuberculose do período 1997-2000. Novamente, selecionou-se um modelo com cinco componentes. Foram classificados, do primeiro ao último componente, 19, 21, 37, 7 e 7, respectivamente. Verifica-se que a variável renda >10 não foi considerada significativa para os componentes 1, 2 e 3. As variáveis *Gini* e *IDH* não foram consideradas significativas para o terceiro componente.

Tabela 5.7: Regressão mista para os dados de morbidade por tuberculose no estado do Rio de Janeiro no período 1997-2000

Covariável	Componente				
	1	2	3	4	5
Intercepto	11,4	10, 5	$9,5 \times 10^{-3}$	$-1,0 \times 10^{3}$	$4,8 \times 10^{4}$
	(2,3)	(3,6)	($6,4 \times 10^{2}$)	($1,7 \times 10^{-3}$)	($3,7 \times 10^{-2}$)
IDH	-8,6	-10,4	$4,2 \times 10^{2}$*	$-3,6 \times 10^{3}$	$-1,4 \times 10^{5}$
	(2,7)	(4, 5)	($7,1 \times 10^{2}$)	($2,5 \times 10^{-3}$)	($2,9 \times 10^{-2}$)
Gini	-3,2	-0,1	$-6,0 \times 10^{2}$*	$5,4 \times 10^{3}$	$9,2 \times 10^{4}$
	(2,3)	(2,0)	($5,4 \times 10^{2}$)	($2,7 \times 10^{-3}$)	($4,8 \times 10^{-2}$)
Densdemog	0,6	0,8	$1,5 \times 10^{2}$	$7,9 \times 10^{2}$	$4,3 \times 10^{3}$
	(0,1)	(0,2)	($2,0 \times 10^{1}$)	($1,4 \times 10^{-4}$)	($1,2 \times 10^{-3}$)
Tx9700	-17,7	51, 1	$6,7 \times 10^{3}$	$3,8 \times 10^{4}$	$1,0 \times 10^{4}$
	(9,7)	(12,3)	($1,6 \times 10^{3}$)	($2,3 \times 10^{-2}$)	($1,2 \times 10^{-1}$)
Pmais10	11,0*	-12,2*	$2,2 \times 10^{3}$*	$8,3 \times 10^{3}$	$2,0 \times 10^{5}$
	(5,8)	(12,5)	($1,4 \times 10^{3}$)	($2,0 \times 10^{-3}$)	($5,0 \times 10^{-2}$)
Nppd	0,6	-0,5	$4,2 \times 10^{1}$	4,3 e $\times 10^{2}$	$-2,4 \times 10^{3}$
	(0,1)	(0,1)	($1,8 \times 10^{1}$)	($1, 5 \times 10^{-4}$)	($5,8 \times 10^{-4}$)

*variável não significativa para o componente

Na Figura 5.17 é apresentado o resultado do modelo de regressão mista ajustado aos dados de morbidade por tuberculose. Foram classificados no último componente os municípios do Rio de Janeiro, Duque de Caxias, Nova Iguaçu, Belford Roxo, Nilópolis, Rio Bonito e Angra dos Reis. Verifica-se da figura que os municípios de Campos dos Goytacazes, Cachoeiras de Macacu, Queimados, São Gonçalo, São João de Meriti,

Maricá e Niterói. O parâmetro λ variou entre 0,000723 e 0,005128 e o parâmetro π, entre 0,40 e 0,08.

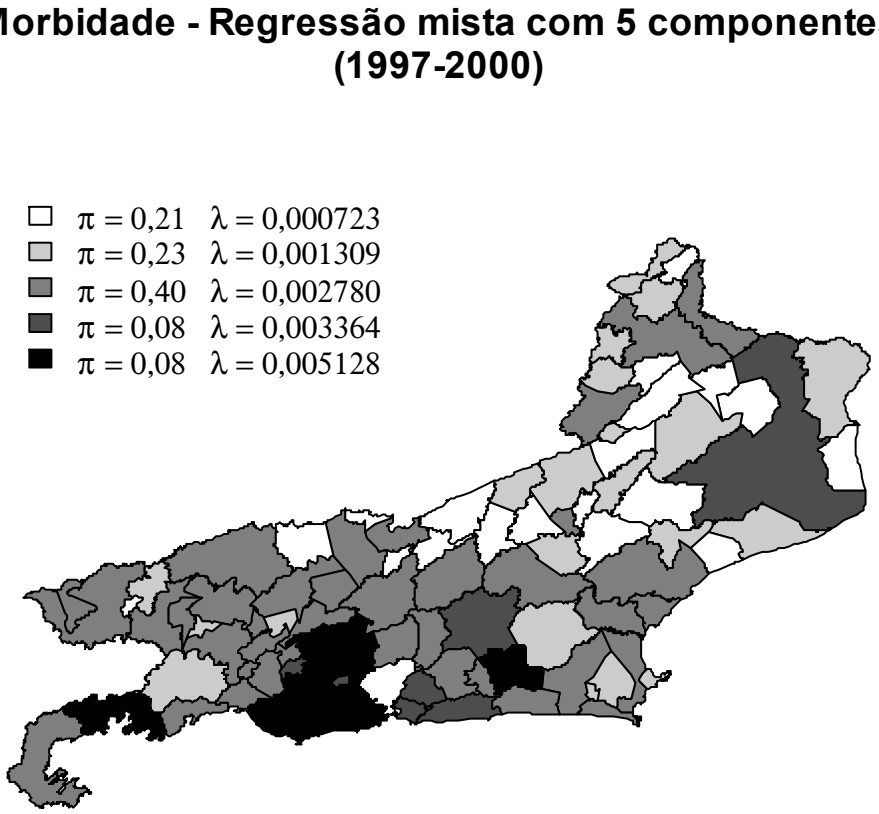


Figura 5.17: Regressão mista para os dados de morbidade por tuberculose no estado do Rio de Janeiro no período 1999-2000

Nesta seção foram apresentados os resultados do modelo de regressão de mista. Por meio desta técnica pode-se verificar, para cada um dos componentes do modelo (grupos de risco), quais covariáveis foram estatisticamente significativas.

Capítulo 6: Discussão

O presente trabalho teve como objetivo principal investigar a aplicação de técnicas para a classificação de áreas quanto ao risco de doenças, sendo investigados, em particular, os modelos de misturas finitas. Com essa finalidade foi escolhida como aplicação a tuberculose, por se tratar de um grave problema de Saúde Pública no Brasil e, em especial, no estado do Rio de Janeiro.

Basicamente, os principais interesses da Saúde Pública na classificação e mapeamento de áreas são: (1) a busca dos fatores etiológicos; (2) utilização dos mapas como ferramenta de auxílio à vigilância, fornecendo o verdadeiro risco das áreas; e (3) identificação de áreas de elevado risco, de forma que ações possam ser mais bem direcionadas.

O interesse da Saúde Pública não se concentra apenas nas áreas de elevado risco. As áreas de baixo risco também devem ser investigadas, pois podem fornecer informações importantes sobre locais onde o tratamento e/ou o sistema de varredura/busca por casos é eficiente, servindo de exemplo para outras áreas.

Foram utilizados dados relativos à mortalidade e à morbidade por tuberculose nos 91 municípios do estado do Rio de Janeiro. Os dados de mortalidade abrangeram o período 1997 à 2002 e os dados de morbidade o período 1997 à 2000.

A análise exploratória dos dados de mortalidade mostrou: (1) que entre 4,4% dos municípios não registraram óbitos no período; (2) que a maior parte dos municípios (aproximadamente 50%) registrou menos de 10 casos no período; (3) que aproximadamente 20% dos municípios registraram entre 11 e 100 casos; (4) que poucos municípios registraram mais de 100 casos (3%).

Com relação aos dados de morbidade verificou-se situação semelhante, isto é, a maior parte dos municípios registrou poucos casos e alguns municípios registraram elevado número de casos. O Rio de Janeiro foi o município com maior número de casos registrados no primeiro (18.358) e segundo (11.083) períodos. No caso da morbidade,

outros municípios registraram mais de 1000 casos. Todos eles se localizam na Região Metropolitana do estado.

Ao se analisar a situação dos municípios com relação à doença é fundamental que o tamanho da população seja levado em consideração. Isto foi feito ao se calcular a razão da mortalidade/morbidade. Ao se considerar o tamanho da população, pôde-se verificar que alguns municípios permaneceram entre aqueles identificados como de situação mais grave e que outros, anteriormente não identificados como áreas de risco, passaram a sê-lo. Situação inversa também pôde ser observada, isto é, municípios anteriormente classificados como áreas de risco tiveram sua classificação alterada com a inclusão do tamanho populacional.

A análise exploratória espacial foi realizada considerando-se três aspectos principais: a visualização dos dados, a verificação da tendência e a verificação de dependência espacial. Mapas coropléticos foram construídos no intuito de verificar a existência de estrutura espacial nos conjuntos de dados. Para a detecção de efeitos de larga escala, foram ajustadas médias móveis, enquanto na avaliação dos efeitos de pequena escala, medidas de autocorrelação global foram obtidas. Na construção dos mapas coropléticos foram adotados como critérios de divisão de classes os quintis da distribuição e o resultado da aplicação da técnica de k-médias.

A principal dificuldade observada na construção dos mapas foi a escolha dos intervalos a serem adotados. Várias opções foram testadas, como: (1) construção de intervalos de igual amplitude; (2) construção de intervalos utilizando quantis da distribuição (quartis, quintis, sextis, decis); e (3) construção de intervalos usando o método de k-médias para classificação das áreas. A busca por critérios que permitissem a visualização dos diferentes cenários existentes no estado mostrou que, dentre as alternativas implementadas, a construção de intervalos usando o método de k-médias se mostrou, inicialmente, a opção mais adequada.

O emprego dos quintis da distribuição para a construção do mapa mostrou um resultado pobre, uma vez que uniu em seu último intervalo todos os municípios com razão de mortalidade superior a 0,92 casos. Desta forma, foram agrupados no último intervalo da distribuição dos quintis municípios com razão de mortalidade entre 0,92 e 2,04 casos, isto é, realidades bem diferentes em um mesmo grupo.

Com base nos resultados obtidos na análise exploratória, decidiu-se formar cinco classes por meio da técnica de k-médias. O número de classes foi escolhido de forma a permitir a distribuição homogênea dos dados pelos grupos formados, a estes correspondendo desde municípios sem casos registrados até municípios com um número elevado de casos.

No que se refere aos resultados, pôde-se observar que a adoção dos dois critérios para a construção dos intervalos de classe, quintis e k-médias, gerou mapas bem diferentes, o que era esperado. No caso dos quintis, apesar de se poder visualizar com um pouco mais de detalhamento os 75% inferiores, municípios bem diferentes foram unidos no último intervalo. No caso das k-médias o objetivo era agrupar conforme a similaridade em cinco aglomerados ou classes de risco. O resultado encontrado se revelou mais adequado para mostrar a situação no estado. A visualização dos mapas de morbidade (Figura 5.5), quando se empregou o critério dos quintis, foi semelhante ao mapa empregando k-médias (Figura 5.6).

Comparando-se os resultados quando se empregou a mortalidade e a morbidade verifica-se que estes foram bastante semelhantes, ou seja, municípios com grande quantidade de casos estão associados a maior número de óbitos, o que seria esperado. Caso isso não ocorresse poder-se-ia pensar que, em alguns municípios, o tratamento tem maior índice de sucesso (devido ao menor abandono do tratamento, a um programa de controle mais eficaz, à existência real do tratamento supervisionado).

A aplicação da técnica de alisamento média móvel para a razão da mortalidade mostra que, no período 1997-2002, os 25% superiores encontram-se majoritariamente na Região Metropolitana, com focos na Região das Baixadas Litorâneas. Observa-se, contudo, uma melhora da situação no Norte e Noroeste do estado. Este resultado mostra o efeito da média móvel, que suaviza a superfície considerando os vizinhos.

Apesar dos mapas alisados produzirem, em média, um mapa mais estável, deve-se ter cuidado para que áreas de risco aumentado não deixem de ser identificadas. Isto pode ocorrer quando não existe uma relação de compromisso adequada entre a sensibilidade e a especificidade obtidas com o método. No caso das técnicas de alisamento, o grau de alisamento determinará o ponto de corte entre a alta sensibilidade e a alta especificidade. Este ponto de corte é importante, uma vez que uma medida

sensível mas não específica pode gerar falsos alarmes positivos, enquanto que uma medida não sensível pode deixar de detectar áreas de risco aumentado (ELLIOTT e WARTENBERG, 2004).

Com relação à autocorrelação observou-se, para os dados de mortalidade e de morbidade, significância estatística. Portanto, a distribuição no espaço não é aleatória e há dependência espacial.

Outra dificuldade encontrada na construção dos mapas se refere à escolha da medida a ser apresentada. O uso de razões de mortalidade e morbidade é freqüente. Desta forma, fazem-se necessários questionamentos sobre a utilização da razão de mortalidade/morbidade pardonizada (SMR). Muitas das pequenas áreas também apresentam pequenas populações, o que leva uma instabilidade na obtenção de estimativas. Isto ocorre quando pequenas alterações no número de casos podem levar a uma mudança acentuada na taxa e, conseqüentemente, a uma classificação equivocada do intervalo de risco do município em questão. Como mostrado por CLAYTON e KALDOR (1987) e CLAYTON *et al.* (1993), a utilização da SMR para a análise dos dados de contagem de pequenas áreas pode ser tecnicamente incorreta, sendo potencialmente enganosos os mapas construídos a partir dessas estimativas. O problema reside na suposição de que o número de casos ocorridos em determinado período de tempo assume uma distribuição de Poisson. No entanto, quando existem riscos diferentes entre as áreas, os dados observados não apresentam uma distribuição de Poisson. Nesta situação, o mapa da distribuição do SMR reflete apenas parcialmente as diferenças no risco, uma vez que a variabilidade amostral do estimador de risco (que é proporcional ao tamanho populacional) pode resultar em taxas muito elevadas para áreas com pequenas populações, empobrecendo a capacidade de análise a partir dessas informações.

Os resultados da análise espacial dos dados de mortalidade e morbidade por tuberculose nos municípios do estado do Rio de Janeiro permitiram verificar que a distribuição espacial dos eventos não segue uma distribuição aleatória, existindo áreas de maior concentração dos mesmos. Além disso, pôde-se observar, por meio dos mapas, quais são os municípios de maior risco para a tuberculose. O passo seguinte deste trabalho consistiu no ajuste de modelos de mistura finita, de forma a obter a classificação das áreas quanto ao risco. Esta técnica tem sido apontada na literatura

como potencialmente interessante para o mapeamento de áreas de risco para doenças (CHANDRASEKARAN *et al.,* 2006; RATTANASIRI, 2004; BÖHNING, 2003; MILITINO *et al.*, 2001; BÖHNING *et al.*, 2000; SCHLATTMANN *et al.,* 1996).

Inicialmente, os modelos de mistura finita foram aplicados para a modelagem dos dados de tuberculose, não tendo sido considerada a presença de covariáveis. A estimação dos componentes deste modelo foi realizada por meio do algoritmo EM, empregando-se a biblioteca FLEXMIX do programa R. Além das estimativas referentes a cada um dos componentes (π_i e λ_i), também foram obtidas as estimativas de máxima verossimilhança e as estimativas das probabilidades *a posteriori* de pertinência de cada observação a cada um dos grupos de risco identificados.

Quanto ao problema da escolha das estimativas iniciais dos parâmetros do modelo de mistura, adotaram-se conjuntos de valores aleatórios. Também foram usadas outras estimativas iniciais, como percentis da distribuição dos dados. A escolha dos valores iniciais dos parâmetros do modelo tem grande importância para a velocidade de convergência do algoritmo e na sua capacidade de identificar máximos globais. Com relação à escolha do número de componentes da mistura, diferentes escolhas foram testadas, variando de 2 a 7. Observou-se que modelos com quatro a seis componentes exibiram bom ajuste aos dados. Quando se utilizou um número maior de componentes, estes não se mostraram mais vantajosos, sendo mais difícil a interpretação da classificação das áreas quanto ao risco.

Com relação aos dados de mortalidade relativos, verifica-se um melhor ajuste dos modelos de mistura com quatro componentes. Para o modelo selecionado, observa-se que o primeiro componente da mistura engloba mais de 50% dos municípios. Foram classificados no penúltimo ou último componente os municípios de Duque de Caxias, Nova Iguaçu, Belford Roxo, São João de Meriti, Niterói, São Gonçalo Itaboraí e Nilópolis. Todos esses municípios pertencem à Região Metropolitana. Para os dados de morbidade, foi selecionado um modelo de mistura com seis componentes. Quanto aos municípios classificados como de maior risco (penúltimo e último componentes da mistura), observou-se que, assim como para a mortalidade, foram identificados aqueles pertencentes à Região Metropolitana. Foram incluídos no penúltimo componente os municípios de Petrópolis (Região Serrana do estado) e o município de Volta Redonda (Região do Médio Paraíba).

Verificou-se que quanto maior o número de componentes, isto é, de distribuições que compunham a mistura, melhor foi o ajuste visual do modelo aos dados e menor foi o valor do BIC. Pode-se observar que tanto para a mortalidade quanto para a morbidade, modelos com quatro a seis componentes se mostraram mais vantajosos para explicar a estrutura de heterogeneidade dos dados de tuberculose. Levando-se em consideração o montante de redução do BIC e a capacidade de interpretação dos resultados, os modelos com mais de sete componentes não se mostraram mais interessantes para explicar a estrutura de risco para tuberculose no estado.

Entre as vantagens do modelo de mistura finita pode-se citar o fato de que o número de componentes não precisa ser previamente definido, podendo ser mais um parâmetro a ser estimado pelo modelo. Isto não ocorre, por exemplo, com o modelo de k-médias que, apesar de ser de simples implementação, requer que se defina *a priori* o número de classes nas quais os dados serão agrupados. Esta característica pode ser desvantajosa quando o conhecimento sobre os possíveis grupos que compõem os dados é escasso. Outra vantagem do modelo de mistura finita é que ele permite a identificação do modelo probabilístico que melhor descreve os dados. Este fato possibilita a realização de inferências a partir do modelo ajustado aos dados. Além disso, o modelo de mistura é flexível e permite a incorporação de covariáveis, o que pode ser interessante para a área da Saúde Pública, avaliar a influência das diferentes covariáveis em cada grupo de risco.

Os modelos de mistura finita representam um método válido para a modelagem da heterogeneidade existente no risco de uma doença em uma população, sendo capazes de identificar populações homogêneas e heterogêneas, ao contrário de outros métodos de mapeamento de doenças. Do ponto de vista da Saúde Pública esta é uma qualidade desejável. Outra característica metodológica dos modelos de mistura finita é o fornecimento do número de subpopulações ou grupos de risco que compõem a população.

Os resultados deste trabalho sugerem que os modelos de mistura finita são adequados para a classificação de regiões e podem ser, portanto, o ponto de partida para a avaliação do risco em mapas de doenças.

A confiabilidade dos modelos de mistura foi investigada por BÖHNING e SCHLATTMANN (1999), em que foram realizadas simulações nas quais a abordagem de mistura finita foi comparada com as abordagens tradicionalmente empregadas na área de mapeamento de doenças, como os percentis da distribuição do SMR ou a abordagem baseada no nível de significância de uma distribuição de Poisson. Foram simulados cenários com vários níveis de risco em populações consistindo de duas ou três subpopulações. O verdadeiro risco de cada área para cada cenário era conhecido. Após a realização de grande número de replicações, o autor verificou que os modelos de mistura classificaram corretamente as áreas em mais de 80% das simulações. Os demais métodos (percentis e nível de significância) apresentaram resultados variando de 15 a 35% de classificações corretas.

Os modelos de mistura finita se mostraram uma técnica interessante e flexível de modelagem e com resultados de fácil interpretação. Atualmente, já existem programas de domínio público como CAMAN (BÖHINING *et al.*, 1998), DISMAPWIN (BÖHINING *et al.*, 2000), e a biblioteca FLEXMIX do programa R (LEISCH, 2004) para aplicação desses modelos, o que pode facilitar e ampliar seu uso.

Os modelos de regressão binomial negativa foram ajustados aos dados com o intuito de verificar, entre as variáveis disponíveis, aquelas que melhor poderiam explicar a situação de mortalidade/morbidade por tuberculose encontrada nos municípios do estado do Rio de Janeiro. Por meio do modelo de regressão, puderam ser identificadas as covariáveis que têm efeito positivo e negativo, estatisticamente significativos, no evento de interesse.

Nesta etapa da modelagem dos dados foram empregadas as seguintes covariáveis: *proporção de famílias com renda maior que dez salários mínimos (renda > 10), número médio de moradores por domicílio (nppd), Índice de Desenvolvimento Humano (IDH), Índice de Gini (Gini), densidade demográfica (densdemog) e a taxa de AIDS (tx9702 e tx9700).* Após a identificação das covariáveis significativas, um novo modelo foi ajustado aos dados.

Verificou-se para os dados de mortalidade, que as covariáveis *Gini* e *tx9702* mostram estatisticamente significativas. A variável *gini* exibiu coeficiente negativo, rindicando efeito protetor e a variável *tx9702*, apresentou coeficiente positivo, indicando

um efeito de aumento do risco. Com relação à morbidade, verificou-se que as variáveis *Gini, IDH* e *tx9700* foram significativas. As variáveis *IDH* e *tx9700* apresentaram coeficientes positivos, significando sua associação com maior morbidade.

Os resultados dos modelos de regressão binomial negativa foram comparados com aqueles obtidos mediante aplicação dos modelos de regressão de mista. Estes modelos foram implementados no programa R, sendo estimados os coeficientes para cada uma das covariáveis, o BIC, o AIC e a log-verossimilhança.

O objetivo da utilização dos dois modelos de regressão foi permitir a comparação entre seus resultados. Na regressão binomial negativa, pôde-se verificar a influência das covariáveis para o estado como um todo. O modelo de regressão de misto, por sua vez, possibilitou a identificação, para cada grupo de risco, da influência das covariáveis. Observou-se que o que pode ser influente para os municípios que compõem um grupo de risco (componente da mistura) pode não ser influente para outro grupo (componente). A detecção dos fatores (covariáveis) associados aos grupos de maior e menor risco pode ser importante para a área da Saúde Pública, permitindo um direcionamento mais adequado dos esforços / recursos para a solução dos problemas de saúde.

O modelo de regressão mista ajustado aos dados de mortalidade do período 1997-2002 possui cinco componentes. Não foram consideradas significativas para nenhum dos cinco componentes as variáveis *renda > 10* e *$n^{o.}$ médio moradores*, tendo sido retiradas do modelo. Foram consideradas significativas as covariáveis, *IDH, Gini, densdemog e tx9702*. Apenas a covariável *Gini* não foi significativa para todos os componentes da mistura. Para o primeiro componente do modelo verificou-se um sinal negativo do coeficiente estimado para a variável *densdemog*, sugerindo, desta forma, um efeito protetor desta variável para este grupo de municípios. Também exibiram sinais negativos as estimativas obtidas para as variáveis *IDH* e *tx9702* (terceiro componente da mistura), *gini* e *tx9702* (quarto componente da mistura). Foram agrupados, respectivamente, do primeiro ao quinto componentes, 41, 34, 5, 5 e 6 municípios. Assim, verifica-se que os dois primeiros componentes são compostos pela grande maioria dos municípios, os quais apresentaram mortalidade por tuberculose reduzida.

A aplicação da regressão mista aos dados de morbidade levou ao ajuste de modelos com cinco componentes. Verificou-se que todas as covariáveis inseridas no modelo foram consideradas significativas para pelo menos um dos componentes, sendo mantida no modelo final selecionado para os dados. Observou-se que para o terceiro componente da mistura, as variáveis *IDH*, *Gini* e *pmais10* não foram estatisticamente significativas. A ausência de significância estatística da variável *pmais10* também foi verificada para o primeiro e o segundo componentes da mistura. Observou-se que uma mesma covariável é fator de risco para alguns componentes e de proteção para outros.

Os modelos ajustados neste trabalho têm seus resultados apresentados em mapas e com isso, pode-se observar se vizinhos foram classificados em grupos semelhantes de risco. Espera-se que isso ocorra uma vez que foi identificada dependência espacial entre as áreas. Esta observação, no entanto, é apenas visual. Efetivamente, a questão da dependência espacial não foi incorporada no modelo. Isto pode ser feito, isto é, pode-se incorporar a dependência espacial ao modelo de mistura. Um exemplo desta aplicação pode ser vista em MILITINO *et al.* (2001) e LAWSON e CLARK (2002).

Com relação ao município do Rio de Janeiro, acredita-se que este deve ser alvo de um modelo de mistura ajustado somente aos seus dados e que considere as ocorrências por bairros ou regiões administrativas, ou, ainda, setores censitários. Para tanto, é necessária a disponibilidade de dados para níveis menores de agregação, o que não foi possível para este trabalho. Além das informações referentes à morbidade e mortalidade por tuberculose em um nível menor de agregação, também seria necessário obter as informações relativas às covariáveis.

A análise e o mapeamento da variação geográfica de taxas de doenças é uma forma de permitir a rápida e imediata identificação de padrões espaciais e aglomerados. Desta forma, desigualdades entre as regiões podem ser identificadas e receber aumento de atenção. A evidência de uma incidência aumentada ou um aglomerado no espaço ou tempo de uma doença pode dirigir a atenção para um fator de risco ambiental e fornecer um aviso do aumento da ocorrência de uma doença. Com a rápida expansão dos serviços e pesquisas em saúde, os pesquisadores podem se beneficiar do rápido acesso às informações chaves disponíveis para pequenas áreas (KELLY, 1999).

A demanda por informação no nível da pequena área tem crescido rapidamente e as vantagens dos modelos de misturas finitas serão mais evidentes onde os dados são escassos e o uso de razões de mortalidade/morbidade leva a estimativas instáveis. Contudo, apesar dos benefícios dos métodos modernos de mapeamento de informações de pequenas áreas, sua introdução e aceitação não é imediata, sendo necessário o desenvolvimento de programas computacionais que realizem tais métodos e que possuam interfaces mais amigáveis para o usuário.

Capítulo 7: Conclusão

O presente trabalho teve por objetivo investigar a utilização dos modelos de mistura finita com e sem covariáveis na identificação de áreas de risco para doenças. Como exemplo de aplicação foram empregados dados de mortalidade e morbidade por tuberculose nos municípios do estado do Rio de Janeiro. O trabalho consistiu das seguintes etapas: (1) análise exploratória; (2) análise exploratória espacial; (3) modelos de mistura finita; (4) regressão binomial negativa; e (5) regressão de Poisson mista.

Com relação às técnicas de análise espacial empregadas neste trabalho, pôde-se concluir que a distribuição da mortalidade e da morbidade no estado do Rio de Janeiro não se distribui espacialmente de maneira aleatória, existindo regiões de dependência espacial. Os mapas coropléticos construídos para a visualização dos dados indicam que há concentração de casos nos municípios que pertencem a Região Metropolitana do estado.

Quanto aos modelos de mistura, verificou-se que modelos com quatro a seis componentes mostraram melhor ajuste aos dados. Desta forma, conclui-se que o estado do Rio de Janeiro pode ser classificado em quatro a seis grupos quanto ao risco para tuberculose. Os modelos de mistura se mostraram uma ferramenta interessante para a modelagem de dados heterogêneos conseguindo captar diferentes situações de risco e sendo capaz de identificá-las de forma adequada. Adicionalmente, conclui-se que esta se trata de uma técnica de fácil interpretação. Quanto à implementação dos modelos, já existem pacotes de domínio público que ajustam modelos de mistura finita.

A aplicação da regressão binomial negativa mostrou ajuste adequado aos dados de mortalidade e morbidade por tuberculose. Este modelo considera que o risco é igual para todos os municípios e identifica as covariáveis com significância estatística para o estado como um todo. Contudo, por meio da análise exploratória e do modelo de mistura ajustado aos dados verificou-se que os dados são heterogêneos, isto é, compostos de várias subpopulações. Assim, a estrutura de risco é melhor definida por um modelo que considere riscos diferentes para cada subpopulação.

A aplicação do modelo de regressão mista aos dados de tuberculose permitiu: (1) refinar o modelo de mistura finita por meio da inclusão de covariáveis; (2) caracterização dos grupos de risco por meio da identificação de covariáveis significativas. Os modelos de regressão mista se mostraram uma alternativa promissora para a modelagem de dados heterogêneos com valores extremos e grande dispersão, permitindo a separação ou classificação das áreas de forma apropriada.

Acredita-se ser necessário explorar ainda mais as técnicas de regressão mista. Habitualmente, na literatura, a seleção dos modelos de mistura ou de regressão mista é realizada empregando-se um critério de decisão como o de Akaike e o BIC. Acredita-se ser interessante, e necessário, fazer a análise dos resíduos destes modelos.

A implementação dos modelos de regressão mista pode ser realizada no programa de domínio público R, onde a biblioteca Flexmix possui funções para ajuste desses modelos. Sua utilização é relativamente simples e os resultados compensadores.

Outra questão observada é a necessidade de incorporação da dependência espacial aos modelos de regressão mista. Este procedimento é descrito na literatura por alguns autores e parece, uma expansão do modelo de mistura bastante interessante. Desta forma, o espaço seria efetivamente incorporado a modelagem.

Conclui-se que os modelos de misturas finitas e a regressão de mista são ferramentas com grande potencial para a classificação de áreas quanto ao risco de doenças, o que pode ser útil para a área da Saúde Pública, pois permitem não só a identificação dos verdadeiros grupos de risco que compõem uma população como também a sua caracterização. A partir de resultados obtidos com o emprego de tais técnicas, é possível realizar uma melhor alocação de recursos e um direcionamento mais adequado de programas e políticas de saúde.

Referências bibliográficas

ANSELIN, L., 1995, "Local indicators of spatial autocorrelation – LISA", *Geographical Analysis*, v. 27, pp. 91-115.

ANSELIN, L., 1996, "The Moran scatterplot as ESDA tool to assess local instability in spatial association". In: Fisher, M., Scholten, H.J., Unwin, D., *Spatial Analytical Perspectives on GIS*. London, Taylor & Francis.

AYUTHYA, R.S., BÖHNING, D., 1995, "Traffic Accident Mapping in Bangkok Metropolis: a Case Study", *Statistics in Medicine*, v. 14, pp. 2445-2458.

BAILEY, T., 2001, "Spatial statistical methods in health", *Cadernos de Saúde Pública*, v. 17, n. 5, pp. 1083-1098.

BAILEY, T.C., GATRELL, A.C., 1995, *Interactive Spatial Data Analysis*. 1ed. Essex, Longman Group Limited.

BEKLAS, K., GALATSANOS, N.P., LIKAS, A., LAGARIS, I.E., 2005, "Mixture Model Analysis of DNA Microarray Images", *IEEE Transactions on Medical Imaging*, v. 34, n. 7, pp. 901-909.

BIVAND, R., 2006, *SPDEP: Spatial dependence: weighting schemes, statistics and models.* R package version 0.3-25. Disponível em: <http://www.R-project.org.>

BÖHNING, D., 2003, "Empirical bayes estimators and non-parametric mixture models for space and time-space disease mapping and surveillance", *Environmetrics*, v. 14, pp. 431-451.

BÖHNING, D., DIETZ, E., SCHLATTMANN, P., 1998, "Recent Developments in Computer-Assisted Analysis of Mixtures", *Biometrics*, v. 54, pp. 525-536.

BÖHNING, D., DIETZ, E., SCHLATTMANN, P., 2000, "Space-time mixture modeling of public health data", *Statistics in Medicine*, v. 19, pp. 2333-2344.

BÖHNING, D., SCHLATTMANN, P., 1999, "Disease Mapping with Hidden Structures Using Mixture Models". In: LAWSON, A., BIGGERI, A., BÖHNING, D., LESAFFRE, E., VIEL, J-F, BERTOLLINI, R. (eds), *Disease Mapping and Risk Assessment for Public Health*, 1. ed., chapter 4, Chichester, Wiley.

BÖHNING, D., SEIDEL, W., 2003, "Editorial: recent developments in mixture models", *Computational Statistics and Data Analysis*, v. 41, pp. 349-357.

BURDEN, R.L., FAIRES, J.D., 1988, *Numerical Analysis*. 4 ed. Boston, PWS-KENT Publishing Company.

CAPPÉ, O., 2001, "H2M: A set of MATLAB/OCTAVE functions for the EM estimation of mixtures and hidden Markov models", Disponível em: <http://www.tsi.enst.fr/~cappe/h2m>. Acesso em: março 2004.

CELEUX, G., SOROMENHO, G., 1996, "An entropy criterion for assessing the number of clusters in a mixture model", *Classification Journal*, v.13, pp. 267-212.

CHANDRASEKARAN, K., ARIVARIGNAN, G., 2006, "Disease mapping using mixture distribution", *Indian Journal of Medical Research*, v. 123, pp. 788-798.

CHOUQUET, C., RICHARDSON, S., BURGARD, M., BLANCHE, S., MAYAUX, M-J., RQUZIOUX, C., COSTAGLIOLA, D., 1999, "Timing of Human Immunodeficiency Virus Type 1(HIV-1) Transmission from Mother to Child: Bayesian Estimation Using a Mixture", *Statistics in Medicine*, v. 18, pp. 815-833.

CIDE – Centro de Informações e Dados do Rio de Janeiro, 2006, *Casos Notificados de AIDS – ano de diagnóstico*, Disponível em: <http://www.cide.rj.gov.br >. Acesso em: 1 julho 2006.

CLAYTON, D.G., BERNARDINELLI, L., MONTOMOLI, C., 1993, "Spatial correlation and ecological analysis", *International Journal of Epidemiology*, v. 22, pp. 1193-1201.

CLAYTON, D.G., KALDOR, J., 1987, "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping", *Biometrics*, v. 43, pp. 671-681.

CLIFF, A.D., ORD, J.K., 1981, *Spatial Processes. Models and Applications*. London, Pion.

DALRYMPLE, M.L., HUDSON, I.L., FORD, R.P.Q., 2003, "Finite Mixture, Zero-Inflated Poisson and Hurdle models with application to SIDS", *Computational Statistics & Data Analysis*, v. 41, pp. 491-504.

DATASUS – Departamento de Informática do SUS, 2006, *Estatísticas vitais – Mortalidade e Nascidos Vivos*, Disponível em: <http://www.datasus.gov.br >. Acesso em: 1 julho 2006.

DEAN, C.B., 1992, "Testing for overdispersion in Poisson and binomial regression models", *Journal of the American Statistical Association*, v. 87, pp. 451-457.

DEMPSTER, A.P., LAIRD, N.M., RUBIN, D.B., 1977, "Maximum likelihood from incomplete data via de the EM algorithm (with discussion)", *Journal of the Royal Statistical Society B*, v. 39, pp. 1-38.

ELLIOTT, P., WARTENBERG, D., 2004, "Spatial Epidemiology: Current Approaches and Future Challenges", *Environmental Health Perspectives*, v. 112, n. 9, pp. 998-1006.

ESM Consultoria, 2003, Atlas do Desenvolvimento Humano no Brasil, versão 1.0.0.

GETIS, A.; ORD, J.K., 1992, "The analysis of spatial association by use of distance statistics", *Geographical Analysis*, v. 24, pp. 189-206.

GHOSH, M., RAO, J.N.K., 1994, "Small Area Estimation: An Appraisal", *Statistical Science*, v. 9, pp. 55-93.

HAYKIN, S., 2001, *Redes Neurais: Princípios e Prática*. 2 ed. Porto Alegre, Bookman.

HIJAR, M.A., 2005, "Tuberculose: desafio permanente", *Cadernos de Saúde Pública*, v. 21, n.2, pp. 348-349.

IBGE - FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2006, *Censo 2000.* Disponível em: <http://www.ibge.gov.br >. Acesso em: 1 julho 2006.

KARLIS, D., XEKALAKI, E., 2003, "Choosing initial values for the EM algorithm for finite mixtures", *Computational Statistics and Data Analysis*, v. 41, pp. 577-590.

KELLY, A., 1999, "Case Studies in Bayesian Disease Mapping for Health Service Research in Ireland". In: LAWSON, A., BIGGERI, A., BÖHNING, D., LESAFFRE, E., VIEL, J-F, BERTOLLINI, R. (eds), *Disease Mapping and Risk Assessment for Public Health*, 1 ed., chapter 27, Chichester, Wiley.

KRAZANOWSKI, W.J., 1998, *An Introduction to Statistical Modelling*. 1 ed. New York, Oxford University Press.

KERR-PONTES, L.R.S., OLIVEIRA, F.A.S., FREIRE, C.A.M., 1997, "Tuberculose associada à AIDS: situação do Nordeste brasileiro", *Revista de Saúde Pública*, v.31, n.4, pp.323-329.

LAGRANGE, P.H., HERRMANN, A.W., 2000, "Mycobacteriosis in the Compromised Host", *Memórias do Instituto Oswaldo Cruz*, v.95, Supl.I, pp.163-170.

LAMBERT, D., 1992, "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing", *Technometrics*, v. 34, n. 1, pp. 1-14.

LAWSON, A., BIGGERI, A., BÖHNING, D., LESAFRE, E., VIEL, J-F., BERTOLLINI, R. (eds), 1999, *Disease Mapping and Risk Assessment for Public Health*. 1 ed. Chichester, Wiley.

LAWSON, A.B., 2005, *Statistical Methods in Spatial Epidemiology*, 1 ed. Chichester, Wiley.

LAWSON, A.B., BIGGERI, A.B., BÖHNING, D., LESAFRE, E., VIEL, J-F., CLARK, A., SCHLATTMANN, P., DIVINO, F., 2000, "Disease mapping models: an empirical evaluation", *Statistics in Medicine*, v. 19, pp. 2217-2241.

LAWSON, A.B., CLARK, A., 2002, "Spatial mixture relative risk models applied to disease mapping", *Statistics in Medicine*, v. 21, pp. 359-370.

LEISCH, F., 2004, "FlexMix: A general framework for finite mixture models and latent class regression in R", *Journal of Statistical Software*, v. 11, n. 8. Disponível em: <http://www.jstatsoft.org/v11/i08/>. Acesso em: 10 setembro 2006.

LIMA, M.M., BELLUOMINI, M., ALMEIDA, M.M.M.B., ARANTES, G.R., 1997, "Co-infecção HIV/tuberculose: necessidade de uma vigilância mais efetiva", *Revista de Saúde Pública*, n.31, v.3, pp.217-220.

LIU, C., WEBER, K., ROBINSON, E., HU, Z., JACOBSON, L.P., GANGE, S.T., 2006, "Assessing the effect of HAART on change in quality os life among HIV-infected women", *AIDS Research and Therapy*, v. 3, n. 6. Disponível em: <http://www.aidsrestherapy.com/content/3/1/6> . Acesso em: 20 outubro 2006.

MARSHALL, R.J., 1991, "Mapping disease and mortality rates using empirical bayes estimators", *Applied Statistics*, v. 40, n. 2, pp. 283-294.

MARTINEZ, W.L., MARTINEZ, A.R., 2002, *Computational Statistics Handbook with MATLAB®*, 1 ed. Florida, Chapman & Hall.

MCLACHLAN, G., PEEL, D., 2001, *Finite Mixture Models*. 1 ed. New York, Wiley.

MILITINO, A.F., UGARTE, M.D., DEAN, C.B., 2001, "The use of mixture models for identifying high risks in disease mapping" *Statistics in Medicine*, v. 20, pp. 2035-2049.

MINISTÉRIO DA SAÚDE, 1998, *Guia de Vigilância Epidemiológica*, 1998, 4 ed. Brasília.

MINISTÉRIO DA SAÚDE, 1999, *Plano Nacional de Controle de Tuberculose*, Brasília.

MINISTÉRIO DA SAÚDE, 2006. *Vigilância Epidemiológica*. Disponível em: <http://portal.saude.gov.br/portal/svs/visualizar_texto.cfm?idtxt=21507. Acesso em: 1 setembro 2006.

MOTA, F.F, SILVA, L.M.V, PAIM, J.S., COSTA, M.C.N., 2003, "Distribuição espacial da mortalidade por tuberculose em Salvador, Bahia, Brasil", *Cadernos de Saúde Publica*, v.19, n.4, pp.915-922.

PEREIRA, M.G., 2005, *Epidemiologia Teoria e Prática*, 8 ed. Rio de Janeiro, Guanabara Koogan S.A.

RATTANASIRI, S., BÖHNING, D., ROJANAVIPART, P., ATHIPANYAKOM, S., 2004, "A mixture model application in disease mapping of malaria", *Southeast Asian J Trop Med Public Health*, v. 35, n. 1, pp. 38-47.

REZAEIAN, M., DUNN, G., ST. LEGER, S., APPLEBY, L., 2004, "The production and interpretation of disease maps: A methodological case study", *Social Psychiatry and Psychiatric Epidemiology*, v. 39, n. 12, pp. 947-954.

RICHARDSON S.; THOMSON A.; BEST N.; ELLIOTT P., 2004, "Interpreting posterior relative risk estimates in disease-mapping studies", *Environmental Health Perspectives*, v. 112, pp. 1016-1025.

RODRIGUES-JR, A., RUFFINO-NETTO, A., CASTILHO, E. A., 2006, "Spatial distribution of *M. tuberculosis*/HIV coinfection in São Paulo State, Brazil, 1991-2001", *Revista de Saúde Publica*, v.40, n.2, pp. 265-270.

RUFFINO-NETTO, A., 2001, "Programa de controle da tuberculose no Brasil: situação atual e novas perspectivas", *Informe Epidemiológico do SUS*, v. 10, n. 3, pp. 129-138.

RUFFINO-NETTO, A., SOUZA, A.M.A.F., 1999, "Reforma do Setor Saúde e Controle da Tuberculose no Brasil", *Informe Epidemiológico do SUS*, v. 8, n. 4, pp. 35-51.

SCHLATTMANN, P., BÖHNING, D., 1993, "Mixture Models and Disease Mapping", *Statistics in Medicine*, v. 12, pp. 1943-1950.

SCHLATTMANN, P., DIETZ, E., BÖHNING, D., 1996, "Covariate adjusted mixture models and disease mapping with the program DISMAPWIN", *Statistics in Medicine*, v.15, pp.919-929.

SCHWARZ, G., 1978, "Estimating the dimension of a model", *Annals of Statistics*, v. 6, pp. 461-464.

SES/RJ, 2003, *Plano Estratégico para o Controle da Tuberculose no Estado do Rio de Janeiro – 2003 a 2005*. Disponível em: http://www.saude.rj.gov.br/Tuberculose/Artigos/plano%20estrat%E9gico%202003-2005.pdf >. Acesso em: 1 de julho de 2006.

SOROMENHO, G., 1993, "Comparing approaches for testing the number of components in a finite mixture model", *Computational Statistics*, v. 9, pp. 65-78.

SOUZA, W. V., 2003, *A epidemiologia da tuberculose em uma cidade brasileira na última década do século XX: uma abordagem espacial*. Tese de D.Sc., ENSP/FIOCRUZ, Recife, PE, Brasil.

SOUZA, W.V., , ALBUQUERQUE, M.F.M., BARCELLOS, C.C., XIMENES, R., , 2000, CARVALHO, M.S., 2005, "Tuberculose no Brasil: construção de um sistema de vigilância de base territorial", *Revista de Saúde Pública*¸ v. 39, n.1, pp. 82-89.

SOUZA, W.V., XIMENES, R., ALBUQUERQUE, M.F.M., LAPA, T.M., PORTUGAL, J.L., LIMA, M.L.C., MARTELLI C.M.T., 2000, "The use of socioeconomic factors in mapping tuberculosis risk areas in a city of northeastern Brazil", *Revista Panamericana de Salud Pública*, v.8, n.6, pp. 403-410.

STROUP, D.F., BROOKMEYER, R., KALSBEEK, W.D., 2004, "Public Health Surveillance in Action: a Framework". In: Stroup, D.F., Brookmeyer, R. (eds) *Monitoring the Health of Populations,* 1 ed., chapter 1, Oxford: Oxford University Press.

THACKER, S.B., 2000, "Historical Development". In: Teusch, S.M., Churchill, R.E. (eds), *Principles and Practice of Public Health Surveillance,* 2 ed., chapter 1, New York: Oxford University Press.

THE MATH WORKS INC., 2002, MATLAB versão 6.5, The Math Works Inc.

THE R CORE DEVELOPMENT TEAM, 2006, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org.>

UGARTE, M.D., IBÁÑEZ, B., MILITINO, A.F., 2004, "Testing for Poisson Zero Inflation in Disease Mapping", *Biometrical Journal*, v. 46, n. 5, pp. 526-539.

VENDRAMINI, S.H.F., SANTOS, M.L.S.G., GAZETTA, C.E., CHIARAVALLOTI-NETO, F., RUFFINO-NETTO, A., VILLA, T.C.S, 2006, "Tuberculosis risks and socio-economic level: a case study of a city in Brazilian south-east, 1998-2004", *International Journal of Tuberculosis and Lung Disease*, v.10, n.11, pp.1231-1235.

VICENTIN G., SANTO, A.H., CARVALHO, M.S., 2002, "Mortalidade por tuberculose e indicadores sociais no município do Rio de Janeiro", *Ciência & Saúde Coletiva*, v. 7, n. 2, pp. 253-263

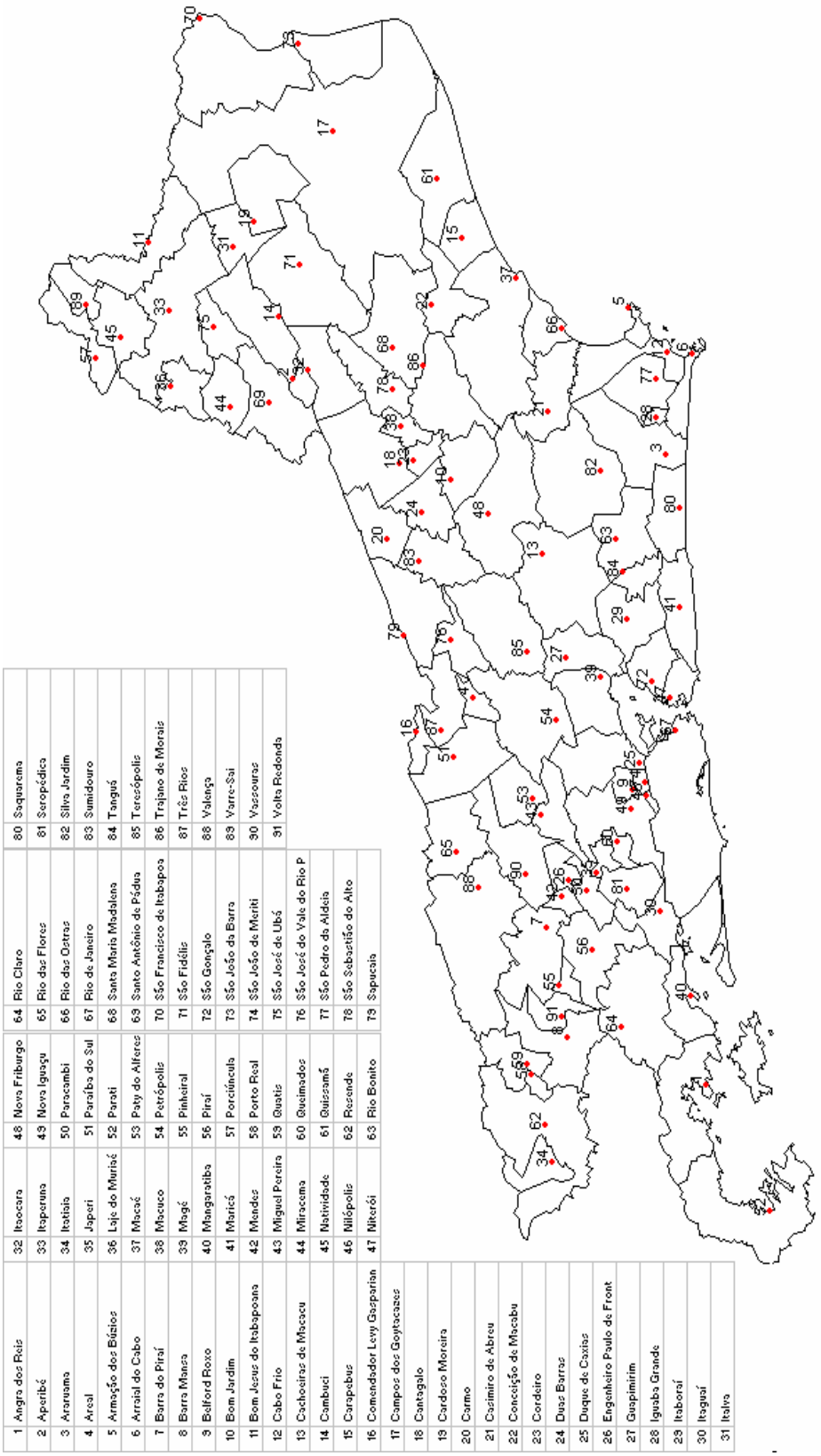
VILLATE, J., IBÁÑEZ, B., CABRIATA, V., PIJOÁN, J.I., TABOADA, J., URKAREGI, A., 2006, "Analysis of latent mixture and mycobacterim avium infection data using mixture models", *BMC Public Health*. Disponível em: <http://www.biomedcentral.com/1471-2458/6/24 > . Acesso em: 20 outubro 2006.

XIAO, J., LEE, A.H., VEMURI, S.R., 1999, "Mixture distribution analysis of length of hospital stay for efficient funding", *Socio-Economic Planning Sciences*, v. 33, pp. 39-59.

YAU, K.K.W., LEE, A.H., NG, A.S.K., 2003, "Finite mixture regression model with random effects: application to neonatal hospital length of stay", *Computational Statistics & Data Analysis*, v. 41, pp. 359-366.

ANEXO I

Mapa do estado do Rio de Janeiro por municípios



1	Angra dos Reis	32	Itaocara	48	Nova Friburgo	64	Rio Claro	80	Saquarema
2	Aperibé	33	Itaperuna	49	Nova Iguaçu	65	Rio das Flores	81	Seropédica
3	Araruama	34	Itatiaia	50	Paracambi	66	Rio das Ostras	82	Silva Jardim
4	Areal	35	Japeri	51	Paraíba do Sul	67	Rio de Janeiro	83	Sumidouro
5	Armação dos Búzios	36	Laje do Muriaé	52	Parati	68	Santa Maria Madalena	84	Tanguá
6	Arraial do Cabo	37	Macaé	53	Paty do Alferes	69	Santo Antônio de Pádua	85	Teresópolis
7	Barra do Piraí	38	Macuco	54	Petrópolis	70	São Francisco de Itabapoa	86	Trajano de Morais
8	Barra Mansa	39	Magé	55	Pinheiral	71	São Fidélis	87	Três Rios
9	Belford Roxo	40	Mangaratiba	56	Pirai	72	São Gonçalo	88	Valença
10	Bom Jardim	41	Maricá	57	Porciúncula	73	São João da Barra	89	Varre-Sai
11	Bom Jesus do Itabapoana	42	Mendes	58	Porto Real	74	São João de Meriti	90	Vassouras
12	Cabo Frio	43	Miguel Pereira	59	Quatis	75	São José de Ubá	91	Volta Redonda
13	Cachoeiras de Macacu	44	Miracema	60	Queimados	76	São José do Vale do Rio P		
14	Cambuci	45	Natividade	61	Quissamã	77	São Pedro da Aldeia		
15	Carapebus	46	Nilópolis	62	Recende	78	São Sebastião do Alto		
16	Comendador Levy Gasparian	47	Niterói	63	Rio Bonito	79	Sapucaia		
17	Campos dos Goytacazes								
18	Cantagalo								
19	Cardoso Moreira								
20	Carmo								
21	Casimiro de Abreu								
22	Conceição de Macabu								
23	Cordeiro								
24	Duas Barras								
25	Duque de Caxias								
26	Engenheiro Paulo de Front								
27	Guapimirim								
28	Iguaba Grande								
29	Itaboraí								
30	Itaguaí								
31	Italva								