

**TÉCNICAS DE PROCESSOS ESPACIAIS E
ESPAÇO-TEMPORAIS COM APLICAÇÕES EM DADOS DE
DENGUE**

Jacqueline Domingues

Dissertação apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” para a obtenção do título de mestre em Biometria.

BOTUCATU
São Paulo – Brasil
Fevereiro – 2017

**TÉCNICAS DE PROCESSOS ESPACIAIS E
ESPAÇO-TEMPORAIS COM APLICAÇÕES EM DADOS DE
DENGUE**

Jacqueline Domingues

Orientador: Prof. Dr. **José Silvio Govone**
Co-orientadora: Profa. Dra. **Vânia Ap. Ceccato**

Dissertação apresentado à Universidade Estadual Paulista “Júlio de Mesquita Filho” para a obtenção do título de mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro – 2017

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Domingues, Jacqueline.

Técnicas de processos espaciais e espaço-temporais com aplicações em dados de dengue / Jacqueline Domingues. - Botucatu, 2017

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: José Silvio Govone

Coorientador: Vânia Aparecida Ceccato

Capes: 90000005

1. Dengue. 2. Análise espacial (Estatística).
3. Epidemiologia - Modelos matemáticos. 4. Biometria.
5. Espaço e tempo.

Palavras-chave: Dengue; Epidemiologia; Estatística espacial; Modelos Espaciais; Modelos espaço-temporais.

MEMBROS DA COMISSÃO JULGADORA DA DISSERTAÇÃO DE MESTRADO DE JACQUELINE DOMINGUES, INTITULADA "TÉCNICAS DE PROCESSOS ESPACIAIS E ESPAÇO-TEMPORAIS COM APLICAÇÕES EM DADOS DE DENGUE", APRESENTADA AO INSTITUTO DE BIOCIÊNCIAS, UNESP, CAMPUS DE BOTUCATU, SÃO PAULO, EM 17 de fevereiro de 2017 .

APROVADA PELA COMISSÃO JULGADORA:

Prof(a) Dr(a) JOSE SILVIO GOVONE

Instituição: IGCE / UNESP

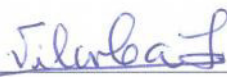
Assinatura:



Prof(a) Dr(a) VILMA MAYUMI TACHIBANA

Instituição: Faculdade de Ciencias e Tecnologia de Presidente Prudente

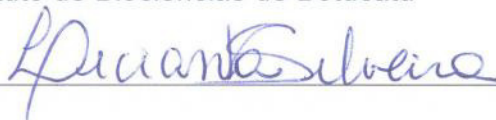
Assinatura:



Prof(a) Dr(a) LICIANA VAZ DE ARRUDA SILVEIRA

Instituição: Instituto de Biociências de Botucatu

Assinatura:



*“Tudo aquilo que sou, ou pretendo ser,
devo a um anjo: minha **mãe**.”*

(Abraham Lincoln)

Agradecimentos

A Deus.

Aos meus pais Fabiana e Agnaldo, minha irmã Joyce; aos meus avós Alécio, Maria Thereza e Maria Furlanetto Domingues (*in memorian*) e ao meu padrinho acadêmico Wesley.

Aos meus amigos.

Ao professor orientador José Silvio Govone pela dedicação e apoio. Ao técnico Jorge Gustavo Falcão pelo georreferenciamento.

À Fundação Municipal de Saúde de Rio Claro – SP.

Ao CEAPLA - Centro de Análise e Planejamento Ambiental.

Aos professores e funcionários do Programa de Pós Graduação em Biometria - Universidade Estadual Paulista “Júlio de Mesquita Filho” - Botucatu.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pelo apoio financeiro.

Sumário

RESUMO	i
SUMMARY	ii
1 Introdução	1
2 Revisão de Literatura	4
2.1 Estatística Espacial	4
2.2 Dados Espaciais	5
2.2.1 Dados em Áreas	6
2.2.2 Matriz de Vizinhaça	7
2.3 Modelo de Regressão Linear Clássico	10
2.4 Autocorrelação Espacial	11
2.4.1 Índice I Global de Moran	12
2.4.2 Índice I_i Local de Moran	15
2.5 Modelos Espaciais	17
2.5.1 Modelo de Defasagem Espacial	17
2.5.2 Modelo de Erro Espacial	21
2.5.3 Conclusões	24
2.6 Modelos Espaço-Temporais	24
2.7 Inferência Bayesiana	26
2.7.1 Teorema de Bayes	27
2.7.2 Modelagem Hierárquica	28
2.8 Modelo Espaço-Temporal: Abordagem I	30
2.9 Modelo Espaço-Temporal: Abordagem II	35
2.10 Modelo Espaço-Temporal: Abordagem III	42
3 Aplicações	48
3.1 A Dengue	48
3.2 Materiais e Métodos	50

3.3	Resultados e Discussão	53
3.3.1	Regressão Linear Clássica	59
3.3.2	Modelo de Defasagem Espacial	63
3.3.3	Modelo de Erro Espacial	66
3.4	Conclusões	72
Referências Bibliográficas		73

Lista de Figuras

1	Região $\mathfrak{R}$ particionada em A_i áreas, $i = 1, \dots, n$, sendo y_n uma observação aleatória referente à área A_n	7
2	Estruturas de vizinhança; Matriz de Contingência.	8
3	Exemplo de vizinhança.	9
4	Teste de permutação aleatória. Fonte: Domingues et al. (2016). .	13
5	Região $\mathfrak{R}$ particionada em 4 áreas. Fonte do exemplo: INPE - Curso de Análise Espacial de Dados Geográficos - Análise de Padrões de Área. Maio/2001.	14
6	Diagrama de Espalhamento de Moran, localização dos quadrantes.	15
7	Tipo de autocorrelação presente. Fonte: Domingues et al. (2016).	16
8	Vizinhos de primeira e segunda ordem, em relação a área 5, respectivamente.	19
9	Representação esquemática do modelo de defasagem espacial. Fonte: Almeida & Lopardi (2010).	20
10	Representação esquemática do modelo de erro espacial. Fonte: Vieira (2009).	22
11	Classificação de risco de ocorrências de dengue para o período da Copa do Mundo de 2014. Fonte: Lowe et al. (2014).	25
12	Ritmo de crescimento do SMR e valores estimados para o coeficiente beta, homens e mulheres, câncer de pulmão, São Paulo e sul do Brasil, 1996 a 2000. Fonte: Castro et al. (2004).	33
13	SMR brutas para homens. São Paulo e sul do Brasil de 1996 a 2000. Fonte: Castro et al. (2004).	34
14	SMR ajustadas pelo modelo, para homens. São Paulo e sul do Brasil de 1996 a 2000. Fonte: Castro et al. (2004).	35
15	(a) Belo Horizonte dividida em seus 9 distritos sanitários; (b) Belo Horizonte dividida em 117 zonas. Fonte: Assunção et al. (2001). .	36
16	Mapas de taxas brutas para leishmaniose (x100.000), por zonas em Belo Horizonte. (a) para o período de 1994/1995; (b) para o período 1995/1996; (c) para o período 1996/1997. Fonte: Assunção et al. (2001). 40	
17	Estimativas bayesianas para a taxa de ocorrência de leishmaniose (x100.000), dividido por zonas em Belo Horizonte, para o período de 1997/98. Fonte: Assunção et al. (2001).	41

18	Estimativas (a) projeções dadas por um modelo linear, (b) taxas brutas e (c) a abordagem II, do modelo quadrático. Fonte: Assunção et al. (2001).	41
19	Áreas em destaque. Fonte: Assunção et al. (2001).	42
20	Faixa etária e taxas específicos de um período de incidência (x 100 000), e o número de casos de morte (entre parênteses) por câncer de pulmão, para o sexo masculino, em Toscana (Itália), 1971-1994. Fonte: Lagazio et al. (2001).	43
21	Mortes por câncer de pulmão para o sexo masculino. Toscana (Itália), 1971-1994. (a) Coorte de nascimento; (b) Por período; riscos relativos e intervalos de confiança de 95%;. Fonte: Lagazio et al. (2001).	46
22	Distribuição espacial dos riscos relativos por coorte de nascimento. Mortes por câncer de pulmão, para o sexo masculino. Toscana (Itália), 1971-1994. Fonte: Lagazio et al. (2001).	47
23	Aedes Aegypt, principal vetor e Aedes Albopictus. Fonte: Ministério da Saúde (2016).	49
24	Aedes aegypt. Fonte: Ministério da Saúde.	49
25	Município de Rio Claro localizado no interior do estado de São Paulo – Brasil. Fonte: https://pt.wikipedia.org/wiki/Usuário(a):Raphael.lorenzeto	51
26	Rio Claro – SP, setores censitários, área urbana.	53
27	(a) Índice global de Moran e (b) Teste de permutação aleatória.	53
28	(a) Índice local de Moran, <i>Lisa cluster map</i> . (b) Níveis de significâncias para as áreas em relação a Dengue no ano de 2011 em Rio Claro–SP.	54
29	(a) Índice local de Moran, <i>Lisa cluster map</i> . (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável pop	56
30	(a) Índice local de Moran, <i>Lisa cluster map</i> . (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável alf	56
31	(a) Índice local de Moran, <i>Lisa cluster map</i> . (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável agua	57

32	(a) Índice local de Moran, <i>Lisa cluster map</i> . (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável lixo	57
33	(a) Índice local de Moran, <i>Lisa cluster map</i> . (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável renda	58
34	(a) Índice local de Moran, <i>Lisa cluster map</i> . (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável morad	59
35	(a) Índice de Moran: Resíduos. (b) Teste de permutação.	61
36	Processo para regressão espacial. Fonte: Anselin (2005).	63
37	(a) Índice de Moran para os resíduos do modelo de defasagem espacial. (b) Teste de permutação.	66
38	(a) Índice de Moran para os resíduos do modelo de erro espacial. (b) Teste de permutação.	69
39	(a) Distribuição dos casos observados de dengue em Rio Claro – SP, 2011 e (b) Distribuição dos casos de dengue previsto pelo modelo de defasagem espacial em Rio Claro – SP, 2011.	71
40	Distribuição espacial dos resíduos.	71

Lista de Tabelas

1	Variáveis socioeconômicas.	52
2	Regressão Linear Clássica,	60
3	Diagnóstico Regressão Linear Clássica.	61
4	Diagnóstico de Dependência Espacial.	62
5	Regressão: Modelo de Defasagem Espacial.	64
6	Regressão: Modelo de Defasagem Espacial, novo ajuste.	65
7	Modelo de defasagem espacial.	65
8	Diagnóstico Modelo de Defasagem Espacial.	66
9	Regressão: Modelo de Erro Espacial.	67
10	Regressão: Modelo de Erro Espacial, novo ajuste.	68
11	Modelo de erro espacial.	68
12	Diagnóstico Modelo de Erro Espacial.	69
13	Critérios de comparação.	69

TÉCNICAS DE PROCESSOS ESPACIAIS E ESPAÇO-TEMPORAIS COM APLICAÇÕES EM DADOS DE DENGUE

Autora: JACQUELINE DOMINGUES

Orientador: Prof. Dr. JOSÉ SILVIO GOVONE

Co-orientadora: Profa. Dra. VÂNIA AP. CECCATO

RESUMO

Recentemente, métodos estatísticos para estudar variáveis distribuídas no espaço e/ou no tempo vêm sendo desenvolvidos, isto se deve aos avanços tecnológicos que permitem a construção de modelos de estrutura complexa, com intuito de descrever melhor a realidade. O presente trabalho abordou o caso dos modelos bayesianos hierárquicos espaço-temporais. Além dos modelos espaço-temporais, dois modelos espaciais, que incorporam a autocorrelação espacial presente nos dados, foram definidos, sendo eles o modelo de defasagem e de erro espacial. Através deles foi possível analisar a distribuição da dengue e relacioná-la com variáveis socioeconômicas para o ano de 2011 em Rio Claro – SP. Diferentes abordagens espaço-temporais, baseadas na literatura recente para dados de área, foram analisadas e compõem as três técnicas apresentadas. Estes modelos tem por objetivo o mapeamento de doenças, como conhecer o seu comportamento ao longo de anos, identificar regiões de maior risco e capturar a estrutura espacial e temporal. Cada técnica apresenta uma particularidade como variação do tempo de forma linear, quadrática ou por coortes de nascimento.

Palavras-chave: *Dengue; epidemiologia; estatística espacial; modelos espaciais; modelos espaço-temporais.*

TECHNIQUES OF SPATIAL AND SPATIO-TEMPORAL PROCESSES WITH APPLICATIONS IN DENGUE DATA

Author: JACQUELINE DOMINGUES

Adviser: Prof. Dr. JOSÉ SILVIO GOVONE

Co-Adviser: Profa. Dra. VÂNIA AP. CECCATO

SUMMARY

Recently, statistical methods to study distributed variable in space and/or on time has been developed, this is due to technological advances that allow the construction of model with complex structure, to better describe the reality, as the case of Bayesian hierarchical model. In addition to the spatio-temporal models, two spatial models incorporating spatial autocorrelation present in the data set were defined, being them spatial lag model and spatial error model, through them was possible to analyze the dengue distribution and relate it with socioeconomic variables for 2011 in Rio Claro - SP. Different approaches spatio-temporal, based on recent literature for areal data were analyzed and form the three techniques presented. These models aims at diseases mapping such as knowing their behavior over the years, identify areas with high risk and capture the spatial and temporal structure. Each technique presents a particularity, for example the time variation that can be analysed by different ways as linear, quadratic or by birth cohorts.

Keywords: *Dengue; epidemiology; spatial statistics; spatio models; spatio-temporal models.*

1 Introdução

Muito se tem discutido, recentemente, sobre o grande desenvolvimento em relação aos métodos estatísticos para estudar variáveis distribuídas no espaço e/ou no tempo, com o objetivo de se fazer previsões espaciais, espaço-temporais, estudar correlações, independência, entre outros. Os métodos estatísticos permitem, baseados em dados amostrais, fazer previsões das variáveis em estudo, em pontos da área de distribuição da variável ou do tempo.

O início deu-se nas geociências há cerca de 50 anos, porém hoje, o campo de aplicações é muito grande, pois são diversas as variáveis desta natureza, em diferentes áreas do conhecimento, como: epidemiologia, climatologia, economia, sociologia, agronomia, ciências ambientais e geociências.

Recentemente, ocorreu certa fusão da estatística espacial com os métodos autoregressivos temporais clássicos, e o estudo passou a ser espaço-temporal; como interessantes aplicações pode-se citar Lowe et al. (2011), Lowe et al. (2014), que analisaram a evolução temporal da distribuição espacial de dengue em regiões do Brasil, considerando dados dos últimos anos. Tais trabalhos, ainda em continuidade, são realizados através de uma parceria internacional da Fundação Oswaldo Cruz.

Modelos espaço-temporais para mapeamento de doenças, utilizando dados de área, se tornaram muito populares em epidemiologia. Frequentemente, o atributo de interesse em estudos deste tipo é o número de casos de uma determinada doença, e o objetivo dos modelos é estimar o risco relativo da doença em cada área e período de tempo.

Os métodos de análise espacial na saúde coletiva vêm sendo usados principalmente em estudos ecológicos, na detecção de aglomerados espaciais ou espaço-temporais, na avaliação e monitoramento ambiental, e aplicados ao planejamento e avaliação de uso de serviços de saúde (Carvalho & Souza-Santos 2005).

Analisar as regiões segundo o uso ou acesso, dos serviços de saúde, permite comparar as diferentes estratégias de obtenção de assistência, conforme a complexidade do procedimento e as alternativas existentes, bem como a área de cobertura dos serviços e a responsabilidade destes sobre a qualidade do serviço prestado (Carvalho & Souza-Santos 2005).

No entanto, não é apenas na saúde que modelos espaço-temporais são aplicados, existem diversas outras áreas, como em Ceccato (2005), que analisou a relação da criminalidade com espaço e tempo para a cidade de São Paulo – SP; pouco se sabe sobre as variações espaço-temporais de eventos criminais e, em particular para países em desenvolvimento como é o caso. Logo, estudos como este contribuem para o desenvolvimento de novas técnicas e aperfeiçoamento das existentes.

O estudo relacionado a doenças tem se tornado uma ferramenta importante para as autoridades. Segundo Kuhn et al. (2005), a identificação precoce de uma epidemia de doenças infecciosas é um importante primeiro passo para a implementação de intervenções eficazes para controlar a doença e reduzir a mortalidade e morbidade na população.

A dengue, por exemplo, é um dos principais problemas de saúde pública no mundo. A Organização Mundial da Saúde (OMS) estima que 2,5 bilhões de pessoas, ou seja, 2/5 da população mundial estão sob risco de contrair dengue. No Brasil, o quadro epidemiológico atual caracteriza-se pela grande distribuição do *Aedes aegypti* em todas as regiões e uma complexa dinâmica de dispersão do vírus, com circulação simultânea dos quatro sorotipos DENV 1, DENV 2, DENV 3 e DENV 4, de acordo com COSEMS/SP (2012).

Em consequência disso, a dengue é classificada como uma doença séria que pode levar a morte, necessitando assim de atenção especial. Estudos relacionados a dengue, que procurem descrever seu comportamento, relação com variáveis sociais, econômicas, ou características climáticas e espaciais particulares; que busquem identificar regiões com alta incidência da doença, por exemplo, podem se tornar uma ferramenta para os planejadores de saúde, com intuito de combatê-la e controlá-la, procurando evitar epidemias/surtos da doença.

Os objetivos do presente trabalho são o estudo de técnicas espaciais e espaço temporais de análise de dados e a aplicação das mesmas a dados de ocorrências de dengue na área urbana de Rio Claro, SP.

Através das aplicações, de modelos espaciais, buscou-se conhecer a distribuição espacial da dengue no município de Rio Claro–SP, de tal modo que seja

possível, através delas, identificar *clusters*, padrões, relações, ou seja, estudar como a dengue se distribui nas áreas. Os resultados obtidos buscaram identificar regiões de maior incidência e com quais variáveis está relacionada.

2 Revisão de Literatura

2.1 Estatística Espacial

A estatística espacial é um ramo da estatística que permite analisar a localização espacial de eventos de interesse, ou seja, fenômenos que estão distribuídos no espaço. Através da estatística espacial é possível descrever ou explicar eventos que ocorrem no espaço, como seu comportamento, relação com outros fenômenos, fazer inferências sobre estes comportamentos e relações, identificar padrões, além da percepção visual da distribuição espacial do problema de interesse (Bailey & Gattrel, 1995). De acordo com Haining (2003), a análise espacial representa um conjunto de técnicas e modelos que usam explicitamente a localização espacial.

É de grande utilidade traduzir estes padrões encontrados, com considerações objetivas e mensuráveis como nos seguintes casos, de acordo com Câmara et al. (2004).

- i. Epidemiologistas coletam dados sobre ocorrência de doenças. A distribuição dos casos de uma doença forma um padrão no espaço? Existe associação com alguma fonte de poluição? Evidência de contágio? Variação no tempo?
- ii. Deseja-se investigar se existe alguma concentração espacial na distribuição de roubos. Roubos ocorrem em determinadas áreas, estão correlacionados com características socioeconômicas dessas áreas?
- iii. Geólogos desejam estimar a extensão de um depósito mineral em uma região, com base em amostras. Pode-se usar amostras para estimar a distribuição do mineral na região?

Na área da saúde, o estudo espacial pode modelar a realidade presente, afirma Santos & Souza (2007); modelos são simplificações da realidade usa-

dos para entender um sistema, estudar seu funcionamento, buscar causas de fenômenos, avaliar intervenções podendo assim prever desfechos.

Estudos que envolvem doenças, onde geralmente a variável a ser modelada é a taxa de mortalidade ou risco, estão se desenvolvendo atualmente na área espacial. Werneck & Struchiner (1997) já afirmavam a importância de relacionar a distribuição de doenças com espaço e tempo, sendo estes um dos temas mais explorados em epidemiologia atualmente, indicando que a localização espacial pode ser um fator relevante na explicação da ocorrência de determinados fenômenos.

Investigações de efeitos denominados de vizinhança apontam para os “riscos associados com a estrutura social e ecológica de vizinhança, possibilitando intervenções inovadoras no nível da comunidade”. Padrões de mortalidade ou morbidade, propagação de epidemias, transmissão sexual de doenças ou a transferência de comportamentos ou valores não podem ser explicados sem uma abordagem que além de considerar os grupos, estude o espaço e o tempo, afirmam Carvalho & Souza-Santos (2005).

Considerando que a estatística espacial é um ramo da estatística que estuda métodos científicos para a coleta, descrição e análise de dados que possuam coordenadas geográficas, além da visualização, os dados espaciais podem ser classificados em diferentes categorias.

2.2 Dados Espaciais

Os dados espaciais podem ser classificados em quatro categorias, de acordo com Assunção (2001), e essa categorização diz respeito à natureza estocástica da observação, sendo elas:

1. Dados de processos pontuais;
2. Dados de superfícies aleatórias;
3. Dados de área;
4. Dados de interação espacial.

Em cada uma das categorias citadas, para a classificação de dados espaciais, existem técnicas estatísticas diferentes para descrever e analisar os dados.

Existem também, alguns efeitos espaciais que podem ser observados na região em estudo; é possível encontrar situações de heterogeneidade espacial e autocorrelação ou dependência espacial, segundo Serrano & Valcarce (2000). Estes efeitos podem ser descritos por:

- i. **Heterogeneidade espacial:** ocorre quando os dados utilizados para explicar um mesmo fenômeno são de unidades espaciais muito distintas.
- ii. **Autocorrelação ou dependência espacial:** ocorre em consequência da existência de uma relação entre o que acontece em uma determinada região do espaço e o que acontece em outra região. Este fenômeno pode ser entendido como uma situação em que observações próximas no espaço possuem valores similares. Neste caso, o desafio seria medir este grau de associação espacial entre as observações de uma ou mais variáveis.

No presente trabalho, será considerado para o estudo, dados em áreas.

2.2.1 Dados em Áreas

A análise espacial para dados deste tipo configura-se em métodos de análise cuja localização está associada a áreas delimitadas por polígonos, por exemplo, eventos de interesse em municípios, bairros, setores censitários, onde não se dispõe da localização exata do evento e sim de valores para cada área em que ocorreram.

Considere uma região $\mathfrak{R}$ particionado em n áreas A_i , $i = 1, \dots, n$, onde $\mathfrak{R} = \bigcup_{i=1}^n A_i$ e $\bigcap_{i=1}^n A_i = \phi$. Em cada área A_i é feita uma observação aleatória Y_i . Com isso, defina $\{Y(A_i) : A_i \in A_1, A_2, \dots, A_n\}$ um conjunto de variáveis aleatórias indexadas por sub-regiões fixas $A_1, \dots, A_n$ na região $\mathfrak{R}$. Em Bailey & Gatrell (1995), generaliza-se as variáveis aleatórias $Y(A_i)$ como Y_i e denota como valores observados y_i . A figuras 1 exemplifica dados em áreas.

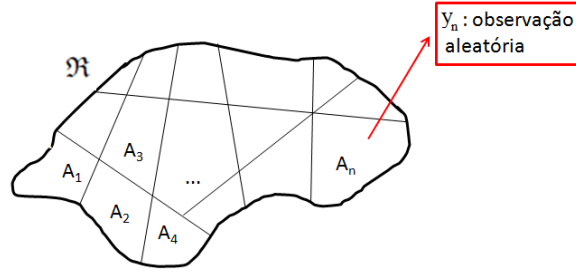


Figura 1: Região $\mathfrak{R}$ particionada em A_i áreas, $i = 1, \dots, n$, sendo y_n uma observação aleatória referente à área A_n .

Com o intuito de definir o grau de associação presente entre as áreas A_i e A_j quaisquer em $\mathfrak{R}$, um conceito importante é acrescentado, que é a matriz de vizinhança ou como também é conhecida, matriz de pesos espaciais, sendo ela responsável por representar o peso ou proximidade espacial entre as áreas A_i e A_j .

2.2.2 Matriz de Vizinhança

A matriz de pesos espaciais, ou matriz de adjacências, busca expressar a estrutura espacial dos dados. Dado um conjunto de n áreas $\{A_1, \dots, A_n\}$, constrói-se uma matriz W de ordem $n \times n$, onde cada um dos elementos w_{ij} representa uma medida de proximidade/distância entre A_i e A_j . Ao associar a matriz de vizinhança com dependência espacial, pressupõe-se que o grau de dependência seja negativamente relacionado com a distância (Câmara et. al, 2004). Em outras palavras, assume-se que a intensidade da dependência espacial diminui à medida que a distância entre as unidades aumenta. Define-se vizinhança de duas maneiras, sendo elas:

1. **Contiguidade:** medidas de contiguidade necessitam de informações sobre o tamanho e forma das unidades regionais; simplificando, pode-se associar contiguidade com regiões que possuem fronteira em comum, sendo que essas fronteiras podem ser definidas de três maneiras distintas. A figura 2 exhibe as três possíveis estruturas de vizinhança.

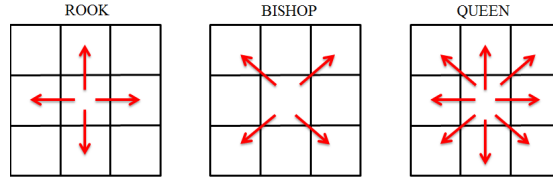


Figura 2: Estruturas de vizinhança; Matriz de Contingência.

- (i) **ROOK/TORRE**: considera como vizinhas as unidades que possuem fronteiras comuns, em que a unidade vizinha é definida da forma $w_{ij} = 1$, caso contrário $w_{ij} = 0$.
 - (ii) **BISHOP/BISPO**: considera como vizinhas as unidades que possuem vértices comuns, em que a unidade vizinha é definida da forma $w_{ij} = 1$, caso contrário $w_{ij} = 0$.
 - (iii) **QUEEN/RAINHA**: considera como vizinhas as unidades que possuem fronteiras ou vértices comuns, em que a unidade vizinha é definida da forma $w_{ij} = 1$, caso contrário $w_{ij} = 0$.
2. **Distância**: neste tipo de matriz de vizinhança considera-se um raio de distância ou número de vizinhos mais próximos, os k primeiros com $k \in \mathbb{N}$. Os principais critérios definidos são:
- (i) $w_{ij} = 1$, se o centróide de A_i está a uma determinada distância δ de A_j ; caso contrário $w_{ij} = 0$. Um problema neste tipo de matriz pode ser o surgimento de várias ilhas.
 - (ii) Para k vizinhos mais próximos, $k \in \mathbb{N}$, $w_{ij}(k) = 0$ se $i = j$; $w_{ij}(k) = 1$ se $d_{ij}(k) \leq \delta_i(k)$, e $w_{ij}(k) = 0$ se $d_{ij}(k) > \delta_i(k)$, onde d_{ij} representa a distância, medida entre os centróides das áreas A_i e A_j ; $\delta_i(k)$ é um valor crítico definido previamente de forma que seja um critério para considerar as áreas como vizinhas à área A_i .

Uma possível dúvida seria qual das matrizes é mais apropriada para determinado problema. Sugere-se analisar a natureza do problema a ser modelado para escolher o critério adequado.

Exemplo: Matriz de Contingência.

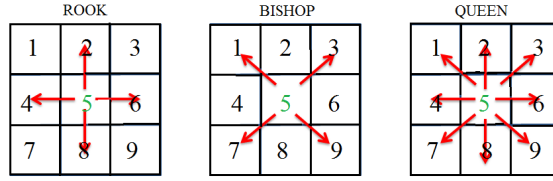


Figura 3: Exemplo de vizinhança.

A figura 3 apresenta um esquema de vizinhança para os três critérios apresentados. Através do critério ROOK, que considera vizinho apenas as regiões que compartilham de fronteira comum, note que a área 5 possui como vizinhos as regiões 2,4,6 e 8. O critério BISHOP considera vizinho apenas as regiões que possuem vértice comum, neste caso a área 5 possui como vizinhos as regiões 1,3,7 e 9 e por fim pelo critério QUEEN, a área 5 considera seus vizinhos regiões que possuem vértice ou fronteira em comum, sendo eles 1,2,3,4,6,7,8 e 9. A matriz de contingência utilizando o critério Rook neste caso, é dada por:

$$W = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

A matriz W pode ainda ser normalizada, desta forma a soma de suas linhas deve ser igual a 1, $\sum_{j=1}^9 w_{ij} = 1$.

$$W = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \end{bmatrix}$$

Esta padronização garante que $0 \leq w_{ij} \leq 1$, pode acontecer que a matriz W padronizada não seja simétrica, mesmo W sendo.

No entanto, para que o estudo espacial seja considerado, convém analisar em quais situações a modelagem espacial deve ser considerada ao invés da clássica. Sendo assim, apresenta-se as suposições do modelo de regressão linear clássico.

2.3 Modelo de Regressão Linear Clássico

A análise de regressão é uma ferramenta estatística que utiliza a relação entre duas ou mais variáveis de modo que a variável resposta Y , dependente, possa ser explicada por uma ou mais variáveis explicativas X , independentes.

O modelo de regressão linear clássico permite estimar a relação existente entre uma variável dependente e um conjunto de variáveis explicativas e é definido segundo Waller & Gotway (2004) por:

$$Y = X\beta + \varepsilon. \quad (1)$$

satisfazendo hipóteses, sendo elas $\varepsilon \sim N(0, \sigma^2)$, e $Cov(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j$, onde,

- ▷ $Y_{n \times 1}$: vetor variável resposta;
- ▷ $X_{n \times k}$: matriz de variáveis explicativas;
- ▷ $\beta_{k \times 1}$: vetor de parâmetros;
- ▷ $\varepsilon_{n \times 1}$: vetor de erros.

Contudo, o modelo de regressão linear clássico não incorpora em sua estrutura componentes espaciais. Em consequência disso, se as observações representarem pontos no espaço, as suposições afirmam que os valores observados na área A_i são independentes das observações feitas na área A_j , implicando $E(\varepsilon_i \varepsilon_j) = E(\varepsilon_i)E(\varepsilon_j) = 0$, sendo que, em contextos espaciais, essas suposições são fortes.

De acordo com Lesage & Pace (2009), em contraste com o modelo linear clássico, a existência de dependência espacial afirma que os valores observados na área A_i dependem dos valores nas áreas vizinhas, ou seja, dependem de observações em áreas próximas.

Portanto, se as suposições do modelo linear clássico não forem satisfeitas, como alternativa apresenta-se os modelos de defasagem espacial e de erro espacial, sendo eles denominados *Spatial Autoregressive Models* - (SAR) e *Conditional Autoregressive Models* - (CAR) respectivamente. Estes modelos incorporam a dependência existente nos dados. Porém, antes das definições formais dos modelos citados, é necessário conhecer a existência de uma estrutura espacial.

2.4 Autocorrelação Espacial

A estrutura de dependência entre os valores observados nas várias áreas do fenômeno em estudo é analisada pela função de autocorrelação espacial. A autocorrelação mede a correlação da variável consigo mesma no espaço. A correlação de uma variável com ela mesma, medida no mesmo local, assume valor igual a 1 (um). Entretanto, a correlação de uma variável com ela mesma cujos dados são amostrados em áreas vizinhas, assume um valor entre -1 e $+1$. Quanto mais próximo de 1 (um), maior a semelhança entre vizinhos, o valor 0 (zero) indica inexistência de correlação, e valores negativos indicam dessemelhança, segundo Santos & Souza (2007).

A função de correlação usual compara valores de duas variáveis. A função de autocorrelação consiste na correlação do valor do indicador comparado aos valores do mesmo indicador em áreas consideradas vizinhas.

De acordo com Câmara et al. (2004), um aspecto fundamental da análise exploratória espacial é a caracterização da dependência espacial, mostrando como os valores estão correlacionados no espaço. Neste caso, as ferramentas utilizadas para indicar a presença ou não de autocorrelação espacial são os índices global e

local de Moran.

O indicador global de autocorrelação espacial, fornece um único valor como medida da associação espacial para todo o conjunto de dados, o que é útil na caracterização da região de estudo como um todo. O índice local é mais apropriado, quando se dispõe de grande número de áreas, resultantes por exemplo de escalas espaciais detalhadas, sendo a natureza dos processos envolvidos com provável existência de diferentes regimes de correlação espacial em diferentes sub-regiões. Para evidenciar estes regimes espaciais, pode-se utilizar os indicadores locais de autocorrelação espacial e o gráfico de dispersão/espalhamento de Moran (Câmara et. al, 2004). Todas estas estatísticas dependem da definição e estrutura de vizinhança adotada.

Em outras palavras, a autocorrelação espacial busca analisar quanto o valor observado de determinada variável de interesse em uma área está associada com o valor observado desta mesma variável em áreas consideradas vizinhas. Através da primeira lei da geografia, descrita por Tobler (1979), afirma que *“Tudo está relacionado com todo o resto, mas coisas próximas estão mais relacionadas do que coisas distantes”*, expressa bem a ideia de autocorrelação.

2.4.1 Índice I Global de Moran

O índice de especificação mais comumente utilizado para autocorrelação espacial é derivado de uma estatística desenvolvida por Moran (1948). De uma forma geral, o índice de Moran serve como um pré-teste, onde a hipótese nula é de independência espacial, Ward & Gleditsch (2008). O índice global de Moran definido em Bailey & Gratell (1995) é dado por:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{(\sum_{i=1}^n (y_i - \bar{y})^2 (\sum_{i \neq j} w_{ij}))}. \quad (2)$$

- ▷ y_i : representa o valor da variável Y na área A_i ;
- ▷ $\bar{y}$: média da variável Y ;
- ▷ w_{ij} : refere-se ao elemento na linha i e coluna j da matriz de vizinhança $W_{n \times n}$ e n o número de observações.

O índice I varia entre -1 e 1 onde,

$$\begin{cases} I = 0, & \text{indica ausência de correlação espacial;} \\ I \rightarrow 1, & \text{indica dependência forte e positiva;} \\ I \rightarrow -1, & \text{indica dependência forte e negativa.} \end{cases}$$

Uma vez calculado, é importante estabelecer sua validade estatística. Uma possibilidade para testar a significância do resultado encontrado é um teste de pseudo-significância. Neste caso são geradas diferentes permutações dos valores observados associados as áreas e recalculam-se as estatísticas. Esse processo é repetido por n vezes, n suficientemente grande, gerando assim uma distribuição de referência. Se o valor do índice I calculado originalmente corresponder a um extremo da distribuição simulada através das permutações, então o valor de I possui significância estatística (Anselin, 2005).

A figura 4 apresenta um teste de permutação aleatória para a estatística I de Moran, exibindo o número de permutações e o nível de pseudo-significância no canto superior esquerdo, bem como o valor da estatística I calculado (0,4365), sua média teórica ($E[I] = -0,0051$), e a média e o desvio padrão da distribuição empírica, sendo a distribuição de referência (roxo) e a estatística I (barra amarela), (Anselin, 2005). Neste caso, como o índice I calculado corresponde a um extremo da distribuição simulada, o valor I possui significância estatística.

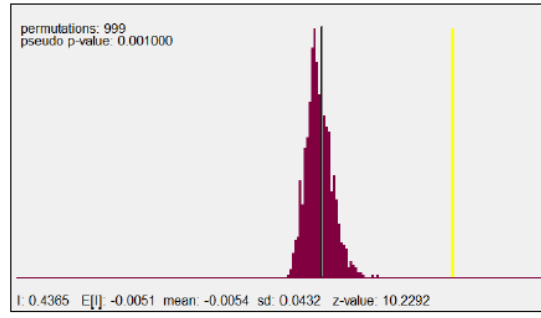


Figura 4: Teste de permutação aleatória. Fonte: Domingues et al. (2016).

Exemplo - Índice I global de Moran

Considere uma região $\mathfrak{R}$ dividida em quatro áreas $\{A_1, A_2, A_3, A_4\}$, tais que $\mathfrak{R} = \bigcup_{i=1}^4 A_i$ e $\bigcap_{i=1}^4 A_i = \emptyset$ se $i \neq j$, conforme a figura 5.

Em cada umas das áreas A_i observou-se um valor para a variável, referente ao evento de interesse.

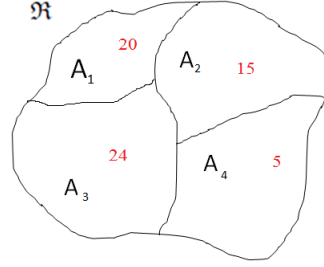


Figura 5: Região $\mathfrak{R}$ particionada em 4 áreas. Fonte do exemplo: INPE - Curso de Análise Espacial de Dados Geográficos - Análise de Padrões de Área. Maio/2001.

A matriz de vizinhança, calculada através do critério Queen, é dada por:

$$W = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$\text{padronizando, tem-se } W = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

$$\text{A média } \bar{y} = \frac{20 + 15 + 24 + 5}{4} = 16.$$

$$\text{A variância } \sigma^2 = \frac{\sum_{i=1}^4 (y_i - \bar{y})^2}{4} = 50,5.$$

$$\text{O desvio-padrão } \sigma = \sqrt{\sigma^2} = \sqrt{50,5} = 7,11.$$

Com o intuito de facilitar os cálculos a equação (2) pode ser simplificada normalizando os dados.

$$z_i = \frac{(y_i - \bar{y})}{\sigma},$$

com isso $z_{A_1} = 0,563$, $z_{A_2} = -0,141$, $z_{A_3} = 1,126$ e $z_{A_4} = -1,548$. O índice I passa a ser dado por:

$$I = \frac{\sum_{i=1}^4 \sum_{j=1}^4 w_{ij} z_i z_j}{\sum_{i=1}^4 z_i^2} = \frac{-0,914}{4} = -0,228.$$

Para estimar a significância de I , como descrito anteriormente, se o índice I encontrado corresponder a um extremo da distribuição simulada, então trata-se de evento com significância estatística, podendo concluir se existe ou não autocorrelação espacial entre as áreas para a variável de interesse.

2.4.2 Índice I_i Local de Moran

Com o intuito de examinar padrões em maior detalhe, o índice local de Moran produz um valor específico para cada área, permitindo assim a identificação de *clusters* espaciais. O índice local de Moran referente a área A_i em particular pode ser expresso por I_i :

$$I_i = \frac{(y_i - \bar{y}) \sum_{j=1}^n w_{ij} (y_j - \bar{y})}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}. \quad (3)$$

Câmara et al. (2004) afirmam que essa técnica permite visualizar graficamente, através do diagrama de espalhamento de Moran (*Moran Scatter Plot*), o grau de similaridade entre vizinhos. Coloca-se no eixo x o valor do indicador da área A_i normalizado e no eixo y o valor da estatística de Moran somente entre os vizinhos de A_i (média dos vizinhos). Cada quadrante corresponde a um tipo diferente de correlação espacial; a figura 6 representa o diagrama de espalhamento de Moran.

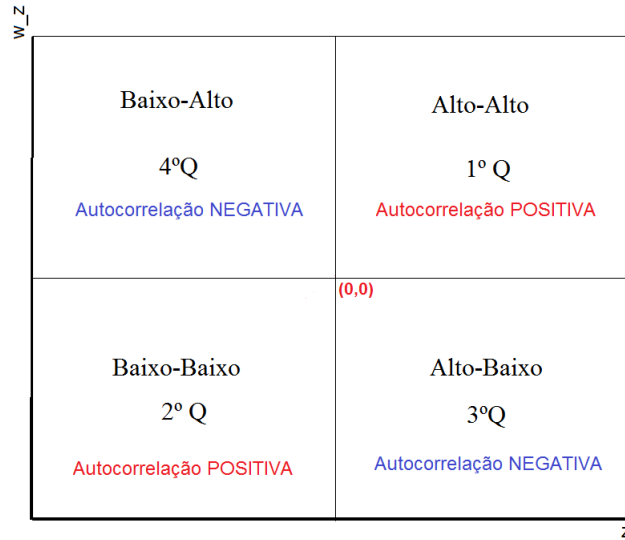


Figura 6: Diagrama de Espalhamento de Moran, localização dos quadrantes.

Os quatro quadrantes são definidos por: Q1 alto-alto (*high-high*) e Q2 baixo-baixo (*low-low*) para autocorrelação espacial positiva, no sentido que uma localização possui vizinhos com valores semelhantes, ou seja, os pontos localizados nestes quadrantes, indicam áreas onde o valor medido do atributo se assemelha a média dos vizinhos; Q4 baixo-alto (*low-high*) e Q3 alto-baixo (*high-low*) para autocorrelação espacial negativa, no sentido que uma localização possui vizinhos com valores distintos, ou seja, as áreas localizadas nestes quadrantes podem ser vistas como extremos, uma vez que não seguem o padrão observado em seus vizinhos.

De acordo com Anselin (2005), as áreas classificadas nos quadrantes alto-alto e baixo-baixo, com autocorrelação espacial local positiva indicam aglomerados espaciais, enquanto o alto-baixo e baixo-alto locais, autocorrelação espacial local negativa, são de valores atípicos espaciais, *outliers*.

Um exemplo, é a região de Rio Claro – SP, dividida por setores censitários, para casos notificados de dengue no ano de 2011. Neste estudo foi detectada presença de autocorrelação espacial e através do índice local de Moran, detectou-se autocorrelação espacial positiva tanto alto-alto quanto baixo-baixo, como mostra a figura 7, Domingues et al. (2016).

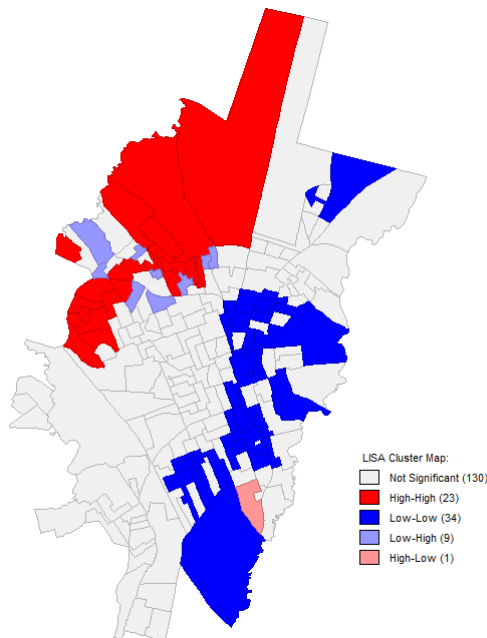


Figura 7: Tipo de autocorrelação presente. Fonte: Domingues et al. (2016).

Estes indicadores fazem parte da análise exploratória de dados espaciais (AEDE), sendo esta uma metodologia para estudar padrões e associações

presentes nos dados. A AEDE é formada por um conjunto de técnicas que permitem descrever distribuições espaciais, identificar localizações atípicas (*outliers*), aglomerados, *clusters*, entre outros comportamentos presentes ou não nos dados (Câmara et. al, 2004).

Contudo, se for constatada a presença de autocorrelação espacial nos dados e houver interesse fazer uma análise de regressão é conveniente considerar modelos espaciais que incorporem em sua estrutura componentes espaciais.

2.5 Modelos Espaciais

De acordo com Câmara et al. (2004), a inclusão explícita de efeitos espaciais em modelos de regressão pode ser feita de diferentes formas. A classe de modelos de regressão espacial mais simples, chamados de modelos com efeitos espaciais globais, supõe que é possível capturar a estrutura de correlação espacial num único parâmetro, que é adicionado ao modelo de regressão tradicional. Um estudo interessante envolvendo dengue e variáveis socioeconômicas pode ser visto em Almeida et al. (2009).

2.5.1 Modelo de Defasagem Espacial

O modelo com defasagens espaciais¹ atribui a autocorrelação espacial à variável resposta Y . Seu uso é apropriado quando os valores observados de Y em uma área A_i são diretamente influenciados pelos valores de Y observados em seus vizinhos.

Se Y não influenciado diretamente pelo seus vizinhos, e existir algum agrupamento espacial que influencie o seu valor para A_i e seus vizinhos e esta característica é omitida ou não observável, convém considerar um modelo alternativo com erros espacialmente correlacionados, que será discutido posteriormente, Ward & Gleditsch (2008).

O modelo de defasagem espacial é definido de acordo com Anselin (2003) por:

¹O efeito de defasagem espacial refere-se a interação espacial entre a variável dependente Y , onde passa a existir um efeito de vizinhança, influenciando diretamente a variável em questão.

$$Y = \rho WY + X\beta + \varepsilon. \quad (4)$$

Assim,

$$\begin{aligned} Y - \rho WY &= X\beta + \varepsilon \Leftrightarrow (I - \rho W)Y = X\beta + \varepsilon \Leftrightarrow \\ Y &= (I - \rho W)^{-1}X\beta + (I - \rho W)^{-1}\varepsilon. \end{aligned}$$

Usualmente representado por;

$$(I - \rho W)Y = X\beta + \varepsilon. \quad (5)$$

- ▷ $Y_{n \times 1}$: variável dependente;
- ▷ $X_{n \times k}$: variáveis independentes;
- ▷ $\beta_{k \times 1}$: coeficientes de regressão;
- ▷ $\varepsilon_{n \times 1}$: erros aleatórios com média zero e variância $\sigma^2 I$;
- ▷ $W_{n \times n}$: matriz de vizinhança espacial ou matriz de ponderação espacial;
- ▷ ρ : coeficiente espacial autoregressivo;

com $E(Y_i Y_j) \neq 0$, que implica que a observação y na área A_i está correlacionada com a observação y na área A_j e WY indica a dependência espacial em Y .

Se o parâmetro ρ for significativo, então parte da variação total de Y é explicada pela dependência de cada observação dos seus vizinhos. O parâmetro ρ mensura o efeito médio da variável dependente Y , em relação a sua vizinhança espacial. Condições são impostas sobre a matriz $(I - \rho W)$, pois ela necessita ser invertível e positiva definida.

Em Lesage (2008), o modelo em questão pode ser interpretado como indicador do valor esperado de cada observação y_i que dependerá do valor médio de $X\beta$ adicionado a uma combinação linear dos valores resultantes das observações vizinhas ponderadas pelo parâmetro ρ .

Para expressar a simultaneidade do processo autorregressivo espacial, afim de explorar ainda mais a natureza deste, a expressão abaixo exhibe $(I - \rho W)^{-1}$ como uma série infinita, dada por

$$(I - \rho W)^{-1} = I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots + \dots \quad (6)$$

com isso a equação (4) pode ser expressa por:

$$Y = X\beta + \rho W X\beta + \rho^2 W^2 X\beta + \dots + \dots + \varepsilon + \rho W \varepsilon + \rho^2 W^2 \varepsilon + \dots + \dots \quad (7)$$

onde W^k representa a matriz de vizinhança W de ordem k , sendo vizinhos de primeira ordem definidos em W , segunda ordem em W^2 , terceira ordem em W^3 e assim por diante.

Utilizando o critério de vizinhança Rook, por exemplo, os vizinhos de segunda ordem (W^2), considera como vizinhança os vizinhos de seus vizinhos, ou seja:

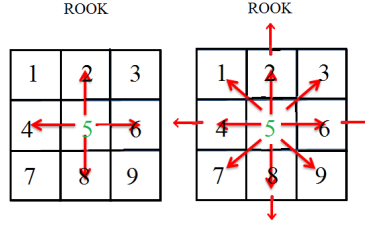


Figura 8: Vizinhos de primeira e segunda ordem, em relação a área 5, respectivamente.

A área 5, tem como vizinhos de segunda ordem as áreas 1, 3, 7, 9 além dos vizinhos do 2, 4, 6 e 8 (figura 8). E assim para as demais ordens de vizinhanças.

Voltando a matriz WY , ela indica a dependência espacial em Y , ou seja, $WY = (\overline{y}_1, \overline{y}_2, \dots, \overline{y}_n)^T$ onde:

$$\begin{aligned}\overline{y}_1 &= w_{11}y_1 + w_{12}y_2 + \dots + w_{1n}y_n \\ \overline{y}_2 &= w_{21}y_1 + w_{22}y_2 + \dots + w_{2n}y_n \\ \overline{y}_3 &= w_{31}y_1 + w_{32}y_2 + \dots + w_{3n}y_n \\ &\dots \\ \overline{y}_n &= w_{n1}y_1 + w_{n2}y_2 + \dots + w_{nn}y_n\end{aligned}$$

Logo $\overline{y}_i$ é uma média ponderada das observações nas áreas vizinhas à A_i , o peso dado a área A_j é w_{ij} ; se A_i e A_j não forem vizinhos, o peso é zero.

Exemplo: O preço de uma casa em determinada região depende do preço das casas vizinhas. O preço não leva em consideração somente as características estruturais e locacionais do imóvel, mas também são influenciados pelos preços dos imóveis vizinhos. A introdução do termo de defasagem espacial WY , como variável explicativa, serve como representante para as variáveis independentes omitidas ou não observadas que estão correlacionadas com as características locacionais.

A figura 9, apresenta um esquema do modelo espacial com defasagem, representando a interação espacial da variável Y em relação as outras variáveis, nota-se que a variável dependente Y interage com a mesma variável em outra área.

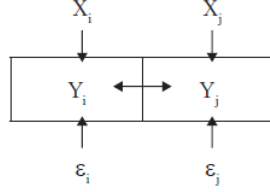


Figura 9: Representação esquemática do modelo de defasagem espacial. Fonte: Almeida & Lopardi (2010).

De acordo com Ywata & Albuquerque (2011), a estimação dos parâmetros no modelo com defasagem espacial, via mínimos quadrados ordinários, produz estimativas inconsistentes, uma vez que o vetor de lags espaciais WY é correlacionado com o vetor de resíduos ε . Pode-se utilizar máxima verossimilhança, a partir da hipótese de que o vetor de resíduos ε possui distribuição normal multivariada com vetor de média nula e matriz de covariância $\sigma^2 I$. Neste caso, Y tem distribuição normal multivariada, visto que é combinação linear de variáveis aleatórias normais. A função de logaritmo da verossimilhança é dada por:

$$\ln(L(\beta, \sigma, \rho|Y, X)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln(I - \rho W) - \frac{1}{2\sigma^2}\varepsilon^T\varepsilon = \quad (8)$$

$$-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln(I - \rho W) - \frac{1}{2\sigma^2}(Y - \rho WY - X\beta)^T(Y - \rho WY - X\beta).$$

n representa o número de dados da amostra, $\ln$ o logaritmo natural, σ^2 a variância do modelo e as demais variáveis têm a mesma representação da equação definida anteriormente.

2.5.2 Modelo de Erro Espacial

Este tipo de modelo de regressão espacial, com parâmetros globais, considera que os efeitos espaciais são um ruído, ou perturbação, ou seja, fator que precisa ser removido (Câmara et. al, 2004). Neste caso, os efeitos da autocorrelação espacial são associados ao termo de erro ε , definido por Bailey & Grattel (1995), sendo o modelo dado por:

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ \varepsilon &= \rho W\varepsilon + U. \\ Y &= X\beta + (I - \rho W)^{-1}U. \end{aligned} \tag{9}$$

- ▷ $W\varepsilon$: erro com efeito espacial;
- ▷ U : erros aleatórios com média zero e variância $\sigma^2 I$;
- ▷ ρ : coeficiente autoregressivo.
- ▷ $Y_{n \times 1}$: variável dependente;
- ▷ $X_{n \times k}$: variáveis independentes;
- ▷ $\beta_{k \times 1}$: coeficientes de regressão;
- ▷ $W_{n \times n}$: matriz de vizinhança espacial ou matriz de ponderação espacial;

com $E(U) = 0$, $E(UU^T) = \sigma^2 I$, $E(\varepsilon_i \varepsilon_j) \neq 0$, isso implica que o erro atribuído à área A_i esta correlacionado com o erro na área A_j .

Este modelo é também chamado de “Condicional Autorregressivo” - CAR. A hipótese nula para a não existência de autocorrelação é que $\rho = 0$, ou seja, o termo de erro não é espacialmente correlacionado.

Neste modelo, onde a $E(\varepsilon_i \varepsilon_j) \neq 0$, a covariância tem a seguinte estrutura, segundo Bailey & Grattel (1995).

$$\begin{aligned} E(\varepsilon \varepsilon^T) &= E[(I - \rho W)^{-1} U U^T ((I - \rho W)^{-1})^T] = \\ &= (I - \rho W)^{-1} E(U U^T) ((I - \rho W)^{-1})^T = \\ &= (I - \rho W)^{-1} \sigma^2 I ((I - \rho W)^{-1})^T = \\ &= \sigma^2 ((I - \rho W)^{-1} ((I - \rho W)^T)^{-1}) = \end{aligned} \tag{10}$$

$$\sigma^2((I - \rho W)^T(I - \rho W))^{-1}.$$

desde que $(I - \rho W)$ seja invertível. Na estrutura de variância e covariância (10), cada área é correlacionada com todas as outras áreas, em maior grau para as mais próximas.

Note que diferentemente do modelo de defasagem espacial, a atuação espacial está presente nas variáveis omitidas do modelo, assim os efeitos espaciais são um ruído. Isto ocorre porque nem sempre se consegue modelar todas as características da região em estudo que possam influenciar as áreas vizinhas, como heterocedasticidade, variáveis não observáveis, ausência de variáveis explicativas, entre outros. A figura 10 apresenta o esquema de interação entre as variáveis das áreas A_i e A_j .

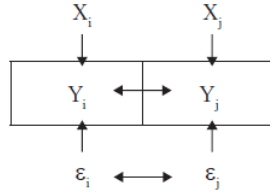


Figura 10: Representação esquemática do modelo de erro espacial. Fonte: Vieira (2009).

Análogo ao modelo com defasagem espacial, os estimadores dos parâmetros serão dados a partir da maximização da função logaritmo da verossimilhança, sendo ela dada por:

$$\ln(L(\beta, \sigma, \rho|Y, X)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln(I - \rho W) - \frac{1}{2\sigma^2}\varepsilon^T(I - \rho W)^T(I - \rho W)\varepsilon =$$

$$-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln(I - \rho W) - \frac{1}{2\sigma^2}(Y - X\beta)^T(I - \rho W)^T(I - \rho W)(Y - X\beta). \quad (11)$$

n representa o número de dados da amostra, $\ln$ o logaritmo natural, σ^2 a variância do modelo de erro e as demais variáveis têm a mesma definição da equação definida anteriormente.

Os modelos com este tipo de estrutura, em relação a dependência espacial, modelos estes denominados CAR, consideram que a observação em cada área não é fixa, e sim uma variável aleatória a qual o valor é condicionado pelo valor em seus vizinhos, ou seja, depende das áreas vizinhas de forma linear e autorregressiva. O uso de efeitos dos modelos CAR, quando incorporado a estrutura de

dependência, auxilia na solução do problema de flutuação aleatória².

Estrutura dos modelos Condicionais Autorregressivos (CAR)

Seja $\mathfrak{R}$ uma região, particionada em n áreas, $\{A_1, \dots, A_n\}$, tais que $\cup_{i=1}^n A_i = \mathfrak{R}$ e $\cap_{i=1}^n A_i = \emptyset$. Seja y_i o valor observado dos casos de determinada doença na área A_i , e deseja-se modelar $y_1, y_2, \dots, y_n$. Se existe interesse em utilizar informações provenientes de áreas vizinhas, com intuito de prever o comportamento de cada área, o modelo mais utilizado para dados de área, são os condicionais autorregressivos (Arab et. al, 2006).

O modelo CAR é especificado por um conjunto de distribuições condicionais; neste caso, se supõe que a observação de uma área tenha distribuição normal com os seguintes parâmetros:

$$y_i | y_j, \tau_i^2 \sim N \left(\mu_i + \sum_{j \in N_i} c_{ij} (y_j - \mu_j), \tau_i^2 \right). \quad (12)$$

onde N_i contém todas as áreas consideradas vizinhas a área A_i , $i, j = 1, \dots, n$, $E(y_i) = \mu_i$, τ_i^2 é a variância condicional, c_{ij} são constantes tais que $c_{ii} = 0$ para $i = 1, \dots, n$. A matriz $C = (c_{ij})$ de ordem $n \times n$, é definida de acordo com o grau de vizinhança entre as áreas A_i e A_j .

Seja $A = (A_{ij})$ de ordem $n \times n$, binária, tal que $a_{ij} = 1$ se as áreas A_i e A_j forem vizinhas, denota-se por $(i \sim j)$, e $a_{ii} = 0$. Defina $W = (w_{ij})$ tal que $w_{ij} = a_{ij}/a_i$, onde $a_i = \sum_j a_{ij} = d_i$ representando o número total de áreas vizinhas a área A_i . Por fim, define-se $C = \rho_c W$ e $\tau_i = \sigma_c/d_i$, sendo ρ_c o parâmetro de correlação espacial do modelo CAR. Desta forma, o modelo CAR assim construído, define uma distribuição conjunta, válida, para $y = (y_1, \dots, y_n)$ dada pela normal multivariada:

$$y \sim N(\mu, (I - C)^{-1}M). \quad (13)$$

onde $\mu = (\mu_1, \dots, \mu_n)^T$, $C = (c_{ij})_{n \times n}$, $M = \text{diagonal}[\tau_1^2, \dots, \tau_n^2] = \sigma_c^2 \text{diagonal}[d_1^{-1}, \dots, d_n^{-1}]$, $(I - C)$ invertível e $(I - C)^{-1}M$ positiva definida.

²O termo significa que o indicador está variando aleatoriamente, sem qualquer padrão explicável.

A implementação do modelo CAR é conveniente em uma configuração bayesiana hierárquica, devido a sua estrutura condicional explícita (Arab et. al, 2006).

2.5.3 Conclusões

Os dois modelos apresentados nesta seção, modelo de defasagem e de erro espacial, são de relevância na área espacial, sendo também aplicados em diversas áreas, como em economia, criminalidade, morbidade, epidemiologia, entre outras. Estes modelos possibilitam detectar valores altos e baixos de determinado evento de interesse, além da visualização de *clusters* espaciais, explorar associações entre fatores que influenciam em determinada doença, como no exemplo da dengue em Almeida et al. (2009), que relaciona a ocorrência com fatores socioeconômicos. No entanto, com interesse em analisar dados com características não só espaciais, acrescenta-se aspecto temporal aos mesmos, e o estudo passa a ser espaço-temporal. Na análise espaço-temporal o interesse passa a ser avaliar evolução temporal de valores extremos, baixo ou alto, comparar mudanças ao longo do tempo, de uma ou mais variáveis, examinar padrões de uma doença, além de tantos outros ramos possíveis, passando a existir técnicas cada vez mais complexas.

2.6 Modelos Espaço-Temporais

Muitas vezes, existe interesse em adicionar variáveis temporais aos modelos espaciais, como verifica-se em Lowe et al. (2011), que afirmam existir vantagens ao fazer uso de previsões climáticas considerando os diferentes meses ou anos, no desenvolvimento de um sistema de alerta para epidemias de dengue, visto que a dengue segue um padrão sazonal no Brasil. A identificação precoce de uma epidemia de doenças infecciosas é um importante passo para a implementação de intervenções eficazes para controlar a doença e reduzir a mortalidade, Lowe et al. (2011). Se as autoridades sanitárias responsáveis forem notificadas a tempo, medidas de controle podem ser postas em prática.

No Brasil, o período de maior incidência de casos ocorre entre Janeiro e Maio, quando o clima é mais quente e úmido Lowe et al. (2011). Logo, adicionar

variáveis climáticas pode ser relevante para o desenvolvimento de um modelo mais apropriado.

Agregação na dimensão temporal, como certos padrões sazonais, endêmicos ou epidêmicos, pode contribuir para o esclarecimento dos mecanismos responsáveis pela geração dos casos de doenças, afirma Werneck & Struchiner (1997).

Um estudo interessante feito em Lowe et al. (2014) apresenta resultados de um modelo espaço-temporal de alerta para regiões, no caso, para 12 capitais do Brasil, no ano em que ocorreu um grande evento esportivo, a Copa do Mundo de 2014. Neste ano o país passava por uma situação crítica, com grandes surtos de dengue. O modelo indicou em três categorias, áreas com baixo, médio e alto risco da doença. Este trabalho fez uso da inferência bayesiana ao invés da abordagem clássica. A figura 11, ilustra as regiões de interesse, classificadas em uma das três categorias.

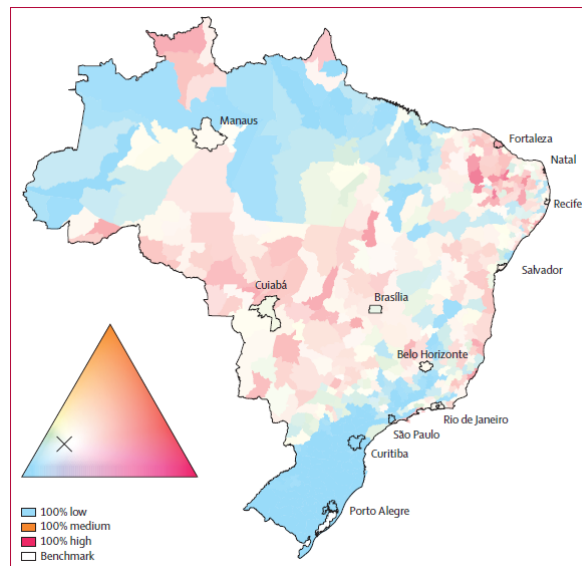


Figura 11: Classificação de risco de ocorrências de dengue para o período da Copa do Mundo de 2014. Fonte: Lowe et al. (2014).

A seguinte situação pode representar a relevância de se considerar o estudo espaço-temporal. Sejam duas regiões, $\mathfrak{R}$ e $\mathfrak{L}$ com mesma população em risco, analisadas durante dois períodos de tempo, t e $t + 1$. No período de tempo t foram observados 10 casos de uma determinada doença na região $\mathfrak{R}$ e 5 na região $\mathfrak{L}$ e no período $t + 1$ foram observados 5 casos na região $\mathfrak{R}$ e 10 na $\mathfrak{L}$. Uma análise puramente espacial entre as duas regiões, desconsiderando o tempo, não mostrará associação com o espaço, pois as duas regiões registraram 15 casos. Se considerar uma análise puramente temporal, não será observado diferença entre

os períodos de tempo, efeito temporal, já que nos dois períodos, observou-se 15 casos.

Ao analisar o problema de forma puramente espacial ou temporal, conclui-se que a ocorrência de casos independe da região e do tempo em que ocorreu. Porém, existe uma interação espaço-temporal que precisa ser considerada, concluindo que as regiões $\mathfrak{R}$ e $\mathfrak{L}$ apresentam diferentes padrões de ocorrência de casos no período de tempo. Este simples exemplo pode ser estendido a um maior número de áreas e períodos de tempo (Vieira, 2006).

Devido ao grande avanço computacional, a inferência bayesiana vem se destacando em aplicações espaço-temporais, quando comparada a inferência clássica, pois se mostra eficiente para modelos com estruturas consideradas complexas. Em virtude do que foi mencionado, conceitos sobre inferência bayesiana serão tratados a seguir.

2.7 Inferência Bayesiana

Na inferência bayesiana, nem toda informação sobre o parâmetro é proveniente da amostra, parte dela está contida na amostra que é representada pela função logaritmo de verossimilhança, e outra parte é devida ao conhecimento que o pesquisador possui antes de observar os dados. Esse conhecimento pode ser apresentado através de uma distribuição de probabilidade para o parâmetro, denominada distribuição *a priori* do parâmetro, Ehlers (2003).

A informação que se tem sobre uma quantidade de interesse θ é fundamental na estatística. O verdadeiro valor de θ é desconhecido e a ideia é tentar reduzir este desconhecimento. Além disso, a intensidade da incerteza a respeito de θ pode assumir diferentes graus. Do ponto de vista Bayesiano, estes diferentes graus de incerteza são representados através de modelos probabilísticos para θ . Logo, é natural que diferentes pesquisadores possam ter diferentes graus de incerteza sobre θ , especificando modelos distintos. Sendo assim, não existe distinção entre quantidades observáveis e os parâmetros de um modelo estatístico, todos são considerados quantidades alatórias (Ehlers, 2003), sendo apenas quantidades observáveis consideradas fixas, diferentemente da inferência clássica, que considera os parâmetros como quantidades fixas e desconhecidas às quais se aproxima, pelo processo de inferência.

2.7.1 Teorema de Bayes

Considere uma quantidade de interesse desconhecida θ . O conhecimento que se tem de θ , resumido probabilisticamente através de $p(\theta)$, pode ser atualizado observando-se uma quantidade aleatória X relacionada com θ . A distribuição amostral $p(x|\theta)$ define esta relação. De acordo com Ehlers (2003), a ideia de que, após observar $X = x$, a quantidade de informação sobre θ aumenta é bastante intuitiva e o teorema de Bayes é a regra de atualização utilizada para quantificar este aumento de informação,

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(\theta, x)d\theta}. \quad (14)$$

Note que $1/p(x)$ não depende de θ , funciona como uma constante normalizadora de $p(\theta|x)$. Para um valor fixo de x , a função $l(\theta; x) = p(x|\theta)$ fornece a plausibilidade ou verossimilhança de cada um dos possíveis valores de θ , enquanto $p(\theta)$ é chamada distribuição *a priori* de θ . Estas duas fontes de informação, *priori* e verossimilhança, são combinadas levando à distribuição *a posteriori* de θ , $p(\theta|x)$ (Ehlers, 2003). Assim, a forma usual do teorema de Bayes é dada por:

$$p(\theta|x) \propto l(\theta; x)p(\theta). \quad (15)$$

ou seja,

$$\text{distribuição } a \text{ posteriori} \propto \text{verossimilhança} \times \text{distribuição } a \text{ priori}$$

A estimação do parâmetro é realizada com a utilização dessas duas informações conjuntamente, através do teorema de Bayes, e tem como resultado a distribuição *a posteriori* do parâmetro, ou seja, é a distribuição final encontrada para o parâmetro e que será utilizada nas inferências a serem realizadas. Em grande parte dos problemas não é possível obter a forma analítica das distribuições *a posteriori* dos parâmetros de interesse (Vieira, 2006).

Outro conceito relevante é referente aos *hiperparâmetros*. De acordo com Ehlers (2003), a partir do conhecimento que se tem sobre um parâmetro θ , pode-se definir uma família paramétrica de densidades. Neste caso, a distribuição *a priori* é representada por uma forma funcional, cujos parâmetros devem ser especificados de acordo com este conhecimento. Estes parâmetros indexadores da família de distribuições *a priori* são chamados de hiperparâmetros para distingui-los dos parâmetros de interesse θ .

Esta abordagem em geral, facilita a análise e o caso mais importante é o de *prioris conjugadas*. A ideia é que as distribuições *a priori* e *a posteriori* pertençam à mesma classe de distribuições e assim a atualização do conhecimento que se tem de θ envolve apenas uma mudança nos hiperparâmetros. Neste caso, o aspecto sequencial do método Bayesiano pode ser explorado definindo-se apenas a regra de atualização dos hiperparâmetros já que as distribuições permanecem as mesmas (Ehlers, 2003).

Modelos para dados com características espaciais e/ou espaço-temporais, possuem uma estrutura considerada complexa, pois precisam capturar a dependência espacial e/ou espaço-temporal existentes nos dados. Os modelos com estas características possuem, geralmente, uma formulação bayesiana hierárquica, afirmam Best et al. (2005).

Em virtude do que foi mencionado, será introduzido a seguir, conceitos básicos sobre modelagem hierárquica.

2.7.2 Modelagem Hierárquica

O principal objetivo na caracterização rigorosa de fenômenos naturais é a estimativa e previsão destes processos, bem como os parâmetros que os regulam. Assim, uma situação capaz de acomodar relações complexas entre os dados e modelos, incorporando várias fontes de incerteza é necessária (Arab et al. 2006).

Os Modelos hierárquicos são decompostos em uma série de níveis ligados por regras simples de probabilidade, assim, assumem um quadro aplicável capaz de acomodar incerteza e grande conhecimento *a priori*. Após a introdução de modelos hierárquicos, o desenvolvimento do método MCMC (Markov Chain Monte Carlo) possibilitou uma grande investigação, tanto teórica quanto aplicada, utilizando/desenvolvendo os modelos hierárquicos. A modelagem hierárquica é baseada na probabilidade de distribuição conjunta de uma coleção de variáveis aleatórias serem decompostas em uma série de modelos condicionais. Por exemplo, se A, B, C são variáveis aleatórias, pode-se fatorar de tal forma que:

$$P(A \cap B \cap C) = P(A | B \cap C)P(B | C)P(C). \quad (16)$$

No caso de modelos espaciais e espaço-temporais, a distribuição conjunta descreve o comportamento do processo em todas as áreas (e, eventualmente, todos os tempos). A distribuição conjunta, lado esquerdo da equação (16), é dificilmente determinada para processos complexos.

Note que, a escolha para esta decomposição é baseada no que se conhece sobre o processo em estudo, e que suposições podem simplificá-las, por exemplo, a independência condicional: $P(A | B \cap C) = P(A | B)$. Muitas vezes, é mais fácil expressar modelos condicionais, lado direito da equação (16), do que os modelos conjuntos completos. Sendo assim, ao modelar processos complicados, o modelo hierárquico pode ser decomposto em três etapas:

1. Modelo das observações com os parâmetros;
2. Modelo dos parâmetros com seus hiperparâmetros;
3. Modelo dos hiperparâmetros.

De acordo com Rossi et al. (2005), os níveis podem ser descritos por;

1. Especificação da função de verossimilhança: $p(y_i|\theta_i)$;
2. Especificação da função de ligação: $p(\theta_i|\tau)$;
3. Especificação das *prioris* dos parâmetros: $p(\tau|a)$.

onde y_i são os valores observados do processo de interesse, θ_i vetor de parâmetros, τ hiperparâmetros e a *prioris* atribuídas para os hiperparâmetros.

De acordo com Arab et al (2006), a primeira etapa está relacionada com o processo de observação ou modelo dos dados, a qual especifica a distribuição dos dados, dado o processo de interesse e os parâmetros que os descrevem. A segunda etapa descreve o processo, condicionado aos parâmetros. Finalmente, o último estágio, modela a incerteza nos parâmetros, a partir de ambos os estágios de dados e do processo. Note que, em cada uma dessas etapas podem ter muitas outras sub-etapas, por exemplo, a inclusão de um quarto nível, que corresponda à especificação das *prioris* de hiperparâmetros.

1. Especificação da função de verossimilhança;
2. Especificação das funções de ligação;
3. Especificação das *prioris* dos parâmetros;
4. Especificação das *prioris* dos hiperparâmetros.

O objetivo é estimar a distribuição do processo em estudo e os parâmetros, através dos dados. Os métodos bayesianos são adequadas para estimar em tais configurações hierárquicas. Usando uma abordagem Bayesiana, a distribuição *a posteriori* (isto é, a distribuição conjunta do processo e parâmetros de acordo com os dados) é obtido através do teorema de Bayes.

$$[\text{processo, parâmetros} \mid \text{dados}] \propto [\text{dados} \mid \text{processo parâmetros}] \times [\text{processo} \mid \text{parâmetros}] \times [\text{parâmetros}]$$

Os métodos bayesianos permitem quantificar a incerteza em todos os aspectos do problema (dados, processo, parâmetros) e a distribuição *posteriori* reflete a quantificação de incerteza. Uma desvantagem ao se trabalhar com modelos hierárquicos é a exigência de habilidade computacional (Arab et al. 2006).

Como mencionado anteriormente, o uso do MCMC no modelo hierárquico levou a uma revolução, em que os modelos realistas (ou seja, complicados) puderam ser considerados. Isto é especialmente evidente na análise de processos espaciais e espaço-temporais.

Com intuito de apresentar diferentes técnicas e aplicações de modelos espaço-temporais, foram escolhidas três diferentes abordagens, os modelos possuem semelhanças em suas estruturas, porém cada um tem sua particularidade, como variação temporal, por exemplo, a forma de variação/inclusão do tempo na análise, sendo feita de forma linear, polinomial ou por coortes de nascimento.

2.8 Modelo Espaço-Temporal: Abordagem I

A abordagem I, baseia-se no modelo proposto por Castro et al. (2004), onde o objetivo principal consiste em analisar os padrões espaço-temporais do câncer de pulmão em quatro estados brasileiros, sendo eles Rio Grande do Sul, Santa Catarina, Paraná e São Paulo, durante o período de 1996 a 2000. Os valores observados foram obtidos junto ao Ministério da Saúde e as unidades geográficas utilizadas foram microrregiões definidas pelo IBGE. Foi considerado um modelo bayesiano que permite interação espaço-temporal, ajustado através do software WinBugs.

Um procedimento clássico para lidar espacialmente com taxas de mortalidade de doenças consiste em mapear as Razões de Mortalidade Padronizadas (*Standardized Mortality Ratio* – SMR). Essas razões são as estimativas de máxima

verossimilhança do risco relativo de mortalidade da doença nas áreas. Considerando que a mortalidade da doença siga o modelo de Poisson,

$$[Y_i|E_i, \xi_i] \sim \text{Poisson}(E_i \times \xi_i). \quad (17)$$

E_i representa o vetor de valores esperados para cada área A_i . $E = (E_1, E_2, \dots, E_n)$, conhecido, os valores esperados foram calculados através da técnica de padronização indireta segundo sexo e faixa etária (a padronização permite que seja feita comparações em igualdade de condições), ξ_i é o risco relativo de óbito na área A_i (a ser modelado), sendo n áreas de uma região $\mathfrak{R}$ e Y_i o número de casos na área A_i , para $i = 1, \dots, n$.

Supondo independência entre as ocorrências, o estimador usual de ξ_i é a SMR, dada pelo estimador de máxima verossimilhança, sendo $\lambda_i = E_i \times \xi_i$.

$$L(\lambda_i; y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}; i = 1, \dots, n. \quad (18)$$

$$l(\lambda_i; y_i) = \ln(L(\lambda_i; y_i)) = -\lambda_i + y_i \ln(\lambda_i) - \ln(y_i!)$$

$$\frac{\partial l(\lambda_i; y_i)}{\partial \lambda_i} = -1 + \frac{y_i}{\lambda_i} = 0.$$

logo $\hat{\lambda}_i = y_i$, então:

$$E_i \times \hat{\xi}_i = y_i \implies \hat{\xi}_i = \frac{y_i}{E_i} = SMR_i. \quad (19)$$

Com intuito de realizar um estudo espaço-temporal, introduz-se o componente tempo na análise. Adiciona-se a matriz Y a variável tempo, com isso a matriz Y possui elementos da forma y_{it} , representando o número de óbitos na área A_i e no tempo t . Os elementos E_{it} , representam, o número esperado de óbitos nas microrregiões/áreas nos anos em estudo, e ξ_{it} os riscos relativos de ocorrência de óbitos por câncer de pulmão, sendo definida de forma similar à Y e E . O novo modelo definido é da forma:

$$[Y_{it}|E_{it}, \xi_{it}] \sim \text{Poisson}(E_{it} \times \xi_{it}). \quad (20)$$

Para acompanhar a evolução das taxas, ξ_{it} , de uma mesma área ao longo dos anos e fazer projeções para os próximos anos, é proposto modelar o logaritmo dos riscos relativos, $\theta_{it} = \ln(\xi_{it})$, ao invés dos riscos diretamente.

Porém, ao se observar E_{it} , quando este número for pequeno, o que geralmente acontece para regiões pequenas, a variabilidade do estimador de ξ_{it} pode

ser muito grande. Isto induz a um erro de interpretação, devido ao resultado da estimação e ao valor pequeno atribuído a E_{it} , fato que chamará erroneamente atenção no mapa. Buscando evitar a instabilidade de estimativas, apenas com base nas taxas brutas, deseja-se que as estimativas para determinada área utilize informação sobre eventos em áreas vizinhas, pois devem ter riscos semelhantes. Proposto por Clayton & Kaldor (1987) e utilizado em Castro et al. (2004), uma alternativa para lidar com esse problema seria modelar o risco relativo através de procedimento bayesiano, pois leva a estimativas mais precisas.

O modelo utilizado por Castro et al. (2004) é definido como:

$$\theta_{it} = \log(\xi_{it}) = \alpha_i + \beta_i t. \quad (21)$$

com $i=1,\dots,157$; sendo 157 o número de microrregiões/áreas estudadas e $t=-2,\dots,2$ representando neste caso os 5 anos (1996 a 2000). O parâmetro α_i representa o logaritmo do risco relativo da área A_i no ano $t = 0$, ou seja, 1998.

Microrregiões com $\alpha < 1$ possuíam em 1998, um risco menor que o valor médio de toda região estudada, e as microrregiões com $\alpha > 1$ possuíam em 1998, um risco mais elevado que o risco médio de toda região. O parâmetro β_i representa o ritmo de crescimento ou decrescimento anual do logaritmo referente ao risco relativo na área A_i .

$$\begin{cases} \beta < 0, \implies \text{decrescimento} \\ \beta > 0, \implies \text{crescimento} \end{cases}$$

Para os parâmetros α e β foram escolhidas distribuições *a priori* espaciais que permitem que áreas próximas no espaço tenham valores semelhantes, ou seja, foi adotado para α e β a estrutura CAR. Os modelos condicionais autoregressivos são modelos hierárquicos espaciais mais utilizados para dados de área, Arab et al. (2006).

O modelo permitiu investigar as taxas ao longo de cinco anos para cada área da região em estudo.

Com a figura 12, as conclusões em Castro et al. (2004) afirmam que percentual de crescimento das SMR no período de cinco anos pode ser visto nos mapas; também são apresentados as estimativas para os coeficientes beta, que representam o ritmo de crescimento ou decrescimento anual, e neste caso não houve um gradiente espacial nítido para este ritmo. Para homens, as taxas se concentram na faixa de -5 a 5 , indicando que a maior parte da região não possui crescimento ou decrescimento importante das taxas, as mulheres apresentam

áreas com maiores índices de crescimento (beta estimado) quando comparado aos homens. O modelo foi estimado para homens e mulheres separadamente.

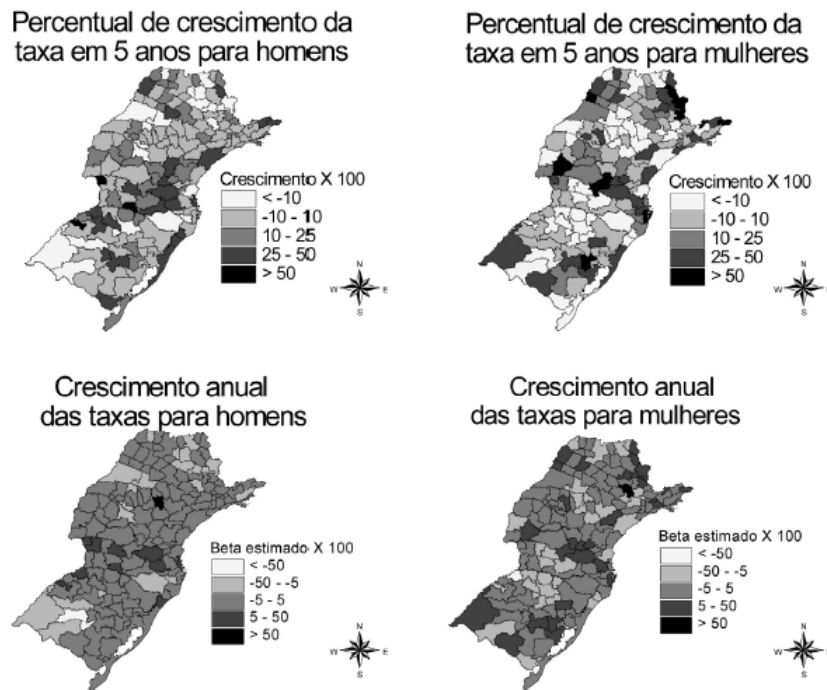


Figura 12: Ritmo de crescimento do SMR e valores estimados para o coeficiente beta, homens e mulheres, câncer de pulmão, São Paulo e sul do Brasil, 1996 a 2000. Fonte: Castro et al. (2004).

As figuras 13 e 14, apresentam respectivamente, as razões de mortalidade padronizada (SMR) em todas microrregiões, para cada ano estudado, em relação aos homens, e as SMR ajustada pelo modelo, em todas microrregiões, para cada ano estudado, em relação aos homens.

O padrão espacial encontrado para os homens, mostra que as microrregiões localizadas ao norte e ao sul da região estudada possuíam um SMR maior, e as localizadas ao centro possuem SMR menor. Uma observação feita, indica que a concentração das maiores SMR estão nas microrregiões próximas à fronteira com Uruguai e Argentina, países que apresentam alta taxa de mortalidade por câncer de pulmão, o que poderia ser explicado por padrões similares de tabagismo, porém, esta variável não foi incluída na análise. O estudo mostrou que não houve um gradiente temporal único para o risco relativo de óbito por câncer de pulmão, sendo que algumas áreas apresentam diminuição e outras crescimento das razões, isto pode estar relacionado a mudanças no padrão de tabagismo ao longo dos anos que antecederam o período estudado.

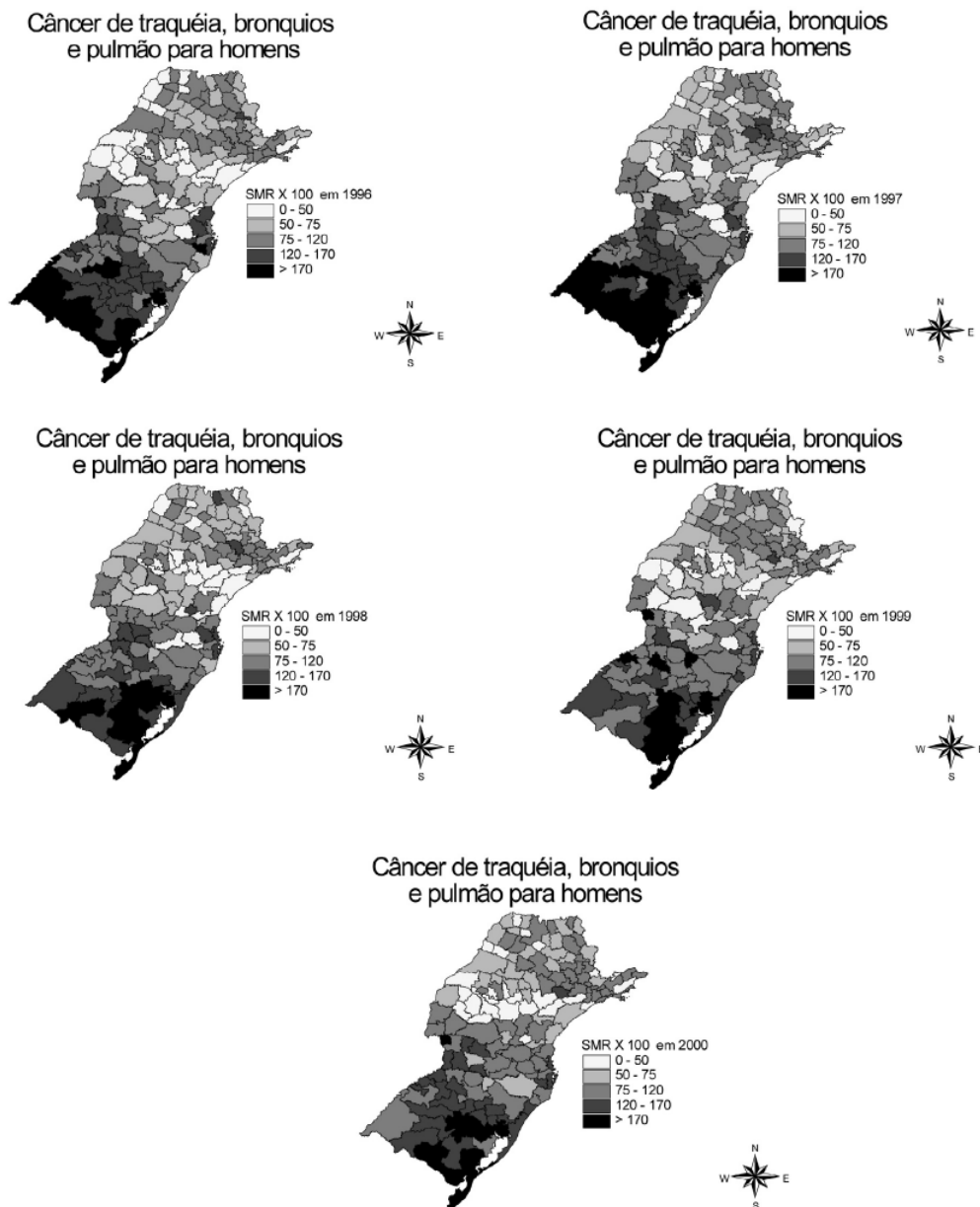


Figura 13: SMR brutas para homens. São Paulo e sul do Brasil de 1996 a 2000. Fonte: Castro et al. (2004).

O artigo traz também a relevância do estudo, onde estes resultados podem ser utilizados pelos planejadores de saúde para elaborar políticas públicas voltadas para a redução do tabagismo e para a melhoria da abordagem do câncer de pulmão, centradas nas micro-áreas que tiveram maior mortalidade, tanto do ponto de vista do maior crescimento temporal quanto das áreas que apresentaram as maiores razões de mortalidade. Um trabalho similar pode ser visto em Prado & Castilho (2009) onde é feito um estudo espaço-temporal sobre a aids no

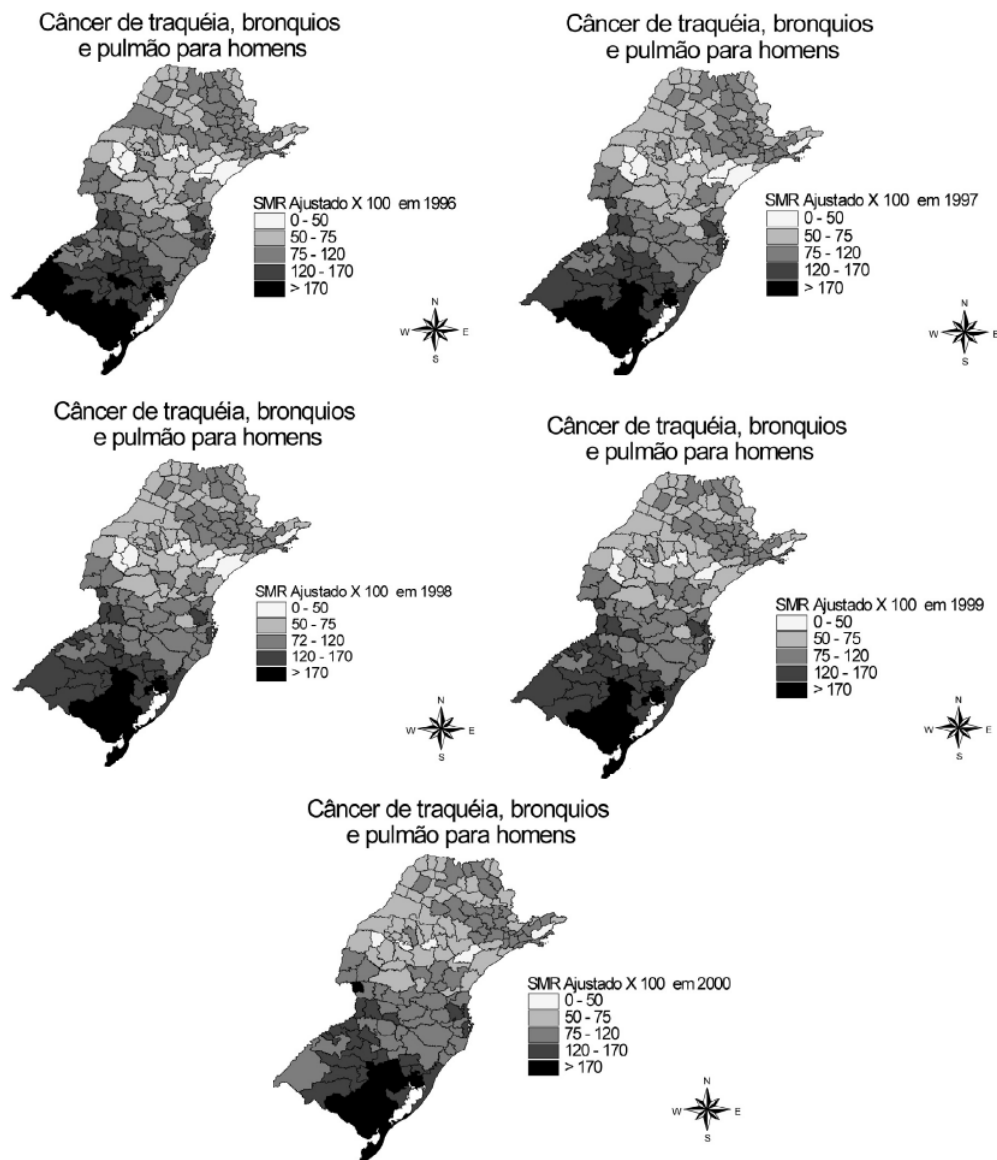


Figura 14: SMR ajustadas pelo modelo, para homens. São Paulo e sul do Brasil de 1996 a 2000. Fonte: Castro et al. (2004).

município de São Paulo, acrescentando ainda informações sobre o IDH.

2.9 Modelo Espaço-Temporal: Abordagem II

A segunda abordagem, baseia-se no modelo proposto por Assunção et al. (2001), onde é feito uma análise de casos de leishmaniose visceral humana³

³Leishmaniose visceral: zoonose caracterizada como doença de caráter eminentemente rural.

com base nos registros de saúde pública em Belo Horizonte – MG, no período de 1994 a 1997. O modelo inclui interação espaço-tempo e adota uma abordagem bayesiana hierárquica, como em Lagazio et al. (2001).

A cidade de Belo Horizonte foi dividida em 117 áreas; e em relação ao tempo foi considerado o período de 12 meses, início em 01/07 e fim em 30/06, para os anos 1994/1995, 1995/1996 e 1996/1997. Foram registrados os casos em cada área e em cada período de tempo; o conjunto de dados foi considerado satisfatório, visto que a leishmaniose é uma doença visceral séria que requer assistência médica, com isso acredita-se que na maioria dos casos, houve procura dos serviços de saúde.

O objetivo principal consistiu em mapear os casos de leishmaniose em Belo Horizonte, para estimar espacialmente o risco de ocorrência, como também sua evolução ao longo do tempo, (1994-1997).

Belo Horizonte, foi dividida em nove *regiões de saúde*, distritos sanitários-DS (figura 15 (a)), estes DS foram particionados de acordo com sua área de abrangência, chamados de zonas, onde estas zonas podem ser consideradas homogêneas. Belo Horizonte, possui um total de 117 zonas (figura 15 (b)), e os casos de leishmaniose foram classificados de acordo com a zona em que ocorreu e a data do primeiro sintoma.

Neste estudo, as informações foram compostas pelo número de casos da doença e o total da população para cada uma das 117 zonas, nos períodos de 1994/95, 1995/96 e 1996/97.

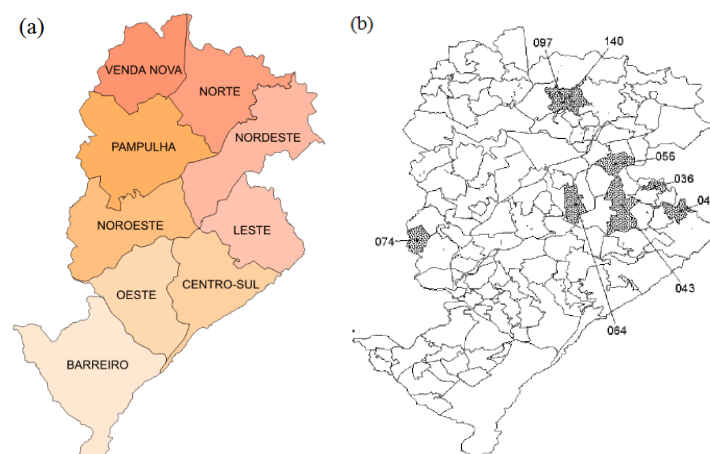


Figura 15: (a) Belo Horizonte dividida em seus 9 distritos sanitários; (b) Belo Horizonte dividida em 117 zonas. Fonte: Assunção et al. (2001).

Assumindo que Y , casos da doença, segue uma distribuição de Poisson:

$$[Y_{it}|P_{it}, \xi_{it}] \sim \text{Poisson}(P_{it} \times \xi_{it}). \quad (22)$$

- ▷ Y_{it} : contagem de casos da doença na área A_i e no tempo t , $i = 1, \dots, 117$ e $t = 1, 2, 3$;
- ▷ P_{it} : população na área A_i e no tempo t , $i = 1, \dots, 117$ e $t = 1, 2, 3$;
- ▷ ξ_{it} : taxa de incidência da doença na área A_i e no tempo t , $i = 1, \dots, 117$ e $t = 1, 2, 3$;
- ▷ Considere que P_{it} e ξ_{it} são independentes.

Decidiu-se trabalhar com o logaritmo da taxa, ao invés dela diretamente;
 $\theta_{it} = \ln(\xi_{it})$.

Para analisar a evolução temporal das taxas, adotou-se um modelo polinomial, de segundo grau, específico para cada área, isto permite que a aceleração e desaceleração das taxas sejam capturadas, desde que seja usado para projeções a curto prazo, afirmam os autores. Assim não se tem problemas quanto a escolha do modelo, visto que projeções a longo prazo para este tipo devem ser evitadas. Com isso,

$$\theta_{it} = \ln(\xi_{it}) = \alpha_i + \beta_i(t - 1) + \gamma_i(t - 1)^2. \quad (23)$$

para $i = 1, \dots, 117$ e $t = 1, 2, 3$. onde,

- ▷ α_i representa o logaritmo da taxa de ocorrência na área A_i no primeiro ano, (1994/95);
- ▷ $\beta_i + \gamma_i(2t - 1)$ representa o incremento no logaritmo da área A_i quando passa do período t para o $t + 1$;

No tempo t ,

$$\begin{aligned} \ln(\xi_{it}) &= \alpha_i + \beta_i(t - 1) + \gamma_i(t - 1)^2 \\ &= \alpha_i + \beta_i t - \beta_i + \gamma_i t^2 - 2t\gamma_i + \gamma_i. \end{aligned}$$

No tempo $t + 1$,

$$\ln(\xi_{it}) = \alpha_i + \beta_i t + \gamma_i t^2.$$

com isso o incremento do logaritmo da taxa na área A_i quando se passa do tempo t para o tempo $t + 1$ é dado por

$$\beta_i + \gamma_i(2t - 1).$$

$\triangleright \alpha, \beta$ e γ denotam os vetores que contém todos os α_i, β_i e γ_i .

Vale ressaltar que os parâmetros/coeficientes do polinômio são específicos para cada área, ou seja, em uma área particular será possível analisar se houve crescimento, decrescimento ou estabilidade das taxas ao longo dos anos.

Na abordagem bayesiana, o conhecimento *a priori* sobre os parâmetros α, β e γ , podem ser tomados de modo a admitir que áreas consideradas vizinhas possuam valores semelhantes; esta distribuição *a priori* não indica a localização de valores altos ou baixos no mapa, pelo contrário, ela apenas faz com que os valores sejam dependentes e tendam a ser semelhantes se forem vizinhos. As *prioris* adotadas, que satisfazem as restrições para os parâmetros são dadas por, sendo τ_α, τ_β e τ_γ os hiperparâmetros respectivamente.

$$\begin{aligned} f(\alpha_1, \dots, \alpha_{117} | \tau_\alpha) &\propto \exp \left(\frac{-\tau_\alpha}{2} \sum_{i \sim j} (\alpha_i - \alpha_j)^2 \right) \\ f(\beta_1, \dots, \beta_{117} | \tau_\beta) &\propto \exp \left(\frac{-\tau_\beta}{2} \sum_{i \sim j} (\beta_i - \beta_j)^2 \right) \\ f(\gamma_1, \dots, \gamma_{117} | \tau_\gamma) &\propto \exp \left(\frac{-\tau_\gamma}{2} \sum_{i \sim j} (\gamma_i - \gamma_j)^2 \right) \end{aligned} \quad (24)$$

onde $i \sim j$ se as áreas A_i e A_j forem vizinhas. Note que, ao fazer com que áreas vizinhas possuam valores similares, as somas dos expoentes diminuem, aumentando assim o valor da densidade.

Foi assumido que α, β e γ são condicionalmente independentes nos hiperparâmetros τ_α, τ_β e τ_γ . Com isso a distribuição *a priori* conjunta é obtida pelo produto das *prioris* definidas em (24).

Denote por α_{-i} o vetor α sem a coordenada i , da mesma forma defina β_{-i} e γ_{-i} ; defina também $\overline{\alpha_i}$ como a média dos n_i vizinhos da área A_i , analogamente para $\overline{\beta_i}$ e $\overline{\gamma_i}$.

As distribuições definidas em (24) implicam em uma distribuição *priori* condicional para α_i, β_i e γ_i , expressas respectivamente por:

$$\begin{aligned} (\alpha_i | \alpha_{-i}) &\sim N(\overline{\alpha_i}, 1/(\tau_\alpha n_i)) \\ (\beta_i | \beta_{-i}) &\sim N(\overline{\beta_i}, 1/(\tau_\beta n_i)) \\ (\gamma_i | \gamma_{-i}) &\sim N(\overline{\gamma_i}, 1/(\tau_\gamma n_i)) \end{aligned} \quad (25)$$

Do modo como foi definido as *prioris* condicionais, a estrutura espacial entre as áreas é considerada, visto que o valor de um parâmetro está centrado em

torno da média dos valores de seus vizinhos. As *prioris* definidas em (24) são geralmente chamadas de condicionais autorregressivas - CAR. O grau de semelhança a ser considerado entre os vizinhos é controlado pelos hiperparâmetros τ_α , τ_β e τ_γ , onde estes assumem distribuições gamas, independentes, mais especificamente:

$$\begin{aligned}\tau_\alpha &\sim \Gamma(a, b) \\ \tau_\beta &\sim \Gamma(c, d) \\ \tau_\gamma &\sim \Gamma(e, f)\end{aligned}\tag{26}$$

onde a , b , c , d , e , e f são conhecidos e definidos de forma a serem vagos, já que não existe qualquer conhecimento prévio sobre os parâmetros, de modo a ter variância grande e valor esperado adequado. Neste caso, os autores assumiram para a , b , c , d , e , e f , valor igual a 0.001.

Após obter os dados Y_{it} , o conhecimento sobre os parâmetros α , β e γ é atualizado através da distribuição *a posteriori* dada por:

$$f_{posteriori}(\alpha, \beta, \gamma, \tau_\alpha, \tau_\beta, \tau_\gamma | Y) \propto L(Y | \alpha, \beta, \gamma, \tau_\alpha, \tau_\beta, \tau_\gamma) \times f_{priori}(\alpha, \beta, \gamma, \tau_\alpha, \tau_\beta, \tau_\gamma)\tag{27}$$

sendo $L(Y | \alpha, \beta, \gamma, \tau_\alpha, \tau_\beta, \tau_\gamma)$ a função de verossimilhança e $f_{priori}(\alpha, \beta, \gamma, \tau_\alpha, \tau_\beta, \tau_\gamma)$ a *priori* conjunta, dada pelo produto das expressões em (24) pela função densidade dos hiperparâmetros.

A distribuição *a posteriori* não pode ser obtida algebricamente, porém, é através dela que se faz inferências sobre o logaritmo do risco de ocorrência e do risco ξ_{it} propriamente. Quanto aos resultados da análise, os mapas resultantes foram considerados satisfatórios, e quando comparados aos de taxas brutas, figura 16, apresentaram uma suavização.

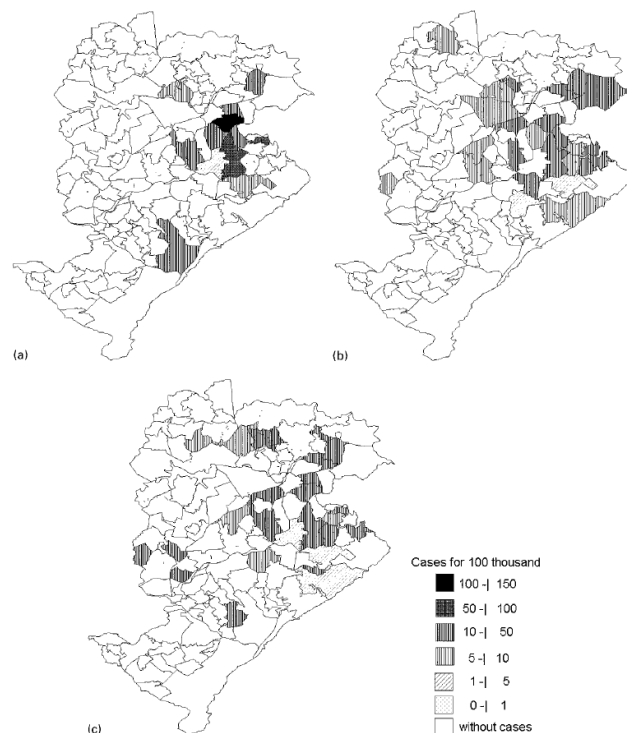


Figura 16: Mapas de taxas brutas para leishmaniose (x100.000), por zonas em Belo Horizonte. (a) para o período de 1994/1995; (b) para o período 1995/1996; (c) para o período 1996/1997. Fonte: Assunção et al. (2001).

O modelo de taxas brutas apresenta grande variabilidade entre as áreas; o modelo apresentado foi útil para auxiliar numa possível solução do problema, visto que as informações obtidas do modelo podem ser uma ferramenta útil para o combate a leishmaniose, possibilitando que os órgãos de saúde responsáveis foquem em regiões de maior risco. Deve-se considerar que a leishmaniose visceral, mais recentemente, vem se expandindo para áreas urbanas de médio e grande porte e se tornou crescente problema de saúde pública no país e em outras áreas do continente americano, sendo uma endemia em expansão geográfica. É uma doença sistêmica, caracterizada por febre de longa duração, perda de peso, astenia, adinamia e anemia, dentre outras manifestações. Quando não tratada, pode evoluir para óbito em mais de 90% dos casos segundo o Portal da saúde (2016). A figura 17 apresenta o mapa de estimativas para o período de 1997/98.

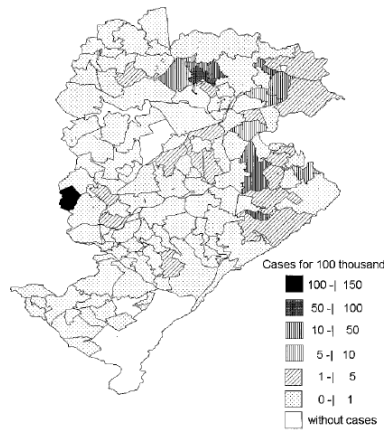


Figura 17: Estimativas bayesianas para a taxa de ocorrência de leishmaniose ($\times 100.000$), dividido por zonas em Belo Horizonte, para o período de 1997/98. Fonte: Assunção et al. (2001).

A figura 18, apresenta as estimativas encontradas para três modelos distintos, onde (a) projeções dadas por um modelo linear, (b) taxas brutas e (c) a abordagem II, do modelo quadrático, em quatro períodos $t = 0, 1, 2, 3 \equiv (94/95, 95/96, 96/97, 97/98)$.

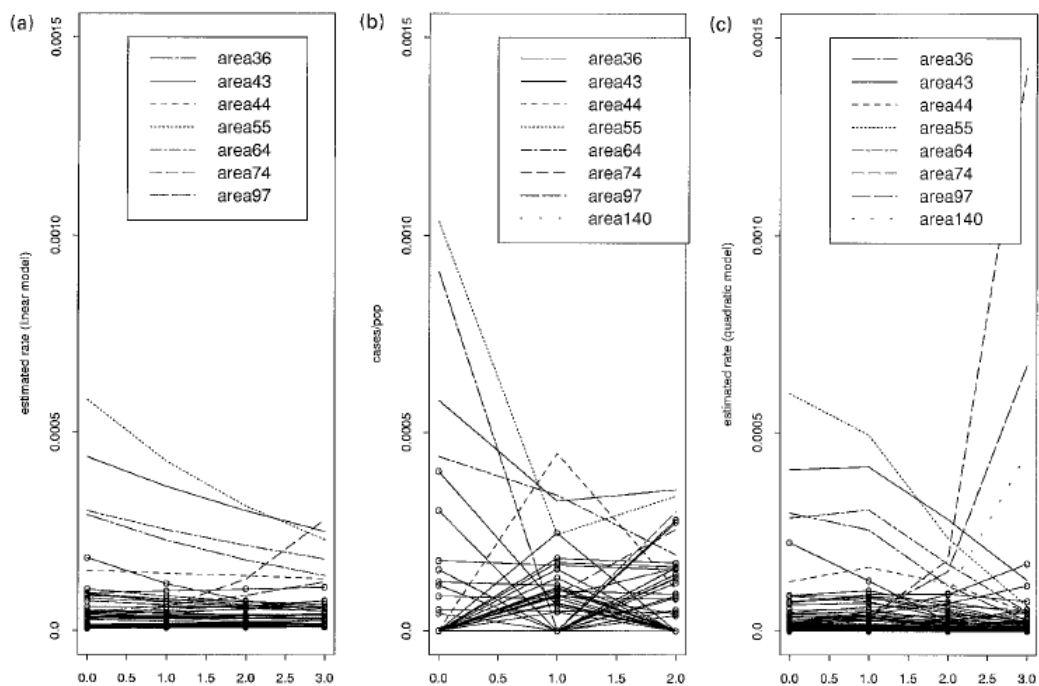


Figura 18: Estimativas (a) projeções dadas por um modelo linear, (b) taxas brutas e (c) a abordagem II, do modelo quadrático. Fonte: Assunção et al. (2001).

As estimativas foram realizadas apenas para 40 zonas que apresentaram

no mínimo um caso nos três períodos de tempo, sendo 18(a) Estimativa bayesiana, através do modelo, linear para o log das taxas. 18(b) Taxas brutas e 18(c) Estimativa bayesiana, através do modelo quadrático (abordagem II). Em 18(a) e 18(c) o tempo 3 corresponde as projeções para 1997/98. Tempo 0 refere-se ao período 1994/95, Tempo 1 refere-se ao período 1995/96 e o tempo 2 ao período 1996/97.

Comparando as figuras 18(a) e 18(c), nota-se que, em geral, as projeções dada pelo modelo quadrático são mais *pessimistas* em relação as do modelo linear, exceto para a área 74, que mostra um rápido aumento das taxas de ocorrência de leishmaniose. O modelo quadrático assim como o linear destacam as mesmas áreas, e algumas adicionais, tais como as áreas 97 e 74, estas duas não são próximas as áreas em que a doença foi introduzida. Ao analisar a área 74 em particular, esta se localiza do lado oposto do município, este aumento pode indicar um novo foco da doença. A figura 19 auxilia na localização das áreas estimadas pelo modelo, citadas na figura 18(a),(b) e (c).

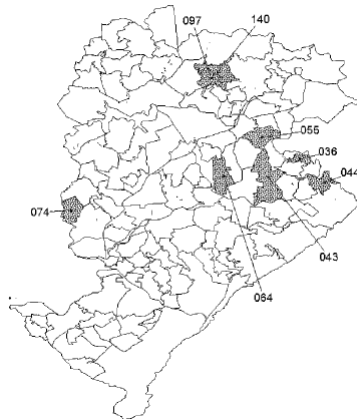


Figura 19: Áreas em destaque. Fonte: Assunção et al. (2001).

2.10 Modelo Espaço-Temporal: Abordagem III

A abordagem III, baseia-se em Lagazio et al. (2001), onde os autores estudaram a variação espacial e temporal para dados de morte por câncer de pulmão em Toscana - Itália, para o sexo masculino durante o período de 1971-1994, através de um modelo bayesiano hierárquico espaço-temporal. A população anual, em média, desta região, possui cerca de 3,5 milhões de habitantes e foram

observados cerca de 33197 casos de morte por câncer de pulmão em 287 municípios da região de Toscana. As mortes são correspondentes para cada município, classificadas em 14 classes de idade (20-24,...,80-84,85 ou mais), as primeiras quatro classes de idade foram omitidas, também foi considerado 5 períodos de tempo (1971-74,...,1990-94). Para a análise espaço-temporal foram considerados 7 coortes de nascimento, sendo eles (1900-04,...,1930-34), que correspondem a pessoas com idades entre 35 e 69 anos, isto para o início dos estudos. A figura 20 apresenta os coortes de nascimento considerados, em negrito, para analisar o risco de óbito por câncer em uma região pré-definida. Vale ressaltar, que existem doenças importantes, que podem ser melhor compreendidas pela análise de coorte.

Age	Period				
	1971-74	1975-79	1980-84	1985-89	1990-94
20-24	0.0000(0)	0.7292(4)	0.1637(1)	0.0000(0)	0.1559(1)
25-29	0.8460(4)	1.0021(6)	0.7065(4)	1.2868(8)	0.2919(2)
30-34	2.2939(11)	2.2017(13)	1.1816(7)	2.2805(13)	0.9526(6)
35-39	5.6904(27)	7.4150(45)	7.1708(43)	4.3454(26)	5.7414(33)
40-44	21.6110(106)	22.7643(135)	19.4246(116)	19.7504(117)	14.3560(86)
45-49	43.2554(218)	44.6608(273)	47.6611(278)	46.0781(273)	32.1820(190)
50-54	75.6624(342)	91.1210(568)	108.2916(642)	85.2005(489)	76.1884(446)
55-59	128.5167(467)	150.7185(780)	187.7351(1112)	180.1408(1031)	143.1949(802)
60-64	190.3393(816)	229.6426(1030)	247.4448(1225)	292.4443(1633)	259.1923(1413)
65-69	279.5167(960)	319.2440(1500)	370.0820(1445)	352.7771(1548)	394.7582(2018)
70-74	286.0588(714)	385.8787(1324)	432.8713(1672)	468.2044(1562)	456.7301(1740)
75-79	264.5848(385)	344.5893(743)	443.8429(1092)	495.5151(1449)	521.9005(1378)
80-84	145.4959(128)	270.1233(293)	407.8348(517)	489.3647(756)	487.9225(971)
85 or more	129.8553(57)	153.7664(92)	257.5429(165)	345.5318(269)	365.5075(385)

Figura 20: Faixa etária e taxas específicas de um período de incidência (x 100 000), e o número de casos de morte (entre parênteses) por câncer de pulmão, para o sexo masculino, em Toscana (Itália), 1971-1994. Fonte: Lagazio et al. (2001).

Os períodos de tempo considerados para estudo, nesta abordagem, apresentam três possíveis tipos de análise, por idade, período e efeito de coorte.

As taxas de mortalidade por qualquer doença, assim como sua variação, podem sofrer impacto de fatores diversos. Por exemplo, a idade da população pode ter um efeito sobre a mortalidade por doenças crônicas. Eventos ocorridos em determinados períodos podem influenciar simultaneamente todas as faixas etárias (efeito período). Finalmente, certos fatores afetam uma geração e promovem mudanças distintas nas taxas de magnitude em sucessivos grupos de idade e sucessivos períodos. Esse é o chamado efeito da coorte, que permite avaliar o

risco associado aos hábitos e às exposições de longa duração, segundo Meira et al. (2015).

Para entender os efeitos desses fatores, modelos chamados de APC (age-period-cohort) são utilizados. Esses modelos permitem desagregar os três efeitos: idade, período e coorte de nascimento, identificando qual deles teve maior impacto na evolução das taxas de incidência e mortalidade das doenças ou agravos à saúde, Meira et al. (2015).

Sendo assim, cada efeito busca analisar a incidência da doença em estudo de forma diferente, onde:

1. **Efeitos de Idade:** Busca-se analisar a incidência de determinada doença, por exemplo, devido ao aumento da idade na população em estudo.
2. **Efeitos de Período:** Busca-se analisar a incidência de determinada doença, por exemplo, que são devidas ao período de tempo na qual a população em estudo vive.
3. **Efeito de Coorte:** Busca-se analisar a incidência de determinada doença, por exemplo, devido às características comuns de um grupo de pessoas da população, reunidas em um ponto no tempo, ao mesmo tempo em suas vidas, que tem portanto a mesma idade. (Ex: Coorte de nascimento, todos os nascidos em 1900).

Quanto ao modelo proposto por Lagazio et al. (2001), foi definido uma distribuição de Poisson com média $\xi_{it} \times E_{it}$, sendo E_{it} o valor esperado dos casos, ξ_{it} o risco relativo da doença para a área A_i e t -ésimo coorte ($i = 1, \dots, N; t = 1, \dots, T$) e Y_{it} o número de casos observados da doença. ($Y_{it} \sim Poisson(\xi_{it} \times E_{it})$). Foi utilizado o logaritmo do risco relativo, logo:

$$\ln(\xi_{it}) = \nu_i + u_i + \theta_t + \psi_{it}. \quad (28)$$

onde ν_i e u_i são, respectivamente, os termos de variabilidade espacial não estruturadas e estruturadas. O componente não estruturado permite que características locais não sejam compartilhadas pelas áreas vizinhas, ou seja, esses efeitos são particulares de cada área A_i e podem ser modelados como independentes. Quanto à parte estruturada espacialmente, é definida com o intuito de considerar áreas vizinhas semelhantes. Existem variáveis observáveis ou não que afetam os riscos à saúde que possuem esta característica, ou seja, áreas próximas tendem a ser

mais similares. θ_t é o efeito do t -ésimo coorte de nascimento e ψ_{it} é o efeito de coorte para uma área específica (parâmetro de interação).

As distribuições *a priori* adotadas para ν_i, u_i, θ_t e ψ_{it} , foram definidas por uma normal multivariada com vetor de média zero e matriz de precisão $\tau_\nu K_\nu, \tau_u K_u, \tau_\theta K_\theta$ e $\tau_\psi K_\psi$. Neste caso, é adequado utilizar a distribuição Gamma como *priori* para $\tau_\nu, \tau_u, \tau_\theta$ e τ_ψ , de forma que admita uma grande dispersão (média 1 e variância 1000) que serão assumidas para os hiperparâmetros descritos.

Deste modo, é necessário definir a estrutura das matrizes K_ν, K_u e K_θ de forma apropriada para os diferentes efeitos. Em particular, para a heterogeneidade espacial, não estruturada K_ν , define-se $K_\nu = I_d$ e o componente espacial estruturado, K_u é modelado por um CAR, onde os elementos adotados definem uma estrutura na qual a vizinhança é considerada.

A matriz K_θ , que traz o efeito de coorte adotado, é estruturada por um passeio aleatório de primeira ordem com incrementos gaussianos independentes, ou seja, esta estrutura faz com que os efeitos de vizinhos em relação ao tempo tendam a ser semelhantes.

Quanto aos termos de interação, estes permitem variação espacial do efeito do tempo ou a variação temporal dos efeitos específicos da área (espacial). Diferentes especificações são possíveis para o termo de interação, dependendo das suposições sobre a sua estrutura de dependência. Dentro do modelo adotado por Lagazio et al. (2001) foi assumido que os termos de interação são estruturados no espaço e no tempo, deste modo, a estrutura da matriz K foi definido como o produto de Kronecker⁴ entre as matrizes de K_u e K_θ . Mais informações sobre as matrizes de precisão K podem ser encontrados em Pinheiro (2009), Clayton (1996) e Lagazio et al. (2001).

Lagazio et al. (2001) afirmam que para o estudo/modelagem de doenças, é necessário que se tenha uma plausibilidade biológica, pois, ela deve servir como orientação principal para a escolha e adequação de um modelo. Para a estimação dos parâmetros foi utilizado o software WinBugs através de uma amostragem Gibbs⁵ de forma a obter a distribuição *posteriori*.

Conclusões sobre a distribuição da doença se mostraram interessantes,

⁴Em matemática, o produto de Kronecker, denotado por $\otimes$, é uma operação em duas matrizes de tamanho arbitrário, resultando em uma matriz de bloco.

⁵Amostragem de Gibbs é um algoritmo para gerar uma sequência de amostras da distribuição conjunta de probabilidades de duas ou mais variáveis aleatórias quando as distribuições condicionais são consideradas conhecidas.

por exemplo, enquanto a análise por período aponta para um nivelamento de risco para a doença em estudo; o modelo que considera coortes de nascimento, sugere uma diminuição acentuada do risco relativo para pessoas nascidas depois de 1930 como mostra a figura 21.

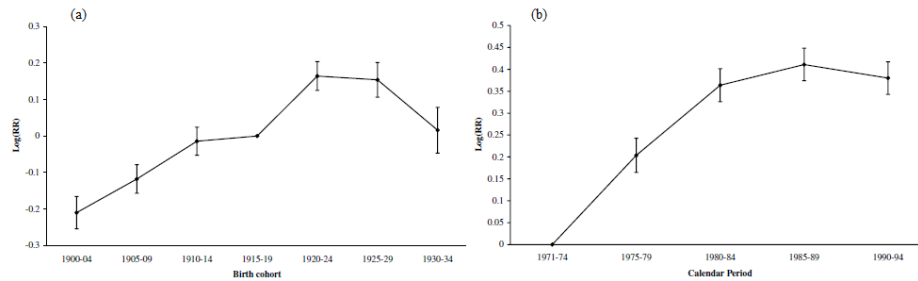


Figura 21: Mortes por câncer de pulmão para o sexo masculino. Toscana (Itália), 1971-1994. (a) Coorte de nascimento; (b) Por período; riscos relativos e intervalos de confiança de 95%;. Fonte: Lagazio et al. (2001).

A figura 22 exibe a distribuição espacial do risco relativo de morte por câncer de pulmão, através dos coortes considerados, de 1900-04 à 1930-34 estimados pelo modelo apresentado (28).

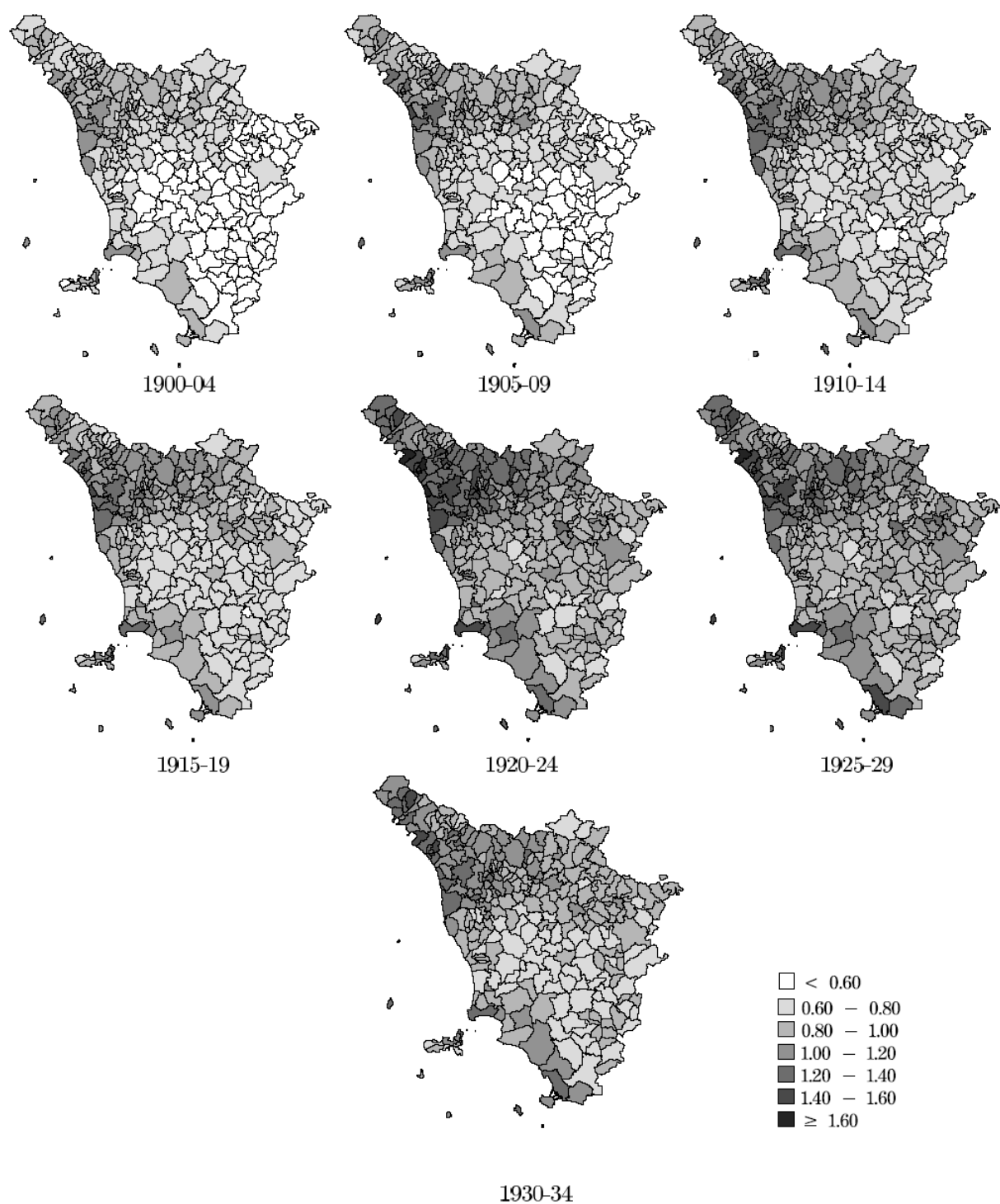


Figura 22: Distribuição espacial dos riscos relativos por coorte de nascimento. Mortes por câncer de pulmão, para o sexo masculino. Toscana (Itália), 1971-1994. Fonte: Lagazio et al. (2001).

3 Aplicações

Nesta seção, foi desenvolvido um estudo espacial para a Dengue, onde obteve-se junto à Fundação Municipal de Saúde do município de Rio Claro – São Paulo, dados referentes aos casos notificados de dengue no ano de 2011. O estudo teve por objetivo analisar a distribuição da doença no município de Rio Claro, dividindo-o por setor censitário, além de analisar a relação da doença com fatores socioeconômicos.

3.1 A Dengue

A dengue é caracterizada como uma doença febril aguda, que pode apresentar um amplo espectro clínico, pois enquanto a maioria dos pacientes se recupera após evolução clínica leve e autolimitada, uma pequena parte progride para doença grave, sendo atualmente a mais importante arbovirose que afeta o ser humano, constituindo-se em sério problema de saúde pública no mundo. A dengue pode ser transmitida por duas espécies de mosquitos, *Aedes aegypti* e *Aedes albopictus*, que picam durante o dia e a noite, (Ministério da Saúde, 2016).

Os transmissores de dengue, principalmente o *Aedes aegypti*, proliferam-se dentro ou nas proximidades de habitações (casas, apartamentos), em recipientes onde se acumula água limpa (vasos de plantas, pneus velhos, cisternas, entre outros), sendo que a principal forma de transmissão é pela picada do mosquito *Aedes aegypti*. Há registros, também, de transmissão vertical (gestante - bebê) e por transfusão de sangue. Atualmente, existem quatro tipos diferentes de vírus da dengue: DENV 1, DENV 2, DENV 3 e DENV 4, afirma o Ministério da Saúde (2016).

A figura 23 apresenta os dois mosquitos transmissores, *Aedes aegypti* e *Aedes albopictus*, respectivamente.



Figura 23: *Aedes Aegypti*, principal vetor e *Aedes Albopictus*. Fonte: Ministério da Saúde (2016).

A figura 24 apresenta os estágios do mosquito transmissor, dividido em quatro fases, sendo elas, ovo, larva, pupa e adulto.

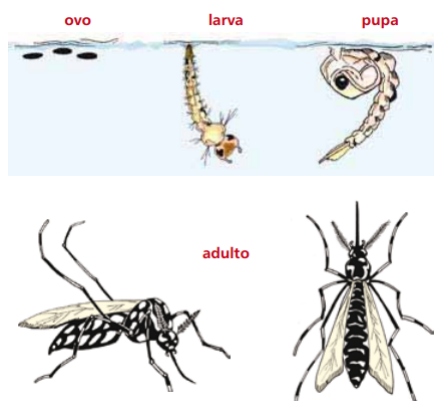


Figura 24: *Aedes aegypti*. Fonte: Ministério da Saúde.

Em relação aos sintomas, depois que o ser humano recebe a picada do mosquito com o vírus, estes se manifestam normalmente do 3º ao 15º dia. Esse período é chamado de incubação, sendo o tempo médio de duração da doença de cinco a seis dias.

A infecção por dengue pode ser assintomática, leve ou causar doença grave, levando à morte. Normalmente, a primeira manifestação da dengue é a febre alta (39° a 40°C), de início abrupto, que geralmente dura de 2 a 7 dias, acompanhada de dor de cabeça, dores no corpo e articulações, prostração, fraqueza, dor atrás dos olhos, erupção e coceira na pele. A forma grave da doença inclui dor abdominal intensa e contínua, vômitos persistentes, sangramento de mucosas, entre outros sintomas.

Não existe tratamento específico para dengue, o único tratamento serve

apenas para aliviar os sintomas; recomenda-se procurar um serviço de saúde mais próximo quando os sintomas aparecem, pois o uso de alguns medicamentos pode agravar a doença. Logo, a única forma de prevenção é acabar com o mosquito, mantendo o domicílio sempre limpo, eliminando os possíveis criadouros. Nos anos recentes a dengue se consolidou como um dos maiores desafios de saúde pública no Brasil, a epidemiologia da doença apresentou alterações importantes, destacando-se o maior número de casos e hospitalizações, com epidemias de grande magnitude, Ministério da Saúde (2016).

Em virtude dos fatos mencionados, estudos relacionados com a doença que buscam descrever seu comportamento e auxiliem a identificar regiões com alta incidência tem se tornado ferramentas úteis para os planejadores de saúde. As aplicações presentes neste trabalho, consistindo em registros de dengue georreferenciados fornecidos pela Fundação Municipal de Saúde de Rio Claro-SP, que serão descritos a seguir.

3.2 Materiais e Métodos

O município de Rio Claro – São Paulo localizado a 22°24'39"S e 47°33'39"O, possuía 186.253 mil habitantes em 2010 e uma área territorial de 498,422km², dividido em 273 setores censitários, segundo o Censo Demográfico de 2010.

Setor censitário é uma unidade territorial de coleta das operações censitárias, definido pelo IBGE, com limites físicos identificados, em áreas contínuas e respeitando a divisão político-administrativa do Brasil, IBGE (2016). A figura 25 apresenta o mapa do estado de São Paulo - Brasil e em destaque o município de Rio Claro.

De acordo com Santos & Souza (2007), ao fazer uso da divisão por setores censitários a população pode ser considerada mais semelhante do ponto de vista socioeconômico e demográfico. Ao trabalhar com áreas pequenas, como os setores censitários, informações socioeconômicas são obtidas facilmente, visto que o Censo Demográfico de 2010 disponibiliza dados por setor; porém é mais complicado obter dados como o número de casos de uma doença ou de óbitos. Além disso, quanto menor for a área, mais raro se tornam os eventos observáveis, óbito por determinada doença, por exemplo, o que origina uma flutuação aleatória dos indicadores e desta decorre vários outros problemas de interpretação.

O termo flutuação aleatória, de acordo com Santos & Souza (2007), signi-



Figura 25: Município de Rio Claro localizado no interior do estado de São Paulo – Brasil. Fonte: [https://pt.wikipedia.org/wiki/Usuário\(a\):Raphael.lorenzeto](https://pt.wikipedia.org/wiki/Usuário(a):Raphael.lorenzeto).

fica que o indicador está variando aleatoriamente, sem qualquer padrão explicável, ou seja, se em uma área rural nasceram 5 crianças em um dado ano e uma faleceu no primeiro ano de vida, a mortalidade infantil seria $1/5$, ou seja, 200 por 1000 nascidos vivos. Entretanto, um epidemiologista sensato, antes de divulgar número tão grave, lembraria que entre 5 recém-natos, um óbito pode ser simplesmente ocasionado por algum acaso, não sendo justo julgar as condições de saúde do município somente por este dado, pois a população no denominador do indicador é muito pequena com isso o estimador fica instável, flutuando aleatoriamente.

As aplicações, especificamente, buscaram relacionar a dengue com fatores socioeconômicos e as variáveis utilizadas no estudo foram obtidas pelo Censo Demográfico de 2010, realizado pelo IBGE. Todas as variáveis possuem informações para cada setor censitário em particular.

Através dos modelos espaciais definidos na seção 2.5, sendo eles modelo de defasagem e erro espacial, buscou-se analisar a relação da dengue com variáveis socioeconômicas. As variáveis foram obtidas pelo Censo Demográfico de 2010, realizado pelo IBGE, todas referentes a setores censitários do município. A tabela 1 apresenta as variáveis consideradas no estudo.

Variáveis Socioeconômicas	Notação
População Residente	<i>pop</i>
Alfabetizados	<i>alf</i>
Abastecimento de Água	<i>agua</i>
Coleta de Lixo	<i>lixo</i>
Renda mensal	<i>renda</i>
Média de Moradores	<i>morad</i>

Tabela 1: Variáveis socioeconômicas.

Descrição:

- ▷ *pop*: Número total de moradores em domicílios particulares permanentes ou população residente em domicílios particulares permanentes;
- ▷ *alf*: Número total de pessoas responsáveis alfabetizadas com mais de 10 anos de idade;
- ▷ *agua*: Número total de domicílios particulares permanentes com abastecimento de água de poço ou nascente na propriedade, com abastecimento de água da chuva armazenada em cisterna ou com outra forma de abastecimento de água.
- ▷ *lixo*: Número total de domicílios particulares permanentes com lixo queimado na propriedade, com lixo enterrado na propriedade, com lixo jogado em terreno baldio ou logradouro, com lixo jogado em rio, lago, mar, ou com outro destino do lixo.
- ▷ *renda*: Valor do rendimento nominal médio mensal das pessoas responsáveis por domicílios particulares permanentes (com e sem rendimento).
- ▷ *morad*: Média de moradores por domicílio.

Para esta análise, foram considerados os casos notificados de dengue no ano de 2011, neste ano foram registrados 1754 casos, porém, devido a falta de informações, foi possível georreferenciar apenas 1324 casos, os quais foram considerados para o estudo. A figura 26 apresenta o município de Rio Claro – SP dividido em 197 setores censitários; neste caso, foram consideradas apenas as áreas urbanas.



Figura 26: Rio Claro – SP, setores censitários, área urbana.

3.3 Resultados e Discussão

Com intuito de identificar presença ou não de autocorrelação espacial da doença em relação aos setores censitários de Rio Claro – SP, calculou-se o índice global de Moran. A figura 27(a) apresenta o diagrama de dispersão de Moran para os casos de dengue. Considerou-se o critério de vizinhança Rook de primeira ordem para determinar a matriz de pesos espaciais, este mesmo critério foi utilizado em toda análise. As análises foram realizadas através do software GeoDA.

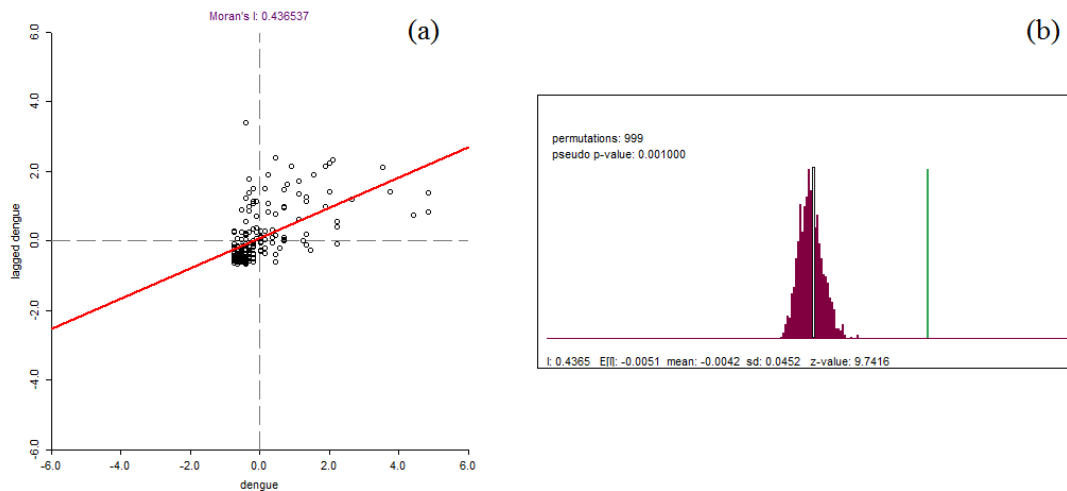


Figura 27: (a) Índice global de Moran e (b) Teste de permutação aleatória.

O valor encontrado para o índice global foi $I=0,44$ possuindo significância estatística com valor $p=0,001$. A figura 27(b) apresenta resultados do teste de permutação aleatória com 999 permutações. Sendo assim, concluiu-se que os dados possuem aglomerados e consequentemente, um estudo espacial sobre a dengue é adequado.

Verificou-se a existência de diferentes regimes espaciais para a região através do índice local de Moran que classifica cada área em um dos quatro quadrantes, sendo eles, Alto-Alto, Alto-Baixo, Baixo-Alto e Baixo-Baixo, além de identificar se possuem autocorrelação positiva ou negativa.

A figura 28 exhibe as classificações para os casos de dengue de 2011 em relação a Rio Claro – SP.

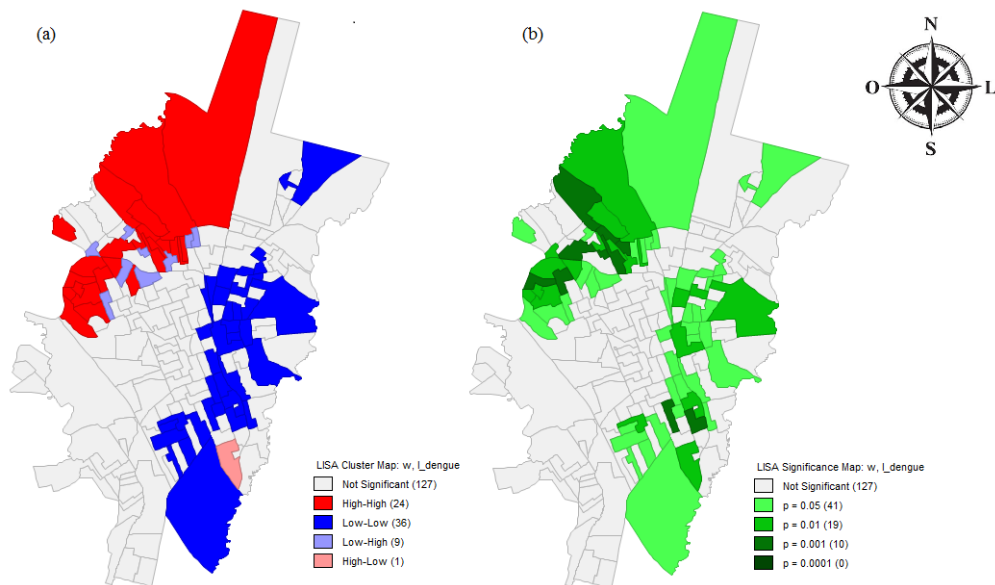


Figura 28: (a) Índice local de Moran, *Lisa cluster map*. (b) Níveis de significâncias para as áreas em relação a Dengue no ano de 2011 em Rio Claro-SP.

Áreas destacadas em vermelho escuro foram classificadas no quadrante Alto-Alto, ou seja, as áreas apresentam altos índices de ocorrência de dengue e seus vizinhos também e áreas destacadas em azul escuro foram classificadas no quadrante Baixo-Baixo, ou seja, áreas que apresentam baixos índices de ocorrência de dengue e seus vizinhos também, ambos classificados como autocorrelação espacial positiva. Regiões em vermelho claro ou azul claro indicam *outliers espaciais*, ou seja, áreas com baixos índices rodeadas por áreas com altos índices ou áreas com altos índices rodeadas por baixos índices, ambos classificados como autocorrelação negativa.

Através do *Lisa cluster map* é possível identificar a existência de *clusters* espaciais para a dengue em relação a região de Rio Claro – SP. Analisando a região de Rio Claro – SP como um todo, à noroeste e norte observa-se *clusters* de alta incidência da doença e à leste, sudeste e sul identifica-se *clusters* de baixa incidência.

Segundo Costa (1998), o crescimento urbano propicia grande fonte de indivíduos suscetíveis e infectados concentrados em áreas restritas. Este fato, associado às condições precárias de saneamento básico, moradia inadequada, fatores culturais e educacionais, são favoráveis para a proliferação do mosquito transmissor da dengue.

Levando em consideração esses aspectos, as variáveis escolhidas para o estudo *pop*, *alf*, *agua*, *lixo*, *renda* e *morad* buscam descrever a situação social e econômica do município.

Com intuito de conhecer suas distribuições no espaço e identificar os regimes espaciais existentes, para cada uma deles foi calculado o índice local de Moran.

Em relação a variável *pop*, foram detectados regimes de autocorrelação positiva e negativa para os dados, de forma mais expressiva para autocorrelação positiva, possuindo significância estatística, figura 29. Encontrou-se *clusters* espaciais para as áreas à oeste, noroeste, algumas regiões centrais e região sul. Conclui-se que a maior concentração da população residente se localiza a noroeste do mapa e menor nas regiões centrais.

Em relação a variável *alf* foram detectados regimes de autocorrelação positiva e negativa para os dados, possuindo significância estatística, figura 30. Identificou-se *clusters* espaciais para as áreas à leste e oeste da região, porém o número de áreas foi pequeno. Neste caso o índice *I* global foi igual a -0.01 indicando que esta variável se distribui aleatoriamente na região.

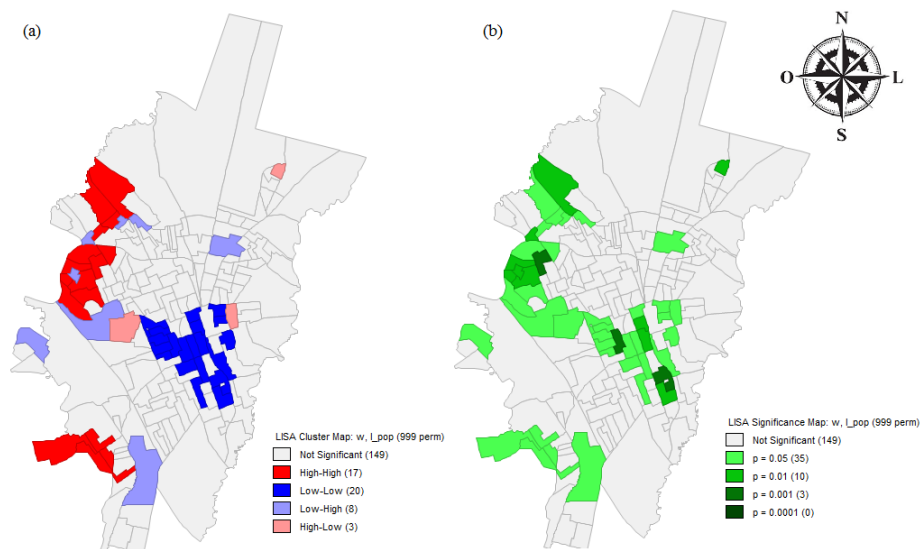


Figura 29: (a) Índice local de Moran, *Lisa cluster map*. (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável *pop*.

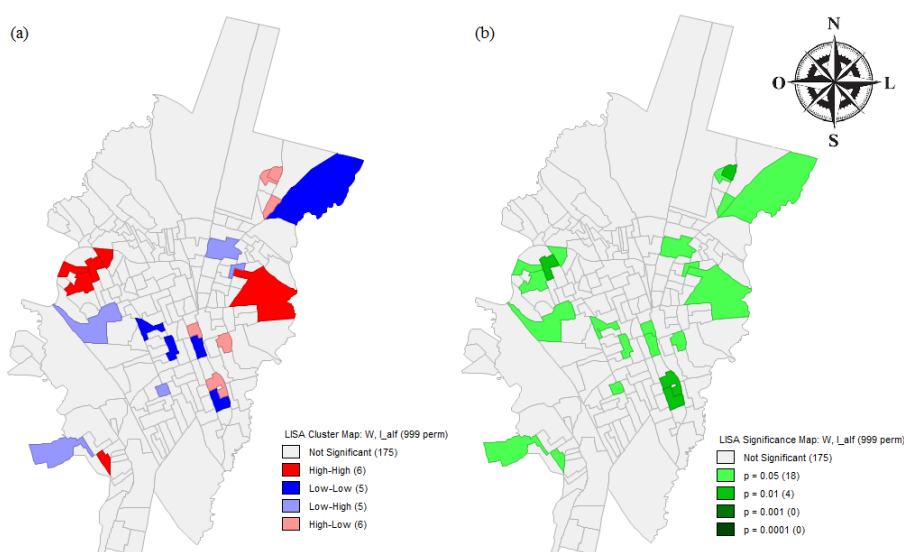


Figura 30: (a) Índice local de Moran, *Lisa cluster map*. (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável *alf*.

Em relação a variável *agua* foram detectados regimes de autocorrelação positiva e negativa para os dados, com significância estatística, figura 31. Identificou-se *clusters* espaciais para as áreas ao norte, sul e algumas regiões centrais. Nas regiões centrais a variável possui valores baixos, para regiões ao norte e sul valores altos indicando grande concentração de abastecimento de água diferente do convencional, como poços, cisternas, entre outros.

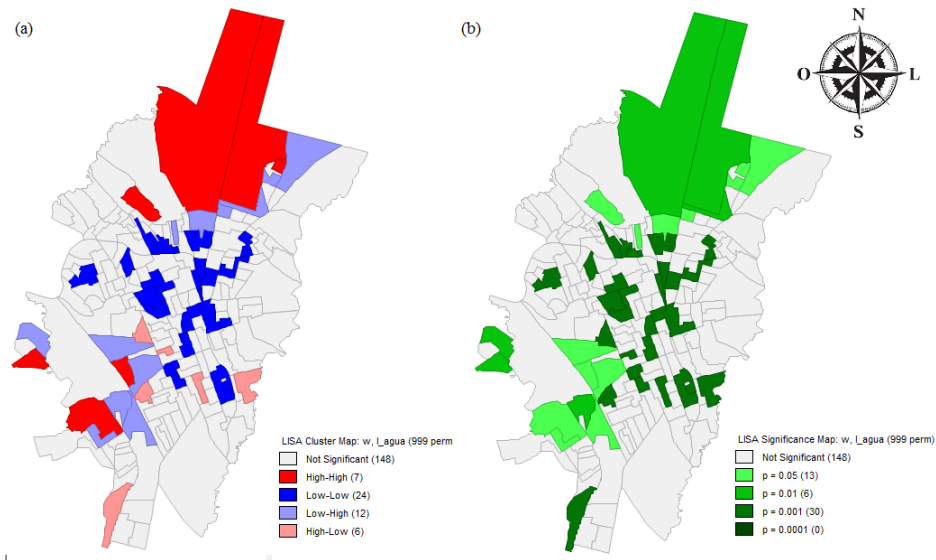


Figura 31: (a) Índice local de Moran, *Lisa cluster map*. (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável *agua*.

Em relação a variável *lixo* foram detectados regimes de autocorrelação positiva e negativa para os dados, possuindo significância estatística, figura 32. Identificou-se *clusters* espaciais para as áreas ao norte e algumas regiões centrais e a leste. As regiões centrais e a leste apresentam valores baixos e a as áreas ao norte, valores altos.

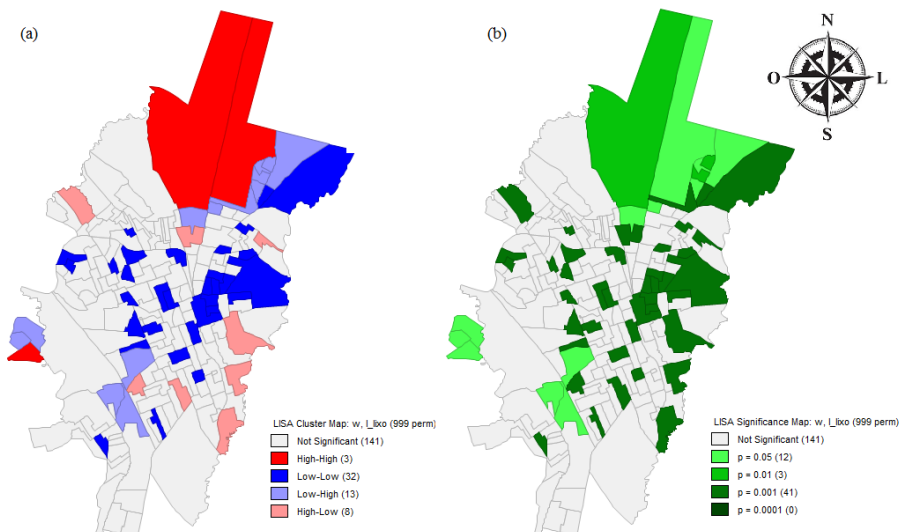


Figura 32: (a) Índice local de Moran, *Lisa cluster map*. (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável *lixo*.

Em relação a variável **renda** foi detectado regimes de autocorrelação positiva e negativa para os dados, possuindo significância estatística, figura 33. Identificou-se *clusters* espaciais para as áreas à noroeste, sudeste e algumas regiões centrais. À noroeste existe uma grande concentração de famílias com baixa renda média mensal, diferentemente da região sudeste e central onde concentra-se famílias com maior renda.

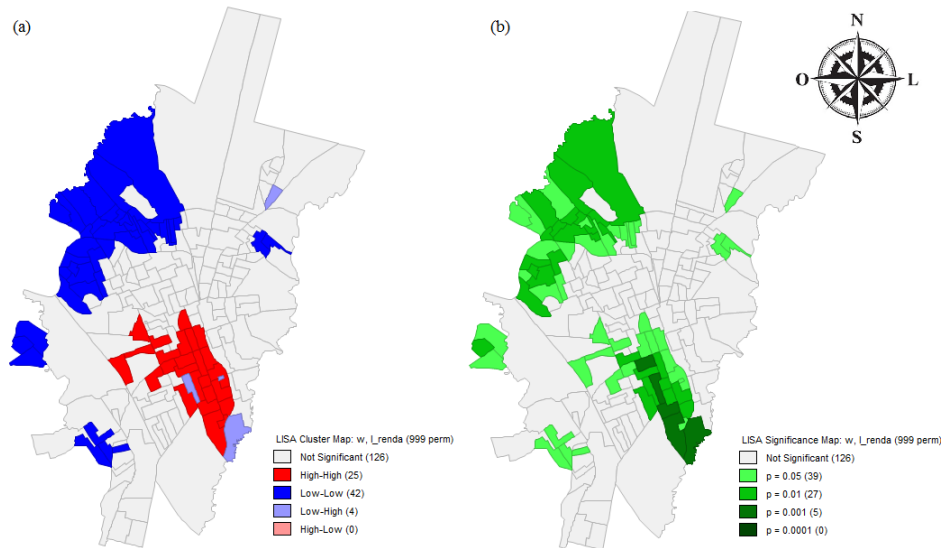


Figura 33: (a) Índice local de Moran, *Lisa cluster map*. (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável **renda**.

Em relação a variável **morad** foi detectado regimes de autocorrelação positiva e negativa para os dados, possuindo significância estatística, figura 34. Identificou-se *clusters* espaciais para as áreas ao norte, noroeste e regiões centrais da região. No centro a variável possui valores baixos e para o norte e noroeste, valores altos.

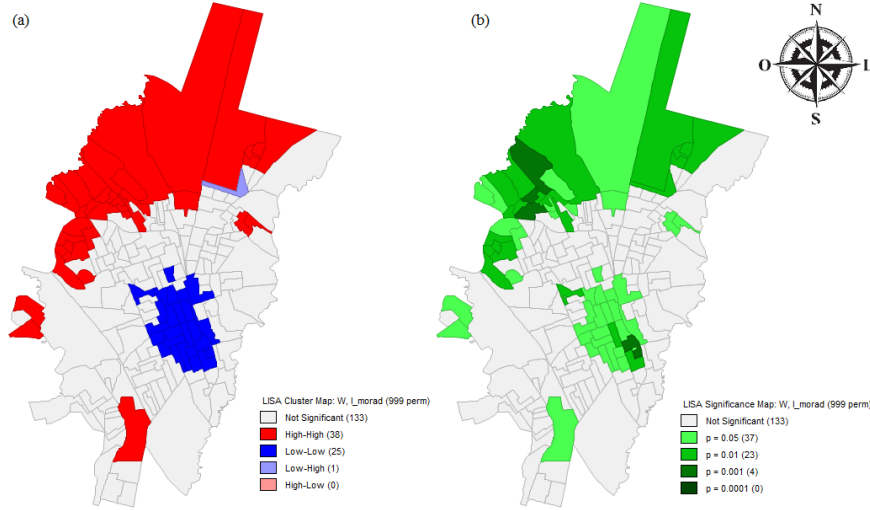


Figura 34: (a) Índice local de Moran, *Lisa cluster map*. (b) Níveis de significâncias para as regiões classificadas em um dos quatro quadrantes presentes. Variável *morad*.

Portando, constatada a concentração espacial de dengue e das variáveis socioeconômicas, verificou-se a relação da doença com estas variáveis através dos modelos de regressão linear clássica, de defasagem e erro espacial, afim de estimar os casos de dengue para a região de Rio Claro – SP.

3.3.1 Regressão Linear Clássica

A análise de regressão linear clássica, busca explicar a relação entre duas ou mais variáveis de modo que a variável resposta, dependente, possa ser explicada por uma ou mais variáveis explicativas, independentes. Sejam:

- ▷ Y : contagem dos casos de dengue, variável dependente;
- ▷ X : matriz que contém as variáveis socioeconômicas *pop*, *alf*, *agua*, *lixo*, *renda* e *morad*, variáveis independentes;
- ▷ β : coeficientes de regressão;
- ▷ ε : vetor de erros.

Como já mencionado, ao considerar o modelo de regressão linear clássico, suposições sobre os erros devem ser válidas, sendo elas, $\varepsilon \sim N(0, \sigma^2)$ e $Cov(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j$. O modelo é definido por:

$$Y = X\beta + \varepsilon. \quad (29)$$

Utilizando o software GeoDa, a regressão resultou em:

Regressão				
Variável	Estimativa	Erro-padrão	Estatística-t	<i>p</i>
<i>Constante</i>	0,99274	5,15192	0,19269	0,84741
<i>pop</i>	0,03103	0,00858	3,61577	0,00038
<i>alf</i>	-0,08811	0,02759	-3,19326	0,00165
<i>agua</i>	1,07973	0,37409	2,88631	0,00435
<i>lixo</i>	-1,08538	0,46279	-2,34533	0,02004
<i>renda</i>	7,5e-005	0,00075	0,09981	0,92058
<i>morad</i>	0,92662	1,75451	0,52814	0,59802
R^2	0,265			
LIK	-684			
AIC	1382			

Tabela 2: Regressão Linear Clássica,

Neste caso as variáveis *renda* e *morad* não foram estatisticamente significante ao nível de 5%, assim como o termo constante. Ao analisar as suposições básicas para um modelo de regressão linear clássico, como normalidade e homocedasticidade para os termos do erro seguem os resultados.

Para testar se os erros são normalmente distribuídos, o teste utilizado é o de Jarque-Bera, Anselin (2005), que tem como hipótese nula a normalidade, sendo assim, se o valor observado em *p* for menor do que 5% (<0,05) rejeita-se a normalidade, se *p* for maior que 5% (>0,05) a normalidade não é rejeitada. Neste caso o teste indicou que os erros não seguem uma distribuição normal.

Os testes para heterocedasticidade espacial são Breusch-Pagan e Koenker-Basset, Anselin (2005). O teste de Breush-Pagan tem como hipótese nula que as variâncias dos erros são iguais, homocedásticas, e hipótese alternativa a heterocedasticidade; no modelo em questão rejeita-se a hipótese nula ao nível de 5%, pois o valor *p* encontrado é 0,0000 <0,05. O teste Koenker-Basset também testa a homocedasticidade dos erros, e neste caso o valor *p* encontrado 0,00025 <0,05 rejeitando também a hipótese nula, dos erros serem homocedásticos. Sendo assim, é constatada a heterocedasticidade espacial. Os resultados são apresentados na tabela 3.

Outro fato importante para ser observado é a existência de autocorrelação espacial nos resíduos, isto indica que o uso de modelos espaciais deve ser mais

Diagnóstico de Regressão			
Multicolinearidade	42,699729		
Teste de Normalidade para os Erros	Graus de Liberdade	Valor	p
Jarque-Bera	2	308,2056	0,00000
Diagnóstico para Heterocedasticidade	Graus de Liberdade	Valor	p
Breusch-Pagan	6	91,1428	0,00000
Koenker-Bassett	6	25,6899	0,00025

Tabela 3: Diagnóstico Regressão Linear Clássica.

apropriado. A estatística de Moran é altamente significativa, sugerindo um problema com a autocorrelação espacial.

A figura 35(a) exibe o diagrama de dispersão de Moran para os resíduos, e indica autocorrelação.

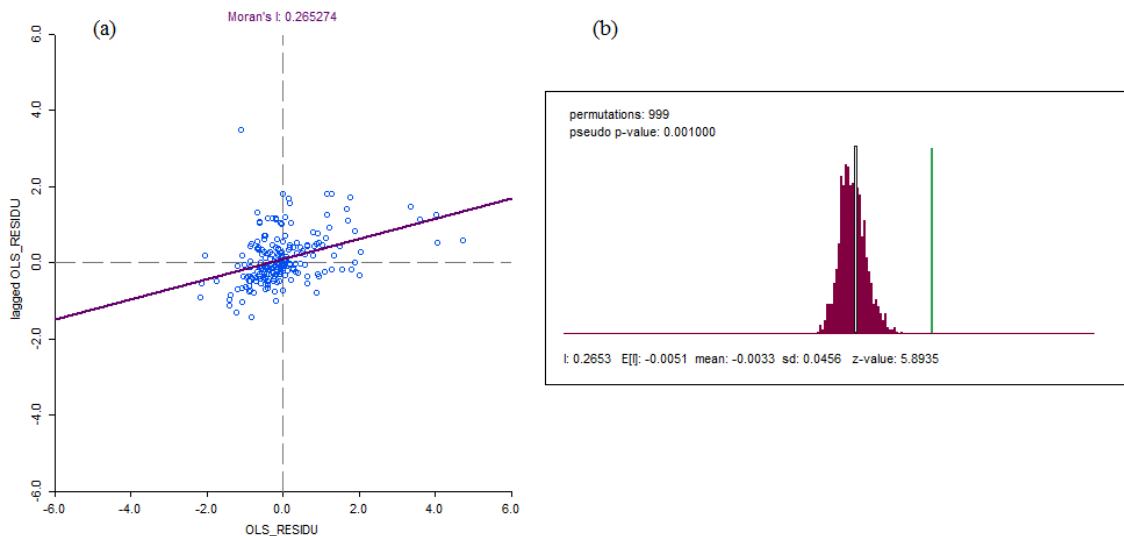


Figura 35: (a) Índice de Moran: Resíduos. (b) Teste de permutação.

O índice de Moran global encontrado para os resíduos foi de $I=0,27$ possuindo significância estatística, figura 35(b), com valor $p=0,001$ para 999 permutações.

Embora a estatística I de Moran tenha grande poder na detecção de especificações erradas no modelo, e não apenas na autocorrelação espacial, não é utilizada para sugerir qual modelo alternativo deve ser usado. Para isso, usa-se as estatísticas dos testes do multiplicador de Lagrange (LM)⁶, (Anselin, 2005).

⁶O teste do Multiplicador de Lagrange, foi utilizado para comparar o Modelo de Regressão Linear Clássico com os modelos de Defasagem Espacial e de Erro Espacial.

Dependência Espacial			
Matriz de contingência W		Critério Rook	
Diagnóstico - Testes	Graus de Liberdade	Valor	p
Índice de Moran (erro)	0,2653(Índice)	6,2042	0,0000
Multiplicador de Lagrange LM (lag)	1	41,4958	0,0000
LM Robusto (lag)	1	7,7055	0,0055
Multiplicador de Lagrange LM (erro)	1	33,9761	0,0000
LM Robusto (erro)	1	0,1858	0,6665

Tabela 4: Diagnóstico de Dependência Espacial.

Neste caso, foram exibidos, conforme tabela 4, quatro estatísticas diferentes para o multiplicador de Lagrange: os dois primeiros (LM-Lag e LM-Lag Robusto) tomam o modelo de defasagem espacial como alternativo. Os outros dois (LM-Erro e Robust LM-Erro) referem-se ao modelo de erro espacial como a alternativo. Vale ressaltar que os testes LM Robustos devem ser considerados quando as versões padrão (LM-Lag ou LM-Erro) são significativas. Quando não estiverem, as propriedades das versões robustas podem não ser válidas.

Conforme a tabela 4, as estatísticas LM-Lag e LM-Erro foram altamente significativas. A rejeição da hipótese nula por ambas as estatísticas é uma situação comumente encontrada na prática, podendo então considerar as formas robustas das estatísticas. Sendo assim, de acordo com o esquema apresentado em Anselin (2005) deve-se continuar com o processo de regressão conforme figura 36. Sendo mais adequado para o ajuste, o modelo de defasagem espacial.

Para fins de comparação, também foi ajustado aos dados o modelo de erro espacial.

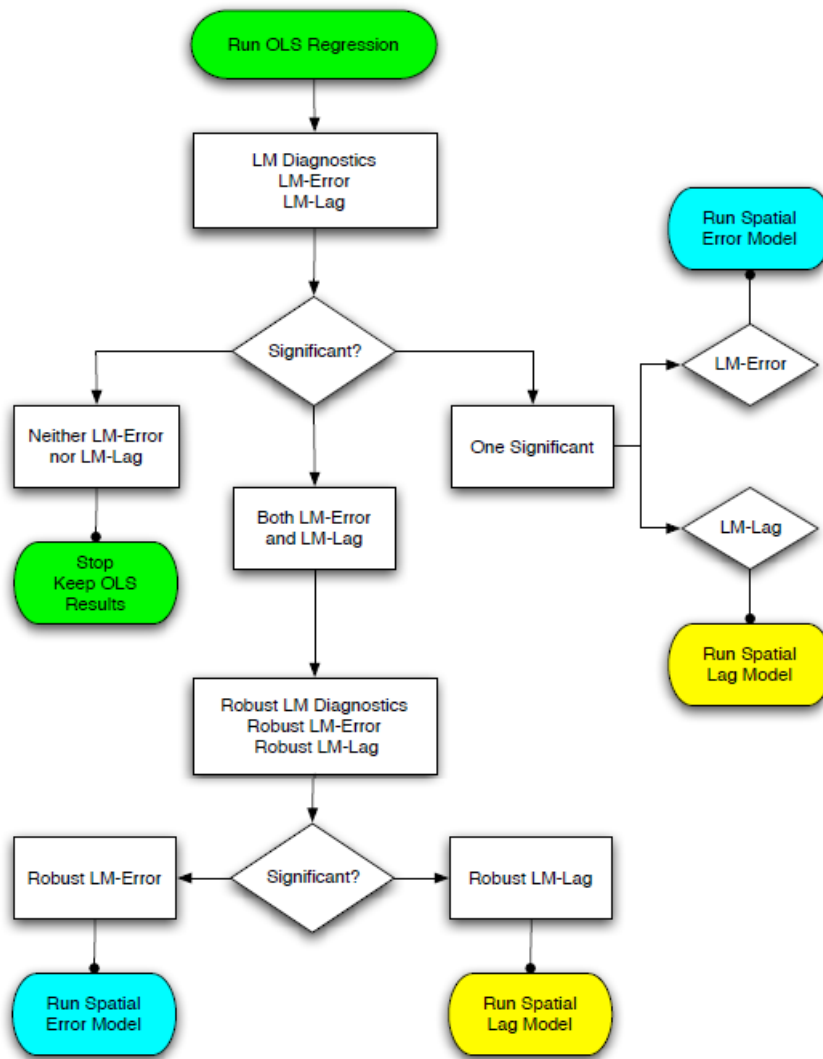


Figura 36: Processo para regressão espacial. Fonte: Anselin (2005).

3.3.2 Modelo de Defasagem Espacial

O modelo com defasagens espaciais atribui a autocorrelação espacial à variável resposta Y , definido por:

$$Y = \rho WY + X\beta + \varepsilon. \quad (30)$$

- ▷ Y : contagem dos casos de dengue, variável dependente;
- ▷ X : matriz que contém as variáveis socioeconômicas *pop*, *alf*, *agua*, *lixo*, *renda* e *morad*, variáveis independentes;

- ▷ W : matriz de vizinhança baseada no critério rook de contingência de primeira ordem;
- ▷ ρ : coeficiente espacial autoregressivo;
- ▷ β : coeficientes de regressão.

Utilizando o software GeoDa, os valores encontrados para a regressão estão apresentados na tabela 5, a estimação foi realizada via máxima verossimilhança. Neste caso, as variáveis *renda*, *morad* e o termo constante também não foram considerados significativos ao nível de 5%, sendo assim, ajustou-se outra regressão, agora sem a variável *renda* e *morad*, a fim de obter melhores resultados (tabela 6).

Regressão				
Variável	Coeficiente	Erroo-Padrão	z	p
ρ	0,42937	0,08256	5,20101	0,00000
<i>Constante</i>	-1,46788	4,60768	-0,31857	0,75005
<i>pop</i>	0,01795	0,00789	2,27728	0,02277
<i>alf</i>	-0,04765	0,02526	-1,88666	0,05921
<i>agua</i>	1,13631	0,33449	3,39712	0,00068
<i>lixo</i>	-1,09545	0,41385	-2,64698	0,00812
<i>renda</i>	0,00030	0,00066	0,44925	0,65325
<i>morad</i>	0,66517	1,57059	0,42351	0,67192
R^2	0,39095			
LIK	-669,282			
AIC	1381,03			

Tabela 5: Regressão: Modelo de Defasagem Espacial.

Regressão				
Variável	Coefficiente	Erro-Padrão	<i>z</i>	<i>p</i>
ρ	0,42788	0,08241	5,19240	0,00000
<i>Constante</i>	0,90847	1,78841	0,50798	0,61147
<i>pop</i>	0,01861	0,01786	3,73527	0,00019
<i>alf</i>	-0,04921	0,01786	-2,75515	0,00587
<i>agua</i>	1,14806	0,33242	3,45268	0,00055
<i>lixo</i>	-1,08007	0,39755	-2,71676	0,00659
R^2	0,38928			
LIK	-669,622			
AIC	1351,240			

Tabela 6: Regressão: Modelo de Defasagem Espacial, novo ajuste.

Logo, o modelo adotado possui as variáveis apresentadas na tabela 7.

Variáveis	Coefficiente	Probabilidade
ρ	0,4279	0,0000
<i>pop</i>	0,0186	0,0001
<i>alf</i>	-0,0492	0,0059
<i>agua</i>	1,1481	0,0006
<i>lixo</i>	-1,0801	0,0066

Tabela 7: Modelo de defasagem espacial.

Em relação ao diagnóstico, a primeira parte corresponde ao teste de Breusch-Pagan para heterocedasticidade nos termos de erro, o valor 84,4342 é significativo, indicando que a heterocedasticidade ainda é um problema.

O teste de verossimilhança compara o modelo nulo (regressão linear clássica) com o modelo alternativo (de defasagem espacial). O valor 29,0930 confirma que o coeficiente autorregressivo espacial foi significativo. Os resultados obtidos são apresentados na tabela 8.

Diagnóstico de Regressão			
Diagnóstico para Heterocedasticidade	Graus de Liberdade	Valor	p
Breusch-Pagan	4	84,4342	0,00000
Diagnóstico: Dependência Esp, W	Graus de Liberdade	Valor	p
Likelihood Ratio Test	1	29,0930	0,00000

Tabela 8: Diagnóstico Modelo de Defasagem Espacial.

Como utilizou-se modelos de regressão espacial, espera-se que os resíduos não sejam autocorrelacionados espacialmente, indicando assim que o modelo capturou todo o padrão espacial. De fato, calculado o índice global de Moran ($-0,041$) o teste de permutação aleatória constatou a não existência de autocorrelação espacial nos resíduos do modelo de defasagem espacial, conforme a figura 37(a) e 37(b).

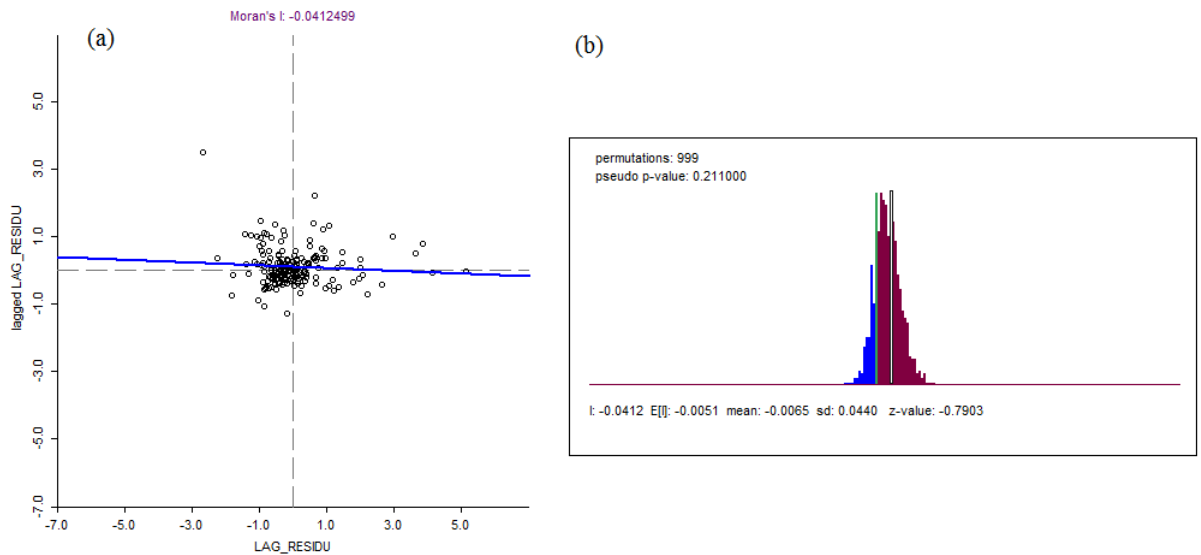


Figura 37: (a) Índice de Moran para os resíduos do modelo de defasagem espacial. (b) Teste de permutação.

3.3.3 Modelo de Erro Espacial

Este tipo de modelo de regressão espacial, considera que os efeitos da autocorrelação espacial estejam associados ao termo de erro ε . O modelo é definido por:

$$Y = X\beta + \varepsilon, \quad (31)$$

$$\varepsilon = \lambda W\varepsilon + U.$$

- ▷ Y : contagem dos casos de dengue;
- ▷ X : matriz que contém as variáveis socioeconômicas *pop*, *alf*, *agua*, *lixo*, *renda* e *morad*, variáveis independentes;
- ▷ W : matriz de vizinhança baseada no critério Rook de contingência de primeira ordem;
- ▷ λ : coeficiente espacial autoregressivo;
- ▷ β : coeficientes de regressão;
- ▷ U : erros aleatórios com média zero e variância $\sigma^2 I$.

Utilizando o software GeoDa, os valores encontrados para a regressão estão apresentados na tabela 9, a estimação foi realizada via máxima verossimilhança. Note que, as variáveis *renda*, *morad* também não foram consideradas significativas ao nível de 5%, bem como o termo constante. Com isso, ajustou-se uma nova regressão sem as variáveis *renda* e *morad*, com intuito de obter melhores resultados, conforme tabela 10.

Regressão				
Variável	Coeficiente	Erro-Padrão	z	p
λ	0,44081	0,08648	5,09747	0,00000
<i>Constante</i>	-0,98712	4,75807	-0,20746	0,83565
<i>pop</i>	0,02302	0,00874	2,63519	0,00841
<i>alf</i>	-0,06357	0,02805	-2,26600	0,02345
<i>agua</i>	0,98153	0,31943	3,07272	0,00212
<i>lixo</i>	-0,88191	0,39199	-2,24978	0,02446
<i>renda</i>	0,00038	0,00008	0,49941	0,61749
<i>morad</i>	1,32079	1,60605	0,82239	0,41086
R^2	0,380513			
LIK	-671,2892			
AIC	1379,56			

Tabela 9: Regressão: Modelo de Erro Espacial.

Regressão				
Variável	Coefficiente	Erro-Padrão	z	p
λ	0,43422	0,08701	4,99042	0,00000
<i>Constante</i>	3,34401	1,87523	1,78325	0,07454
<i>pop</i>	0,02516	0,00621	4,05428	0,00005
<i>alf</i>	-0,06939	0,02102	-3,30140	0,00096
<i>agua</i>	0,99278	0,31947	3,10764	0,00189
<i>lixo</i>	-0,84602	0,37869	-2,23041	0,02548
R^2	0,375813			
LIK	-671,89794			
AIC	1353,800			

Tabela 10: Regressão: Modelo de Erro Espacial, novo ajuste.

Logo, o modelo adotado possui as seguintes variáveis, de acordo com a tabela 11.

Variáveis	Coefficiente	Probabilidade
<i>pop</i>	0,0251612	0,00005
<i>alf</i>	-0,0693965	0,00096
<i>agua</i>	0,9927800	0,00189
<i>lixo</i>	-0,8460190	0,02548
λ	0,4342260	0,00000

Tabela 11: Modelo de erro espacial.

Em relação ao diagnóstico, o teste de Breusch-Pagan para heterocedasticidade nos termos de erro, obteve-se o valor 90,6699 significativo, indicando que a heterocedasticidade ainda é um problema. O teste de verossimilhança que compara o modelo nulo (regressão linear clássica) com o modelo alternativo (de erro espacial) apresentou valor igual a 24,5405 confirmando que o coeficiente autotregressivo espacial foi significativo para o modelo, tabela 12.

Contudo, assim como no modelo de defasagem espacial, espera-se que os resíduos não sejam autocorrelacionados espacialmente, indicando que o modelo capturou todo o padrão espacial. Através do índice global de Moran e teste de permutação aleatória, constatou-se a não existência de autocorrelação espacial nos resíduos do modelo de erro espacial, conforme a figura 38(a).

Diagnóstico de Regressão			
Diagnóstico para Heterocedasticidade	Graus de Liberdade	Valor	p
Breusch-Pagan	4	90,6699	0,00000
Diagnóstico: Dependência Esp, W	Graus de Liberdade	Valor	p
Likelihood Ratio Test	1	24,5405	0,00000

Tabela 12: Diagnóstico Modelo de Erro Espacial.

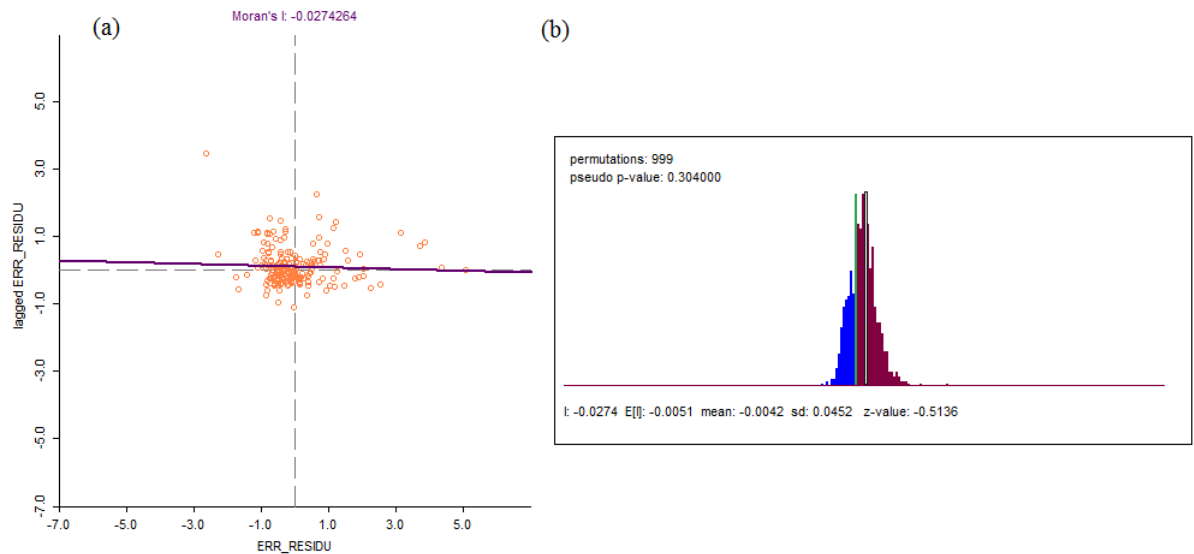


Figura 38: (a) Índice de Moran para os resíduos do modelo de erro espacial. (b) Teste de permutação.

O índice de Moran global encontrado para os resíduos foi de $I=-0,03$ com valor $p=0,3040$ para 999 permutações, figura 38(b).

A escolha do modelo apropriado, do ponto de vista estatístico, é extremamente importante na análise, busca-se o modelo mais parcimonioso. O critério de seleção utilizado nesta análise foi o Critério de Informação de Akaike (AIC) e o Logaritmo da Verossimilhança (LIK), (Anselin, 2005). A tabela 13 apresenta os três modelos ajustados e os critérios de comparação.

Modelos	LIK	AIC
Clássico	-684,145	1380,290
Defasagem Espacial	-669,662	1351,240
Erro Espacial	-671,898	1353,800

Tabela 13: Critérios de comparação.

Analisando os valores de AIC e LIK obtidos para os modelos, o modelo de defasagem espacial possui o menor AIC dentre os três e maior valor para o LIK, indicando que o mesmo deve ajustar melhor os dados de dengue para Rio Claro – SP.

Quanto às variáveis escolhidas, observou-se maior incidência de dengue em áreas com alta densidade populacional, áreas esta que, em geral, são as mais desiguais, do ponto de vista social. Essa associação pode ser explicada pela desigualdade social, regiões centrais apresentam menor densidade comparada a regiões periféricas. As variáveis que melhor explicaram a incidência de dengue foram densidade populacional, alfabetização, abastecimento de água e lixo inapropriados, indicando que condições socioeconômicas podem afetar a saúde.

Ao analisar o sinal dos coeficientes, nota-se que a variável relacionada a alfabetização apresenta sinal negativo, indicando que quanto maior o número de famílias com responsáveis alfabetizados, menor o número de casos de dengue, bem como para a variável que descreve o descarte de lixo; porém ao analisar a cidade de Rio Claro–SP como um todo, grande parte da cidade possui coleta de lixo apropriada, sendo assim, poucas áreas possuem lixo descartado erroneamente, talvez esta seja uma justificativa para a relação negativa entre lixo e dengue.

As variáveis densidade populacional e abastecimento de água apresentam valores positivos, são relacionadas com a dengue indicando que áreas com alta densidade populacional e abastecimento de água não convencional apresentam condições mais propensas para ocorrência de dengue, sendo estas regiões localizadas à noroeste e norte de Rio Claro–SP, que são áreas industriais do município.

A figura 39(a) apresenta a distribuição dos dados observados de dengue (2011) e a figura 39(b) apresenta os casos preditos pelo modelo de defasagem espacial.

Ao analisar os resíduos do modelo de defasagem espacial, é possível observar através do mapa de desvios, figura 40, onde o modelo menos aproxima das observações reais, seja superestimando ou subestimando-as, além de observar que possuem uma distribuição aleatória.

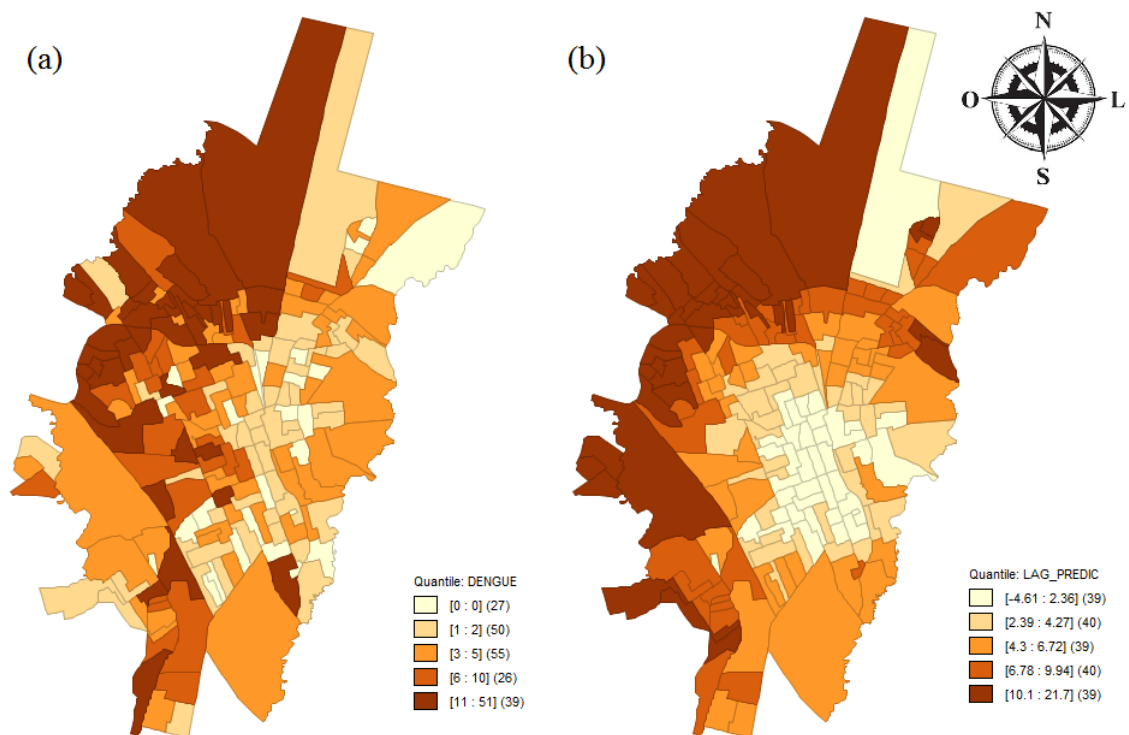


Figura 39: (a) Distribuição dos casos observados de dengue em Rio Claro – SP, 2011 e (b) Distribuição dos casos de dengue previsto pelo modelo de defasagem espacial em Rio Claro – SP, 2011.

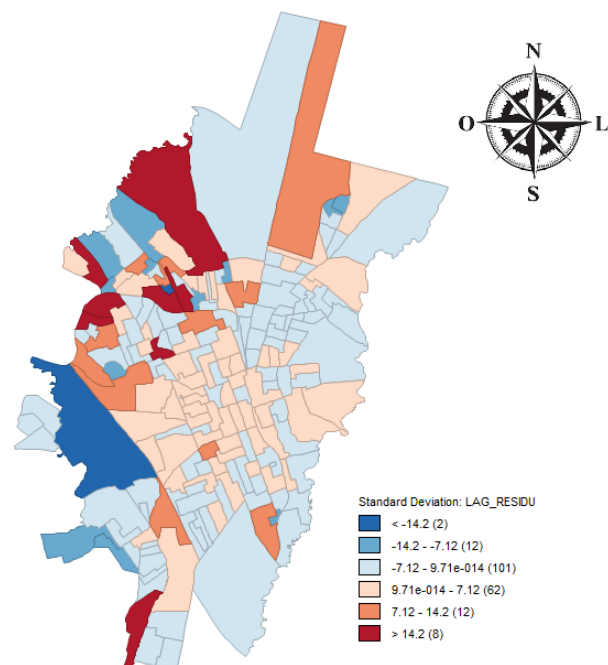


Figura 40: Distribuição espacial dos resíduos.

3.4 Conclusões

Portanto, através das aplicações, identificou-se regiões com maior incidência de casos de dengue e também quais variáveis socioeconômicas estão relacionadas com esta incidência. Com intuito de verificar a autocorrelação espacial, aplicou-se o índice de Moran global e local e, neste caso o índice global indicou autocorrelação entre as áreas, detectando agrupamentos e áreas de transição para as variáveis em estudo. E através do índice de Moran local, foram localizados *clusters* e *outliers* espaciais.

Ajustou-se três modelos de regressão: regressão linear clássica, de defasagem e de erro espacial, propondo um modelo que pudesse relacionar a dengue com as variáveis socioeconômicas escolhidas para o estudo. Na regressão linear clássica os resultados obtidos foram considerados insatisfatórios devido a quebra das suposições usuais além de não incorporar em sua estrutura a dependência espacial, visto que a análise espacial dos resíduos mostrou haver tendência espacial. Sendo assim, os modelos com efeitos globais que incorporam estrutura espacial foram analisados, e se mostraram mais eficientes quando comparados à regressão linear clássica. Ao comparar os dois métodos alternativos com a regressão clássica, os melhores resultados foram obtidos pelos modelos que incorporam a dependência espacial, indicando que o acréscimo do parâmetro espacial auxilia significativamente na explicação das variáveis.

O modelo que melhor se ajustou aos dados foi o modelo de defasagem espacial indicando a região noroeste da cidade de Rio Claro – SP como a de maior incidência, o que pode ser uma ferramenta útil para as autoridades de saúde responsáveis, que podem tratar esta área com maior ênfase e auxiliar nas campanhas de prevenção para anos futuros.

Em futuras pesquisas, sugere-se incorporar a variável tempo, explorando os modelos espaço-temporais e aplicando a ocorrências de dengue de outros anos, de modo a verificar como a distribuição espacial se comporta nos diferentes anos.

Referências

- [1] Almeida A.S.; Medronho R.A.; Valencia L.I.O. *Análise espacial da dengue e o contexto socioeconômico no município do Rio de Janeiro, RJ*. Rev Saúde Pública 2009;43(4):666-73.
- [2] Almeida M.A.S.; Lopardi A.L.S. *Crime contra o patrimônio na região metropolitana de Belo Horizonte: uma análise espacial*. Vianna Sapiens, Juiz de Fora v.1n.2-Out/2010-issn 21773726.
- [3] Anselin L. Chapter fourteen - *Spatial Econometrics; A Companion to Theoretical Econometrics*. Edited by Badi H. Baltagi (2003).
- [4] Anselin L. *Exploring Spatial Data with GeoDa™ : A Workbook*. *Spatial Analysis*. Laboratory Department of Geography - University of Illinois, Urbana-Champaign Urbana, IL 61801. Revised Version, March 6, 2005.
- [5] Anselin L., Ibnu S. and Youngihn K. *GeoDa: An Introduction to Spatial Data Analysis*. Geographical Analysis 38 (1), 5-22. (2006).
- [6] Arab A., Hooten M.B., Wikle C.K. *Hierarchical Spatial Models*. Department of Statistics, University of Missouri-Columbia. June 2006.
- [7] Assunção R. *Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia*. Associação Brasileira de Estatística, 1-131. (2001).
- [8] Assunção R.M.; Reis I.A. and Oliveira C.L. *Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model*. STATISTICS IN MEDICINE Statist. Med. 2001; 20:2319–2335 (DOI: 10.1002/sim.844).
- [9] Bailey C.T. and Grattel C.A. *Interactive Spatial Data Analysis*. (1995).

- [10] Best N.; Richardson S. and Thomson A. - *A comparison of Bayesian spatial models for disease mapping*. Stat Methods Med Res February 2005 14: 35-59.
- [11] Câmara G.; Carvalho M.S.; Cruz O.G.; Correa V. *Análise espacial de áreas*. In. Embrapa Cerrados. Análise espacial de dados geográficos. Planaltina, DF: Embrapa; 2004. p.157-209.
- [12] Carvalho, M.S. and SOUZA-SANTOS, R. Análise de dados espaciais em saúde pública: métodos, problemas, perspectivas. Cad. Saúde Pública [online]. 2005, vol.21, n.2 [cited 2016-09-09], pp.361-378.
- [13] Castro M.S.M.; Vieira V.A.; Assunção R.M. *Space-time analysis of lung cancer mortality in Southern Brazil*. Rev. Bras. Epidemiol. Vol. 7, Nº 2, 2004.
- [14] Ceccato V. *Homicide in São Paulo, Brazil: Assessing spatial-temporal and weather variations*. Institute of Criminology University of Cambridge - UK. Journal of Environmental Psychology, 25:249-360, 2005.
- [15] Clayton D. *Generalized linear mixed models In: Gilks WR, Richardson S, Spiegelhalter DJ, eds. Markov Chain Monte Carlo in practice* (1996). London: Chapman and Hall, 275–301.
- [16] Clayton D. E., Kaldor J. *Empirical Bayes estimates of agestandardized relative risks for use in disease mapping*. Biometrics 1987; 43: 671-81.
- [17] COSEMS/SP, *EM DEFESA DO SUS - 14/11/2012 15:22:22. Jornal do COSEMS/SP - Conselho de Secretários Municipais de Saúde do Estado de São Paulo “Dr. Sebastião de Moraes”*. Edição nº133, Novembro de 2012.
- [18] Costa A.I.P. *Distribuição espacial da dengue e determinantes socioeconômicos em localidade urbana no Sudeste do Brasil*. Rev. Saúde Pública, 32 (3): 232-6, 1998.
- [19] Domingues J., Almeida M.F.F., Brancaglione V.A., Govone J.S. *Análise da Autocorrelação Espacial dos setores censitários de Rio Claro-SP para os casos de dengue, através do Índice de Moran*. Rbras - 2016 - Salvador – Bahia.
- [20] Ehlers R.S. *Introdução a Inferência Bayesiana*. Versão Revisada em junho de 2003.

- [21] Haining R.P. *Spatial data analysis: theory and practice* / Robert Haining. First published 2003. Printed in the United Kingdom at the University Press, Cambridge.
- [22] IBGE - *Instituto Brasileiro de Geografia e Estatística*
- [23] Portal da Saúde – Ministério da Saúde – www.saude.gov.br.
- [24] INPE: *Instituto Nacional de Pesquisas Espaciais*.
- [25] Kuhn K., Campbell-Lendrum D., Haines A., Cox J., Corvalan C., Anker M. *Using climate to predict infectious disease epidemics. World Health Organization*, Geneva, 55pp. (2005).
- [26] Lagazio C.; Dreassi E. and Biggeri A. *A hierarchical Bayesian model for space-time variation of disease risk*. Statistical Modelling 2001; 1: 17–29.
- [27] Lawson A. *Statistical methods in spatial epidemiology*. Sussex: John Wiley and Sons. (2001).
- [28] LeSage James P. *Revue d'économie industrielle* 123 (3e trimestre 2008).
- [29] LeSage James P, Pace K.R. - *Introduction to Spatial Econometrics*. (2009).
- [30] Lowe et al. *Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil*. Computers & Geosciences 37 (2011) 371–381.
- [31] Lowe et al. *Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts*. (2014). 14:619–26. Published Online. May 17, 2014. S1473-3099(14)70781-9.
- [32] Lunn D.J., Thomas A., Best N. and Spiegelhalter D. (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.
- [33] MINISTÉRIO DA SAÚDE – Secretaria de Vigilância em Saúde e Secretaria de Atenção à Saúde. *O Agente Comunitário de Saúde no controle da dengue*. Brasília/DF 2009.
- [34] Meira K.C., Guimarães R.M., Santos J., Cabrelli R. *Análise de efeito idade-período-coorte na mortalidade por câncer de mama no Brasil e regiões*. Rev Panam Salud Publica. 2015;37(6):402–8.

- [35] Moran, P.A.P. The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* 37, 243- 251. (1948).
- [36] Prado R.R.; Castilho E.A. *A epidemia de aids no Estado de São Paulo: uma aplicação do modelo espaço-temporal bayesiano completo*. *Revista da Sociedade Brasileira de Medicina Tropical* 42(5):537-542, set-out, 2009.
- [37] Pinheiro L.C. *Alguns Aspectos de Modelos Espaço-Temporais*. Dissertação - Departamento de Estatística da Universidade Federal de Minas Gerais.
- [38] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- [39] Rossi P.E.; Allenby G.M.; McCulloch R. *Bayesian Statistic and Marketing*. John Wiley & Sons,Ltd, 2005.
- [40] Santos S.M.; Souza W.V. Organizadores. Fundação Oswaldo Cruz. *Introdução à Estatística Espacial para a Saúde Pública - Brasília : Ministério da Saúde*, 2007.120 p. : il. – (Série B. Textos Básicos de Saúde) (Série Capacitação e Atualização em Geoprocessamento em Saúde; 3).
- [41] Serrano R.M.; Valcarce E.V.; *Técnicas econométricas para el tratamiento de datos espaciales: la econometría espacial*. Barcelona, 2000.
- [42] Tobler W.R., 1979. *Cellular geography. Philosophy in geography*, pp. 379-386.
- [43] Vieira D.M.*Análise espaço-temporal na estimação de taxas de incidência/mortalidade*. Dissertação PPG em Medicina: Epidemiologia. UFRGS, Porto Alegre-RS 2006.
- [44] Vieira R.S. Crescimento econômico no estado de São Paulo: uma análise espacial [online]. São Paulo: Editora UNESP; São Paulo: Cultura Acadêmica, 2009. 103 p. ISBN 978-85-7983-013-6.
- [45] Waller L.A, Gotway C.A. *Applied spatial statistics for public health data*. WILEY-INTERSCIENCE. A John Wiley & Sons, INC, Publication. (2004).
- [46] Ward, M.D. & Gleditsch, K.S. (2008). *Quantitative Applications in the Social Sciences: Spatial regression models*: SAGE Publications Ltd doi: 10.4135/9781412985888.

- [47] Werneck, G.L.; Struchiner, C.J. *Estudos de agregados de doença no espaço-tempo: conceitos, técnicas e desafios*. Cad. Saúde Públ., Rio de Janeiro, 13(4):611-624, out-dez, 1997.
- [48] Ywata C.A.X.; Albuquerque P.H.M. *Métodos e Modelos em Econometria Espacial, uma Revisão*. Rev. Bras. Biom., São Paulo, v.29, n.2, p.273-306, 2011.