LOURENÇO MANUEL

MODELOS DE REGRESSÃO LINEAR COM EFEITOS ESPACIAIS NA ANÁLISE DA MORTALIDADE INFANTIL

LAVRAS – MG
2011

LOURENÇO MANUEL

MODELOS DE REGRESSÃO LINEAR COM EFEITOS ESPACIAIS NA ANÁLISE DA MORTALIDADE INFANTIL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para obtenção do título de Mestre.

Orientador

Dr. João Domingos Scalon

LAVRAS –MG
2011

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Manuel, Lourenço.
Modelos de regressão linear com efeitos espaciais na análise da mortalidade infantil / Lourenço Manuel. – Lavras : UFLA, 2011.
82 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2011.
Orientador: João Domingos Scalon.
Bibliografia.

1. Modelo espacial autorregressivo. 2. Mortalidade neonatal. 3. Dados de área. 4. Predição em modelos espaciais. 5. Saúde pública. I. Universidade Federal de Lavras. II. Título.

CDD – 519.536

LOURENÇO MANUEL

MODELOS DE REGRESSÃO LINEAR COM EFEITOS ESPACIAIS NA ANÁLISE DA MORTALIDADE INFANTIL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para obtenção de título de Mestre.

APROVADA em 20 de dezembro de 2011

Dr. Renato Ribeiro de Lima	UFLA
Dr. Denismar Alves Nogueira	UNIFAL

Dr. João Domingos Scalon

Orientador

LAVRAS –MG
2011

Aos meus filhos, Kaid Gauss e Klaus Milvan, pelo imensurável amor que sinto por eles.

Ao meu pai Manuel Uasse pelo amor, carinho e ensinamentos.

A minha falecida irmã Laura Manuel que em vida sempre me impulsionou na carreira estudantil.

Aos meus irmãos, Luísa Manuel, Carlota Manuel, Fernando Manuel, Adelina Manuel e Álvaro Manuel, que apesar da distância demonstraram sempre o seu amor, carinho e amizade.

Aos meus sobrinhos, Bernardo Massango, Resalda Massango, Charla Dimbane e Shelsia José, pelo amor e carinho.

Em especial a minha querida mãe Marta Xavier Manhisse

DEDICO

AGRADECIMENTOS

Em primeiro lugar a Deus pela saúde e em segundo a todos que de forma direta e indireta contribuíram para a efetivação deste trabalho.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Ciências Exatas (DEX) pela oportunidade concedida para a realização do mestrado.

Ao Professor Doutor João Domingos Scalon pela orientação, amizade e pelas contribuições científicas inestimáveis dadas para o enriquecimento deste trabalho.

Aos Professores Denismar Alves Nogueira e Renato Ribeiro de Lima pelas contribuições dadas para efetivação do trabalho.

Aos professosres do Departamento de Ciências Exatas pelos conhecimentos transmitidos durante esta caminhada.

Aos colegas e amigos do DEX, Marcelo Ribeiro, Edmundo Caetano, Juliano Bortolini, Adriana Costa, dentre outros, pela amizade e convivência que tivemos.

Aos amigos Lourena Arone, Marques Donça, Suluza Gafur, Nair Costa, Moisés Otameh, e outros aqui não citados, pela amizade e convivência.

À Universidade Eduardo Mondlane (UEM) e ao Departamento de Economia Agrária (DEA) por terem homologado a licença de formação para o nível de mestrado.

Ao CNPq pela concessão da bolsa de estudos, essencial para esta conquista.

RESUMO

A mortalidade infantil é uma das principais preocupações de muitos governos em programas de saúde pública por constituir um dos indicadores de avaliação de qualidade de vida. Neste trabalho estudou-se a distribuição espacial da ocorrência da mortalidade infantil na área urbana do município de Alfenas, MG, no período de 2000 a 2004, usando técnicas de análise espacial de áreas. A principal variável de interesse foi o número de óbitos com menos de um ano de idade e como variáveis explicativas foram consideradas, o número de mulheres em idade fértil, o número de mulheres em idade de risco gestacional, o número de mulheres analfabetas, a renda mensal da mulher, a renda mensal do homem, o número de residências com mais de seis moradores e a densidade demográfica do setor censitário. A dependência espacial entre as observações da mortalidade infantil foi avaliada através das estatísticas global e local de Moran. Na modelagem dos dados foram ajustados o modelo de regressão clássico, o modelo espacial autoregressivo (SAR) e o modelo de erro espacial (CAR). Verificou-se que, o número de mulheres em idade fértil e a renda mensal da mulher, são as covariáveis que exercem influência sobre os modelos, e que o parâmetro rho *(ρ)* que mede a dependência espacial nos modelos SAR e CAR foi negativo e significativamente diferente de zero, isto é, os valores da mortalidade infantil em áreas vizinhas tendem a ser dissimilares entre si. Pelo critério de informação de Akaike *(AIC)*, o modelo SAR foi considerado o melhor modelo e usou-se este modelo para identificar as áreas de maiores ocorrências da mortalidade infantil.

Palavras-chave: Modelo espacial autoregressivo. Dados de área. Predição em modelos espaciais. Mortalidade neonatal. Saúde pública.

ABSTRACT

Infant mortality is one of the main worries of many governments in programs of public health for constituting an important indicator used to evaluate the quality of life. In this work it was studied the spatial distribution of infant mortality in the urban area of Alfenas, MG, during the period of 2000 to 2004, using techniques of spatial lattice analysis. The main variable of interest was the number of deaths with less than one year old and the independent variables were, the number of women in fertile age, the number of women in age of gestational risk, the number of illiterate women, the monthly income of the woman, the monthly income of the man, the number of residences with more than six inhabitants and the demographic density of the census sector. The spatial dependence of observations of infant mortality was evaluated through global and local Moran indexes. In the modeling of the data, three models were adjusted, namely, the classic regression model, the spatial autoregressive model (SAR) and the conditional autoregressive model (CAR). It was verified that, the number of women in fertile age and the monthly income of the woman, are the variables having explanatory power on the models. It was verified that, the parameter rho *(ρ)* which measure the spatial dependence in SAR and CAR models, was negative and significantly different from zero, that is, the values of infant mortality in neighboring areas tend to be dissimilar between them. The Akaike information criteria *(AIC)* showed that the SAR model presented better goodness of fit than the other two models, and this model was used to identify areas with high risk of infant mortality.

Keywords: Spatial autoregressive model. Spatial lattice data. Prediction in spatial models. Neonatal mortality. Public health.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A taxa de mortalidade infantil é um dos mais importantes indicadores utilizados para medir a qualidade de vida de uma dada população numa região, estado ou país. Ela indica o número de óbitos com menos de um ano de idade por cada mil nascidos vivos, numa determinada área geográfica durante um período de tempo. A taxa de mortalidade infantil dá uma ideia da estimativa do risco de um nascido vivo morrer durante o seu primeiro ano de vida.

Em geral, valores altos da taxa de mortalidade infantil, estão aliados a níveis precários de assistência à saúde, baixas condições de vida e fraco desenvolvimento socioeconômico.

Numa determinada área geográfica, a taxa de mortalidade infantil tende a apresentar padrões diferenciados de uma zona para outra. Assim, a análise desse tipo de evento deve ser feita considerando a distribuição espacial do fenômeno usando técnicas de estatística espacial.

As técnicas de estatística espacial compreendem essencialmente a análise de dados de processos pontuais, dados de superfície aleatória e a análise de dados em áreas.

A análise de dados espaciais em áreas é mais adequada quando lidamos com eventos agregados por municípios, bairros ou setores censitários, onde não se dispõe da localização exata dos eventos, mas de um valor por área (DRUCK et al., 2004).

Essas técnicas de análise visam identificar padrões de distribuição do evento segundo alguma dependência espacial, isto é, procura-se avaliar a existência de autocorrelação espacial e identificar as possíveis variáveis explicativas dessa ocorrência.

Em problemas de saúde pública sobre a mortalidade infantil, o uso das técnicas de análise espacial, é de primordial importância, pois o mapeamento da

mortalidade infantil, permite uma compreensão do carácter geográfico da ocorrência do fenômeno e possibilita a identificação das áreas com maiores taxas. Segundo Andrade e Szwarcwald (2001), essas áreas, irão merecer especial atenção para delimitar estratégias de saúde pública em municípios, regiões do estado ou mesmo bairros de uma cidade.

O presente trabalho teve como objetivo principal a análise espacial da mortalidade infantil na área urbana do município de Alfenas usando técnicas de análise espacial de áreas. A principal variável de interesse foi o número de óbitos com menos de um ano de idade, e um conjunto de sete covariáveis. A modelagem dos dados foi feita usando o modelo de regressão múltipla e os modelos de regressão com efeitos espaciais globais, nomeadamente o modelo espacial autoregressivo (SAR – "*spatial autoregressive*") e o modelo de erro espacial (CAR – "*Conditional autoregressive*").

1.1 Objetivos específicos

a) Avaliar a presença da dependência espacial entre as observações da mortalidade infantil.
b) Ajustar um modelo autoregressivo espacial que melhor descreve o inter-relacionamento entre a mortalidade infantil e as diferentes covariáveis.
c) Deduzir equações de predição para cálculos dos valores preditos em modelos de regressão ajustados com transformação na variável dependente.
d) Identificar áreas nas quais há elevadas ocorrências de mortalidade infantil com base no(s) modelo(s) ajustado(s).

2 REFERENCIAL TEÓRICO

Este capítulo aborda cinco temas principais: a mortalidade infantil, análise de dados de área, o modelo de regressão linear, os modelos de regressão espacial e o cálculo dos valores preditos médios em modelos de regressão linear com transformação raíz quadrada na variável dependente.

2.1 Mortalidade infantil

A taxa de mortalidade infantil é definida como a razão entre o número de crianças mortas com menos de um ano de idade, nascidas vivas, pelo número de nascidos vivos num determinado período de tempo. Normalmente, essa taxa é multiplicada por 1000, para indicar a proporção das crianças que perderiam a vida numa situação de mil nascimentos.

Segundo Duarte (2007), as taxas de mortalidade infantil são classificadas em função da proximidade ou distância de valores já alcançados em sociedades mais desenvolvidas. Em geral, são consideradas altas, taxas maiores que 50‰, médias aquelas entre 20‰ e 50‰ e baixas as menores do que 20‰.

Dados mais recentes publicados pela Cia Word Factbook (2011) mostram que Singapura é o país que apresenta a taxa de mortalidade infantil mais baixa, ocupando a 1ª posição no *ranking* mundial com uma taxa de 2‰. Angola figura como o país que apresenta a taxa mais alta com cerca de 170‰, enquanto que o Brasil possui uma taxa de mortalidade infantil de 21‰, ocupando a 92ª posição no *ranking* mundial.

No Brasil, os dados de óbitos e nascimentos necessários para o cálculo da mortalidade infantil são originários de duas fontes principais: o Ministério da Saúde, através do Sistema de Informação de Mortalidade (SIM) e do Sistema de

Informação sobre Nascidos Vivos (SINASC), e o Instituto Brasileiro de Geografia e Estatística (IBGE), responsável pelas estatísticas do registro civil.

Caldeira et al. (2005) afirmam que a mortalidade infantil, também pode ser avaliada segundo os seus componentes neonatal e pós-neonatal. A neonatal compreende o número de crianças mortas durante os primeiros 27 dias de vida em cada mil nascidos vivos, e a pós-neonatal os óbitos ocorridos entre o 28º dia de vida até 11 meses e 29 dias de idade. O componente neonatal pode ser dividido ainda em precoce (0 a 6 dias) e tardio (7 a 27 dias).

Vários são os fatores apontados como causas da mortalidade infantil, como as condições biológicas maternas e infantis, a idade da mãe, intervalo entre os partos, prematuridade, baixo peso ao nascer, etc. As condições ambientais estão ligadas a existência dos serviços de saúde e da acessibilidade da população a eles, o abastecimento de água potável e saneamento básico adequados. Finalmente, as relações sociais que organizam a vida das pessoas tais como moradia, trabalho, renda e nível de informação (DUARTE, 2007).

Andrade e Szwarcwald (2001) num estudo sobre a mortalidade neonatal precoce no município de Rio de Janeiro, encontraram que os aglomerados que possuíam as taxas mais elevadas, eram explicados pelas características socioeconômicas das mães, refletidas tanto na ausência do acompanhamento pré-natal, como no acesso dificultado à assistência ao parto.

Abreu e Vasconcelos (1998) num estudo sobre fatores socioeconômicos ligados a mortalidade infantil, chegaram a conclusão de que fatores como o baixo nível de escolaridade das mães, baixa renda, desemprego e saneamento básico deficiente, contribuem substancialmente para o aumento da taxa de mortalidade infantil.

Autores como Menezes (1996) e Caldeira et al. (2005) afirmam que apesar das condições de vida desfavoráveis em alguns países do mundo, a taxa de mortalidade infantil, tanto a neonatal assim como a pós-neonatal, de uma

forma geral, têm demonstrado uma tendência de declínio em todo mundo, assim como no Brasil. Entretanto, a evolução de decréscimo não é homogênea no país, com um padrão que obedece às desigualdades regionais e sobretudo as diferenças das condições socioeconômicas (MENEZES et al., 1998).

Duarte (2007) afirma que a redução da mortalidade infantil no Brasil apresentou maior tendência de declínio, principalmente às custas da redução dos óbitos no período pós-neonatal por fatores fundamentalmente ligados à melhoria das condições de saneamento básico. Em contrapartida, as taxas de mortalidade no período neonatal, apresentam tendência de declínio muito lento, em virtude da permanência dos elevados níveis de mortalidade por fatores ligados à gestação e ao parto.

2.2 Estatística espacial

A estatística espacial é um ramo da estatística que estuda métodos científicos para a coleta, descrição, visualização e análise de dados que possam ser modelados como processos estocásticos, definidos como conjunto de variáveis aleatórias $\{Y(s_i): s_i \in A \subset R^2\}$, em que $Y(s_i)$ é a variável aleatória na coordenada s_i e A corresponde a região de estudo (ASSUNÇÃO, 2001).

A grande pecularidade da estatística espacial é que a informação espacial do fenômeno que está sendo analizado é incorporada nas análises.

O grande desenvolvimento computacional recente é uma das razões que levaram ao crescimento da área espacial na estatística. As técnicas de estatística espacial são aplicadas nas mais variadas áreas de conhecimento como ciências agronômicas, em análise de experimentos agrícolas, estudos ecológicos de comunidade de plantas, estudos de mineração, geologia, economia, epidemiologia, etc.

Segundo Cressie (1993), a tipologia dos dados espaciais divide-se em três áreas:

a) dados de processos pontuais;
b) dados de superfície aleatória (geoestatística);
c) dados de áreas.

Dados de processos pontuais - são aqueles em que se conhece a localização exata dos eventos como as residências de indivíduos doentes num dado município. Estes dados são representados por pontos normalmente numa superfície bidimensional. O grande objetivo nesse tipo de análise é identificar o padrão de ocorrência dos eventos, podendo ser classificado em agregado, regular ou aleatório.

Dados de superfície aleatória (também chamada de geoestatística) - são caracterizados pela continuidade espacial da variável aleatória de interesse. A variável de interesse $Y(s_i)$, pode ser a concentração de um determinado minério numa região. Nessa região, são feitas amostragens em alguns pontos e posteriormente estima-se o valor da variável aleatória $Y(s_i)$ em pontos não amostrados produzindo mapas de distribuição espacial da variável.

Dados de áreas - são caracterizados por constituirem valores agregados de uma determinada variável numa região ou mapa que encontra-se dividido em sub-áreas. Neste trabalho são abordados os métodos de análise inerentes a esse tipo de dados.

2.3 Análise de dados de área

Segundo Assunção (2001), os dados de área referem-se a um mapa particionado em áreas contíguas e disjuntas e, em cada uma delas medem-se

uma ou mais variáveis aleatórias $Y(s_i)$ e possivelmente covariáveis de interesse, que supostamente afetam a distribuição de probabilidade de $Y(s_i)$. Em outras palavras, a análise espacial de áreas é adequada nas situações em que a localização do evento a ser analisado está associado a áreas delimitadas por polígonos.

Druck et al. (2004) afirmam que esse caso ocorre com muita frequência quando lidamos com eventos agregados por municípios, bairros ou setores censitários, onde não se dispõe da localização exata dos eventos, mas de um valor por área. Alguns desses dados são contagens, como exemplo, o número de óbitos, partos, doenças transmissíveis por município, geralmente disponibilizados pelo Ministério e Secretarias de Saúde.

Embora o valor da variável de interesse esteja associado com toda a área e não com um ponto particular, associa-se este valor com um ponto específico dentro da área. Esse ponto é denominado por centroide, que corresponde ao centro de massa do polígono que delimita a área (ASSUNÇÃO, 2001).

Na prática, é comum associar esse ponto com a sede, no caso em que as áreas correspondem a municípios.

De acordo com Druck et al. (2004), a forma usual de apresentação de dados agregados por áreas é o uso de mapas coloridos, com o padrão espacial do fenômeno a ser analisado. Isso permite por um lado, uma visualização clara sobre como o fenômeno se “comporta” no espaço, dando a possibilidade de identificar áreas de maior ou menor magnitude de ocorrência desse evento. Por outro lado permite identificar áreas com características semelhantes ou diferentes desse fenômeno.

2.3.1 Estrutura da dependência espacial

Uma das suposições básicas feitas na estatística clássica é que as observações de uma variável aleatória são independentes. Em situações em que a variável aleatória de interesse encontra-se espacializada, a suposição de independência entre as observações pode não ser verdadeira, isto é, valores de uma variável aleatória em distâncias menores, tendem a ser mais parecidos do que observações em distâncias maiores (CRESSIE, 1993). Assim, surge uma necessidade de incorporar nas análises esse grau de similaridade (dissimilaridade) entre as observações.

Segundo Waller e Gotway (2004), na análise de dados de área, esse grau de similidaridade ou essa dependência espacial é avaliada atráves da autocorrelação espacial que pode ser medida através do índice de Moran. A aplicação desse índice depende da definição de uma matriz de vizinhança ou matriz de proximidade espacial.

2.3.1.1 Matriz de proximidade espacial

Segundo Werneck (2008) a matriz de proximidade espacial ou de vizinhança $(\boldsymbol{W})$ representa a estrutura da dependência espacial de uma variável aleatória em dados de áreas, ou seja, ela indica o grau de proximidade ou não entre observações.

Vários são os critérios usados para definir a matriz de vizinhança. Dado um conjunto de n áreas $\{A_1,..,A_n\}$, constrói-se a matriz $\boldsymbol{W}$ $(n \times n)$, onde cada um dos elementos w_{ij} representa uma medida de proximidade entre A_i e A_j.

Assunção (2001) aborda diferentes critérios utilizados na obtenção de $\boldsymbol{W}$, tais como:

a) w_{ij} = 1, se a área A_i compartilha de mesma fronteira com a área A_j *(i ≠ j)*, w_{ij} = 0 caso contrário;

b) w_{ij} = 1, se o centroide de A_j dista menos que "*k*" quilômetros do centroide de A_i e w_{ij} = 0, caso contrário;

c) $w_{ij} = L_{ij}/L_i$, onde L_{ij} é o comprimento da fronteira entre A_i e A_j e L_i é o perímetro de A_i.

Autores como Waller e Gotway (2004) e Druck et al. (2004) recomendam a normalização das linhas da matriz ***W*** dividindo cada elemento da matriz pelo total da linha e assim $w_i = \sum_{j=1}^{n} w_{ij} = 1$. Portanto, os pesos w_{ij} associados com a área A_i somam um.

As Figuras 1 e 2 ilustram, respectivamente, o exemplo da construção da matriz ***W*** usando como critério a fronteira entre as áreas no mapa de Moçambique, bem como a normalização dessa matriz nas linhas.

	M	G	I	S	Ma	Z	T	NP	NS	CD
M	0	1	0	0	0	0	0	0	0	0
G	1	0	1	0	1	0	0	0	0	0
I	0	1	0	1	1	0	0	0	0	0
S	0	0	1	0	1	1	1	0	0	0
Ma	0	1	1	1	0	0	1	0	0	0
Z	0	0	0	1	0	0	1	1	1	0
T	0	0	0	1	1	1	0	0	0	0
NP	0	0	0	0	0	1	0	0	1	1
NS	0	0	0	0	0	1	0	1	0	1
CD	0	0	0	0	0	0	0	1	1	0

Figura 1 Matriz de proximidade espacial usando como critério a fronteira entre as áreas no mapa de Moçambique, onde: M – Maputo, G – Gaza, I – Inhambane, S – Sofala, Ma - Manica, Z – Zambézia, T – Tete, NP – Nampula, NS – Niassa e CD – Cabo Delgado

	M	G	I	S	Ma	Z	T	NP	NS	CD
M	0	1	0	0	0	0	0	0	0	0
G	1/3	0	1/3	0	1/3	0	0	0	0	0
I	0	1/3	0	1/3	1/3	0	0	0	0	0
S	0	0	1/4	0	1/4	1/4	1/4	0	0	0
Ma	0	1/4	1/4	1/4	0	0	1/4	0	0	0
Z	0	0	0	1/4	0	0	1/4	1/4	1/4	0
T	0	0	0	1/3	1/3	1/3	0	0	0	0
NP	0	0	0	0	0	1/3	0	0	1/3	1/3
NS	0	0	0	0	0	1/3	0	1/3	0	1/3
CD	0	0	0	0	0	0	0	1/2	1/2	0

Figura 2 Matriz de proximidade espacial normalizada nas linhas, onde: M – Maputo, G – Gaza, I – Inhambane, S – Sofala, Ma - Manica, Z – Zambézia, T – Tete, NP – Nampula, NS – Niassa e CD – Cabo Delgado

2.3.1.2 Autocorrelação espacial

A autocorrelação espacial é uma medida que indica o grau de similaridade (dissimilaridade) de uma mesma variável no espaço. A ideia básica da autocorrelação espacial é determinar um indicador que mede a relação de uma determinada variável no espaço, isto é, como essa variável se comporta numa determinada região geográfica.

Segundo Druck et al. (2004), a autocorrelação espacial pode ser medida através do índice global e local de Moran, o índice de Geary e o variograma. Dentre esses indicadores, o índice de Moran é o mais usado em análise de dados de áreas.

a) Índice de Moran (indicador global)

Essa é a medida de autocorrelação espacial mais utilizada em análises espaciais de áreas. Almeida et al. (2008) afirmam que esta medida incorpora a similiaridade entre valores de uma determinada variável avaliada em áreas localizadas a uma distância pré-definida.

Segundo Waller e Gotway (2004), o índice de Moran é calculado por:

$$\hat{I} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \qquad (1)$$

em que:
n - é o número de áreas ou de observações;
Y_i - é a variável aleatória na área *i*;
Y_j - é a variável aleatória na área *j*;
$\bar{Y}$ - é a média amostral da variável aleatória em toda região ($\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$);
w_{ij} – são os elementos da matriz de proximidade espacial normalizada nas linhas.

Um valor positivo da estatística de Moran, indica que valores da variável em áreas vizinhas tendem a ser similares entre si (autocorrelação espacial positiva, podendo formar "*clusters*"). Já valores negativos indicam dissimilaridade entre os valores dessa variável em áreas vizinhas, ou seja, ausência de um padrão entre áreas vizinhas (SANTOS; SOUSA, 2007).

O valor esperado da estatística de Moran na ausência de autocorrelação espacial, conforme Cliff e Ord (1981) é dado por:

$$E[\hat{I}] = \frac{-1}{n-1}, \qquad (2)$$

que se aproxima de zero quando *n* aumenta.

Diferente de outros coeficientes de correlação, como o de Pearson e de Spearman, cujo valor do coeficiente encontra-se no intervalo [-1;1], Waller e Gotway (2004) afirmam que o índice de Moran pode assumir qualquer valor no conjunto dos reais. Porém, na maior parte dos casos ele encontra-se no intervalo [-1;1].

Uma vez calculado o índice de Moran, é importante fazer inferência sobre ele. De uma forma geral, o índice de Moran presta-se a um teste cuja hipótese nula é de independência espacial, nesse caso, seu valor seria zero (GRIFFITH, 2010).

Segundo Cliff e Ord (1981), o índice de Moran segue assintoticamente uma distribuição Normal com média e variância dadas pela equações 2 e 3, respectivamente.

$$Var[\hat{I}] = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n-1)(n+1)S_0^2} - \left(\frac{1}{n-1}\right)^2, \qquad (3)$$

com $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, $S_1 = {}^1/_2 \sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2$,
$S_2 = \sum_{i=1}^{n}(w_{i+} + w_{+j})^2$, com $w_{i+} = \sum_{j=1}^{n} w_{ij}$ e $w_{+j} = \sum_{i=1}^{n} w_{ij}$.

Assim, a significância da estatística de Moran pode ser avaliada com base no teste de Wald cuja estatística é dada por:

$$z = \frac{\hat{I} - E[\hat{I}]}{\sqrt{Var[\hat{I}]}}, \qquad (4)$$

em que $\hat{I}$ é dado pela equação (1), $E[\hat{I}]$ é dado pela equação (2) e $Var[\hat{I}]$ é dado pela equação (3).

O valor de z obtido na equação 4, corresponde a um determinado quantil da distribuição normal padronizada, que corresponde a um determinado *valor-p*. O índice de Moran será considerado significativamente diferente de zero se o *valor-p* for inferior ao nível nominal de significância previamente estabelecido.

É importante salientar ainda que o índice de Moran só deve ser aplicado se o processo estocástico $\{Y(s_i): s_i \in A \subset R^2\}$ apresenta estacionariedade de segunda ordem (SANTOS; SOUSA, 2007). De acordo com Cressie (1993), um processo estocástico é dito estacionário de segunda ordem se:

a) $E[Y(s_i)] = \mu$, ou seja a média do processo em qualquer ponto da região não depende da posição (s_i);
b) $E[Y^2(s_i)] = const \ \forall s_i$. Isto implica que a variância do processo é constante em toda a região de estudo e que a $COV[Y(s_i), Y(s_j)] = f(d_{ij})$ é uma função da distância entre s_i e s_j $\forall i \neq j$.

Porém, Mondini e Chiaravalloti (2008) afirmam que a estatística global de Moran dá apenas uma ideia global sobre a presença da autocorrelação espacial presente na variável, não identificando as áreas que são mais similares entre si, ou seja, o índice global de Moran não indica o conjunto de áreas que podem formar "*clusters*". Assim, observa-se a necessidade de evidenciar também um indicador local de autocorrelação espacial denominado "LISA".

b) Índice local de autocorrelação espacial ("LISA")

No índice local de autocorrelação espacial, cada unidade de espaço (área) é caracterizada por um único valor de índice. Segundo Pouliou e

Elliott (2009), esse índice indica a contribuição desse local no índice global da autocorrelação espacial (Moran), medida em todas as *n* áreas.

De acordo com Druck et al. (2004), o "LISA" é calculado por:

$$\hat{I}_i = \frac{Y_i \sum_{j=1}^{n} w_{ij} Y_j}{\sum_{j=1}^{n} Y_j^2}, \qquad (5)$$

em que:

$\hat{I}_i$ - é o índice de autocorrelação espacial na área *i*;

Y_i - é a variável aleatória na área *i*;

Y_j - é a variável aleatória na área *j*;

w_{ij} – são os elementos da matriz de proximidade espacial normalizada nas linhas.

Werneck (2008) afirma que esse índice é uma medida de autocorrelação espacial entre o valor de uma variável numa determinada área e os valores de seus vizinhos, permitindo evidenciar padrões significativos de associação espacial.

Em outras palavras pode-se dizer que o "LISA" indica em que medida os valores de alguma observação são similares ou diferentes das observações vizinhas. Isto permite ao índice $\hat{I}_i$ estar associado com cada unidade espacial (área), e requer que haja especificação da matriz de vizinhança ou de proximidade espacial (SANTOS; SOUSA, 2007).

Segundo Almeida et al. (2008), igualmente ao índice global de Moran, a estatística "LISA" não está estritamente limitada ao intervalo [-1; 1], mas seu valor se afasta de zero à medida que aumenta o grau de correlação positiva ou negativa. Na ausência de dependência espacial, esse valor é próximo de zero.

Depois de obtido o índice local de autocorrelação espacial é necessário fazer inferência sobre ele, isto é, avaliar a sua significância. Flahaut et al. (2002)

afirmam que uma vez avaliada a significância estatística “LISA”, é útil gerar um mapa indicando o grupo de áreas que podem formar “*clusters”* na região.

Santos e Sousa (2007), ressaltam que outra forma de identificar áreas com associação espacial é feita usando o diagrama de espalhamento de Moran. Nesse diagrama são colocados no eixo das abscissas os valores normalizados da variável *Z*, e no eixo das ordenadas a média dos seus vizinhos *WZ*, conforme ilustra o exemplo na Figura 3.

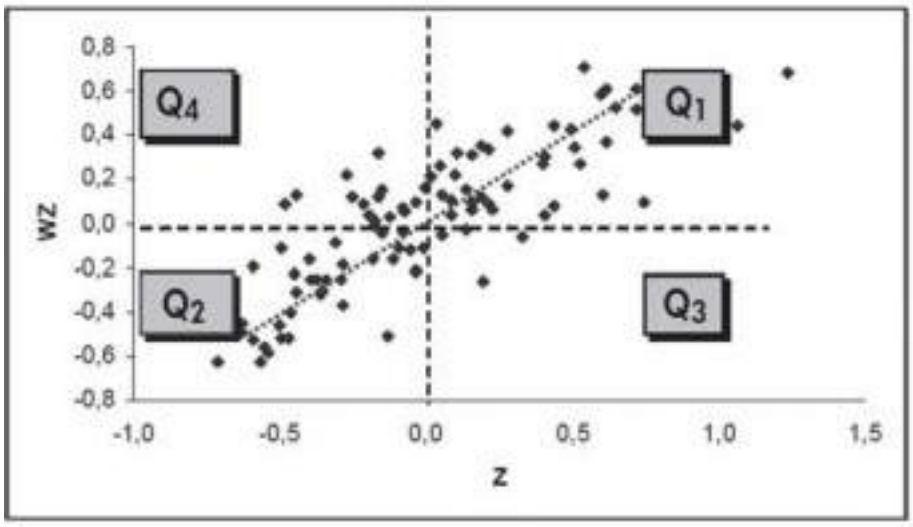

Figura 3 Exemplo de diagrama de espalhamento de Moran
Fonte: Druck et al. (2004)

Na Figura 3 os quadrantes Q_1 (valores positivos e médias positivas) e Q_2 (valores negativos e médias negativas) indicam áreas de associação espacial positiva, no sentido que uma localização possui vizinhos com valores semelhantes. Os quadrantes Q_3 (valores positivos e médias negativas) e Q_4 (valores negativos e médias positivas) indicam locais de associação espacial negativa, no sentido que uma localização possui vizinhos com valores distintos (SANTOS; SOUSA, 2007).

2.3.2 Estimação de indicadores (Taxas)

Na análise de dados de áreas, a variável aleatória de interesse $Y(s_i)$ normalmente representa um valor agregado da área A_i. Na maior parte dos casos, esses valores correspondem a dados agregados de contagens de determinados fenômenos, como por exemplo o número de óbitos com menos de um ano de idade, número de casos de uma determinada doença como a dengue, etc. Se o objetivo for analisar taxas de ocorrência desses fenômenos (razão entre o número de casos observados pelo total da população exposta ao fenômeno), o uso das taxas brutas levariam a conclusões erradas.

Por exemplo, ao comparar taxas de mortalidade infantil de cidades de um determinado Estado, as cidades com menor número de nascimentos, mesmo com baixos casos de ocorrência de mortes, indicarão valores altos de taxas, que podem levar a interpretações erradas para aquela cidade.

De acordo com Assunção (2001), em análise de taxas deve-se reestimar uma taxa mais próxima do risco real ao qual a população está exposta.

Segundo Druck et al. (2004), considera-se que a taxa "real" θ_i não é conhecida e dispõe-se de uma taxa observada $t_i = \frac{Y_i}{V_i}$ em que Y_i é o número de mortes observadas na i-ésima área e V_i é o número de nascimentos na i-ésima área. A estimação da taxa "real" é feita usando o estimador bayesiano que supõe que θ_i é uma variável aleatória com média μ_i e variância σ_i^2.

De acordo com Santos e Sousa (2007) o melhor estimador bayesiano é dado pela combinação linear entre a taxa observada t_i e a média μ_i , isto é,

$$\hat{\theta}_i = K_i t_i + (1 - K_i)\hat{\mu}_i, \qquad (6)$$

em que $\hat{\theta}_i$ é o estimador da taxa corrigida na i-ésima área e o fator K_i é dado por:

$$K_i = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + \hat{\mu}_i/V_i}, \quad (7)$$

em que:

$\hat{\sigma}_i^2$ - é o estimador da variância na i-ésima área;

$\hat{\mu}_i$ - é o estimador da média na i-ésima área;

V_i – é o número de nascimentos na i-ésima área.

Para encontrar os parâmetros μ_i e σ_i^2 da distribuição de θ_i, supõe-se que a distribuição da variável aleatória θ_i é a mesma em todas as áreas, e por isso todas as médias e variâncias são iguais. Assim, μ e σ^2 podem ser obtidos a partir dos dados. A média (μ) e a variância (σ^2) são estimados a partir das taxas observadas, ou seja,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} V_i} \quad (8)$$

e

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} V_i (t_i - \hat{\mu})^2}{\sum_{i=1}^{n} V_i} - \frac{\hat{\mu}}{\bar{V}}, \quad (9)$$

em que $\bar{V}$ é o número de nascimentos médio ($\bar{V} = \frac{\sum_{i=1}^{n} V_i}{n}$).

Contudo, vale salientar que o melhor estimador bayesiano parte da hipótese que a distribuição da variável aleatória θ_i é a mesma para todas as áreas. Em algumas situações, como em análises socioeconômicas, as características das populações estudadas são muito heterogêneas, mostrando que essa hipótese nem sempre é realista. Assim, uma alternativa seria o uso dos estimadores bayesianos completos que tornam possível resolver o problema, através da utilização de

técnicas de simulação baseadas em cadeias de Markov e Monte Carlo (MCMC) (DRUCK et al., 2004).

2.4 Modelos de regressão linear

Neste tema são abordados o modelo clássico de regressão e os modelos de regressão com efeitos espaciais.

2.4.1 Modelo clássico de regressão

Um modelo de regressão é uma ferramenta estatística que descreve a variação que ocorre numa variável aleatória *Y*, denominada variável dependente ou resposta, em função de um conjunto de outras variáveis denominadas variáveis independentes ou covariáveis *X*, mais uma compontente aleatória, ε (DRAPPER; SMITH, 1998).

Quando o modelo apresenta apenas uma covariável ele é denominado modelo de regressão simples, e quando apresenta duas ou mais variáveis independentes, o modelo é chamado de modelo de regressão múltipla. Segundo Drapper e Smith (1998), os modelos de regressão simples e múltiplos podem ser lineares ou não lineares. Um modelo é não linear em seus parâmetros se, pelo menos, uma de suas derivadas em relação aos parâmetros é dependente de, pelo menos, um desses parâmetros.

Charnet et al. (1999) afirmam que o objetivo geral de uma análise de regressão é quantificar a relação entre uma variável dependente e um conjunto de variáveis explicativas. Esta dissertação concentra-se apenas nos modelos lineares.

Na sua forma matricial, o modelo de regressão linear clássico (também chamado modelo Gauss-Markov ordinário), conforme Drapper e Smith (1998) é descrito por:

$$Y = X\beta + \varepsilon, \tag{10}$$

ou seja,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ .... \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & .... & X_{1p-1} \\ 1 & X_{21} & X_{22} & .... & X_{2p-1} \\ 1 & X_{31} & X_{32} & .... & X_{3p-1} \\ .... & .... & .... & .... & .... \\ 1 & X_{n1} & X_{n2} & & X_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ .... \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ .... \\ \varepsilon_n \end{bmatrix}$$

em que:

Y - é um vetor *(n* x *1)* de observações da variável dependente;

X - é uma matriz *(n* x *p)* com *p-1* variáveis explicativas;

β - é um vetor *(p* x *1)* com os coeficientes de regressão;

ε - é um vetor *(n* x *1)* de erros aleatórios, ou resíduos, que se supõe seguir uma distribuição normal, com média zero e variância constante, isto é, $\varepsilon \sim N(0, I\sigma^2)$.

2.4.1.1 Estimação dos parâmetros no modelo clássico de regressão

A estimação dos parâmetros num modelo Gauss Markov ordinário, pode ser feita usando o método dos mínimos quadrados ordinários ou pelo método de máxima verossimilhança. O método dos mínimos quadrados ordinários consiste

em minimizar a soma de quadrados dos resíduos. Um resíduo é a diferença entre o valor observado *Y* e o valor estimado *E[Y] = Xβ*, isto é,

$$\varepsilon = Y - X\beta. \tag{11}$$

A soma dos quadrados dos resíduos *(SQE)* é definida por:

$$SQE = (Y - X\beta)'(Y - X\beta). \tag{12}$$

A minimização da soma dos quadrados dos resíduos, consiste em derivar a equação (12) em relação ao vetor de parâmetros β e igualar a zero e resolver o sistema de equações resultantes. O estimador de mínimos quadrados ordinários obtido é dado por:

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{13}$$

Pode-se demonstrar (CHARNET et al., 1999) que o estimador não viciado da variância (σ^2) é dado por:

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n - p}, \tag{14}$$

em que *n* é o número de observações e *p* é o número de parâmetros conforme pode ser visto em (10).

2.4.1.2 Análise das pressuposições sobre a variável aleatória erro

Em modelos de regressão linear, várias pressuposições são feitas para tornar os modelos ajustados válidos. Tais pressuposições são feitas sobre o

termo erro e consistem basicamente em assumir que os erros sejam independentes, homocedásticos e que seguem uma distribuição normal com média zero e variância constante, isto é, $\varepsilon \sim N(0, I\sigma^2)$.

Após ajustado o modelo de regressão, é necessário avaliar se tais pressuposições não foram violadas, para que as inferências feitas com base no modelo ajustado, sejam válidas. Caso esses pressupostos sejam violados, as inferências que possam ser feitas com base no modelo ajustado não serão válidas.

a) Pressuposição de normalidade

A hipótese de normalidade dos resíduos da regressão pode ser avaliada usando gráficos de probabilidade normal ou através de testes de normalidade, como o teste de Shapiro-Wilk, o teste de Jarque–Bera, dentre outros.

Alguns autores como Drapper e Smith (1998) e Charnet et al. (1999) afirmam que de uma forma geral a falta de normalidade não constitui problema pois os testes de *t* e *F* são robusto ao desvio da normalidade. Uma atenção à falta de normalidade será dada nos casos em que a distribuição dos dados for excessivamente assimétrica. Nesssas situações, para solucionar o problema de falta de normalidade pode-se aplicar uma transformação logaritmo ou raíz quadrada nos dados.

b) Pressuposição de homocedasticidade

A homocedasticidade dos resíduos, é uma das mais importantes pressuposições feitas neste tipo de modelos. Uma vez ajustado o modelo é de extrema importância que seja avaliada esta pressuposição, pois ela é que irá

garantir a eficiência dos testes de hipóteses, pois a estimativa da variância populacional (quadrado médio do resíduo) será não viesada.

Para avaliar a pressuposição da homocedasticidade dos resíduos podem ser usados vários testes estatísticos nomeadamente o de Hartley, Levene, Barttlet, Breusch-Pagan, dentre outros.

Segundo Anselin (2005), o teste de Breusch-Pagan, sob a hipótese de nulidade avalia a homocedasticidade dos resíduos. Esse teste, pertence a classe dos testes de multiplicadores de Lagrange e sua estatística é baseada numa "regressão auxiliar" dada por:

$$\varepsilon^2 = X\alpha + \gamma\,, \tag{16}$$

e a estatística do teste é definida por:

$$BP = nR^{*2}, \tag{17}$$

em que:

n – é o número de observações;

ε – vetor de resíduos da regressão;

X – é a matrix de incidência;

R^{*2} – Coeficiente de determinação da regressão auxiliar;

α- parâmetros a estimar na regressão auxiliar;

γ – resíduos da regressão auxiliar.

Taamouti e Dufour (2009) afirmam que sob a hipótese de nulidade, o teste de Breusch-Pagan segue uma distribuição de qui-quadrado com *p-1* graus de liberdade - $\chi^2_{(p-1)}$, em que *p* é o número de parâmetros do modelo.

Assim, em modelos de regressão, deseja-se que essa estatística seja menor que o quantil da distribuição qui-quadrado com *p* graus de liberdade a um nível de significância previamente estabelecido, e assim não violar a presuposição da homocedasticidade dos resíduos. Porém, na prática o valor da estatística de *BP* pode ser maior que o quantil de qui-quadrado. Assim, para corrigir o problema de heterocedasticidade deve-se usar os mínimos quadrados ponderados na estimação.

c) Pressuposição da independência

O princípio da independência é uma das mais importantes pressuposições feitas na estatística clássica. Num modelo de regressão linear, a independência dos resíduos pode ser avaliada usando o método gráfico ou através de teste de hipótese. Os testes para avaliar a independência dos resíduos são os testes de Durbin - Watson e o teste de Breusch GodFrey.

Uma das exigências do modelo descrito na equação (10) é que os erros devem ser independentes. No entanto, para o caso de dados espaciais, onde está presente a dependência espacial, é muito pouco provável que a hipótese padrão de independência dos erros seja verdadeira. Druck et al. (2004) afirmam que no caso mais comum os resíduos continuam apresentando a autocorrelação espacial presente nos dados.

Na situação dos dados espaciais, quando está presente a autocorrelação espacial, as estimativas do modelo devem incorporar essa estrutura espacial, uma vez que a dependência entre as observações altera o poder explicativo do modelo (DRUCK et al., 2004).

Segundo Cressie (1993), uma classe de modelos que incorpora a autocorrelação espacial existente entre as observações são o modelo espacial

autoregressivo (SAR ou "*lag model*") e o modelo condicional autoregressivo (CAR ou "*error model*").

2.4.2 Modelo espacial autoregressivo (SAR)

Em séries temporais, a classe de modelos autoregressivos, indica a variação que ocorre numa variável aleatória *Y* no instante de tempo *t* em função das observações passadas dessa mesma variável (MORETIN; TOLOI, 2006). No modelo SAR, Waller e Gotway (2004), afirmam que variação que ocorre na variável aleatória *Y* na posição (s_i), além de ser devida ao efeito das covariáveis, também é descrita pela variável de interesse em outros locais $Y(s_j)$, o que pode ser verificado através da autocorrelação espacial presente entre as observações.

Segundo Druck et al. (2004), no modelo SAR ou *lag model*, a autocorrelação espacial é incorporada num único parâmetro do modelo que é adicionado ao modelo de regressão clássica. Nesse caso, a autocorrelação espacial é incorporada na variável dependente.

O modelo espacial autoregressivo é descrito conforme Anselin (1999) por:

$$Y = X\beta + \rho WY + \varepsilon\,, \tag{18}$$

em que:

Y – é um vetor de observações *(nx1)* nas *n* áreas;

W - é a matriz de proximidade espacial;

X - é uma matriz *(n* x*p)* com *p-1* variáveis explicativas, medidas nas *n* áreas;

β - é o vetor de parâmetros *(p* x *1)*;

ρ - é o coeficiente espacial autoregressivo;

ε - é um vetor *(n* x *1)* de erros aleatórios não correlacionados que seguem uma distribuição normal com média zero e variância constante, isto é, $\varepsilon \sim N(0, I\sigma^2)$.

A hipótese nula para a não existência de autocorrelação espacial é que $\rho = 0$. A ideia básica neste modelo é incorporar a autocorrelação espacial como componente do modelo. Pode-se observar que na ausência de autocorrelação espacial ($\rho = 0$), o modelo espacial autoregressivo (equação 18) é o próprio modelo de regressão linear geral (equação 10).

2.4.2.1 Estimação dos parâmetros no modelo SAR

A estimação dos parâmetros no modelo SAR pode ser feita pelo método de máxima verossimilhança. Esse método consiste em maximizar a função de verossimilhança que dá a probabilidade máxima de obter a amostra já observada.

Uma das representações do modelo SAR apresentada por Anselin (1999) é dada por:

$$Y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon. \tag{19}$$

No modelo apresentado na equação (19), como a variável aleatória Y é uma combinação linear de variáveis aleatórias normais, então Y seguirá uma distribuição normal com média dada por:

$$E[Y] = (I - \rho W)^{-1} X\beta, \tag{20}$$

e a matriz de variâncias e covariâncias será definida por:

$$Var[Y] = \sum_Y = \sigma^2 (I - \rho W)^{-1} (I - \rho W')^{-1}, \tag{21}$$

ou seja, $Y \sim N((I-\rho W)^{-1}X\beta, \sum_Y)$ em que os elementos fora da diagonal principal da matrix $\sum_Y$ representam a autocorrelação espacial na variável aleatória *Y*.

Segundo Anselin (1999), para o caso do modelo espacial autoregressivo ou "*lag model*" representado por (18), o logarítmo da função de verossimilhança é definido como:

$$ln(L(\beta,\sigma,\rho|Y,X) = -{}^{n}/_{2}\, ln(2\pi) - {}^{n}/_{2}\, ln(\sigma^2) + ln|I-\rho W| \\ -(1/2\sigma^2)(Y-\rho WY-X\beta)'(Y-\rho WY-X\beta). \qquad (22)$$

Para encontrar os estimadores de máxima verossimilhança, deriva-se a equação (22) em relação aos parâmetros e iguala-se a zero, resolvendo o sistema de equações resultantes (conhecidas como equações de Escore). Porém, as equações de Escore não retornam uma solução fechada, exigindo que sejam usados métodos iterativos para obter as estimativas de máxima verossimilhança. Isto pode ser feito usando métodos de aproximação numérica, como o de Gauss-Newton ou o algorítmo de Newton-Raphson.

Bivand, Pebesma e Gomes (2008) afirmam que no modelo SAR, primeiro estima-se o coeficiente espacial autoregressivo (ρ) que maximiza (22) e posteriormente são estimados os outros parâmetros usando a solução dos mínimos quadrados generalizados (equação 24). A estimação do parâmetro ρ consiste em explorar a decomposição da matriz $|\boldsymbol{I}-\rho\boldsymbol{W}|$ em termos dos autovalores da matriz $\boldsymbol{W}$. Assim, tem-se:

$$ln(|I-\rho W|) = ln[\prod_{i=1}^{n}(1-\rho\omega_i)] \qquad (23)$$

em que ω_i são os autovalores da matriz $\boldsymbol{W}$. O parâmetro ρ na equação (23) é estimado através de métodos iterativos. Obtido o $\hat{\rho}$, este é substituído no estimador de mínimos quadrados generalizados dado por:

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y, \tag{24}$$

e o estimador da variância é definido por:

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})'V^{-1}(Y - X\hat{\beta})}{n - p}, \tag{25}$$

em que $V = (I - \hat{\rho}W)^{-1}(I - \hat{\rho}W')^{-1}$ reflete a autocorrelação espacial residual.

2.4.3 Modelo de erro espacial ("*Conditional autoregressive*" – *CAR*)

Nesse tipo de modelo, considera-se que os efeitos espaciais são ruídos, ou perturbações, ou seja, termo que precisa ser removido. Nesse caso, os efeitos da autocorrelação espacial estão associados ao termo erro (DRUCK et al., 2004).

O modelo de erro espacial conforme Anselin (1999) é dado por:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ \varepsilon &= \rho W\varepsilon + u \end{aligned} \tag{26}$$

em que:

Y – é um vetor de observações *(nx1)* nas *n* áreas;

W - é a matriz de proximidade espacial;

X - é uma matriz *(n* x*p)* com *p-1* variáveis explicativas;

β - é vetor de parâmetros *(p* x *1)*;

ρ - é o coeficiente espacial autoregressivo;

u - é a componente do erro não correlacionada que se supõe seguir uma distribuição normal com média zero e variância constante, isto é, $u \sim N(0, I\sigma^2)$.

A hipótese nula para a não existência de autocorrelação espacial é que $\rho = 0$, ou seja, o termo erro não é espacialmente correlacionado. Pode-se observar que na ausência de autocorrelação espacial ($\rho = 0$), o modelo de erro espacial (equação 26) é o próprio modelo de regressão linear geral (equação 10).

2.4.3.1 Estimação dos parâmetros no modelo CAR

A estimação dos parâmetros no modelo de erro espacial, através do método de máxima verossimilhança, é feita de forma similar à apresentada no modelo SAR.

Anselin (1999) mostra que uma das parametrizações do modelo CAR é dada por:

$$Y = X\beta + (I - \rho W)^{-1}u, \text{ com } u \sim N(0, I\sigma^2). \tag{27}$$

Nesse modelo, como a variável aleatória *Y* é uma combinação linear de variáveis aleatórias normais, então *Y* terá distribuição normal com média e matriz de variâncias e covariâncias definidas por:

$$E[Y] = X\beta \tag{28}$$

e

$$Var[Y] = \sum_Y = \sigma^2(I - \rho W)^{-1}(I - \rho W')^{-1}, \tag{29}$$

ou seja, $Y \sim N(X\beta, \sum_Y)$, em que os elementos fora da diagonal principal da matrix $\sum_Y$, representam a autocorrelação espacial na variável aleatória *Y*.

Como a variável aleatória *Y* segue uma distribuição normal com média e matriz de variâncias e covariâncias dadas pelas equações (28) e (29),

respectivamente, então o logarítmo da função de verossimilhança, conforme Anselin (1999), será dado por:

$$ln(L(\beta,\sigma,\rho|Y,X) = -{}^{n}/_{2}\, ln(2\pi) - {}^{n}/_{2}\, ln(\sigma^2) + ln|I - \rho W| \\ -(1/2\sigma^2)(Y - X\beta)'(I - \rho W)'(I - \rho W)(Y - X\beta). \qquad (30)$$

Para obter os estimadores de máxima verossimilhança, deriva-se a equação (30) em relação aos parâmetros e iguala-se a zero, resolvendo o sistema de equações resultantes. Porém, esse sistema de equações não é linear, exigindo que sejam usados métodos iterativos para obter as estimativas de máxima verossimilhança.

De acordo com Bivand, Pebesma e Gomes (2008), igualmente ao caso do modelo SAR, no modelo CAR, estima-se primeiro o coeficiente espacial autoregressivo (ρ) que maximiza a equação (30) e em seguida são estimados os outros parâmetros usando a solução dos mínimos quadrados generalizados. A estimação do parâmetro ρ é obtida pela equação (23) por um processo iterativo. Obtido o $\hat{\rho}$, este é substituído na equação (24) obtendo-se as estimativas dos parâmetros. Obtido o $\hat{\beta}$, este é substituído na equação (25) para encontrar a estimativa da variância.

2.4.4 Avaliação da necessidade de uso dos modelos espaciais

Na análise de dados de áreas, uma das formas de escolha para ajuste de um modelo espacial, é avaliar a presença da dependência espacial dos resíduos do modelo Gauss Markov ordinário (ANSELIN, 2005). A justificativa para uso dessa forma de análise, é que se o coeficiente espacial autoregressivo nos modelos espaciais for nulo ($\rho = 0$), estes modelos "transformam-se" num modelo clássico de regressão linear.

A presença da dependência espacial dos resíduos, pode ser analisada calculando a estatística de Moran dos resíduos, conforme Waller e Gotway (2004), ou seja,

$$\hat{I}_{res} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\varepsilon' W \varepsilon}{\varepsilon' \varepsilon}, \tag{31}$$

em que ε dado por (11), representa o vetor de resíduos do modelo de regressão clássica, W é a matriz de proximidade espacial, w_{ij} são os elementos da matriz de proximidade espacial e n o número de áreas.

O índice de Moran dos resíduos segue assimptoticamente uma distribuição normal com média e variância dadas pela equações (32) e (33), respectivamente (CLIFF; ORD, 1981).

$$E[\hat{I}_{res}] = \frac{n tr[(X'X)^{-1}X'WX]}{(n-p)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}, \tag{32}$$

em que p é o número de parâmetros e tr[.] é o traço da matriz, e

$$Var[\hat{I}_{res}] = \frac{n^2\{S_1(n-p) + 2tr(G^2)(n-p) - tr(F)(n-p) - 2[tr(G)]^2\}}{(S_0)^2(n-p)^2(n-p+2)} \tag{33}$$

com $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, $S_1 = {}^1/_2 \sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2$,
$F = (X'X)^{-1}X'(W + W')^2X$ e $G = (X'X)^{-1}X'WX$

A avaliação da significância do índice de Moran dos resíduos pode ser feita usando o teste de Wald cuja a estatística é apresentada na equação (4).

De acordo com Anselin (2005), a significância do índice de Moran, em um modelo de regressão linear, para além do teste de Wald, também pode ser avaliada usando o teste de razão de verossimilhança e os testes de multiplicadores de Lagrange.

Os testes de multiplicadores de Lagrange, não só permitem avaliar a presença da dependência espacial como também identificam o modelo espacial a ser ajustado (SAR ou CAR). Alguns desses testes abordados por Anselin (2005) são o "LM–Lag" e "LM–Error". O primeiro teste avalia a necessidade de ajuste de um modelo espacial autoregressivo (SAR) enquanto que o segundo avalia a necessidade de ajuste de um modelo de erro espacial (CAR).

Segundo Anselin (1999), a estatística do teste de multiplicadores de Lagrange para avaliar a necessidade de ajuste de um modelo SAR é dada por:

$$LM_{lag} = [\frac{n\varepsilon' WY}{\varepsilon'\varepsilon}]^2 \frac{1}{D} \qquad (34)$$

com $D = [(WX\hat{\beta})'(I - X(X'X)^{-1}X')(WX\hat{\beta})/\hat{\sigma}^2]$

em que:

Y – é um vetor (*nx1*) de observações;

n - é o número de observações;

ε – é um vetor (*nx1*) dos resíduos da regressão definido por (11);

W – é a matriz de proximidade espacial;

I – é uma matriz identidade;

X - é uma matriz *(n* x*p);*

$\hat{\beta}$ – é o estimador dos parâmetros definido por (13);

$\hat{\sigma}^2$ – é o estimador da variância definido por (14).

A estatística dada em (34), segue assintoticamente uma distribuição qui-quadrado com 1 grau de liberdade, $\chi^2_{(1)}$ (ANSELIN, 1999). Assim, se a estatística do teste dada em (34) for superior ao quantil da $\chi^2_{(1)}$ a um nível nominal de significância previamente estabelecido, deve-se ajustar um modelo SAR.

Para o ajuste de um modelo CAR, a estatística do teste de multiplicadores de Lagrange conforme Anselin (1999) é dada por:

$$LM_{err} = [\frac{n\varepsilon' W\varepsilon}{\varepsilon'\varepsilon}]^2 \frac{1}{tr[W^2 + W'W]} \qquad (35)$$

em que:

n - é o número de observações;

ε – é um vetor (*nx1*) dos resíduos da regressão;

W – é a matriz de proximidade espacial;

tr[.] – é o traço da matriz.

Segundo Anselin (1999), a estatística dada em (35), também segue assintoticamente uma distribuição $\chi^2_{(1)}$. Assim, se o valor desta estatística for superior ao quantil da $\chi^2_{(1)}$ a um nível nominal de significância previamente estabelecido, deve-se ajustar um modelo CAR.

2.5 Cálculo dos valores preditos médios para os modelos de regressão linear com transformação na variável dependente

O cálculo de valores médios preditos para os modelos de regressão linear clássico é bem conhecido na literatura (CHARNET et al., 1999). Para o caso de algumas transformações do tipo logaritmo e raíz quadrada na variável

dependente, existem algumas propostas para obter as predições médias nesse modelo (MILLER, 1984).

Os valores preditos para o modelo de regressão linear clássico ajustado com transformação na variável dependente podem ser determinados usando a proposta apresentada por Miller (1984). Para o caso de uma transformação raíz quadrada na variável dependente, a dedução da equação de predição é apresentada a seguir:

Dado o modelo $Y = \beta_0 + \beta_1 X + \varepsilon$, com $\varepsilon \sim N(0, \sigma^2)$. Se a variável dependente sofreu uma transformação raíz quadrada tem-se:

$$Y^{0.5} = \beta_0 + \beta_1 X + \varepsilon$$

Elevando ao quadrado ambos os membros da equação obtém-se:

$$Y = (\beta_0 + \beta_1 X + \varepsilon)^2$$

Aplicando a esperança em ambos os membros nessa equação tem-se:

$$E[Y] = E[(\beta_0 + \beta_1 X + \varepsilon)^2]$$

O segundo membro desta equação pode ser escrito como:

$$E[(\beta_0 + \beta_1 X + \varepsilon)^2] = Var[\beta_0 + \beta_1 X + \varepsilon] + \{E[\beta_0 + \beta_1 X + \varepsilon]\}^2$$

Que é equivalente a:

$$E[(\beta_0 + \beta_1 X + \varepsilon)^2] = \sigma^2 + (\beta_0 + \beta_1 X)^2$$

Deste modo obtem-se:

$$E[Y] = \sigma^2 + (\beta_0 + \beta_1 X)^2$$

Assim, após a transformação raíz quadrada, o estimador de *E[Y|X]* é dado por:

$$\hat{Y} = (\widehat{\beta_0} + \widehat{\beta_1} X)^2 + \hat{\sigma}^2 \quad (36)$$

em que o $\hat{\beta}_i$ e $\hat{\sigma}^2$ são as estimativas dos parâmetros do modelo e a variância, respectivamente (MILLER, 1984).

Para o caso dos modelos de regressão espacial, não foram encontradas, metodologias para obter os valores preditos médios no modelo ajustado com transformação na variável dependente. Assim, nesta dissertação procurou-se desenvolver equações de predição em modelos de regressão espacial ajustados com transformação raíz quadrada na variável dependente, usando as mesmas analogias apresentadas por Miller (1984). Este desenvolvimento é apresentado no Capítulo 4.2.1.

3 MATERIAL E MÉTODOS

3.1 Local do estudo

No estudo foram analisados dados de 68 setores censitários da área urbana do município de Alfenas, Minas Gerais, que se situa sob as coordenadas geográficas 21º25´45´´ Latitude Sul e 45º56´50´´ Longitude Oeste. Esse município localiza-se a uma altitude de 888 metros acima do nível médio do mar e possui uma área de extensão de cerca de 849,2 Km^2. Segundo o IBGE, o município de Alfenas é composto por 70 setores censitários que possuem residências de diferentes classes sociais (NOGUEIRA, 2008).

Nas Figuras 4 e 5 tem-se a região de estudo dividida em setores censitários indicando as coordenadas dos centroides e os nomes dos bairros, respectivamente. O contorno nessa figura, delimita a área urbana do município que corresponde a região de estudo.

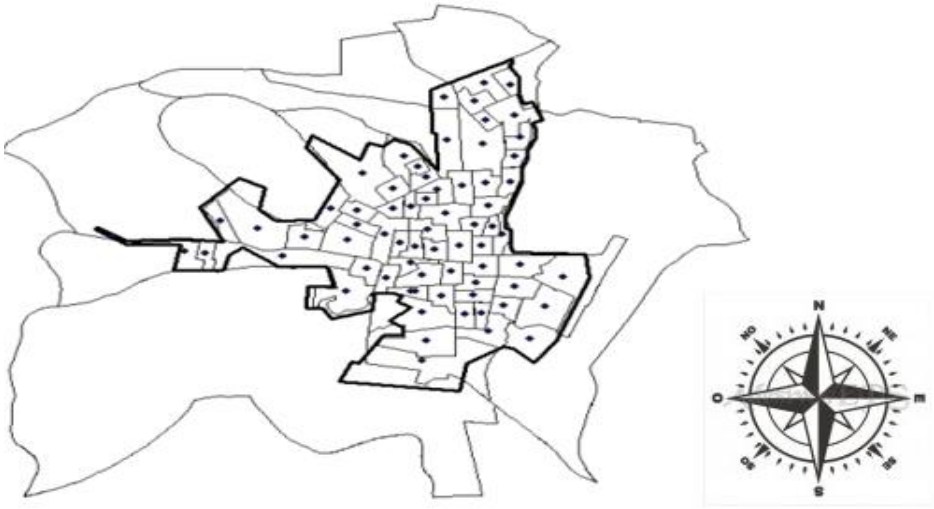

Figura 4 Mapa da área urbana do município de Alfenas exibindo os centroides dos setores censitários.
Fonte: Nogueira (2008)

Figura 5 Cartograma dos bairro da área urbana do município de Alfenas, MG
Fonte: Nogueira (2008)

3.2 Variáveis do estudo

No estudo foi usada uma base de dados contendo informação sobre 68 setores censitários da área urbana do município de Alfenas, MG. A base de dados é composta pelas coordenadas geográficas dos centroides de cada setor censitário mais oito variáveis (ANEXO A). Estas variáveis correspondem a caracterização da área urbana do município no período de 2000 a 2004.

A principal variável de interesse, foi o número de óbitos com menos de 1 ano de idade, e como covariáveis foram consideradas o número de mulheres em idade fértil, o número de mulheres em idade de risco gestacional, o número

de mulheres analfabetas em idade fértil, o número de residências com mais de 6 moradores, a renda mensal da mulher, a renda mensal do homem e a densidade demográfica do setor censitário. Todos os dados foram coletados por Nogueira (2008).

3.3 Análise estatística

As análises estatísticas foram realizadas utilizando os *softwares* R (versão 2.12.1) (R DEVELOPMENT CORE TEAM, 2010) com o pacote *spdep* (versão 0.5-29) (BIVAND et al., 2011), ArcGis (versão 9.3) e Geoda (GEODA CENTER FOR GEOSPATIAL ANALYSIS AND COMPUTATION, 2010).

3.3.1 Análise exploratória

A análise exploratória de todas as variáveis foi feita usando mapas coloridos, com o intuito de avaliar o comportamento espacial das mesmas. Foram determinados os índices global e local de Moran como indicadores da dependência espacial para a variável de resposta. Determinou-se também algumas medidas de posição e dispersão e foram construídos "*box maps*" para verificar possível presença de "*outliers*".

3.3.2 Ajuste dos modelos

A primeira etapa da modelagem, consistiu em fazer uma transformação dos dados na variável resposta. Como esta variável constitui uma contagem, foi realizada uma transformação raíz quadrada, para aproximar a variável resposta a uma distribuição normal (CHARNET et al., 1999).

Ajustou-se inicialmente um modelo de regressão linear clássico entre a variável resposta transformada e todas as covariáveis descritas em 3.2. Como algumas covariáveis apresentaram alta correlação entre si, aplicou-se o procedimento "*stepwise*" com vista a selecionar as covariáveis que têm maior poder explicativo sobre o modelo (CHARNET et al., 1999). Para o caso dos modelos de regressão espacial, aplicou-se o procedimento "*backward*".

As pressuposições básicas do modelo clássico de regressão, nomeadamente, a normalidade, homocedasticidade e independência dos erros, foram avaliados utilizando-se os testes de Shapiro-Wilk, Breusch-Pagan e Durbin Watson, respectivamente.

Também foram efetuados testes de multiplicadores de Lagrange sobre os resíduos do modelo de regressão clássico, com vista a avaliar a necessidade de ajuste de modelos SAR e CAR, dada a suposta presença da dependência espacial entre as observações na variável de resposta.

Comprovada a presença da dependência espacial nos resíduos, foram ajustados os modelos SAR e CAR.

A seleção do melhor modelo foi feita usando os seguintes critérios: coeficiente de determinação ajustado (R^2 ajustado), log verossimilhança e o Akaike (*AIC*).

Para calcular os valores preditos médios nos modelos de regressão espacial com transformação raíz quadrada na variável dependente, foram desenvolvidas equações de predição com base na abordagem apresentada em 2.5. A partir dessas equações, foram calculados os valores preditos médios e construídos mapas, para identificar áreas de maior incidência de mortalidade, ou seja, aquelas em que há elevadas ocorrências de mortalidade infantil.

4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados três resultados principais, nomeadamente, as análises exploratórias de todas as variáveis usadas no estudo, o ajuste dos modelos e são desenvolvidas equações de predição para o cálculo dos valores preditos em modelos de regressão espacial ajustados com transformação raíz quadrada na variável dependente.

4.1 Análise exploratória

Na Figura 6, tem-se a representação gráfica do número de ocorrências de óbitos com menos de um ano de idade, em cada setor censitário na área urbana do município de Alfenas-MG.

Figura 6 Número de casos de mortes de crianças com menos de 1 ano de idade em cada setor censitário na área urbana do município de Alfenas
Fonte: Cartograma do Instituto Brasileiro de Geografia e Estatística - IBGE (2000)

Para esta variável, pode-se notar que as ocorrências de mortes nos setores censitários, variam de zero até sete casos. Com base no "*box map*" (Figura 7), observa-se que não há presença de "*outliers*" e que não existe uma tendência clara de determinado(s) quartis se situarem numa zona específica da área de estudo. Isto, segundo Anselin (2005), mostra que esta variável apresenta estacionariedade de primeira ordem, ou seja, ela não possui tendência.

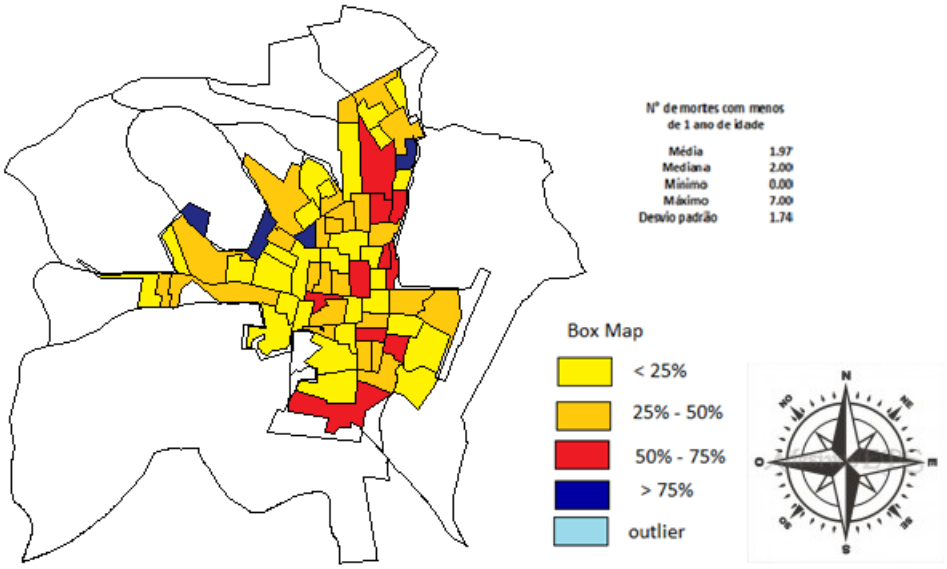

Figura 7 "*Box map*" para o número de crianças mortas com menos de 1 ano de idade na área urbana do minicípio de Alfenas

Na Tabela 1 tem-se a estimativa da estatística global de Moran bem como a avaliação de sua significância através do teste de Wald.

Tabela 1 Estimativa da estatística de Moran e avaliação de sua significância

Índice	**Estimativa**	**Valor esperado**	**Erro padrão**	**z**	**Valor - p**
Moran	-0,0814	-0,0149	0,0511	-1,300	0,9032

O valor da estatística de Moran é negativo, porém estatisticamente não difere de zero, o que significa que para esta variável há ausência de dependência

espacial. Contudo, esta é uma análise global, necessitando que seja feita também uma análise através do índice local de autocorrelação espacial (LISA).

Na Figura 8 está representado graficamente o mapa de significância de índice local de autocorrelação espacial "LISA *map*". Nesse mapa, observa-se que as áreas que apresentam coloração verde, possuem um valor do índice local de autocorelação espacial estatisticamente diferente de zero. Porém, nota-se que não existe uma tendência destas áreas formarem padrões espaciais agregados com seus vizinhos ("*clusters*"). Isso pode ser visualizado através do "LISA *cluster map*" na Figura 9.

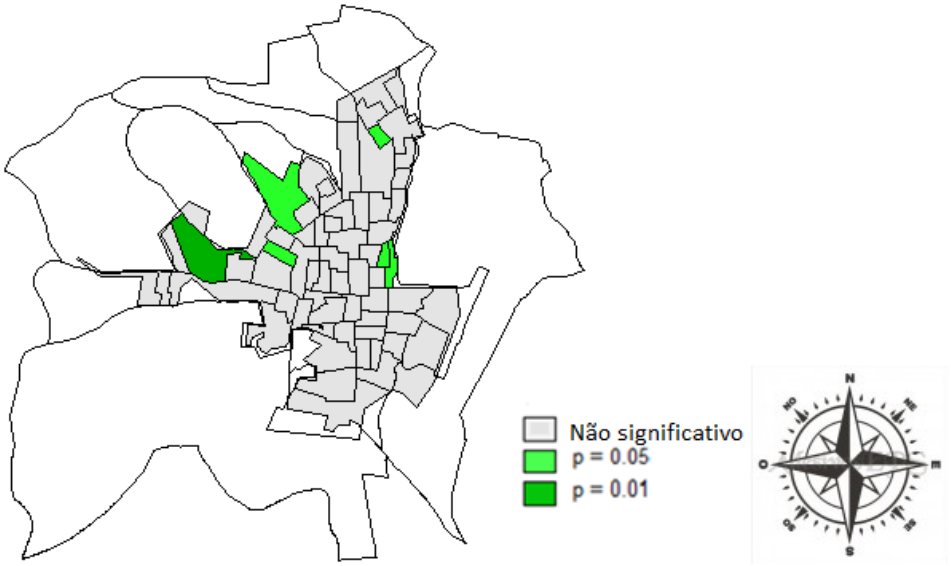

Figura 8 "LISA *map*" para o número de óbitos com menos de 1 ano de idade na área urbana do município de Alfenas

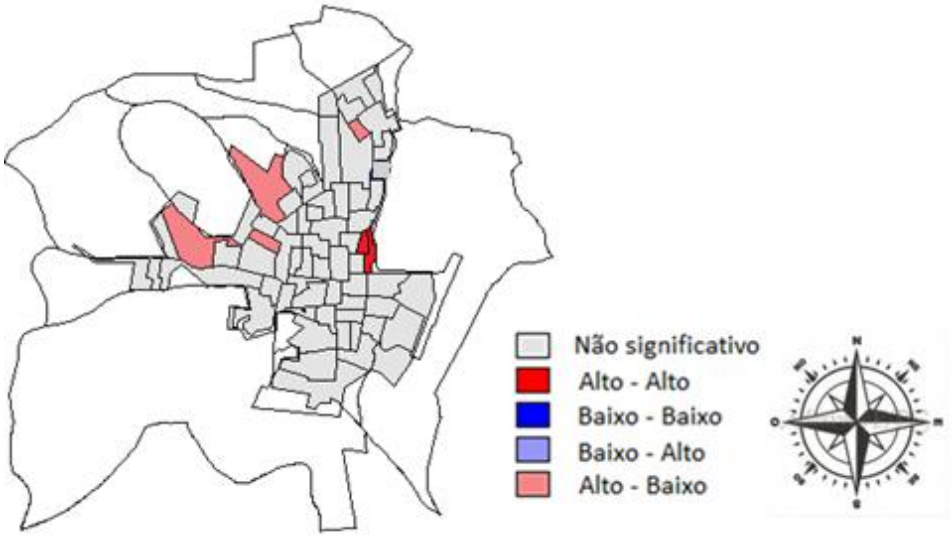

Figura 9 "LISA *cluster map*" para o número de crianças mortas com menos de 1 ano de idade na região urbana do município de Alfenas

Na Figura 9 as áreas que possuem designação "Alto – Alto" e "Baixo – Baixo" correspondem respectivamente, aos quadrantes Q_1 e Q_2 no diagrama de espalhamento de Moran, que são locais onde a autocorrelação espacial é positiva (Figura 3). As designações "Baixo – Alto" e "Alto – Baixo" correspondem aos respectivamente, aos quadrantes Q_3 e Q_4 no diagrama de espalhamento de Moran, que são locais de associação espacial negativa (Figura 3).

Pode-se notar que das seis áreas que apresentaram um "LISA" estatisticamente diferente de zero, duas correspondem ao quadrante Q_1 e as restantes quatro correspondem ao quadrante Q_4. Para o caso das áreas pertecentes ao quadrante Q_1, significa que a dependência espacial entre estas áreas é positiva, isto é, os valores da mortalidade infantil nessas áreas são similares entre si. Para o caso das áreas correspondentes ao quadrante Q_4, a dependência espacial nessas áreas, em relação aos seus vizinhos é negativa, ou seja, os valores da mortalidade infantil nessas áreas, tendem a ser dissimilares em áreas vizinhas.

Observa-se também que, a maior parte das áreas, não apresentou autocorrelação espacial local significativa, isto é, ausência da dependência espacial entre áreas vizinhas. Esta falta de significância do "LISA", em muitas áreas, pode ter contribuído significativamente, para um baixo índice global de Moran, estatisticamente não diferente de zero (Tabela 1).

Nas Figuras 10 a 15, estão representados os mapas da distribuição espacial das covariáveis usadas no presente estudo, assim como algumas estatísticas descritivas.

Para a número de mulheres em idade fértil, verifica-se que existem setores censitários com apenas 4 mulheres em idade fértil e outros setores que atingem 370 mulheres em idade fértil. A maior parte dos setores censitários incluíndo os do centro da cidade, apresentam valores entre 230 a 311 mulheres. Valores maiores que 311 mulheres em idade fértil, são observados em alguns setores que se localizam na zona norte da área de estudo nos bairros Jardim Eunice, Vila Esperança, Jardim São Carlos e Campo Elísio que fazem parte da periferia do município (Figura 10).

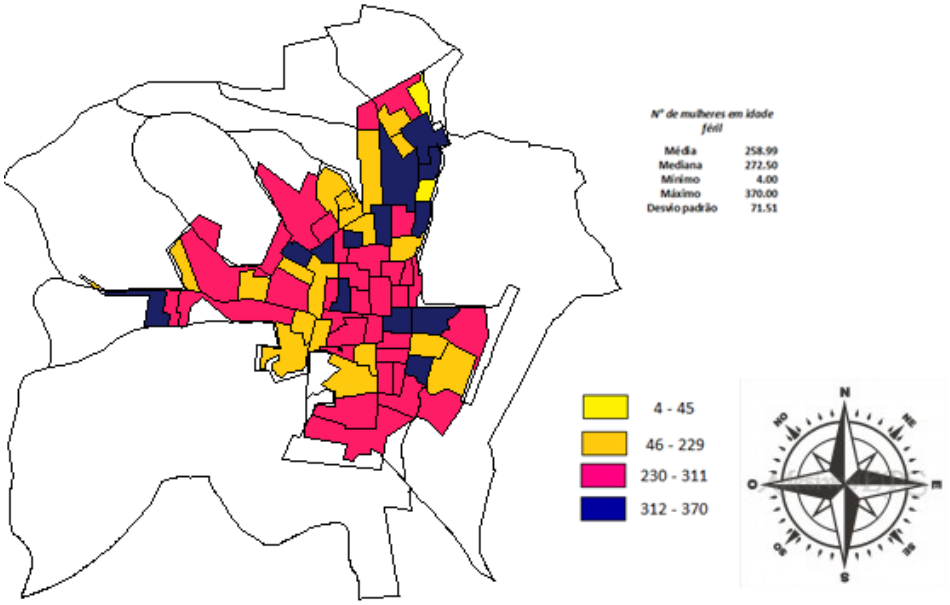

Figura 10 Análise exploratória para o número de mulheres em idade fértil na área urbana do município de Alfenas com base nos setores censitários

Na Figura 11 tem-se o mapa da distribuição espacial do número de mulheres em idade de risco gestacional, isto é, mulheres entre 15 e 19 anos e mulheres com idade compreendida entre 34 e 49 anos de idade. Essa variável apresenta um comportamento espacial similar ao comportamento da variável número de mulheres em idade fértil. Segundo Nogueira (2008), o número de mulheres em idade de risco gestacional, possui informação sobre o risco biológico.

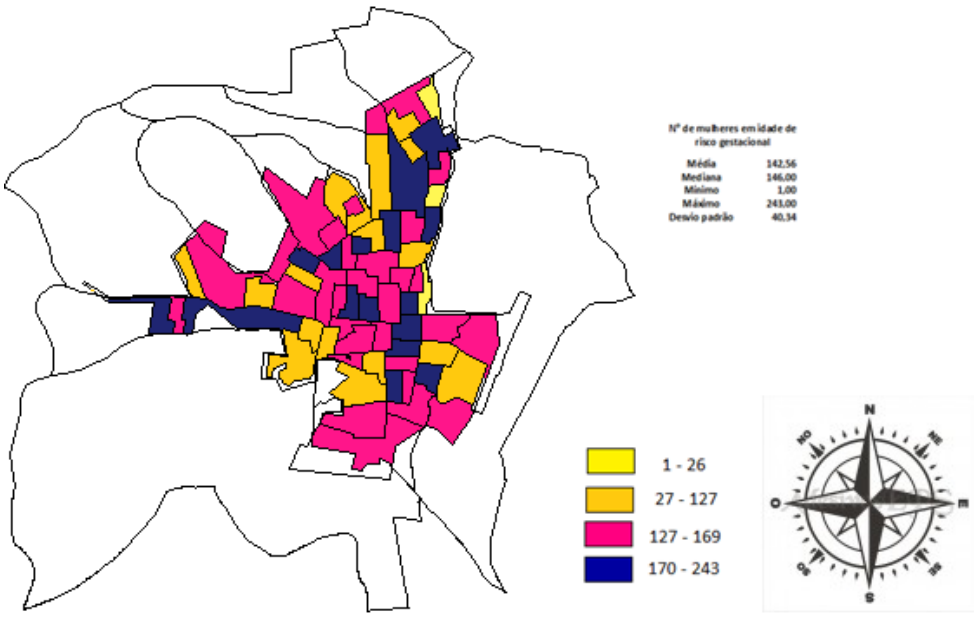

Figura 11 Análise exploratória para o número de mulheres em idade de risco gestacional na área urbana do município de Alfenas

Para a variável número de mulheres em idade fértil analfabetas com idade compreendida entre 20 e 34 anos de idade, que segundo Nogueira (2008) é uma variável que representa o risco social, verifica-se que a maior parte dos setores censitários apresenta valores que variam de 0 a 6 mulheres analfabetas (Figura 12). Dos 68 setores censitários, apenas nove setores possuem valores acima de 7 mulheres analfabetas, que em geral, tendem a localizar-se na zona norte do município, nos bairros Jardim Eunice, Vila Grande, Campo Elísio e Chapada salvo algumas ocorrências em setores censitários próximos do centro da cidade.

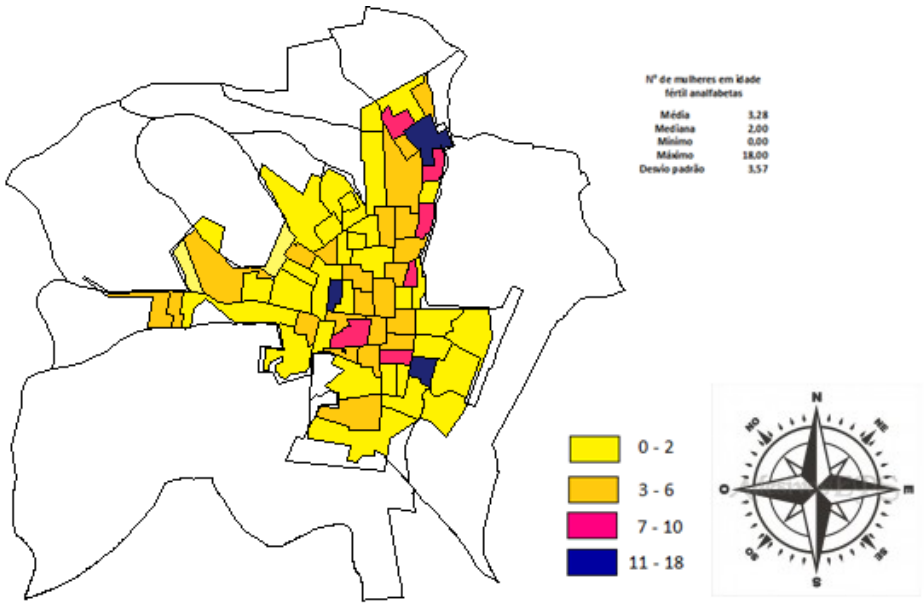

Figura 12 Análise exploratória para o número de mulheres em idade fértil analfabetas na área urbana no município de Alfenas

Na Figura 13 está representado o mapa da distribuição espacial do número de residências com mais de 6 moradores na área urbana do municício de Alfenas. Verifica-se que existem setores censitários com até 63 residências que possuem um agregado familiar composto por mais de 6 indivíduos. De uma forma geral observa-se que maior parte dos setores censitários possuem de 12 a 42 residências com mais de 6 moradores.

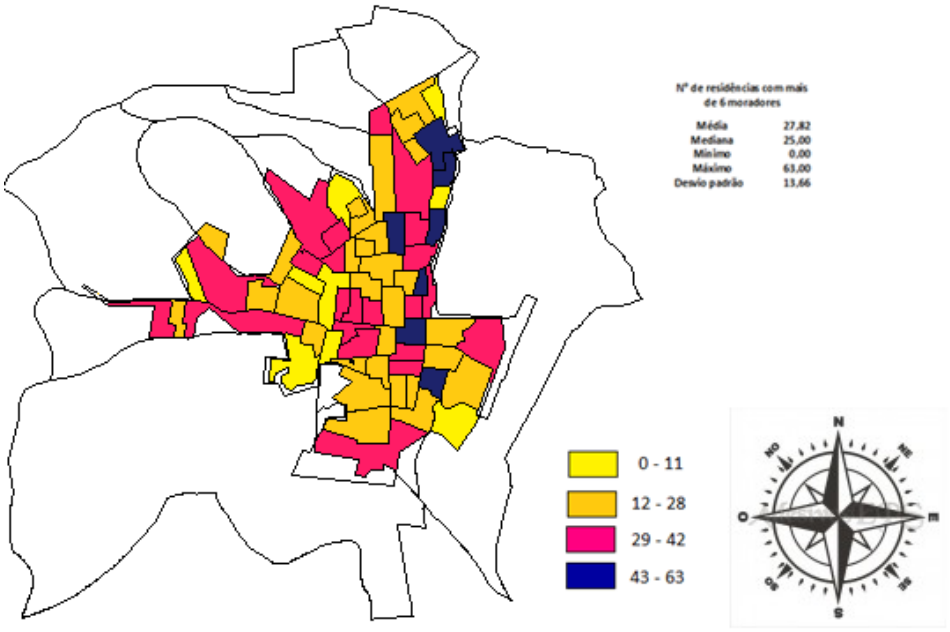

Figura 13 Análise exploratória para o número de residências com mais de 6 moradores na área urbana do município de Alfenas

A distribuição espacial da renda mensal do homem e da mulher encontra-se representada nos mapas das Figuras 14 e 15, respectivamente. Os valores dessas variáveis constituem a soma dos rendimentos mensais dos homens e das mulheres em cada setor censitário. Para a renda mensal dos homens, verifica-se que a maior parte dos setores censitários possuem uma renda que varia de R$ 971,00 a R$ 189.696,00. Dos 68 setores censitários, apenas 12 possuem rendas superiores a R$ 189.696,00, que na sua maior parte, tendem a localizar-se nos arredores do centro da cidade. No que diz respeito a renda mensal das mulheres, observa-se que maior parte dos setores censitários, possuem uma renda mensal superior a R$ 12.911,00 e inferior a R$ 32.118,00. Os setores que possuem uma renda superior a R$ 32.118,00 localizam-se nos arredores do centro da cidade e na região noroeste do município nos bairros Jardim São Paulo, Vale do sol e Jardim Alvorada.

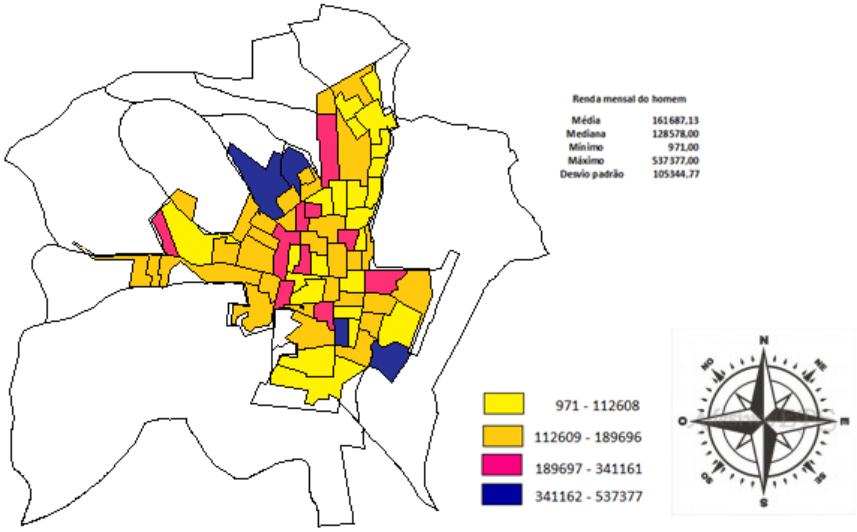

Figura 14 Análise exploratória para a renda mensal do homem em cada setor censitário na área urbana do município de Alfenas

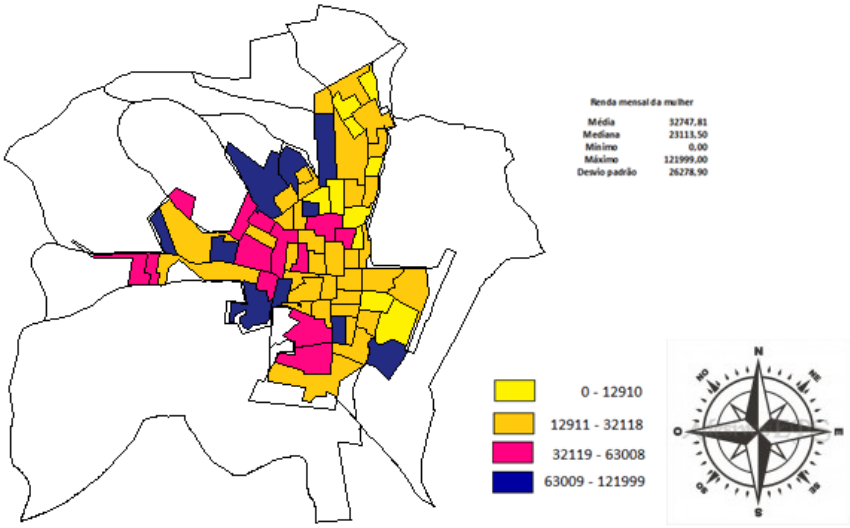

Figura 15 Análise exploratória para a renda mensal da mulher em cada setor censitário na área urbana do município de Alfenas

Para a densidade demográfica do setor censitário, verifica-se que ela varia de 17 até 1396 indivíduos por setor. Os setores censitários localizados no

centro da cidade bem como os que se encontram nas redondezas, tendem a apresentar os maiores valores para esta variável (Figura 16).

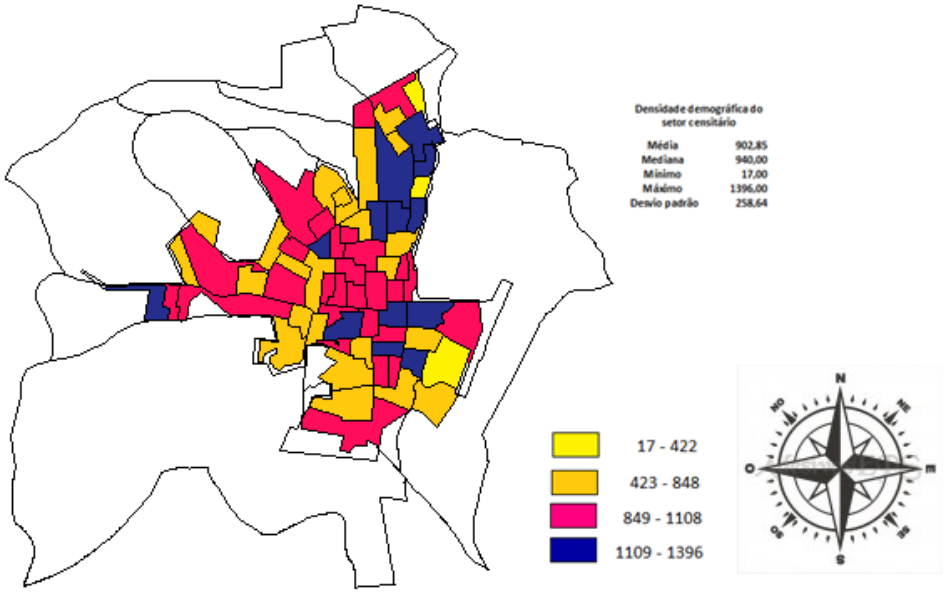

Figura 16 Análise exploratória para a densidade demográfica do setor censitário na área urbana do município de Alfenas

4.2 Ajuste dos modelos

Numa primeira fase da modelagem, ajustou-se o modelo Gauss Markov ordinário usando como variável resposta a raíz quadrada do número de óbitos com menos de 1 ano de idade em função de todas as covariáveis. Pelo procedimento "*stepwise*", verificou-se que as covariáveis que exercem influência sobre a variável dependente foram, o número de mulheres em idade fértil e a renda mensal das mulheres. Para o caso dos modelos de regressão espacial, ao se aplicar o procedimento "*backward*" foram obtidas estas mesmas covariáveis como aquelas que exercem influência sobre a mortalidade infantil.

Na Tabela 2, têm-se as estimativas dos parâmetros do modelo clássico de regressão usando essas duas covariáveis. Para a covariável renda das mulheres, foi obtido um coeficiente negativo, o que significa que quanto maior a renda, menor será o número de óbitos com menos de 1 ano de idade. Para a covariável número de mulheres em idade fértil, verificou-se que esta apresenta uma relação direta com o número de óbitos com menos de 1 ano de idade.

Tabela 2 Estimativas dos parâmetros no modelo clássico de regressão

Covariáveis	**Coeficiente**	**Erro padrão**	**t**	**Valor - p**
Constante	0,06110	0,3093	0,198	0,8440
Renda mensal das mulheres	-8,7E-06	2,9E-06	-3,021	0,0036
N^{o} de mulheres em idade fértil	5,4E-06	1,1E-03	5,091	3,2E-06

R^2 = 0,3714
R^2 ajustado = 0,3521
Log verossimilhança = - 65,44
AIC=132,8

O modelo apresenta um coeficiente de determinação de cerca de 37%, o que significa que, 37% da variabilidade que ocorre nos valores transformados do número de óbitos com menos de um ano de idade, é explicado pela renda mensal das mulheres e pelo número de mulheres em idade fértil.

Os testes de normalidade de Shapiro – Wilk e de heterocedasticidade de Breusch-Pagan, mostraram que as pressuposições de normalidade e homocedasticidade dos resíduos, não foram violadas. Contrariamente a isto, verificou-se pelo teste de Durbin–Watson, que a pressuposição da independência dos resíduos foi violada (Tabela 3).

A autocorrelação presente nos resíduos, mostra que há indícios de suposta presença de autocorrelação espacial entre as observações. Entretanto, para melhorar a qualidade de ajuste do modelo, deve-se incluir a autocorrelação espacial presente, usando um modelo de regressão com efeitos espaciais.

Tabela 3 Avaliação das pressuposições no modelo clássico de regressão

Pressuposição	**Nome do teste**	**Estimativa do teste**	**Valor – p**
Normalidade	Shapiro-Wilk	0,9726	0,1398
Homocedasticidade	Breusch-Pagan	1,9395	0,3792
Independência	Durbin-Watson	2,5736	0,0160*

*significativo a 5%

Na Tabela 4 estão apresentados os resultados de diagnóstico da dependência espacial usando os testes de multiplicadores de Lagrange.

Tabela 4 Estimativa dos testes de multiplicadores de Lagrange

Teste de multiplicador de Lagrange	**Estimativa**	**Valor – p**
LM – Error (CAR)	4,9087	0,0267*
LM – Lag (SAR)	6,3270	0,0119*

*significativo a 5%

Com base nos testes de multiplicadores de Lagrange, verifica-se que os resíduos do modelo clássico de regressão, apresentam dependência espacial ($p < 0,05$). Ambos os testes, mostram significância para ajuste, tanto do modelo espacial autoregressivo, assim como do modelo de erro espacial.

Essa dependência espacial presente nos resíduos era esperada, pois tratando-se de dados espaciais, Druck et al. (2004) afirmam que no caso mais comum os resíduos continuam apresentando a autocorrelação espacial presente nos dados.

As estimativas dos parâmetros nos modelos SAR e CAR, estão representados nas Tabelas 5 e 6, respectivamente.

Tabela 5 Estimativas dos parâmetros no modelo SAR

Covariáveis	**Coeficiente**	**Erro padrão**	**z**	**Valor - p**
Constante	1,01460	0,40784	2,488	0,0128
Renda mensal das mulheres	-9,2E-06	2,6E-06	-3,525	0,0004
N^{o} de mulheres em idade fértil	4,9E-03	9,6E-04	5,114	3,2E-07
Rho	-0,69401	0,22475	-3,090	0,0020
Log verossimilhança	-58,35			
R^2ajustado	0,4427			
AIC	126,70			

Tabela 6 Estimativas dos parâmetros no modelo CAR

Covariáveis	**Coeficiente**	**Erro padrão**	**z**	**Valor - p**
Constante	0,17070	0,27948	0,611	0,54146
Renda mensal das mulheres	-7,3E-06	2,3E-06	-3,142	0,00168
N^{o} de mulheres em idade fértil	4,7E-03	9,7E-04	4,927	8,4E-07
Rho	-0,71343	0,26010	-2,743	0,00608
Log verossimilhança	-59,32			
R^2 ajustado	0,4281			
AIC	128,60			

Em ambos os modelos SAR e CAR, verificou-se que os valores transformados do número de óbitos com menos de um ano de idade, apresentam uma relação inversa com a renda mensal das mulheres. Em relação ao número de mulheres em idade fértil, verificou-se que esta possui uma relaçao direta com a mortalidade infantil. Portanto, espera-se que os setores censitários que apresentem menor renda de mulheres e maior número de mulheres em idade fértil, possuam valores altos de mortalidade infantil, e consequentemente constituam os locais de maior risco.

Alguns autores como Abreu e Vasconcelos (1998) e Andrade et al. (2004) mostraram que é comum associar fatores socioeconômicos com a mortalidade infantil. Abreu e Vasconcelos (1998) apontam a renda da mãe como um dos fatores determinantes da mortalidade infantil, pelo papel que a renda exerce sobre aquisição e utilização de bens e serviços necessários a garantia da qualidade de vida e saúde das populações.

Menezes et al. (1998) pesquisando sobre fatores de risco para a mortalidade perinatal em Pelotas, mostraram que a mortalidade infantil em famílias com renda abaixo de um salário mínimo, era 3 vezes maior do que as famílias com maiores rendas.

Nogueira (2008) em modelagem espacial via inferência bayesiana obteve como um dos fatores determinantes da mortalidade infantil o número de mulheres em idade fértil.

Os valores dos coeficientes das covariáveis nos dois modelos são muito parecidos. O coeficiente espacial autoregressivo nos dois modelos apresentou um valor negativo e estatisticamente diferente de zero. Isto significa que os valores da mortalidade infantil em áreas próximas apresentam uma autocorrelação espacial negativa, ou seja, os valores em áreas que fazem fronteira entre si são dissimilares.

Os testes de normalidade de Shapiro–Wilk e de heterocedasticidade de Breusch-Pagan mostram que as pressuposições de normalidade e homocedasticidade dos resíduos, tanto para o modelo SAR assim como o CAR, não foram violadas (Tabela 7).

Tabela 7 Avaliação das pressuposições nos modelos SAR e CAR

Modelo	Pressuposições	Nome do teste	Estimativa	Valor - p
SAR	Normalidade	Shapiro-Wilk	0,9889	0,8120
	Homocedasticidade	Breusch-Pagan	1,7844	0,4097
CAR	Normalidade	Shapiro-Wilk	0,9739	0,1644
	Homocedasticidade	Breusch-Pagan	2,5785	0,2755

Esses resultados mostram que os dois modelos se ajustam muito bem aos dados. Assim, inferências podem ser feitas usando ambos os modelos.

Pelos critério de qualidade de ajuste dos modelos, verifica-se que o modelo SAR é o melhor modelo, pois, apresentou o menor valor de *AIC* (126,7), maior valor do log verossimilhança (- 58,35) e maior R^2 ajustado (0,4427).

4.2.1 Cálculo dos valores preditos médios em modelos de regressão espacial com transformação raíz quadrada na variável dependente

Para se obter as predições médias em modelos de regressão espacial ajustados com transformação na variável dependente, não foram encontradas, na literatura, metodologias que descrevem que equação de predição deverá ser usada. Porém, para o caso de um modelo de regressão linear simples, Miller

(1984) apresenta algumas propostas para obter as predições médias nesse modelo.

Para o caso do modelo SAR, considerando que a variável resposta sofreu uma transformação do tipo raíz quadrada e usando uma analogia ao que foi proposto por Miller (1984) tem-se:

$$Y^{0.5} = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon$$

Elevando ao quadrado ambos os membros da equação obtém-se:

$$(Y^{0.5})^2 = H^2$$

em que $H = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon$

Aplicando a esperança em ambos os membros nessa equação tem-se:

$$E[Y] = E[H^2],$$

Como o segundo membro dessa equação é composto pelos quadrados dos elementos do vetor *H*, então é equivalente a:

$$E[Y] = diag\{E[HH']\}$$

em que diag[.] corresponde aos elementos da diagonal principal da matriz, assim tem-se:

$E[Y] = diag\{E[(AX\beta + A\varepsilon)(AX\beta + A\varepsilon)']\}$, em que $A = (I - \rho W)^{-1}$.

Aplicando a transposta, obtém-se:

$$E[Y] = diag\{E[(AX\beta + A\varepsilon)\big((X\beta)'A' + \varepsilon'A'\big)]\}$$

que é equivalente a:

$$E[Y] = diag\{E\big[AX\beta(X\beta)'A' + AX\beta\varepsilon'A' + A\varepsilon(X\beta)'A' + A\varepsilon\varepsilon'A'\big]\}$$

Como a esperança da variável aleatória erro é nula, então tem-se:

$$E[Y] = diag\{\big(AX\beta(X\beta)'A' + AE\big[\varepsilon\varepsilon'\big]A'\big)\}$$

que é equivalente a:

$$E[Y] = diag\{AX\beta(AX\beta)' + AI\sigma^2A'\}.$$

Assim, usando a mesma analogia apresentada por Miller (1984), os valores preditos do modelo após a transformação serão obtidos por:

$$\hat{Y} = diag[(I - \hat{\rho}W)^{-1}X\hat{\beta}\big((I - \hat{\rho}W)^{-1}X\hat{\beta}\big)' + \hat{\sigma}^2(I - \hat{\rho}W)^{-1}(I - \hat{\rho}W')^{-1}], \quad (37)$$

em que $\hat{\rho}$, $\hat{\beta}$ e $\hat{\sigma}$, obtidos pelas equações (23), (24) e (25), respectivamente, são as estimativas dos parâmetros do modelo obtidas por máxima verossimilhança usando os dados transformados.

Para o caso do modelo CAR, aplicando o mesmo procedimento demonstrado para o modelo SAR, obtém-se que os valores preditos nesse modelo, após a transformação raíz quadrada, serão dados por:

$$\hat{Y} = diag[X\hat{\beta}\left(X\hat{\beta}\right)' + \hat{\sigma}^2(I - \hat{\rho}W)^{-1}(I - \hat{\rho}W')^{-1}]. \quad (38)$$

Na Figura 17, tem-se a representação gráfica dos valores médios estimados a partir do modelo SAR ajustado. Esses valores foram calculados segundo a equação 37.

Nessa figura, verifica-se que os valores estimados para a mortalidade infantil, indicam que existem setores censitários sem nenhuma ocorrência e outros com até quatro casos. Os setores que não apresentam nenhum caso de ocorrência, localizam-se nos bairros Santa Clara, Jardim Primavera, Colina Parque e Jardim Colina.

Pode-se ver ainda que, a maior parte dos setores censitários, apresenta casos de ocorrência de uma a duas crianças mortas com menos de um ano de idade. Esses valores encontram-se distribuídos em alguns setores censitários que fazem parte do centro da cidade, bem como em alguns bairros arredores da cidade. Em alguns bairros periféricos da cidade como Jardim América, Por do Sol, Jardim Boa Esperança, Vila Teixeira, Jardim Panorama, Vila Santos Reis e Jardim Aeroporto também são observados valores de mortalidade infantil de um a dois casos de ocorrência.

Verifica-se também que cerca de 24 setores censitários, apresentam os maiores casos de ocorrências de mortes, isto é, 3 a 4 crianças mortas com menos de um ano de idade. Alguns desses setores censitários, localizam-se no centro da cidade, e outros na sua maioria, localizam-se nos bairros Jardim Eunice, Vila Esperança, Jardim São Carlos, Campo Elísio, Chapada, Vila Formosa, Jardim

Aeroporto II, Vila Betânia, Jardim Alvorada e Novo Horizonte, que fazem parte da periferia da cidade.

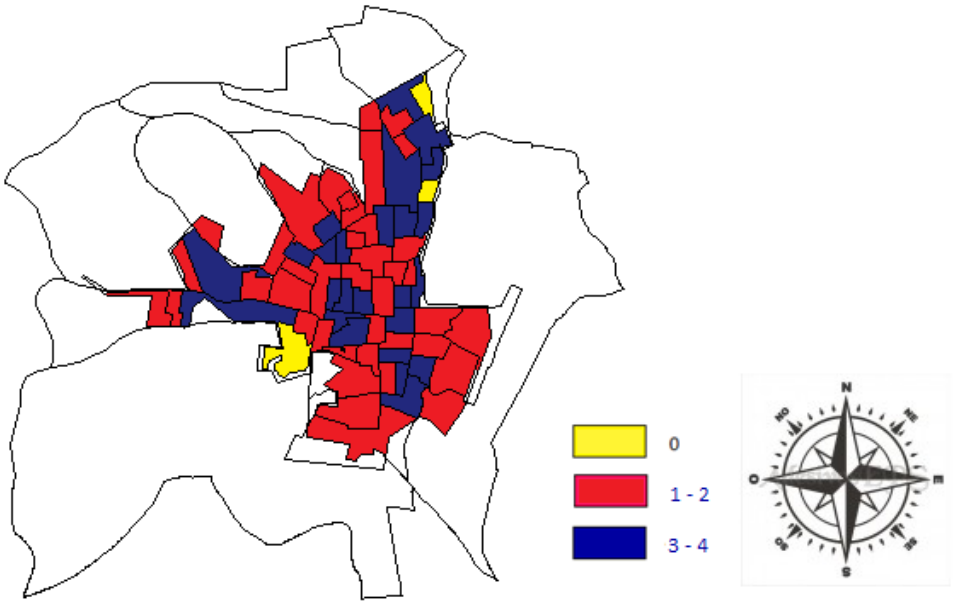

Figura 17 Valores estimados da mortalidade infantil em cada setor censitário usando o modelo SAR

Nogueira (2008) afirma que alguns bairros como Chapada e Vila Betânica, apresentam características socioeconômicas desprivilegiadas em termos de infraestrutura, e são considerados alguns dos bairros mais pobres do município.

Assim, num programa de saúde pública para a redução de ocorrências de mortalidade infantil no município de Alfenas, os bairros que apresentam maiores valores de mortalidade infantil serão considerados como áreas de atuação prioritária.

4.3 Considerações finais

A falta de significância de algumas covariáveis utilizadas no presente estudo, leva a necessidade de inclusão de outras variáveis em trabalhos futuros, como o peso das crianças ao nascer, fecundidade da mãe, prematuridade, número de postos de saúde, distância percorrida até o posto de saúde, número de consultas feitas no período pré-natal, dentre outras.

A estatística global de Moran, detectou ausência da dependência espacial entre as observações da mortalidade infantil, contudo através do índice local de autocorrelação espacial (LISA), verificou-se que algumas observações em determinados setores censitários apresentaram dependência espacial negativa em relação aos seus vizinhos, isto é, os valores de mortalidade infantil nesses setores censitários tendem a ser dissimilares em relação aos setores censitários vizinhos.

Os modelos de regressão com efeitos espaciais, ajustaram-se muito bem aos dados, e permitiram uma melhoria nas inferências, comparativamente ao modelo clássico de regressão. Pelo *AIC* verificou-se que o modelo espacial autoregressivo (SAR), apresentou melhor qualidade de ajuste, em relação ao modelo de erro espacial (CAR) e ao modelo clássico de regressão. A variação que ocorre na mortalidade infantil é explicada pela renda mensal das mulheres, pelo número de mulheres em idade fértil e pela dependência espacial negativa, detectada através do coeficiente espacial autoregressivo, que apresentou um valor negativo. Para trabalhos futuros, pode-se conduzir análises baseadas na análise de taxas, aplicando outros métodos de análise como o índice bayesiano empírico, construção de mapas de probabilidade e mapas bayesianos de doenças.

Através das equações de predição desenvolvidas nesta dissertação para o cálculo dos valores preditos médios nos modelos de regressão espacial ajustados com transformação raiz quadrada, foi possível detectar os setores censitários que

apresentaram maiores valores médios de mortalidade infantil. Alguns localizam-se no centro da cidade e outros em alguns bairros periféricos como Chapada, Vila Betânica, Vila Esperança, Jardim Eunice e Jardim São Carlos. Para trabalhos futuros, pode-se desenvolver novas equações para obter os valores preditos médios em modelos de regressão espacial ajustados com outros tipos de transformação, como a transformação logarítmo. Pode-se também pensar em extender os modelos de regressão espacial para aqueles casos onde a variável dependente pertence a família exponencial, na maneira dos modelos lineares generalizados.

Seria interessante aplicar as equações de predição desenvolvidas nesta dissertação em bancos de dados mais atualizado, para melhorar as inferências obtidas nesse estudo.

5 CONCLUSÕES

Os resultados obtidos nesta dissertação, mostraram que foi possível identificar a presença de dependência espacial da mortalidade infantil e, consequentemente, ajustar um modelo espacial capaz de descrever o inter-relacionamento entre a mortalidade infantil e as diferentes covariáveis. Através das equações de predição desenvolvidas nesta dissertação para o cálculo dos valores preditos médios em modelos de regressão espacial ajustados com transformação raiz quadrada na variável dependente, pode-se também identificar as áreas de maiores ocorrências de mortalidade infantil na cidade de Alfenas-MG.

REFERÊNCIAS

ABREU, C.; VASCONCELLOS, E. Análise dos fatores sócio-econômico-ambientais e a sua interferência na mortalidade infantil na favela da Aldeia. **Vértices**, Campos dos Goitacazes, v. 1, n. 2, p. 1-13, nov. 1998.

ALMEIDA, M. C. D. et al. Dinâmica intra-urbana da epidemia de dengue em Belo Horizonte, Minas Gerais, Brazil, 1996-2002. **Cadernos de Saúde Publica**, Rio de Janeiro, v. 24, n. 10, p. 2385-2395, out. 2008.

ANDRADE, C. L. T. et al. Desigualdades sócio-econômicas do baixo peso ao nascer e da mortalidade perinatal no Município do Rio de Janeiro. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 20, n. 1, p. 44-51, 2004.

ANDRADE, C. L. T.; SZWARCWALD, C. L. Análise espacial da mortalidade neonatal precoce no Município do Rio de Janeiro, 1995-1996. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 17, n. 5, p. 1199-1210, out. 2001.

ANSELIN, L. **Exploring spatial data with GeoDaTM** : a workbook. Urbana: Spatial Analysis Laboratory Department of Geography, University of Illinois, 2005. 245 p.

ANSELIN, L. **Spatian econometrics**. Dallas: University of Texas, 1999. 50 p.

ASSUNÇÃO, R. M. **Estatística espacial com aplicações em epidemiologia, economia e sociologia**. São Carlos: Universidade Federal de São Carlos, 2001. 136 p.

BIVAND, R. S. et al. **Spdep**: spatial dependence: weighting schemes, statistics and models. 2011. Disponível em: <http://CRAN.R-project.org/package=sp dep>. Acesso em: 10 jul. 2011.

BIVAND, R. S.; PEBESMA, E. J; GOMES, V. R. **Applied spatial data analysis with R**. New York: Springer, 2008. 378 p.

CALDEIRA, A. P. et al. Evolução da mortalidade infantil por causas evitáveis, Belo Horizonte, 1984-1998. **Revista de Saúde Pública**, São Paulo, v. 39, n. 1, p. 67-74, jan. 2005.

CHARNET, R. et al. **Análise de modelos de regressão linear com aplicações**. Campinas: Unicamp, 1999. 351 p.

CIA WORLD FACTBOOK. **Taxa de mortalidade infantil - mundo**. 2011. Disponível em: <http://www.indexmundi.com/map/?v=29&l=pt>. Acesso em: 7 nov. 2011.

CLIFF, A. D.; ORD, K. **Spatial processes**: models and applications. London: Pion, 1981.

CRESSIE, N. A. C. **Statistics for spatial data**. New York: J. Wiley, 1993.

DRAPPER, N. R.; SMITH, H. **Applied regression analysis**. New York: J. Wiley, 1998. 407 p.

DRUCK, S. et al. **Análise espacial de dados geográficos**. Brasília: EMBRAPA, 2004.

DUARTE, C. Reflexos das políticas de saúde sobre as tendências da mortalidade infantil no Brasil: revisão da literatura sobre a última década. **Cadernos e Saúde Pública**, Rio de Janeiro, v. 23, n.7, p. 1511-1528, 2007.

FLAHAUT, B. et al. The local spatial autocorrelation and the kernel method for identifying black zones - A comparative approach. **Accident Analysis and Prevention,** Elmsford**,** v. 35, n. 6, p. 991-1004, Nov. 2002.

GEODA CENTER FOR GEOSPATIAL ANALYSIS AND COMPUTATION. Versão 9.5. 2010. Disponível em: <http://www.geodacenter.asu. edu>. Acesso em: nov. 2010.

GRIFFITH, D. A. The moran coefficient for non-normal data. **Journal of Statistical Planning and Inference**, Amsterdam, v. 140, n. 11, p. 2980-2990, Nov. 2010.

MENEZES, A. M. B. et al. Fatores de risco para mortalidade Perinatal em Pelotas, RS. **Revista de Saúde Pública**, Rio de Janeiro, v. 32 p. 209-216, 1998.

MENEZES, A. M. B. Mortalidade infantil em duas coortes de base populacional no Sul do Brasil: Tendências e diferenciais. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 12, p. 79-86, 1996.

MILLER, D. M. Reducting transformation bias in curve fitting. **The American Statistician**, Washington, v. 38, p. 124–126, 1984.

MONDINI, A.; CHIARAVALLOTI, F. N. Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a Brazilian city. **Science of the Total Environment,** Amsterdam, v. 393, n. 2/3, p. 241-248, Apr. 2008.

MORETIN, P. A; TOLOI, C. M. C. **Análise de séries temporais**. 2. ed. São Paulo: E. Blucher, 2006. 538 p.

NOGUEIRA, D. A. **Análise espacial da mortalidade infantil no município de Alfenas, MG**. 2008. 101 p. Tese (Doutorado em Agronomia) - Universidade Federal de Lavras, Lavras, 2008.

POULIOU, T.; ELLIOTT, S. J. An exploratory spatial analysis of overweight and obesity in Canada. **Preventive Medicine,** San Diego, v. 48, n. 4, p. 362-367, Apr. 2009.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing: versão 12.2.1. Vienna: R Foundation for Statistical Computing, 2010. Disponível em: <http://www.R-project.org/>. Acesso em: 12 jun. 2010.

SANTOS, S. M.; SOUSA, W. V. **Introdução à estatística espacial para a saúde pública.** Brasília, DF: Ministério da Saúde, Fundação Oswaldo Cruz, 2007. 123 p.

TAAMOUTI, A.; DUFOUR. Exact optimal inference in regression models under heteroskedasticity and non-normality of unknown form. **Computational Statistics and Data Analysis**, Amsterdam, v. 54, n. 11, p. 2532-2553, 2009.

WALLER, L. A.; GOTWAY, C. A. **Applied spatial statistics for public health data**. Hoboken: J. Wiley, 2004. 518 p.

WERNECK, G. L. Georeferenced data in epidemiologic research. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 13. n. 6, p. 1753-1766, 2008.

ANEXO

ANEXO A Dados utilizados no estudo

Tabela 1A Dados utilizados no estudo

X	Y	morte	Fer	ges	analf	res	rendm	rendh	Dens
401096.8	7632311	0	132	78	1	8	121999	454695	528
398571.8	7630523	2	247	139	4	23	63008	164376	937
399565.9	7630476	2	300	176	2	41	18840	121143	1022
398313.1	7630573	1	331	190	5	31	42841	172467	1146
401331.2	7629442	0	214	124	0	23	58468	150403	789
401331.3	7628543	4	267	146	2	30	21175	84346	1002
401387.3	7628912	1	245	151	3	25	40810	112608	837
400370.8	7629828	0	186	108	1	11	88602	183910	690
400878.2	7630072	1	161	92	0	10	78190	279203	572
400625.7	7630249	0	208	122	4	20	42662	128529	716
399243	7630995	3	251	131	4	33	21532	71481	967
400182.4	7631362	6	240	144	1	16	59746	169376	817
397403	7630850	0	182	107	0	17	59909	164810	618
398778	7631122	0	199	118	0	10	104597	305900	651
401643.8	7632619	0	217	127	2	19	66435	202961	810
400507.2	7631326	2	339	182	4	33	35776	169645	1094
400376.6	7630777	0	243	150	0	23	38441	141245	865
399842.1	7630824	1	191	107	0	23	76065	184532	685
400859.2	7630873	0	202	136	0	11	42843	341161	627
400500.9	7631062	1	157	92	0	5	15195	173317	484
400577.1	7632001	2	310	160	1	35	82955	475765	966
401608.4	7633413	2	293	162	2	30	24386	189696	992
400957	7631358	7	357	199	3	41	32118	145072	1249
400971.9	7631725	1	293	157	2	34	23290	151199	1008
402092.2	7630677	1	298	169	1	33	27495	116394	993
401801.1	7630677	4	249	151	3	28	23718	121024	917
401697.7	7630193	0	260	142	3	24	22557	152307	859
401332.7	7630131	3	306	164	7	36	17980	109155	1146
401235.8	7629829	2	292	153	3	27	15244	162414	1011
401578	7629737	1	220	113	3	24	25734	205676	754
401240.9	7630658	2	336	197	1	31	40662	283385	1081
401182.8	7630362	5	282	138	5	33	14925	91618	1108
401048.5	7630731	3	276	134	13	34	15985	73082	989
401487	7630596	2	298	180	3	42	16276	93816	1093
402369	7629563	4	348	183	18	53	20641	116855	1201
402691.6	7628946	0	250	153	2	11	68743	502703	759
402172.1	7629097	2	263	142	0	24	23759	126436	842
401866.5	7629410	2	311	179	2	24	89992	537377	943
402078.5	7629426	2	272	140	1	25	25970	105096	890
403119.2	7630088	2	283	152	1	30	23922	134845	1020
402893.2	7629552	1	132	80	2	16	8595	88067	422
402099.7	7630287	0	370	209	6	56	15952	81094	1396
401976.5	7629753	4	302	151	10	42	16573	82503	1193
402007.4	7629995	2	294	173	4	42	14493	115327	1071
402581.3	7630320	2	342	155	0	26	27298	206179	1172

“continua”

Tabela 1A “conclusão”

X	Y	morte	fer	ges	analf	res	rendm	rendh	Dens
402502.8	7629918	1	226	121	0	16	9579	119138	707
402340.5	7630882	4	285	116	2	32	15836	126615	1001
401836.6	7631781	2	357	189	5	55	10222	104830	1336
401519.6	7631700	2	187	106	2	23	6615	74988	643
401408.8	7630977	1	284	155	5	22	22232	122144	913
401193.5	7631393	3	302	156	1	18	23510	210998	1024
401630.6	7631270	0	295	156	0	21	37224	172578	948
401380.6	7631538	3	351	243	1	26	70129	251216	1040
402120.7	7633665	2	273	142	0	21	17384	184900	926
401377.8	7631935	3	214	106	1	15	16925	116500	696
401274.6	7632126	1	229	140	1	16	14936	170472	770
402097.9	7632543	5	349	203	6	41	22937	128627	1230
402142.3	7631828	4	309	169	4	40	20649	99636	1130
401994.3	7631049	0	268	146	4	21	34967	279368	845
402172.3	7631411	2	213	106	6	42	8311	60436	848
402220.8	7631024	5	281	146	10	57	12910	75750	1088
402444.5	7631854	5	328	195	10	55	18301	88250	1159
402504.5	7633075	3	350	176	13	63	20152	103006	1358
402131.3	7632990	1	212	102	4	18	7708	71804	762
401992.2	7633334	1	171	79	8	22	2228	56005	627
402576.1	7632669	6	329	159	7	49	16224	108759	1226
402503.1	7632315	0	45	26	1	6	1475	24541	168
402453.5	7633642	0	4	1	3	0	0	971	17

Onde:
morte – número de crianças mortas com menos de 1 ano de idade;
fer – número de mulheres em idade fértil;
ges – número de mulheres em idade de risco gestacional;
analf – número de mulheres analfabetas em idade fértil;
res – número de residências com mais de 6 moradores;
rendm – renda mensal da mulher;
rendh – renda mensal do homem;
dens – densidade demográfica do setor censitário.

APÊNDICE

APÊNDICE – Script das análises

```
library(spdep) # carregar pacote para análises espaciais de áreas
library(SpatialEpi) #carregar pacote para mapear variáveis

#importação dos dados
g=read.dta("c:/Users/zomolas/Documents/2.dta")
g

#importação da matriz de conectividade a partir do geoda
lds=read.gal("c:/Users/zomolas/Documents/mestrado/dissertacao/dados/data/mat
riz2.gal")
lds

#criando o data frame
df=data.frame(xcoord=g$x,ycoord=g$y,mor=g$morte,ida.fert=g$a,risc.gest=g$b
,analf=g$c,res.dom=g$d,rendm=g$e,rendh=g$f,resid=g$g)
df

#ler formatos de polignos
area=readShapePoly("c:/Users/zomolas/Documents/mestrado/dissertacao/mapa/
alfenas.shp")
area

#visualizar área de estudo
plot(area, border="black", axes=TRUE, las=1)

# plotando casos de ocorrência de mortalidade infantil por setor censitário
z=df[,1:2]
text(coordinates(z), labels=df[,3], cex=0.6)

#criando a matriz de vizinhança normalizada, usando a lista de vizinhos
vizin=nb2mat(lds)
vizin

#criando a matriz W
wmat<-mat2listw(vizin)
wmat

#outra forma de criar a matriz W
matriz.w=nb2listw(lds)
```

```
matriz.w

#cálculo da estatística de moran
moran(df[,3], wmat, 68, 68, zero.policy=NULL, NAOK=FALSE)

#testando o índice de moran
moran.test(df$mor, wmat, randomisation=TRUE, zero.policy=NULL,
alternative="greater", rank = FALSE, na.action=na.fail, spChk=NULL,
adjust.n=TRUE)

#####################################################
#####################################################

### ajuste do modelo gauss markov usando todas as covariáveis
reg=lm(sqrt(mor)~ida.fert + risc.gest + analf + res.dom + rendm + rendh+resid,
data=df)
summary(reg)

###precedimento de seleção de covariáveis no modelo
reg.step=step(reg,direction="both")
summary(reg.step)

##modelo gauss markov com as covariáveis seleccionadas
reg1=lm(sqrt(mor)~ida.fert+rendm,data=df)
summary(reg1)

############diaognósticos do modelo###############

shapiro.test(reg1$residual)## normalidade dos resíduos

library(lmtest)## carregando pacote para análise da homocedasticidade dos
resíduos
bptest(reg1)##homocedasticidade dos resíduos pelo teste de Breusch Pagan
library(car)
durbinWatsonTest(reg1)## independência dos resíduos pelo teste de Durbin
Watson
bgtest(reg1)## independência dos resíduos pelo teste de Breusch Godfrey

##diagnóstico de modelo SAR ou CAR pelos testes de multiplicador de
Lagrange
lm.LMtests(reg1, nb2listw(lds), test=c("LMerr", "LMlag","SARMA"))
```

```
######Ajuste do modelo SAR (lag model)
reg3=lagsarlm(sqrt(mor)~ida.fert + rendm, data = df,nb2listw(lds,
style="W"),type="lag", tol.solve=1.0e-11)
summary(reg3)

#####Ajuste do modelo CAR (error model)
reg4=errorsarlm(sqrt(mor)~ida.fert + rendm, data = df,nb2listw(lds,
style="W"),method="eigen", tol.solve=1.4e-11)
summary(reg4)

#####teste de homocedasticidade dos resíduos nos modelos SAR e CAR
bptest.sarlm(reg3)#lag model
bptest.sarlm(reg4)#error model

#####teste de normalidade dos resíduos
shapiro.test(reg3$residual)#lag model
shapiro.test(reg4$residual)#error model

#########################
#predição

est=predict.sarlm(reg3)#valores preditos no modelo SAR
est=as.matrix(est)

#predição no modelo SAR para variável dependente que sofreu transformação
i=matrix(rep(0,4624),68,68)
diag(i)=1
i

a=solve(i+0.69401*vizin)
b=0.31092*a%*%t(a)
as.matrix(b)
c=est%*%t(est)
as.matrix(c)
res=c+b
estimad=round(diag(res))
as.matrix(estimad)

##mapeamento dos valores estimados na área de estudo
mapvariable(estimad,area,ncut=4,nlevels=4,lower=0,upper=5,main="observado"
,xlab="Eastings (km)",ylab="Northings (km)")
```