



**SIMONE SOARES LIMA JARDIM**

**ANÁLISE DO PADRÃO ESPACIAL DA  
MORTALIDADE INFANTIL UTILIZANDO  
GEOESTATÍSTICA**

**LAVRAS – MG  
2011**

**SIMONE SOARES LIMA JARDIM**

**ANÁLISE DO PADRÃO ESPACIAL DA MORTALIDADE INFANTIL  
UTILIZANDO GEOESTATÍSTICA**

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

Orientador

Dr. João Domingos Scalon

Coorientador

Dr. Denismar Alves Nogueira

**LAVRAS – MG  
2011**

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

Jardim, Simone Soares Lima.

Análise do padrão espacial da mortalidade infantil utilizando  
geoestatística / Simone Soares Lima Jardim. – Lavras : UFLA, 2010.  
75 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2010.

Orientador: João Domingos Scalon.

Bibliografia.

1. Geoestatística clássica. 2. Geoestatística baseada em modelos.  
3. Máxima verossimilhança. 4. Saúde pública. 5. Mapas de  
mortalidade infantil. I. Universidade Federal de Lavras. II. Título.

CDD – 526.982

**SIMONE SOARES LIMA JARDIM**

**ANÁLISE DO PADRÃO ESPACIAL DA MORTALIDADE INFANTIL  
UTILIZANDO GEOESTATÍSTICA**

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

APROVADA em 31 de março de 2010.

Dr. Marcelo Silva Oliveira	UFLA
Dr. Denismar Alves Nogueira	UNIFAL

Dr. João Domingos Scalon  
Orientador

**LAVRAS – MG  
2011**

*À minha mãe, Maria de Lourdes, por todos os ensinamentos, pelas constantes orações e por ser um exemplo de mãe e de ser humano a ser seguido. Ao meu pai, Cloves Jair, por me ensinar que nunca se deve sucumbir a um obstáculo, a minha irmã Iracinara, pelos cuidados e pelos exemplos de caráter e profissionalismo.*

*Ao meu esposo, Valdomiro, pela oportunidade de cursar o mestrado e por ter estado ao meu lado todos os dias, me apoiando incondicionalmente.*

*Ao meu mais novo e mais importante presente, meu bebê Davi, incentivo para que eu procure ser a cada dia melhor em todos os aspectos.*

*A todos vocês, amores de minha vida, por tudo que significam para mim.*

**DEDICO**

## AGRADECIMENTOS

A Deus, em primeiro lugar, por ter me permitido começar o mestrado e, em segundo lugar, por ter me dado forças para perseverar diante dos obstáculos, permitindo que eu chegasse até aqui.

Ao meu cunhado Moabe, por ter estado presente nos momentos mais difíceis de minha família, sempre nos apoiando, sendo para mim o irmão que não tive.

Aos meus familiares, que sempre torceram e confiaram em mim, principalmente vó Calu, meus tios Gérson, Marcos, Mita e minha prima Rayne, que acompanharam de perto os meus passos até aqui.

À família do meu esposo, que se tornou minha também e que sempre me deu forças; em especial, à minha cunhada Eliana, pelo incentivo nos momentos em que fraquejei.

Ao “Clube da Luluzinha”, Franciella (Fran), Rosiana (Rosi), Flávia e Deyse, por todo apoio, amizade, companheirismo e solidariedade. Sem vocês, as dificuldades teriam sido muito maiores, vocês foram como irmãs, obrigada!

À minha afilhada Gledsa, pelo carinho, orações e confiança de sempre.

Ao professor Gérson Rodrigues dos Santos, por ter sido presença constante no desenvolvimento do meu trabalho, dando sugestões, ajudando com o programa, tendo paciência com as minhas dúvidas. Enfim, sem a sua ajuda, tudo teria sido muito mais difícil.

Ao professor Dr. Scalon, pela orientação, paciência, confiança e pela sabedoria em transmitir tranquilidade. Você é um exemplo!

Ao professor Dr. Denismar, por tão gentilmente me ceder seu material para que eu desenvolvesse o meu trabalho e por me coorientar.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas e seus professores, por me permitirem crescer profissionalmente e como ser humano.

A Selminha, Edila, Maria, “Josizinha” e “Josizona”, funcionárias do Departamento de Ciências Exatas da UFLA, por sempre me atenderam com gentileza, sendo agradáveis em todos os momentos.

A Cristiani Regina Vieira Santana e Jabnei Gomes Rodrigues, por me apoiarem e incentivarem na conquista desta vitória.

A Adriana, do setor de capacitação de recursos humanos da Secretaria de Estado de Educação de Minas Gerais, pela simpatia e boa vontade com que sempre me atendeu.

## RESUMO

O estudo da mortalidade infantil por meio da estatística espacial tem sua relevância ligada à identificação de regiões de risco com necessidade de intervenção dos órgãos de saúde. Dentre as técnicas de possível utilização encontra-se a geoestatística, que é empregada em trabalhos com dados continuamente distribuídos no espaço. O presente trabalho foi realizado com o objetivo de analisar a distribuição espacial da mortalidade infantil utilizando geoestatística clássica e geoestatística baseada em modelos. Foram utilizados dados de contagem de mortes infantis, por setor censitário, do município de Alfenas, MG, no período de 2000 a 2004. Os parâmetros da geoestatística clássica foram estimados por máxima verossimilhança, enquanto os parâmetros da geoestatística baseada em modelos, assumindo a distribuição de Poisson para os dados, foram estimados por máxima verossimilhança Monte Carlo. O erro absoluto e o critério de Akaike foram utilizados para determinar a metodologia mais adequada para modelar a mortalidade infantil. A análise exploratória mostrou que os dados apresentavam dependência espacial e aderência à distribuição de Poisson. O erro absoluto e o critério de Akaike mostraram que a geoestatística baseada em modelos é mais adequada que a geoestatística clássica para a análise de dados de mortalidade infantil no município de Alfenas. Utilizando a metodologia geoestatística baseada em modelos foram construídos mapas de predição da mortalidade infantil e das duas covariáveis mais correlacionadas com a mortalidade infantil. Pode-se observar a existência de relação entre os mapas das covariáveis e o mapa da mortalidade infantil. Os resultados demonstraram que a geoestatística baseada em modelos é uma metodologia poderosa para a análise da distribuição espacial da mortalidade infantil.

Palavras-chave: Geoestatística clássica. Geoestatística baseada em modelos. Máxima verossimilhança. Saúde pública. Mapas de mortalidade infantil.

## **ABSTRACT**

The study of infant mortality using spatial statistics is important to the identification of risk areas that need intervention by health organizations. Geostatistics, applied for analyzing data continuously distributed in space, is considered one of the most important methods for spatial analysis. The aim of this work was to analyze the spatial distribution of infant mortality using both classical geostatistics and model-based geostatistics. It was used count data of infant deaths located at census sectors in the city of Alfenas - MG from 2000 to 2004. It was used the method of maximum likelihood to find the estimator of the parameters of the classical geostatistics, while the method of Monte Carlo maximum likelihood was used to find the parameters of the model-based geostatistics. The absolute error and the Akaike information criteria were applied to detect the most appropriate methodology for modeling the infant mortality. The exploratory analysis showed that the data presented spatial dependence and goodness-of-fit to the Poisson distribution. Both, the absolute error and the Akaike information criteria showed that model-based geostatistics was more suitable than classical geostatistics for modeling the infant mortality in the city of Alfenas. It was used model-based geostatistics to construct a prediction map of the infant mortality and prediction maps of the two covariates more correlated to the infant mortality. It was observed evidences of relationship between the map of the infant mortality and the maps of the two covariates. The results showed that model-based geostatistics is a powerful methodology for analyzing the spatial distribution of infant mortality.

Keywords: Classical geostatistics. Model-based geostatistics. Maximum likelihood. Public Health. Infant mortality maps.

## LISTA DE FIGURAS

Figura 1	Semivariograma empírico hipotético apresentando os parâmetros $(\sigma^2 + \tau^2)$ (patamar), $\phi$ (alcance), $\tau^2$ (efeito pepita) e $\sigma^2$ (contribuição).....	28
Figura 2	Cartograma dos 68 setores censitários considerados no estudo, com os limites de área e respectivos centroides .....	44
Figura 3	Número de mortes infantis ocorridas na cidade de Alfenas, MG, entre 2000 e 2004, localizadas em cada setor censitário pelo respectivo centroide.....	46
Figura 4	Gráfico de barras do número de localizações amostrais que apresentaram de 0 a 7 casos de mortes infantis na cidade de Alfenas, MG, entre 2000 e 2004 .....	55
Figura 5	Localização pontual das classes (superior esquerdo), valores contra as coordenadas (superior direito e inferior esquerdo) e gráfico com representação espacial dos dados de mortalidade infantil na cidade de Alfenas, MG, ocorrida entre 2000 e 2004....	57
Figura 6	Semivariograma empírico para os dados de mortalidade infantil .....	58
Figura 7	Semivariograma empírico ajustado pelo modelo esférico.....	59
Figura 8	Mapas de predição: covariável “número de mulheres em idade fértil” (à esquerda) e variável “mortalidade infantil” (à direita), para a cidade de Alfenas, MG, entre 2000 e 2004.....	65
Figura 9	Mapas de predição: covariável “mulheres em idade de risco gestacional” (à esquerda) e variável “mortalidade infantil” (à direita) para a cidade de Alfenas, MG, entre 2000 e 2004 .....	65

## LISTA DE QUADROS E FIGURAS

Quadro 1	Covariáveis utilizadas no estudo da mortalidade infantil, selecionadas no banco de dados do IBGE, censo de 2000.....	48
Tabela 1	Análise descritiva para os dados de mortalidade infantil na cidade de Alfenas, MG, referentes aos anos de 2000 a 2004.....	53
Tabela 2	Valores de AIC para os modelos ajustados ao semivariograma empírico gerado para os dados de mortalidade infantil usando geoestatística clássica .....	59
Tabela 3	Valores de AIC para os modelos ajustados aos dados de mortalidade infantil por meio da geoestatística baseada em modelos.....	61
Tabela 4	Valores do erro absoluto e de AIC para a predição de valores usando a geoestatística clássica e a geoestatística baseada em modelos.....	61
Tabela 5	Algumas análises descritivas para as covariáveis representadas pelas letras de A a G descritas na Tabela 1.....	62
Tabela 6	Valores do coeficiente de correlação de Spearman com os respectivos p-valores para a relação entre cada covariável e a variável dependente mortalidade infantil.....	64

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b> .....	13
<b>2</b>	<b>REFERENCIAL TEÓRICO</b> .....	15
<b>2.1</b>	<b>Mortalidade infantil</b> .....	15
<b>2.2</b>	<b>Estatística espacial</b> .....	20
<b>2.3</b>	<b>Geoestatística</b> .....	23
<b>2.4</b>	<b>Geoestatística baseada em modelos</b> .....	35
<b>2.4.1</b>	<b>Modelo linear gaussiano geoestatístico</b> .....	36
<b>2.4.2</b>	<b>Modelos lineares generalizados geoestatísticos</b> .....	36
<b>2.4.2.1</b>	<b>Modelo de Poisson log-linear</b> .....	40
<b>2.5</b>	<b>Estimação de parâmetros para modelos lineares generalizados geoestatísticos não gaussianos</b> .....	40
<b>3</b>	<b>MATERIAL E MÉTODOS</b> .....	43
<b>3.1</b>	<b>Caracterização da área do estudo</b> .....	43
<b>3.2</b>	<b>Dados utilizados e método de amostragem</b> .....	44
<b>3.3</b>	<b>Análise exploratória</b> .....	46
<b>3.4</b>	<b>Geoestatística baseada em modelos: modelo linear generalizado geoestatístico para os dados de mortalidade infantil</b> .....	49
<b>3.5</b>	<b>Escolhendo entre a geoestatística clássica e a geoestatística baseada em modelos como metodologia mais apropriada para o estudo da mortalidade infantil</b> .....	50
<b>3.6</b>	<b>As covariáveis e a variável mortalidade infantil</b> .....	51
<b>3.7</b>	<b>Software</b> .....	51
<b>4</b>	<b>RESULTADOS E DISCUSSÃO</b> .....	53
<b>4.1</b>	<b>Análise exploratória para a variável resposta mortalidade infantil</b> .....	53

<b>4.2</b>	<b>Escolha entre a geoestatística clássica e a geoestatística baseada em modelos na predição da mortalidade infantil .....</b>	<b>60</b>
<b>4.3</b>	<b>Covariáveis e o modelo linear generalizado geoestatístico.....</b>	<b>62</b>
	<b>REFERÊNCIAS.....</b>	<b>68</b>

## 1 INTRODUÇÃO

A estatística espacial vem sendo cada vez mais utilizada no meio científico. Ao se propor a inclusão das coordenadas geográficas no estudo de um determinado fenômeno, há a necessidade de uma técnica que seja capaz de analisar os dados, explicitando o papel dessas coordenadas e isso é o que a estatística espacial se propõe a fazer.

O fato de ser possível incluir as coordenadas geográficas na análise do evento torna o processo mais seguro, já que permite que áreas onde há maior ou menor incidência do fenômeno sejam devidamente identificadas, o que possibilita ações de intervenção direta.

A relevância do tema mortalidade infantil está no fato de ele ser um indicativo da situação de bem-estar social de uma população, o que faz com que os estudos que promovem o conhecimento de regiões onde há uma necessidade maior de intervenção por parte dos meios de saúde auxiliem no que diz respeito à promoção de melhorias sociais para a população ali instalada.

No presente estudo foram utilizadas geoestatística clássica e geoestatística baseada em modelos, sendo esta última, bastante recente, proposta por Diggle, Moyeed e Tawn (1998) e tendo como diferença principal da primeira a explicitação de um modelo estatístico representante dos dados e métodos de inferência estatística baseados na máxima verossimilhança.

Foram utilizados dados de mortalidade infantil do município de Alfenas, MG, referentes aos anos de 2000 a 2004. O objetivo foi modelar o fenômeno da mortalidade infantil no referido município, utilizando as metodologias de geoestatística clássica com estimação por máxima verossimilhança e geoestatística baseada em modelos, assumindo uma distribuição de Poisson para os dados e fazendo estimação de parâmetros por máxima verossimilhança Monte Carlo. Ao final, selecionou-se, através dos critérios de Akaike e do erro

absoluto, uma das duas metodologias como a mais adequada ao estudo da mortalidade infantil.

Dado o fato de que as mortes infantis nem sempre ocorrem de maneira isolada, procurou-se ainda envolver covariáveis no estudo, a fim de estabelecer uma relação entre elas e a mortalidade infantil.

O trabalho dividiu-se em três etapas. Na primeira, fez-se a análise exploratória, em que o semivariograma empírico verificou a dependência espacial entre os dados. Na segunda, efetuou-se a modelagem utilizando geoestatística clássica e geoestatística baseada em modelos, fazendo a seleção da metodologia mais adequada por meio dos critérios de Akaike e do erro absoluto. Na terceira, realizou-se a predição de mortes em locais não amostrados, utilizando-se parâmetros estimados por meio da metodologia selecionada na segunda etapa e a verificação da existência de correlação entre a mortalidade infantil e as covariáveis selecionadas para o estudo.

## **2 REFERENCIAL TEÓRICO**

A mortalidade infantil tem sido um assunto discutido em todo o mundo, desde países pobres até potências mundiais. Essa atenção deve-se à sua relevância como um dos mais usados indicadores do quadro de bem-estar social de uma população, já que o tema abrange serviços de saúde, educação e distribuição de renda, entre outros. Logo, seu estudo se torna de suma importância para o conhecimento da qualidade de tais serviços e, conseqüentemente, possíveis ações de intervenção.

### **2.1 Mortalidade infantil**

Entendem-se por mortalidade infantil, óbitos de crianças menores de um ano, nascidas vivas. Tradicionalmente, divide-se em dois grupos (considerando-se a época de ocorrência dos óbitos): mortalidade neonatal e mortalidade pós-neonatal. A mortalidade neonatal compreende a mortalidade neonatal precoce, que se refere aos óbitos nos seis primeiros dias de vida e a mortalidade neonatal tardia, que diz respeito aos óbitos do 7º ao 27º dia de vida. A mortalidade pós-neonatal se refere aos óbitos do 28º dia de nascimento até o 12º mês de vida (NOGUEIRA, 2008).

A obtenção de dados regulares sobre mortalidade infantil é feita utilizando-se uma importante fonte, o Sistema de Informações sobre Mortalidade (SIM), enquanto os dados sobre nascidos vivos são obtidos através do Sistema de Informações sobre Nascidos Vivos (SINASC).

Os dados do SINASC são provenientes da Declaração de Nascidos Vivos (DN), instrumento padronizado e impresso pelo Ministério da Saúde e distribuído à Diretoria de Informação e Comunicação em Saúde (DICS) e repassado às Diretorias Regionais de Saúde (DIRES) para distribuição entre as

Secretarias Municipais de Saúde (SMS). O SIM tem seus dados provenientes da Declaração de Óbito (DO), impressa pelo Ministério da Saúde e repassada às Secretarias Estaduais de Saúde (SES) e, posteriormente, por estas às Secretarias Municipais de Saúde (SMS), aos Institutos de Medicina Legal, aos Serviços de Verificação de Óbito (SVO) e aos Cartórios do Registro Civil (GOTLIEB; LAURENTI; MELLO JORGE, 2007).

O interesse no estudo da mortalidade infantil está em encontrar formas de intervir positivamente no bem-estar da criança, minimizando ao máximo as ocorrências de mortes, principalmente no primeiro ano de vida, que se constitui no mais importante, visto que, acima desta idade, a criança já tem condições maiores de sobrevivência.

Costa et al. (2001) comentam que as crianças com menos de um ano de idade tornam-se mais ou menos vulneráveis devido às alterações que ocorrem no seu ambiente social e econômico e, ainda, devido às intervenções de saúde. Logo, nessa faixa etária, a mortalidade é considerada um indicador, tanto da situação de saúde quanto das condições em que uma população vive.

Segundo um estudo feito pelo Fundo das Nações Unidas para Infância - UNICEF (2004), dentre 49 nações, as que apresentaram a maior média anual de crescimento entre 1990 e 2000 foram aquelas que em 1980 apresentaram índices de baixa mortalidade infantil e pobreza. Enquanto no mesmo período o contrário ocorreu com economias que tiveram altos índices de mortalidade infantil ou altos índices de pobreza ou ambos.

De acordo com Kozu et al. (2001), em quase todo o mundo, as taxas de mortalidade infantil têm decrescido bastante, mas, em diversos países, os números ainda são altos. Como taxa (ou coeficiente) de mortalidade infantil, considera-se a relação entre o número de óbitos de crianças menores de um ano e o número de nascidos vivos em determinado local e calculado na base de mil nascidos vivos.

Carvalho e Costa (1998) confirmam essa tendência de decréscimo da mortalidade infantil em países como México, Costa Rica, Equador e também no Brasil, embora na maioria dos países da América Latina os coeficientes ainda apresentem-se elevados.

Na Europa, o destaque dado por Laurenti (1987) é para a Suécia que, desde 1920, tem se destacado como o país que apresenta os menores coeficientes de mortalidade infantil, tendo perdido esse posto apenas em 1981, para a Finlândia.

Segundo dados do Instituto Brasileiro de Geografia e Estatística - IBGE (2008), o Brasil tem reduzido consideravelmente as taxas de mortalidade infantil. Observando-se o período de 1990 a 2008, essa taxa de mortalidade (por mil nascidos vivos) passou de 47,0, em 1990, para 23,3, em 2008, com acentuada queda nos anos de 1990 para 2001, quando houve um decréscimo de aproximadamente 62% nessa taxa.

De acordo com Duarte et al. (2002), mesmo havendo uma queda contínua no coeficiente de mortalidade infantil brasileiro, ressaltando o período de 1991 a 1999, quando a taxa (por mil nascidos vivos) passou de 44,1 para 31,8, esse indicador demonstra que, entre as regiões brasileiras, há ainda uma grande diferença no que diz respeito à diminuição desse coeficiente. Na região nordeste, ele é de 52,5, enquanto na região sul não passa de 17,1 e, entre estados, é de 64,0 em Alagoas, para 15,1 no Rio Grande do Sul.

Costa (2008) ressalta a necessidade de intervenção em algumas áreas, visando níveis mínimos para índice de mortalidade infantil. Procurar formas de melhorar a saúde da mulher, a assistência ao pré-natal, ao parto e ao recém-nascido em sala de parto e berçário e, prestar atenção significativa às condições sócio-econômicas da população com baixa renda são de fundamental importância para se atingir esse objetivo.

Simões (2002) também sugere que os níveis de mortalidade infantil poderiam diminuir bastante se fossem organizadas ações e introduzidas medidas de controle e políticas públicas voltadas diretamente para o aperfeiçoamento da atenção às crianças. Dessa forma, diminuiriam as desigualdades sociais, que excluem grande parte da população do acesso à educação, melhoria de renda, saneamento e serviços de saúde básicos relacionados ao pré-natal e seu acompanhamento.

Ainda segundo Simões (2002), outra questão associada ao alto índice de mortalidade infantil é, muitas vezes, a falta do registro de nascimento que permite, em caso da morte da criança, sua contabilização nas estatísticas oficiais, favorecendo os resultados das análises dos perfis epidemiológicos e facilitando a implantação adequada de políticas de saúde.

Em muitos estudos brasileiros, de acordo com Carvalho e Costa (1998), a associação entre mortalidade infantil e variáveis sócio-econômicas tem sido evidenciada, ainda que as conclusões desses estudos sejam postas em risco devido à indisponibilidade de dados mais confiáveis e precisos.

Em estudos no município de Pelotas, RS, sobre fatores de risco para a mortalidade perinatal (óbitos ocorridos a partir da 22ª semana completa de gestação até o 7º dia completo de vida), foi demonstrado que, em famílias com renda mensal abaixo de um salário mínimo, esse tipo de mortalidade era três vezes maior que em famílias com maior renda mensal. Generalizando para a mortalidade infantil precoce (óbitos ocorridos nos seis primeiros dias de vida), o maior risco de morte se relacionou com o nível sócio-econômico das mães. Em uma investigação realizada no município de Santo André, SP, o problema foi o nível de escolaridade inferior ao ensino fundamental completo, que mostrou ser um fator de risco relevante para baixo peso ao nascer (ANDRADE et al., 2004).

Caldwell (1979) enfoca que, dentre os fatores sócio-econômicos que determinam a saúde infantil, a educação da mãe e a renda têm sido considerados

como de suma importância, por serem indicadores de disponibilidade de recursos e conhecimento ou comportamento em relação à saúde da criança.

Visando à redução dos índices de mortalidade infantil no Brasil, uma parte expressiva das iniciativas analisadas corresponde a programas de saúde materno-infantil, como o Projeto Casulo, de São José dos Campos, SP e o Mãe Canguru, do estado de Pernambuco, programas de atendimento a gestantes e recém-nascidos de risco. Também citam-se os programas de saúde da família, de Campina Grande, PB e de Joinville, SC, que são mais abrangentes e incluem ações dirigidas a gestantes e a recém-nascidos. Tais programas e subprogramas estão focados nas políticas sociais, uma vez que tornam as mulheres gestantes e seus filhos grupos de risco (FARAH, 2004).

Keen, Rip e Woods (1987) realizaram um estudo na área metropolitana da Cidade do Cabo, África do Sul, e demonstraram que quanto menor o nível sócio-econômico maior é a taxa de mortalidade infantil. Esse estudo foi realizado com o objetivo de determinar a variação espacial da mortalidade neonatal no subúrbio dentro da cidade. Por meio da análise espacial, constatou-se que variações na taxa de mortalidade infantil eram significativamente correlacionadas com as variações sócio-econômicas.

Nascimento et al. (2007) afirmam que, de forma geral, as análises estatísticas utilizadas nos estudos sobre mortalidade infantil servem para determinar fatores associados ao óbito, tanto no período neonatal quanto no período pós-neonatal.

Segundo Leal e Szwarcwald (1997), a descrição da distribuição espacial dos fenômenos enriquece bastante a análise de dados. A incorporação da dimensão espacial não apenas delimita geograficamente o problema, mas permite fazer uma análise integrada que inclui o estudo da dependência no espaço, percorrendo campos da epidemiologia, da geografia e da estatística.

Ainda de acordo com Leal e Szwarcwald (1997), o desenvolvimento recente da geoestatística veio reforçar sua utilização nos estudos epidemiológicos. Ao se detectar a existência de um componente espacial por meio de instrumental estatístico, a análise passa a verificar os fatores ou processos que geram a não aleatoriedade da distribuição espacial da doença e a identificação de áreas de maior interesse para posteriores investigações.

## **2.2 Estatística espacial**

O termo estatística espacial é empregado para descrever a coleção de métodos estatísticos nos quais a localização espacial desempenha papel explícito na análise de dados (DIGGLE; RIBEIRO JÚNIOR, 2001).

Os métodos utilizados para o estudo na área da saúde, desde o final da década de 1980, têm se caracterizado por destacar a localização espacial do evento e os Sistemas de Informações Geográficas (SIGs) (CARVALHO; SANTOS, 2005).

Para Guimarães (2004), a análise espacial dos dados pode complementar a análise clássica destes, uma vez que, na análise clássica, em geral, as realizações das variáveis aleatórias são independentes. Logo, observações vizinhas não sofrem influência umas das outras, o que não acontece na análise espacial, pois esta considera, em suas estimativas, as correlações existentes entre as observações.

Grande parte dos eventos que ocorrem na natureza ou em meio social apresentam entre si uma relação que depende da distância. De acordo com Cressie (1993), existe dependência em todas as direções e, quando a dispersão na localização dos dados vai aumentando, essa dependência vai ficando mais fraca.

A característica fundamental de uma técnica de análise espacial é que a referência geográfica é utilizada explicitamente no modelo, ou seja, faz-se uso explícito das coordenadas espaciais dos dados no processo de coleta, descrição ou análise dos mesmos (ASSUNÇÃO, 2001).

Segundo Druck et al. (2004), os procedimentos de análise espacial são utilizados para escolher um modelo inferencial que faça uso direto dos relacionamentos espaciais presentes no fenômeno. Isso acontece após ser feita uma análise exploratória dos dados, com apresentação dos mesmos em forma de gráficos e mapas e identificação de padrões de dependência espacial no fenômeno em estudo.

O processo de análise espacial, segundo Bailey e Gatrell (1995), conceitualmente, inclui três etapas básicas inter-relacionadas: a visualização espacial, a análise exploratória e a modelagem dos dados.

Para Rodriguez Yi et al. (2000), a visualização espacial serve para explorar visualmente o dado objeto de análise, por meio da criação de mapas, com o auxílio de banco de dados geográficos; a análise exploratória permite descrever a distribuição espacial dos dados e os padrões de associação espacial existentes ou padrões de agrupamento, logo, é uma etapa considerada como essencial para a modelagem dos dados e a modelagem é aquela na qual os procedimentos de validação estatística e modelos de estimação dos dados serão incluídos. É na fase da modelagem que se afirma, por exemplo, que certo fenômeno pode ser considerado uma variável aleatória e com determinada distribuição de probabilidade.

A classificação mais utilizada para descrever o problema da análise e modelagem espacial considera três tipos de dados, conforme pode ser visto em Bailey e Gatrell (1995), Cressie (1993) e Druck et al. (2004): eventos (ou padrões pontuais), superfícies contínuas (geoestatística) e dados discretos (ou áreas com contagens).

**Eventos ou padrões pontuais:** os fenômenos são expressos por meio de ocorrências identificadas como pontos localizados no espaço. São exemplos desse tipo de dado a localização da ocorrência de casos de doenças e a localização de espécies de plantas.

**Superfícies contínuas (geoestatística):** são fenômenos que se distribuem continuamente em uma região. Usualmente, esse tipo de dado é resultante de levantamento de recursos naturais e que incluem mapas geológicos, topográficos e ecológicos. Um exemplo desse tipo de dado é a medida da concentração de um elemento químico no solo.

**Áreas com contagens:** são fenômenos associados aos dados de levantamentos populacionais, como os censos, e que, originariamente, se referem a indivíduos localizados em pontos específicos no espaço. Normalmente, esses dados são agregados em unidades de análises, usualmente delimitadas por polígonos fechados, tais como setores censitários, municípios e microrregiões.

De acordo com Nascimento et al. (2007), a análise espacial e o mapeamento de eventos são importantes instrumentos para a saúde pública, pois auxiliam nos diagnósticos e também no tratamento, o que favorece a compreensão da ocorrência dos eventos.

Shimakura et al. (2001) desenvolveram um trabalho em Porto Alegre, RS, utilizando métodos de análise espacial, que foram aplicados para mensurar o número de óbitos neonatais e também estabelecer as áreas de risco na referida cidade.

Morais Neto et al. (2001) utilizaram técnicas de análise de dados espaciais para investigar o padrão espacial da mortalidade neonatal e pós-neonatal no espaço urbano de Goiânia, GO e, ainda, produzir mapas que identificassem áreas de risco para os dois componentes da mortalidade infantil.

Nogueira (2008) utilizou as bases teóricas da análise espacial para estudar a relação existente entre a mortalidade infantil e algumas covariáveis no município de Alfenas, MG. Em seu estudo, o autor fez uso dos modelos lineares generalizados geoestatísticos, assumindo uma distribuição binomial para os dados, com  $n$  e  $p$  conhecidos, utilizando, na modelagem, o modelo logístico, via inferência bayesiana. Tomando como base tal estudo, foi que o presente trabalho se desenvolveu.

### 2.3 Geoestatística

A estatística clássica supõe que as variáveis aleatórias são independentes entre si, ou seja, as observações vizinhas não exercem influência umas sobre as outras e as posições relativas das amostras são ignoradas. Na prática, isso, muitas vezes, não acontece e é esse, então, o ponto no qual a geoestatística se inicia.

Segundo Diggle e Ribeiro Júnior (2001), o objetivo principal da geoestatística é modelar a variação espacial contínua, assumindo uma estrutura do tipo  $Y(x)$ ,  $x \in \mathbb{R}^d$ ,  $d = 1, 2$  ou  $3$ , sendo  $Y$  a variável aleatória a ser medida em uma região de interesse em localizações  $x$ . A dimensão espacial utilizada neste trabalho foi  $d=2$ . Druck et al. (2004) ainda acrescentam que a geoestatística procura reconstruir a superfície da qual se retiraram e mediram as amostras.

Os problemas geoestatísticos apresentam, então, as seguintes características:

- a) respostas  $Y_i$  associadas a localizações  $\mathbf{x}_i$ ;
- b) em princípio,  $Y$  poderia ser determinado em qualquer localização  $\mathbf{x}$  da região espacial contínua em estudo,  $A \subset \mathbb{R}^2$ ;

- c)  $\{Y(x) : x \in A \subset \mathbb{R}^2\}$ , chamado processo de medição, está relacionado a um processo estocástico não observado  $\{S(x) : x \in A \subset \mathbb{R}^2\}$ , que é chamado de processo sinal;
- d) objetivos científicos incluem predição de um ou mais funcionais do processo estocástico  $\{S(x) : x \in A \subset \mathbb{R}^2\}$ .

O conceito de geoestatística iniciou-se com Krige (1951 apud VIEIRA, 2000), que concluiu, em seus trabalhos com dados de concentração de ouro, que somente a informação dada pela variância não seria suficiente para explicar o fenômeno em estudo, sendo necessário, portanto, levar em consideração a distância entre as observações.

Matheron (1963), baseado nas observações de Krige, desenvolveu a teoria das variáveis regionalizadas e que contém os fundamentos da geoestatística. Este autor define variável regionalizada como uma função espacial numérica, que varia de um local para outro, com uma continuidade aparente e cuja variação não pode ser representada por uma função matemática simples. Essa continuidade ou dependência espacial pode ser estimada pelo semivariograma.

Segundo Ribeiro Júnior (1995), cada observação carrega não apenas a informação relacionada ao seu valor, mas também à sua posição no espaço, que é expressa por um sistema de coordenadas. Considerando que observações mais próximas têm valores mais parecidos, o que pode ser avaliado por medidas de associação, tem-se, então, que a geoestatística não pode trabalhar com a suposição de independência entre as observações.

A geoestatística, então, calcula estimativas dentro de um contexto regido por um fenômeno natural com distribuição no espaço e supõe que os valores das variáveis regionalizadas são espacialmente correlacionados. Devido a essa característica, tem sido grande a sua aplicação, principalmente para efetuar

estimativas e/ou simulações de variáveis em locais não amostrados (LANDIM; STURARO, 2002).

Um método de interpolação chamado krigagem, cujo nome homenageia o matemático sul-africano D. G. Krige, é utilizado com o objetivo de estimar valores em qualquer posição dentro do campo de estudo, sem tendência e com variância mínima, utilizando a dependência espacial entre amostras vizinhas (VIEIRA, 2000).

Assim, tem-se que a hipótese central de trabalho da geoestatística é o conceito de estacionaridade, que supõe um comportamento homogêneo da estrutura de correlação espacial na região de estudo.

Segundo Oliveira (2003), são três as hipóteses de estacionaridade de uma função aleatória: hipótese de estacionaridade de 1ª ordem, hipótese de estacionaridade de 2ª ordem e hipótese intrínseca.

Uma função aleatória  $Y(x)$  é estacionária de 1ª ordem quando todas as variáveis aleatórias  $Y(x_i)$ ,  $i = 1, \dots, n$  têm uma mesma média  $m$  em todas as localizações da região em estudo  $A \subset \mathbb{R}^2$ , o que pode ser definido como

$$E[Y(x_1)] = E[Y(x_2)] = E[Y(x_3)] = \dots = E[Y(x_i)] = E[Y(x)] = m$$

A estacionaridade de 2ª ordem para uma função aleatória  $Y(x)$  ocorre quando, além de atender à hipótese de estacionaridade de 1ª ordem, para cada par de variáveis aleatórias  $Y(x)$  e  $Y(x+\mathbf{h})$ , a função de covariância  $C(x, x+\mathbf{h}) = C(\mathbf{h})$ , além de existir, é função de  $\mathbf{h}$ , ou seja, para que uma função seja estacionária de 2ª ordem, ela deverá satisfazer a:

- a)  $E[Y(x)] = m, \forall x$  dentro da área em estudo  $A \subset \mathbb{R}^2$ ;
- b)  $C(x, x+\mathbf{h}) = C(\mathbf{h}) = E\{Y(x) Y(x+\mathbf{h})\} - m^2, \forall x$  dentro da área em estudo  $A \subset \mathbb{R}^2$ .

A existência de covariância e, em conseqüência, de uma variância finita dos valores medidos, é pressuposto da hipótese de estacionaridade de 2ª ordem. Neste caso, a expressão para variância será  $\text{Var}[Y(x)] = \text{Cov}(0)$ .

Segundo Oliveira (2003) e Vieira (2000), para alguns fenômenos físicos que apresentam dispersão infinita, esta hipótese de variância finita, considerada muito forte, pode não ser satisfeita, podendo, portanto, ser substituída por uma hipótese menos restritiva, mais fraca, chamada de hipótese intrínseca.

A hipótese intrínseca é definida da seguinte forma:

- a) o incremento entre dois pontos da região em estudo tem esperança nula, ou seja,  $E[Y(x) - Y(x+h)] = 0, \forall \mathbf{h}$ ;
- b) independentemente da localização  $x$  na região em estudo, para qualquer vetor de distância  $\mathbf{h}$ , o incremento tem variância finita, ou seja,  $\text{Var}[Y(x) - Y(x+h)] = E[Y(x) - Y(x+h)]^2 = 2\gamma(\mathbf{h})$ , em que  $2\gamma(\mathbf{h})$  corresponde ao variograma.

Para a aplicação de geoestatística, basta que a hipótese de estacionaridade de 2ª ordem seja satisfeita (VIEIRA, 2000).

Para se descrever quantitativamente a variação no espaço de um fenômeno regionalizado, utiliza-se o semivariograma (VIEIRA, 2000). Isso quer dizer que o semivariograma é a ferramenta por meio da qual a dependência espacial em um fenômeno é quantificada.

O semivariograma é, como já definido anteriormente, dado por Guimarães (2004)

$$\gamma(\mathbf{h}) = \frac{1}{2} E \left\{ [Y(x) - Y(x+h)]^2 \right\},$$

e pode ser estimado por

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [Y(x) - Y(x+\mathbf{h})]^2, \quad (2.1)$$

em que  $N(\mathbf{h})$  é o número de pares de valores medidos e  $Y(x)$  e  $Y(x+\mathbf{h})$  são separados por um vetor distância  $\mathbf{h}$ .

De acordo com e Diggle e Ribeiro Júnior (2001), o semivariograma apresenta como parâmetros o patamar, o alcance e o efeito pepita. Cada um desses parâmetros apresenta características próprias, assim:

- a) o patamar ( $\sigma^2 + \tau^2$ ) representa o valor máximo no qual  $\hat{\gamma}(\mathbf{h})$  se estabiliza, tornando-se, a partir daí, independente à variação de  $\mathbf{h}$ ;
- b) o alcance ( $\phi$ ) é o valor de  $\mathbf{h}$  para o qual essa estabilização ocorre, sendo este um parâmetro importante no estudo da variabilidade espacial, pois é a distância limite na escolha entre utilizar a geoestatística e a estatística clássica, já que, nas distâncias menores que o alcance, as observações amostrais são espacialmente dependentes e usa-se a geoestatística, enquanto nas distâncias maiores que o alcance pode-se, então, utilizar a estatística clássica;
- c) o efeito pepita ( $\tau^2$ ) é o valor positivo do qual  $\hat{\gamma}(\mathbf{h})$  se aproxima quando  $\mathbf{h}$  tende para zero.

Costuma-se chamar  $\sigma^2$  de contribuição, sendo este encontrado pela diferença entre o patamar ( $\sigma^2 + \tau^2$ ) e o efeito pepita  $\tau^2$ .

Na Figura 1, pode-se visualizar um semivariograma com os parâmetros acima citados.

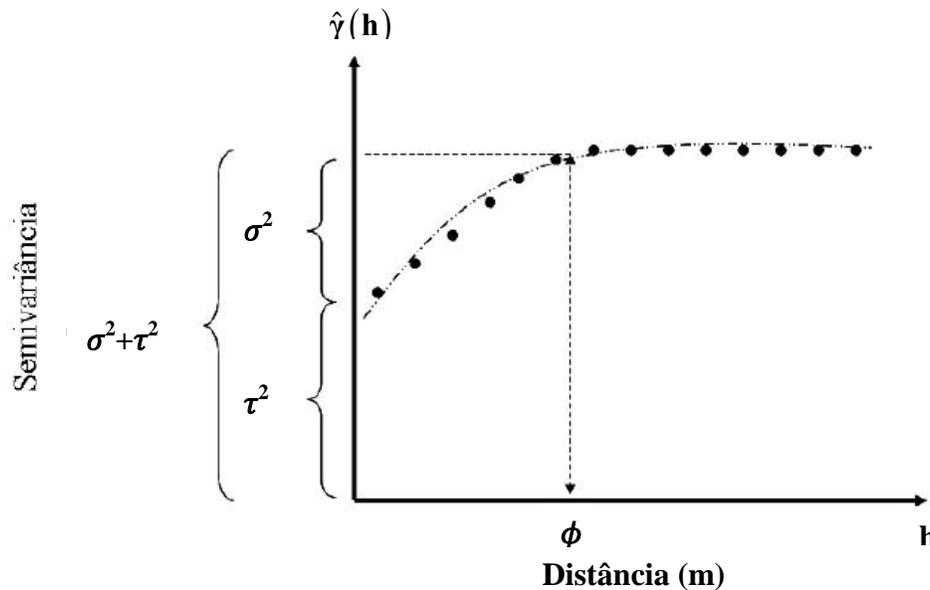


Figura 1 Semivariograma empírico hipotético apresentando os parâmetros ( $\sigma^2 + \tau^2$ ) (patamar),  $\phi$  (alcance),  $\tau^2$  (efeito pepita) e  $\sigma^2$  (contribuição)

O semivariograma da Figura 1 é um exemplo do que se deseja obter ao trabalhar com dados espaciais em geoestatística. Existem vários modelos teóricos que podem ser ajustados aos dados para se obter um semivariograma com características semelhantes às aquelas apresentadas na Figura 1.

O ajuste de modelos teóricos aos semivariogramas experimentais é uma etapa muito importante da análise geoestatística, já que o modelo escolhido permite interpretar a estrutura de correlação espacial a ser utilizada nos procedimentos de inferência de krigagem (BIGGAR; NIELSEN; VIEIRA, 1981), que serão discutidos mais adiante.

A seguir, são dadas as equações para alguns modelos teóricos conhecidos, caracterizados como modelos com patamar. São eles: linear, esférico, exponencial e gaussiano (VIEIRA, 2000).

a) modelo linear

$$\gamma(\mathbf{h}) = \begin{cases} \tau^2 + \frac{\sigma^2}{\phi} \mathbf{h} & ; 0 < \mathbf{h} < \phi \\ \tau^2 + \sigma^2 & ; \mathbf{h} \geq \phi \end{cases} \quad (2.1)$$

b) modelo esférico

$$\gamma(\mathbf{h}) = \begin{cases} \tau^2 + \sigma^2 \left[ \frac{3}{2} \left( \frac{\mathbf{h}}{\phi} \right) - \frac{1}{2} \left( \frac{\mathbf{h}}{\phi} \right)^3 \right] & ; 0 < \mathbf{h} < \phi \\ \tau^2 + \sigma^2 & ; \mathbf{h} \geq \phi \end{cases} \quad (2.2)$$

c) modelo exponencial

$$\gamma(\mathbf{h}) = \tau^2 + \sigma^2 \left[ 1 - \exp\left(-3 \frac{\mathbf{h}}{\phi}\right) \right] ; 0 < \mathbf{h} \leq d, \quad (2.3)$$

em que  $d$  é a máxima distância na qual o semivariograma é definido.

d) modelo gaussiano

$$\gamma(\mathbf{h}) = \tau^2 + \sigma^2 \left[ 1 - \exp\left(-3 \frac{\mathbf{h}}{\phi}\right)^2 \right] ; 0 < \mathbf{h} < d. \quad (2.4)$$

Uma diferença fundamental entre os modelos exponencial e esférico é que o exponencial atinge o patamar apenas assintoticamente, enquanto o modelo esférico o atinge no valor do alcance ( $\phi$ ).

Diggle e Ribeiro Júnior (2000) sugerem o uso da família Matérn para o estudo da estrutura de correlação espacial em um modelo Gaussiano. A função de correlação de Matérn apresenta um parâmetro  $\nu$  que determina a suavização

analítica do sinal  $S(x)$  em um modelo Gaussiano e é chamado de ordem do modelo Matérn. A função de correlação de Matérn é assim definida

$$\rho(\mathbf{h}; \phi, \kappa) = \{2^{\kappa-1} \Gamma(\kappa)\}^{-1} \left(\frac{\mathbf{h}}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{\mathbf{h}}{\phi}\right), \quad (2.5)$$

em que  $\phi$  e  $\kappa$  são parâmetros,  $K_{\kappa}$  é a função de Bessel de ordem  $\kappa$  e  $\mathbf{h}$  é o vetor distância.

Diggle e Ribeiro Júnior (2000) comentam que a família Matérn é válida para  $\phi$  e  $\kappa > 0$ . No caso em que  $\kappa = 0,5$ , a função de correlação corresponde à exponencial, que se apresenta por  $\rho(\mathbf{h}) = \exp(-\mathbf{h}/\phi)$  e, quando  $\kappa \rightarrow \infty$ , a função de correlação é a Gaussiana. Ainda segundo os autores, a vantagem no uso da família Matérn está no fato de o parâmetro  $\kappa$  conseguir controlar a suavidade do processo sinal subjacente.

De acordo com Cressie (1985), para se escolher um modelo adequado, é necessário obter os estimadores dos parâmetros efeito pepita  $\tau^2$ , patamar  $(\sigma^2 + \tau^2)$  e alcance  $\phi$ , empregando, para isso, métodos estatísticos de otimização, tais como mínimos quadrados ordinários, mínimos quadrados ponderados e máxima verossimilhança.

Esses métodos podem, segundo Mello et al. (2005) ser divididos em dois grupos, que são:

- a) ajuste dos modelos ao semivariograma experimental, do qual fazem parte os métodos de mínimos quadrados ordinários (Ordinary Least Squares – OLS) e mínimos quadrados ponderados (Weight Least Squares – WLS);
- b) ajuste de um modelo diretamente aos dados, do qual faz parte o método da máxima verossimilhança (Maximum likelihood – ML).

O método dos mínimos quadrados, tanto ordinários quanto ponderados, consiste em obter os valores dos parâmetros de um modelo que procuram minimizar a soma do quadrado da diferença entre os valores observados e os estimados.

A expressão seguinte representa, de acordo com Mello et al. (2005), a estimativa dos parâmetros do semivariograma por meio do método dos mínimos quadrados ordinários

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^k h_i [g(h_i) - \gamma(h_i; \boldsymbol{\theta})]^2, \quad (2.6)$$

em que  $\boldsymbol{\theta}$  é o vetor de parâmetros estimados de dimensões  $p \times 1$ ,  $g(h_i)$  são as estimativas,  $k$  refere-se ao número de distâncias discretas (*lags*) do semivariograma experimental e  $\gamma(h_i; \boldsymbol{\theta})$  é a semivariância calculada pelo modelo e depende dos parâmetros estimados  $\boldsymbol{\theta}$  e, após minimização, a função  $g(h_i)$  se torna a estimativa  $\hat{\gamma}(h_i)$  de quadrados mínimos.

Ainda segundo Mello et al. (2005), a expressão a seguir é a que representa a estimativa dos parâmetros do semivariograma pelo método dos mínimos quadrados ponderados

$$Q(\boldsymbol{\theta}) = \frac{\sum_{i=1}^k [g(h_i) - \gamma(h_i; \boldsymbol{\theta})]^2}{m(h_i)},$$

em que  $m(h_i)$  é o número de pares de pontos em cada distância no semivariograma.

Segundo os autores, essa é uma tentativa de ponderar a semivariância, já que a ponderação pela variância das estimativas das semivariâncias não é conhecida.

Segundo Myung (2003), uma vez que os dados foram coletados e a função de verossimilhança de um modelo, dados os valores observados, foi

determinada, utiliza-se a estimação por máxima verossimilhança, de modo a se obter os valores das estimativas dos parâmetros da função de verossimilhança com possibilidade máxima de terem produzido os valores observados de  $Y$ .

Este processo de estimação é realizado maximizando-se o logaritmo da função de verossimilhança, o que, segundo Ribeiro Júnior (1995), é conseguido com o uso de procedimentos numéricos iterativos.

Diggle e Ribeiro Júnior (2007) salientam que, no contexto geoestatístico, a estimação por máxima verossimilhança só é simples quando os dados são provenientes de uma distribuição Gaussiana, mas é possível implementar este tipo de estimação em muitas aplicações geoestatísticas em que a variável resposta  $Y$  é contínua espacialmente.

Assim, por meio dos parâmetros, em um conjunto com o modelo ajustado, busca-se quantificar a dependência espacial da variável regionalizada em estudo.

Dos modelos teóricos de correlação espacial, um deverá ser escolhido para dar prosseguimento ao próximo passo, que é a interpolação de valores através da krigagem. Para tal, é necessário que os modelos sejam validados.

Segundo Isaaks e Srivastava (1989), o que os métodos de validação de modelos fazem é comparar valores teóricos do modelo geoestatístico escolhido com valores empíricos obtidos por meio do processo de amostragem. Assim, o melhor modelo de semivariância é escolhido com base na análise desses erros de estimação.

Aqui serão comentados dois dos métodos usados para selecionar modelos: a validação cruzada e o critério de informação de Akaike (Akaike's Information Criterion – AIC).

O método da validação cruzada é uma técnica em que se utilizam os erros de estimativas para comparar os valores medidos com os valores preditos. Funciona da seguinte forma: um valor  $Y(x_i)$  da amostra é temporariamente

suprimido do conjunto de dados amostrais e, então, é feita uma predição por krigagem desse valor, usando-se os valores remanescentes. Logo, obter-se-á  $\hat{Y}(x_i)$ , procedimento que é repetido para todos os valores da amostra (OLIVEIRA et al., 2006).

O erro absoluto (EA), segundo Cressie (1993), serve para medir a magnitude dos erros na unidade da variável. Sendo conhecido o conjunto de valores medidos  $Y(x_i)$  e o de valores preditos  $\hat{Y}(x_i)$ , o erro absoluto na unidade da variável estudada, pode ser definido por

$$EA = \sum_{i=1}^n |Y(x_i) - \hat{Y}(x_i)|,$$

em que  $Y(x_i)$  é o valor observado na localização  $x_i$  e  $\hat{Y}(x_i)$  é o valor estimado para a localização  $x_i$ , sem considerar a observação  $Y(x_i)$ .

Outro método de seleção de modelos bastante utilizado é o Critério de Informação de Akaike, ou AIC. De acordo com Emiliano (2009), esse método fornece uma medida que serve para avaliar a qualidade do ajuste do modelo paramétrico, estimado pelo método da máxima verossimilhança, e é definido por

$$AIC = -2 \log L(\hat{\theta}) + 2(k), \quad (2.8)$$

em que  $\log L(\hat{\theta})$  é o logaritmo natural da verossimilhança maximizada para um modelo particular e  $k$  é o número de parâmetros do modelo particular.

Ainda segundo o autor, dentre vários modelos comparados, serão considerados melhores aqueles que tiverem menores valores de AIC. Sendo assim, se o objetivo for a escolha de um único modelo, o melhor será aquele com menor valor de AIC.

Tendo sido feita a escolha do melhor modelo teórico, segue-se para a interpolação de valores por meio do método chamado krigagem.

Na krigagem utilizam-se as informações do semivariograma para encontrar os pesos ótimos a serem associados às amostras com valores conhecidos que irão estimar pontos desconhecidos. Este método fornece, além dos valores estimados, o erro associado a tal estimação, o que o distingue dos demais algoritmos de interpolação (LANDIM; STURARO, 2002).

Ainda segundo Landim e Sturaro (2002), a krigagem é entendida como uma série de técnicas de análise de regressão que procuram minimizar a variância estimada a partir de um modelo prévio, que leva em consideração a dependência estocástica entre os dados distribuídos no espaço.

Segundo Matheron (1963), o estimador da krigagem ordinária é obtido da seguinte forma:

$$Y^*(x_0) = \sum_{i=1}^N \lambda_i Y(x_i), \quad (2.9)$$

em que  $Y^*(x_0)$  é o atributo da variável, estimado no ponto  $x_0$ ,  $\lambda_i$  são os ponderadores de krigagem e  $Y(x_i)$  é o valor observado da variável  $Y$  no  $i$ -ésimo ponto.

Vale lembrar que, para a krigagem ser ordinária  $\sum_{i=1}^N \lambda_i = 1$ .

Os ponderadores são obtidos pela resolução de um sistema de equações lineares do tipo  $AX=B$ , denominado sistema de krigagem que, segundo Vieira (2000), pode ser escrito da seguinte forma:

$$\begin{bmatrix} \gamma(x_1; x_1) & \cdots & \gamma(x_1; x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n; x_1) & \cdots & \gamma(x_n; x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(x_1; x_0) \\ \vdots \\ \gamma(x_n; x_0) \\ 1 \end{bmatrix}$$

em que  $\gamma(x_n; x_n)$  é a variância espacial da n-ésima amostra com relação a ela mesma,  $\mu$  é o Multiplicador de Lagrange e  $\gamma(x_n; x_0)$  é a variância espacial entre a n-ésima amostra e o ponto  $x_0$  que será estimado.

## 2.4 Geoestatística baseada em modelos

O termo geoestatística baseada em modelos foi criado por Diggle, Moyeed e Tawn (1998) com o objetivo de descrever problemas geoestatísticos fundamentados na aplicação de métodos estatísticos formais, usando um modelo explícito e métodos de inferência estatística baseados na máxima verossimilhança.

Ainda segundo Diggle, Moyeed e Tawn (1998), o modelo geoestatístico proposto por eles descreve a dependência espacial por meio de um componente de erro aleatório e espacialmente correlacionado.

Segundo Diggle e Ribeiro Júnior (2007), algumas das vantagens em se usar geoestatística baseada em modelos está em uma maior eficiência na estimativa de parâmetros desconhecidos e uma avaliação adequada da incerteza nas previsões espaciais.

Os modelos Gaussianos para dados geoestatísticos e os modelos lineares generalizados para dados geoestatísticos são os mais comentados por Diggle e Ribeiro Júnior (2007) neste tipo de abordagem.

### 2.4.1 Modelo linear gaussiano geoestatístico

Considerado por Diggle e Ribeiro Júnior (2007) como modelo de referência em geoestatística, o modelo gaussiano geoestatístico pode ser escrito como

$$Y_i = S(x_i) + Z_i,$$

em que  $\{S(x) : x \in A \subset \mathbb{R}^2\}$  é um processo Gaussiano com média  $\mu$  e variância  $\text{Var}[S(x)] = \sigma^2$  e função de correlação  $\rho(h) = \text{Corr}\{S(x), S(x')\}$ , em que  $h = \|x - x'\|$  é a distância Euclidiana entre as localizações  $x$  e  $x'$  e  $Z_i$ 's são variáveis aleatórias mutuamente independentes e seguem distribuição Normal com média 0 (zero) e variância  $\tau^2$  e, para um conjunto finito de pontos, o vetor aleatório de respostas  $\mathbf{Y}$  segue uma distribuição gaussiana multivariada. Assim,

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \tau^2 \mathbf{I}),$$

em que  $\mathbf{1}$  representa um vetor com  $n$  elementos 1;  $\mathbf{I}$  é a matriz identidade  $n \times n$  e  $\boldsymbol{\Sigma}$  é a matriz  $n \times n$  com o  $(i, j)$ -ésimo elemento  $\rho(h_{ij})$  em que  $h_{ij} = \|x_i - x_j\|$ .

### 2.4.2 Modelos lineares generalizados geoestatísticos

No caso do modelo linear Gaussiano geoestatístico, a variável resposta  $Y$  segue uma distribuição Gaussiana multivariada, mas há casos em que  $Y$  segue outras distribuições. Assim, considerando-se o contexto geoestatístico, o modelo linear Gaussiano não servirá para modelar esses dados.

Em situações em que  $Y$  segue outras distribuições que não seja a Gaussiana, o uso de modelos como os lineares generalizados geoestatísticos se encaixa. Porém, há a necessidade de se explicar o porquê. Logo, uma introdução com os modelos lineares generalizados torna-se necessária.

Segundo Lima, Soares Junior e Viola (2009), os modelos lineares generalizados são uma extensão dos modelos lineares clássicos.

No modelo linear clássico, tem-se

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

em que  $\mathbf{Y}$  é o vetor da variável dependente ou variável resposta, de dimensões  $n \times 1$ ;  $\boldsymbol{\mu} = E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$  é o componente sistemático, sendo  $\mathbf{X}$  a matriz do modelo, de dimensões  $n \times p$  e  $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$  o vetor de parâmetros do modelo e  $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$  é o componente aleatório do modelo com  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $i = 1, 2, \dots, n$ .

Já no modelo linear generalizado, de acordo com Biele e Resende (2002), a variável resposta passa a poder seguir uma distribuição que não é, necessariamente, a normal, podendo, portanto, provir da família exponencial de distribuições.

Segundo Couto, Oliveira e Viola (2008), uma variável aleatória  $Y$  pertence à família exponencial de distribuições quando sua função de probabilidade (f.p) ou função densidade de probabilidade (f.d.p) pode ser escrita na forma

$$f(y_i; \theta, \phi) \propto \exp \left\{ \frac{y_i \theta - b(\theta)}{a(\phi)} + c(y_i, \phi) \right\}$$

em que  $\theta$  é o parâmetro de localização,  $\phi$  é o parâmetro de dispersão e  $a(\cdot)$ ,  $b(\cdot)$  e  $c(\cdot)$  são funções específicas para cada distribuição.

Assim, por exemplo, tomando-se a distribuição de Poisson cuja função de probabilidade, de acordo com Magalhães (2006), é

$$f(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}, \text{ com } y = 0, 1, 2, \dots \text{ e } \lambda > 0,$$

tem-se

$$f(y; \lambda) = \exp \left\{ \ln \left( e^{-\lambda} \right) + y \ln \left( \lambda \right) - \ln \left( y! \right) \right\}$$

$$f(y; \lambda) = \exp \left\{ y \ln \left( \lambda \right) - \lambda - \ln \left( y! \right) \right\}$$

Fazendo  $\theta = \ln(\lambda)$ , tem-se

$$f(y, \theta) = \frac{e^{-\theta} \theta^y}{y!} = \exp\{\theta - y \ln \theta - \ln(y!)\}$$

Obtendo-se, assim, para o caso da distribuição de Poisson,  $a(\theta) = 1$ ,  $b(\theta) = e^{-\theta}$ ,  $c(y, \theta) = -\ln(y!)$ .

Logo, a distribuição de Poisson é um exemplo de distribuição que faz parte da família exponencial de distribuições.

Assim, a caracterização de um modelo linear generalizado, segundo Demétrio (2001) e McCullagh e Nelder (1989) se dá mediante três itens, que são:

a) as variáveis resposta  $Y_i$ ,  $i = 1, \dots, n$ , mutuamente independentes, com média  $\lambda_i = E[Y(x_i)]$ , devem ter f.p. ou f.d.p. proveniente de uma distribuição pertencente à família exponencial na forma canônica;

b) as variáveis explicativas constituem a matriz  $\mathbf{X}$  do modelo e, a partir dessa matriz, define-se o preditor linear  $\eta_i$ , sendo

$$\eta_i = \sum_{j=1}^p x_{ij} \beta_j, \quad i = 1, \dots, n, \text{ e } j = 1, \dots, p, \text{ em que } x_{ij} \text{ designa o valor da}$$

$j$ -ésima variável explicativa para a  $i$ -ésima observação das  $p$  variáveis explicativas  $X_j$  e  $\beta_j$  é um vetor de parâmetros desconhecidos, a serem estimados a partir dos dados;

c) a ligação entre a média e o preditor linear se dá por uma função de ligação  $h(\cdot)$ , monótona e derivável e que transforma  $\lambda_i$  em  $\eta_i$  da

$$\text{seguinte forma: } \eta_i = h(\lambda_i), \text{ logo, } h(\lambda_i) = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n \text{ e } j = 1,$$

$\dots, p$ .

Uma particularidade dos modelos lineares generalizados, como já citado, é que eles trabalham com dados independentes entre si. Assim, quando existe uma estrutura de dependência espacial entre os dados, há a necessidade de se usar outro tipo de modelagem.

Segundo Diggle e Ribeiro Júnior (2007), diversos caminhos têm sido propostos no sentido de estender a classe de modelos lineares generalizados para o trabalho com dados dependentes e, desses, um caminho bastante utilizado é introduzir no preditor linear, um efeito aleatório não observado. Assim, o preditor linear passa a ter a forma  $\eta_i = \sum_{j=1}^p x_{ij}\beta_j + S_i$ ,  $i = 1, \dots, n$ , sendo que  $S = (S_1, \dots, S_n)$  segue uma distribuição multivariada com média zero.

Os  $S_i$ ,  $i = 1, \dots, n$ , são chamados de efeitos aleatórios ou variáveis aleatórias latentes.

Diggle e Ribeiro Júnior (2007) salientam, ainda, que modelos deste tipo, cujas observações são independentes, mas condicionais a valores de um conjunto de variáveis aleatórias latentes  $S_1, \dots, S_n$ , recebem o nome de modelos lineares generalizados mistos.

Nogueira (2008) informa que os modelos lineares generalizados geoestatísticos são um caso particular dos modelos lineares generalizados mistos, em que as variáveis aleatórias latentes  $S_1, \dots, S_n$  derivam de um processo espacial  $S(\cdot) = \{S(x) : x \in A \subset \mathbb{R}^2\}$ .

Então, os modelos lineares generalizados geoestatísticos são um caso particular dos modelos lineares generalizados mistos, em que as variáveis resposta  $Y_i$ ,  $i = 1, \dots, n$  condicionais a  $S(x)$  são mutuamente independentes com esperança  $E[Y_i|S(x)] = \lambda_i$  e preditor linear  $h(\lambda_i) = S(x_i)$ ,  $i = 1, \dots, n$  para uma função de ligação  $h(\cdot)$  conhecida, sendo  $\{h^{-1}(S(x)) : x \in A\}$  o sinal do processo (DIGGLE; RIBEIRO JÚNIOR, 2007).

No caso de presença de covariáveis, segundo Viola (2007), o preditor linear passa a ter a forma  $h(x_i) = \sum_{j=1}^p \beta_j S(x_i)$ ,  $i = 1, \dots, n$  e  $j = 1, \dots, p$ , sendo  $\beta_j$  o vetor de parâmetros da regressão.

Para dados de contagem e dados binomiais, ambos com estrutura de correlação espacial, adotam-se, respectivamente, o modelo de Poisson-log-linear e o modelo logístico (NOGUEIRA, 2008).

#### 2.4.2.1 Modelo de Poisson log-linear

O modelo de Poisson log-linear é um candidato a representar dados de contagem espacialmente referenciados; apresenta função de ligação logarítmica e a distribuição condicional de cada  $Y_i$  é Poisson (DIGGLE; RIBEIRO JÚNIOR, 2007).

Na forma mais simples deste modelo, a variável resposta  $Y_i$  segue uma distribuição de Poisson condicionalmente independente, com esperanças condicionais  $\mu_i$ , em que

$$\log(\lambda_i) = \alpha + S(x_i)$$

e lembrando que  $S(\cdot)$  é um processo Gaussiano estacionário com média zero, variância  $\sigma^2$  e função de correlação  $\rho(h)$ , em que  $h = \|x - x'\|$ .

### 2.5 Estimação de parâmetros para modelos lineares generalizados geoestatísticos não gaussianos

Segundo Diggle e Ribeiro Júnior (2007), a alta dimensionalidade do vetor aleatório  $S = \{S(x_1), \dots, S(x_n)\}$  dificulta a aplicação de métodos baseados em verossimilhança em modelos lineares generalizados geoestatísticos não Gaussianos.

No modelo misto linear generalizado, por pressupor-se que as variáveis resposta  $Y = (Y_1, \dots, Y_n)$  são condicionalmente independentes dado  $S$ , a função de verossimilhança tem uma forma simples, condicional aos valores de um vetor de variáveis aleatórias não observadas  $S$ .

Ainda de acordo com Diggle e Ribeiro Júnior (2007), considerando que  $\theta$  denota os parâmetros que determinam a distribuição condicional de  $Y$  dado  $S$  e a distribuição condicional de  $Y_i$  dados  $S$  e  $\theta$  é  $f_i(y_i; S, \theta)$ . Assim, a função de verossimilhança condicional para  $\theta$  dado  $S$  é

$$L(\theta|S) = \prod_{i=1}^n f_i(y_i | S, \theta) \quad (2.10)$$

A distribuição conjunta de  $S$  com parâmetro  $\phi$  sendo  $g(S; \phi)$ , então, a função de verossimilhança para o modelo misto, baseada em variáveis aleatórias observadas  $Y$ , passa a ser

$$L(\theta, \phi) = \int_S \prod_{i=1}^n f_i(y_i | S, \theta) g(S | \phi) ds,$$

que foi obtida pela marginalização em relação às variáveis aleatórias não observadas  $S$ .

A dificuldade que se encontra em aplicar essa função de verossimilhança a um contexto geoestatístico é que  $S_i = S(x_i)$  são dependentes. Assim, a integral tem a mesma dimensão de  $Y$  e não se pode, então, utilizar métodos convencionais de integração numérica.

Uma possível solução para essa questão é o uso dos métodos de Monte Carlo via Cadeias de Markov (MCMC), em que se utiliza um algoritmo para simular da distribuição condicional dos efeitos aleatórios (CHRISTENSEN; DIGGLE; RIBEIRO JÚNIOR, 2003).

Expressando a função de verossimilhança (2.10) como uma esperança com respeito à distribuição de  $S$ , tem-se

$$L(\theta, \phi) = E \left[ \prod_{i=1}^n f_i(y_i | S, \theta) \right]$$

Logo, pode-se, em princípio, simular várias vezes a partir da distribuição multivariada gaussiana correspondente a  $S$  para qualquer conjunto de valores de  $(\theta, \phi)$  e a esperança pode ser aproximada por uma média de Monte Carlo. Logo,

$$L_{MC}(\theta, \phi) \approx \frac{1}{K} \sum_{k=1}^K \left[ \prod_{i=1}^n f_i(y_i | S_k, \theta) \right],$$

em que  $S_k$  denota a  $k$ -ésima realização simulada do vetor  $S$ .

### **3 MATERIAL E MÉTODOS**

Os dados utilizados neste trabalho foram coletados pelo professor Dr. Denismar Alves Nogueira, da Universidade Federal de Alfenas, na região urbana do município de Alfenas, MG, para comporem o estudo desenvolvido em sua tese de doutorado. As informações relacionadas à caracterização do município, inclusive os cartogramas aqui mostrados, constam em Nogueira (2008).

#### **3.1 Caracterização da área do estudo**

O município de Alfenas, MG, situa-se geograficamente a 21°25'45'' de latitude Sul e 45°56'50'' de longitude Oeste. Localiza-se na bacia hidrográfica do rio São Tomé, que termina no lago de Furnas. Ocupa área de 849,2 km<sup>2</sup> de extensão, com altitude máxima de 888 m acima do nível do mar e é composto por 70 setores censitários.

Dos 70 setores censitários que compõem o município, os setores do distrito industrial e regiões rurais não fazem parte do estudo em questão, devido à sua baixa densidade demográfica. Logo, considerou-se apenas o perímetro urbano.

A Figura 2 representa o cartograma dos 68 setores censitários e os limites da área considerada no estudo, com centroides representando as coordenadas espaciais da informação constante em cada setor censitário.

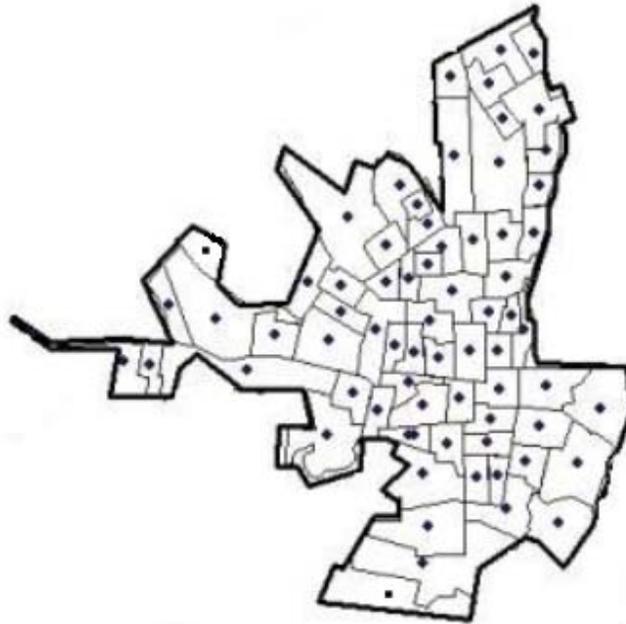


Figura 2 Cartograma dos 68 setores censitários considerados no estudo, com os limites de área e respectivos centroides  
Fonte: Adaptação Nogueira (2008)

Os limites de cada setor censitário foram ignorados para que a geoestatística fosse utilizada, visto que a mesma pressupõe continuidade espacial. Assim, considerou-se a possibilidade de ocorrência da morte em qualquer região da área em estudo.

### **3.2 Dados utilizados e método de amostragem**

Os dados para o estudo são contagens de mortes infantis na cidade de Alfenas, MG, no período de 2000 a 2004. São provenientes do Sistema de Informações sobre Mortalidade (SIM), que por sua vez, recebe tais informações da declaração de óbito (DO). Utilizaram-se, para compor o conjunto de dados,

apenas as contagens de natimortos e óbitos em crianças com menos de um ano de vida.

Um Sistema de Informação Geográfica (SIG) foi utilizado para fornecer as bases cartográficas necessárias ao estudo. Para maiores detalhes, consultar Nogueira (2008).

Para georreferenciar as residências das crianças que vieram a óbito em Alfenas, MG, Nogueira (2008) realizou a comparação entre o endereço constante no SIM e o cadastro de logradouros. As coordenadas geográficas de cada ponto foram, então, encontradas a partir da interpolação da Inscrição Cadastral (IC) dos lotes dos logradouros por meio do SIG, o que permitiu que os eventos fossem visualizados na base cartográfica, em que cada ponto corresponde à localização da residência da criança que veio a óbito.

Como Alfenas é considerada uma cidade pólo, com hospital universitário, espera-se que haja muitas pessoas de outras cidades em busca de atendimento no sistema de saúde desta, embora, para tal, haja a necessidade do preenchimento de um cadastro em que conste, dentre outros, o endereço nesta cidade, fornecido pelo interessado no atendimento, o que faz, muitas vezes, com que seja gerado um endereço que não existe. Assim, o referenciamento utilizando a metodologia acima citada só conseguiu identificar 46% dos locais de residência das crianças que morreram no período utilizado para o estudo (NOGUEIRA, 2008).

Tendo sido obtidas as localizações pontuais das mortes na área em estudo, o próximo passo foi realizar a contagem das mesmas em cada um dos 68 setores censitários e georreferenciá-las pelos centroides.

As contagens georreferenciadas no mapa da área em estudo, já sem os limites dos setores censitários, podem ser observadas na Figura 3.

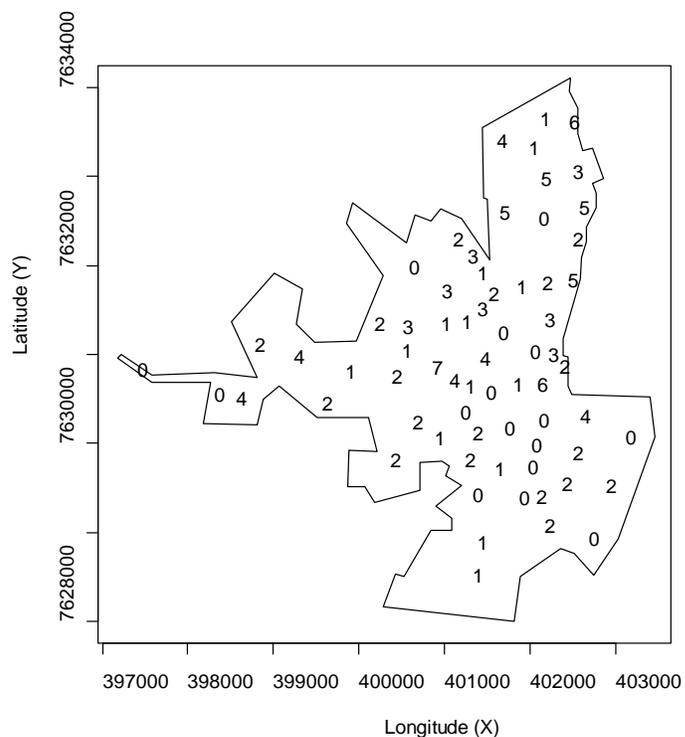


Figura 3 Número de mortes infantis ocorridas na cidade de Alfenas, MG, entre 2000 e 2004, localizadas em cada setor censitário pelo respectivo centroide

Percebe-se, pela Figura 3, que em algumas localizações não houve casos de mortes infantis, enquanto em outras houve casos com 5, 6 e até 7 mortes.

### 3.3 Análise exploratória

No intuito de descrever aspectos importantes do conjunto de dados, utilizou-se a análise descritiva, com o auxílio do gráfico de barras e gráficos para avaliar visualmente a distribuição espacial dos dados e seu comportamento em relação às coordenadas.

Para verificar a aderência dos dados à distribuição de Poisson, utilizou-se o teste de aderência Qui-quadrado (SIEGEL, 1975), cuja estatística é dada por

$$\chi^2_{\text{calculado}} = \sum_{i=1}^k \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}, \text{ para g.l.} = k-p-1, \text{ sendo } f_i \text{ a frequência observada na } i\text{-ésima classe; } \hat{f}_i \text{ a frequência esperada para a } i\text{-ésima classe; g.l. o grau de liberdade; } k \text{ o número de classes da distribuição de frequência e } p \text{ o número de parâmetros estimados. O } \chi^2 \text{ tabelado (} \chi^2_t \text{) foi obtido a partir de tabela específica, a 5\% de significância com 6 graus de liberdade. Para provirem de uma distribuição de Poisson, } \chi^2_{\text{calculado}} < \chi^2_{\text{tabelado}}.$$

Para a variável mortalidade infantil, após uma análise descritiva dos dados, gerou-se um semivariograma ao qual foram ajustados, pelo método da máxima verossimilhança, os modelos esférico e Matérn com kappa 0,5 e 1,0, dos quais o mais bem ajustado foi selecionado pelo critério de Akaike (AIC).

O modelo selecionado nesta etapa permitiu encontrar os parâmetros necessários para prever mortes infantis em locais não amostrados da área de estudo, por meio da krigagem ordinária.

No Quadro 1 observam-se as covariáveis que foram utilizadas no estudo. O interesse em envolvê-las foi o de verificar a existência de uma relação entre elas e a mortalidade infantil na cidade de Alfenas, MG. Tais covariáveis foram selecionadas utilizando-se o banco de dados do IBGE do ano 2000 (NOGUEIRA, 2008).

Quadro 1 Covariáveis utilizadas no estudo da mortalidade infantil, selecionadas no banco de dados do IBGE, censo de 2000

A. Número de mulheres em idade fértil
B. Número de mulheres em idade de risco gestacional
C. Número de mulheres em idade fértil analfabetas
D. Número de residências com mais de 6 moradores
E. Renda mensal da mulher
F. Renda mensal do homem
G. Densidade demográfica do setor censitário

Fonte: Adaptado de Nogueira (2008)

Foi feita uma análise exploratória de cada covariável e obtido o semivariograma empírico de cada uma delas. A cada semivariograma gerado foi ajustado um modelo teórico por meio do método de mínimos quadrados ordinários.

Considerando-se os dados da Tabela 1, ao semivariograma gerado para a covariável A foram ajustados os modelos Gaussiano, Matérn com valores de kappa 0,5; 1,0 e 1,5. Ao semivariograma gerado para a covariável B foram ajustados os modelos Esférico e Matérn, com kappa 0,5; 1,0; 1,5; 2,0; ao da covariável C, os modelos Esférico e Matérn, com kappa 0,5 e 1,0; ao da covariável D, os modelos Esférico, gaussiano e Matérn, com kappa 0,5; 1,0 e 1,5; ao da covariável E, os modelos gaussiano e Matérn, com kappa 0,5; 1,0 e 1,5; ao da covariável F, os modelos Esférico e Matérn, com kappa 0,5; 1,0 e 1,5 e ao da covariável G, os modelos Esférico e Matérn, com kappa 0,5; 1,0; 1,5 e 2,0.

O modelo com melhor ajuste a cada covariável foi selecionado através do método da validação cruzada.

Os parâmetros encontrados foram, então, utilizados na predição de valores em locais não amostrados, para cada covariável, separadamente.

### 3.4 Geoestatística baseada em modelos: modelo linear generalizado geoestatístico para os dados de mortalidade infantil

As respostas  $Y_i$ ,  $i=1, \dots, n$ , medidas nas localizações  $x_i$ ,  $i=1, \dots, n$  correspondem às contagens das mortes infantis na região estudada  $A \subset \mathbb{R}^2$ . Neste caso,  $A$  é a região tomada para estudo no município de Alfenas, MG.

O conjunto de respostas  $Y(\cdot) = \{Y(x): x \in A \subset \mathbb{R}^2\}$ , conhecido por processo de medição (DIGGLE; RIBEIRO JÚNIOR, 2007), contém o número de mortes infantis em cada localização amostrada. Assume-se que os elementos do processo de medição, condicionais ao processo espacial  $S(\cdot)$ , são realizações de variáveis aleatórias Poisson independentes e identicamente distribuídas.

O processo Gaussiano  $S(\cdot)$ , conhecido como processo sinal, apresenta  $E[S(x)] = 0$ ,  $\text{Var}[S(x)] = \sigma^2$  e função de correlação  $\rho(h) = \text{Corr}\{S(x), S(x')\}$ , em que  $h = \|x - x'\|$  é a distância Euclidiana entre  $x$  e  $x'$ .

Assim, tem-se que

$$Y_i | S(x_i) \sim \text{Poisson}(\lambda_i), \quad i = 1, \dots, n,$$

em que a média de mortes infantis na localização  $x_i$  amostrada, condicional a  $S(\cdot)$ , é  $\lambda_i = E[Y_i | S(x_i)]$ .

O modelo de preditores é, então, dado por

$$h(\lambda_i) = \alpha + S(x_i),$$

sendo  $h(\cdot)$  a função de ligação logarítmica dada por

$$\log \lambda_i = \alpha + S(x_i)$$

Assim, sendo o número de mortes infantis  $Y(x_i) | S(x_i) = s$  com distribuição de Poisson com média  $\exp[\alpha + S(x_i)]$ ,  $i = 1, \dots, n$ , tem-se que

$$P[Y(x) = y | S(x) = s] = \frac{e^{-\exp[\alpha + S(x_i)]} \cdot (e^{\exp[\alpha + S(x_i)]})^{y(x_i)}}{y(x_i)!}$$

Para esse modelo, a função de verossimilhança não tem solução explícita. Assim, uma solução possível foi a utilização dos métodos de Monte Carlo via Cadeias de Markov (MCMC). Por meio desse procedimento, obteve-se uma cadeia com 50.000 iterações, tendo sido descartadas as 1.000 primeiras iterações (*burn-in*), sendo armazenada 1 amostra a cada 20 amostras geradas. Logo, obteve-se um total efetivo de 2.500 amostras. As cadeias obtidas para cada parâmetro foram analisadas para verificar a convergência do algoritmo de MCMC.

Para a simulação condicional pelo algoritmo MCMC, os valores iniciais para comporem o modelo foram obtidos da análise geoestatística clássica, em que o modelo esférico foi ajustado ao semivariograma experimental, fornecendo os parâmetros  $\sigma^2 = 2,598$ ;  $\phi = 400,5$ ;  $\tau^2 = 0,38$  e  $\beta = 1,9721$ .

Na modelagem linear generalizada geoestatística, utilizou-se, para a função de correlação, os modelos esférico e Matérn, com kappa 0,5; 1,0; 1,5 e 2,0, dos quais o mais bem ajustado foi selecionado com a ajuda do critério de Akaike (AIC). Os parâmetros estimados pelo modelo selecionado foram, então, utilizados na obtenção de valores preditos e do mapa de predição para a variável mortalidade infantil.

### **3.5 Escolhendo entre a geoestatística clássica e a geoestatística baseada em modelos como metodologia mais apropriada para o estudo da mortalidade infantil**

Com a obtenção dos valores preditos da variável mortalidade infantil pela geoestatística clássica e pela geoestatística baseada em modelos, procurou-se comparar as duas metodologias, de modo a selecionar a que melhor se adequasse aos dados de mortalidade infantil em estudo. Em tal procedimento utilizou-se o erro absoluto e o AIC, que consideraram como mais adequada a

metodologia que forneceu os menores valores tanto para o erro absoluto quanto para o AIC.

### 3.6 As covariáveis e a variável mortalidade infantil

Fez-se uso do teste de correlação de Spearman (SIEGEL, 1975) para verificar se as covariáveis e a variável mortalidade infantil apresentavam-se

correlacionadas. O teste utilizou a equação  $R=1-\frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)}$ , sendo n o número

de pares  $(x_i, y_i)$ ;  $d_i = (\text{postos de } x_i \text{ dentre os valores de } x) - (\text{postos de } y_i \text{ dentre os valores de } y)$ , entendendo-se por posto o termo utilizado para representar o número que se atribui a um item amostral individual, considerando-se a ordem que ele ocupa em uma lista previamente ordenada.

Para compor o teste, foram utilizados os valores preditos encontrados para cada covariável e os valores preditos da variável mortalidade infantil, encontrados por meio da metodologia mais adequada dentre a geoestatística clássica e a geoestatística baseada em modelos.

O intuito deste procedimento foi apenas o de verificar a existência de fatores externos que poderiam influenciar a mortalidade infantil.

### 3.7 Software

O software utilizado para a realização das análises foi o R, versão 2.9.2 (R DEVELOPMENT CORE TEAM, 2009) e, deste, foram utilizados os pacotes geoR (DIGGLE; RIBEIRO JÚNIOR, 2001) e geoRglm (CHRISTENSEN; RIBEIRO JÚNIOR, 2002). As versões do geoR e geoRglm usadas foram, respectivamente, 1.6-27 e 0.8-26.



## 4 RESULTADOS E DISCUSSÃO

A seguir estão apresentados os resultados obtidos no desenvolvimento do trabalho, devidamente discutidos.

### 4.1 Análise exploratória para a variável resposta mortalidade infantil

Na Tabela 1 constam algumas medidas descritivas da variável mortalidade infantil.

Tabela 1 Análise descritiva para os dados de mortalidade infantil na cidade de Alfenas, MG, referentes aos anos de 2000 a 2004

MEDIDAS DESCRITIVAS	VALORES
Número de observações	68
Soma	134
Valor mínimo	0
Valor máximo	7
Média	1,971
Mediana	2
Variância	3,014
Desvio padrão	1,736
Coefficiente de variação	0,881
Coefficiente de assimetria	0,820
Coefficiente de curtose	0,023
1º quartil	1
3º quartil	3

Pode-se observar, pelos dados da Tabela 2, que os dados de mortalidade infantil em estudo são compostos por 68 setores censitários (observações), que somam 134 casos de mortalidade infantil. Os valores mínimo e máximo mostram, respectivamente, a menor e a maior quantidade de mortes infantis observada. Pelo 3º quartil, observa-se que 75% dos casos de mortes infantis ocorreu abaixo de 3, ou seja, a maior concentração de mortes encontra-se em localizações que apresentam de 0 a 3 casos. Considerando-se os valores da

variância e do desvio padrão, os dados apresentam uma variabilidade considerável em relação à média, o que pode ser confirmado pelo coeficiente de variação (CV) de 88,1%. O coeficiente de assimetria de 0,820 e o coeficiente de curtose, de 0,023, sugerem a não simetria dos dados, direcionando o estudo para uma análise diferente da que seria feita, caso os dados tivessem indicação de simetria.

Seria interessante informar a taxa de mortalidade infantil para a cidade de Alfenas, MG, porém, como esta, de acordo com Kozu et al. (2001), é expressa pela relação entre o número de óbitos de crianças menores de um ano e o número de nascidos vivos em determinado local e calculado na base de mil nascidos vivos, há uma impossibilidade em se encontrar essa taxa, já que o estudo abrangeu apenas as mortes infantis para o município em questão.

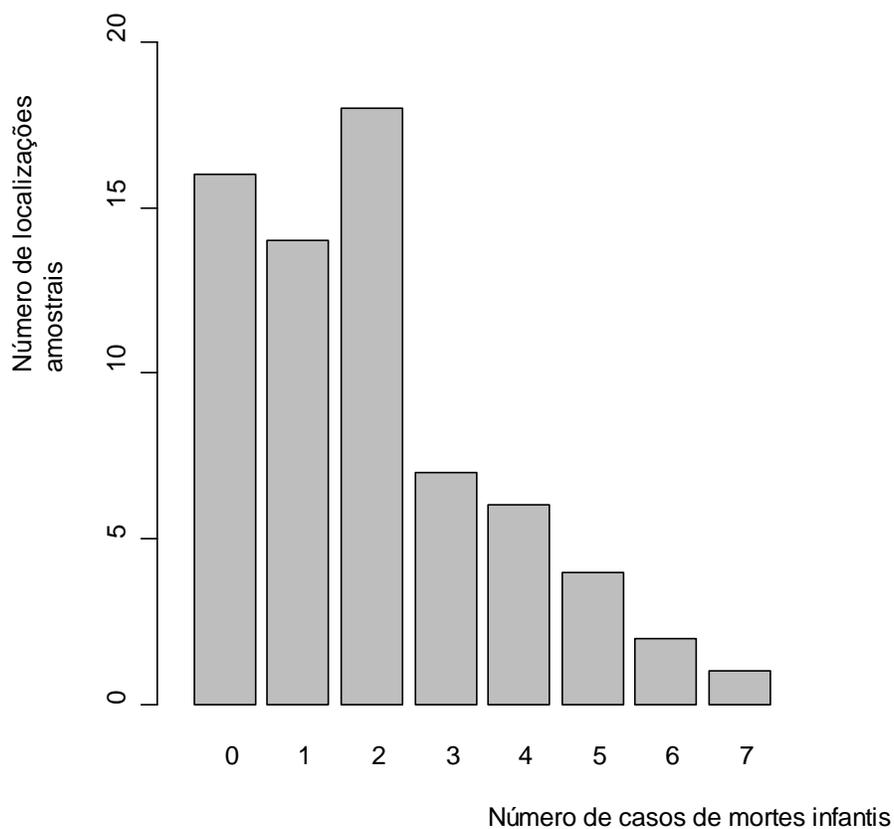


Figura 4 Gráfico de barras do número de localizações amostrais que apresentaram de 0 a 7 casos de mortes infantis na cidade de Alfenas, MG, entre 2000 e 2004

Pelo gráfico de barras da Figura 4, pode-se enxergar uma grande quantidade de localizações com 0 (zero) caso de morte e uma quantidade ainda maior de localizações com 2 casos de morte. Assim, a mediana, como esperado, foi igual a 2, enquanto a média, bastante influenciada pela quantidade de zeros, ficou com valor abaixo de 2.

Pela Figura 4 observa-se, ainda, que, aparentemente, os dados seguem uma distribuição de Poisson, o que foi confirmado pelo uso do teste de aderência de qui-quadrado ( $\chi^2_{\text{calculado}} = 11,6193 < \chi^2_{\text{tabelado}} = 12,5916$ , em que foram utilizados 6 graus de liberdade e nível de significância de 5%).

Pela Figura 5, deseja-se explorar os dados de modo a analisar seu comportamento em relação às coordenadas geográficas, verificando se eles sofrem alguma influência destas.

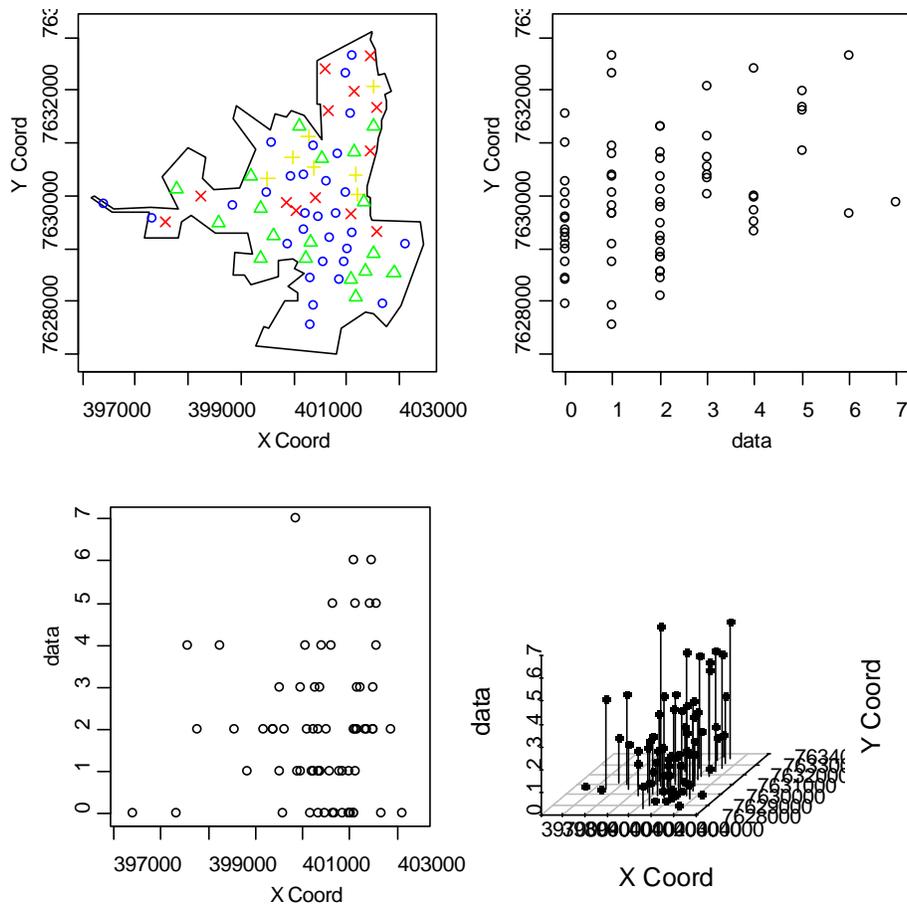


Figura 5 Localização pontual das classes (superior esquerdo), valores contra as coordenadas (superior direito e inferior esquerdo) e gráfico com representação espacial dos dados de mortalidade infantil na cidade de Alfenas, MG, ocorrida entre 2000 e 2004

Pelos gráficos do canto superior direito e inferior esquerdo da Figura 5, pode-se dizer que, aparentemente, não há tendência dos dados com relação às coordenadas.

Para analisar a dependência espacial entre os dados, foi produzido o semivariograma empírico mostrado na Figura 6, em que se utilizou uma distância máxima de 3.700 m para o cálculo das semivariâncias.

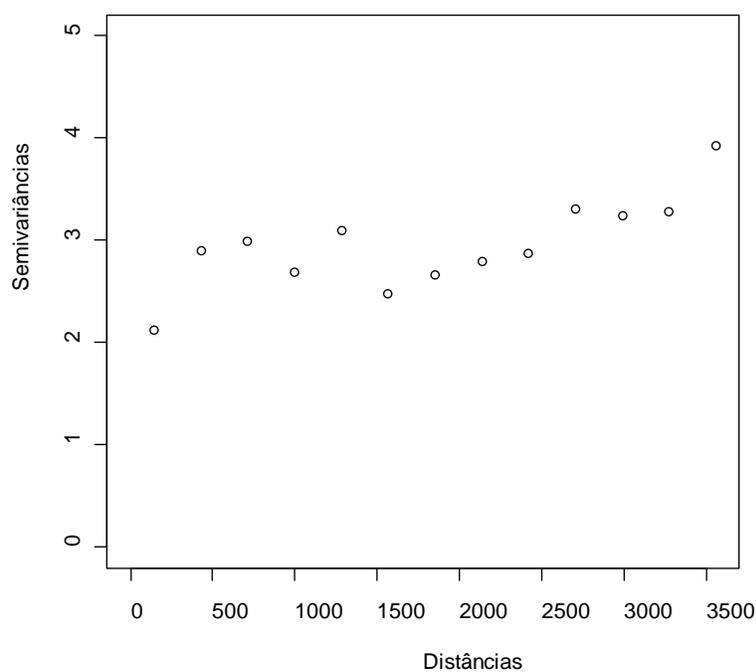


Figura 6 Semivariograma empírico para os dados de mortalidade infantil

Ao semivariograma da Figura 6 foram ajustados os modelos teóricos Matérn, com  $\kappa$  0,5 (modelo exponencial), Matérn, com  $\kappa$  1,0 e o modelo esférico, utilizando-se o método de ajuste da máxima verossimilhança.

A seleção do melhor modelo ajustado baseou-se no critério de Akaike (AIC), cujos valores para cada modelo encontram-se na Tabela 2.

Tabela 2 Valores de AIC para os modelos ajustados ao semivariograma empírico gerado para os dados de mortalidade infantil usando geostatística clássica

MODELOS TEÓRICOS	VALORES DE AIC
Esférico	272,4
Matérn com kappa 0,5	272,9
Matérn com kappa 1,0	272,8

De acordo com os dados da Tabela 3, o modelo teórico que melhor se ajustou aos dados foi o esférico.

Na Figura 7 observa-se o semivariograma empírico para os dados de mortalidade infantil ajustado pelo modelo esférico.

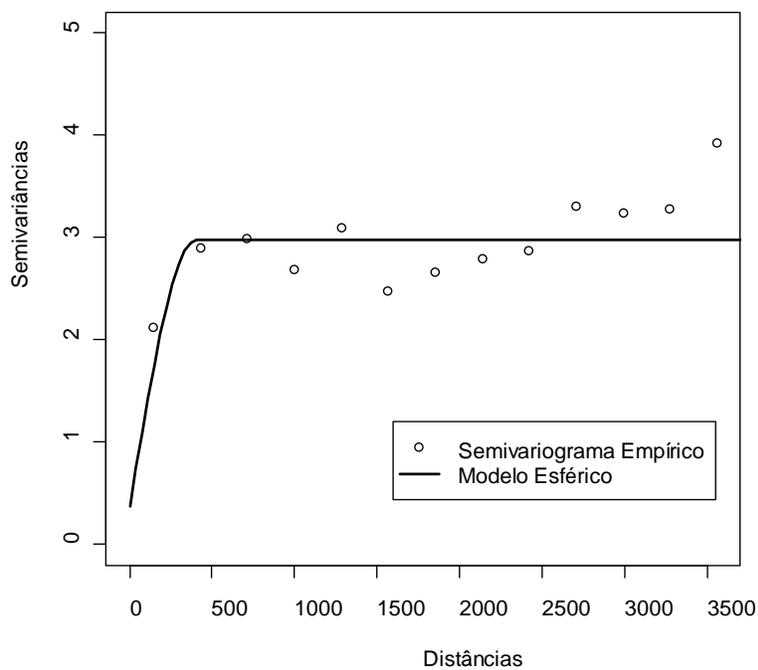


Figura 7 Semivariograma empírico ajustado pelo modelo esférico

O modelo esférico, ajustado ao semivariograma empírico, forneceu os seguintes parâmetros:  $\sigma^2 = 2,598$ ;  $\phi = 400,5$ ;  $\tau^2 = 0,38$ . Assim, tem-se um patamar de 2,978, a partir do qual  $\gamma(h)$  é constante e um alcance da dependência espacial dos dados de 400,5 m, ou seja, pressupõe-se que as mortes ocorridas a uma distância de até 400,5 m são espacialmente dependentes.

Dependência espacial em dados de mortalidade infantil foi detectada também nos trabalhos de Andrade e Szwarcwald (2001) e Morais Neto et al. (2001), tendo, no segundo trabalho, a dependência espacial sido detectada em bairros periféricos de Goiânia, onde habitam populações de baixa renda.

#### **4.2 Escolha entre a geoestatística clássica e a geoestatística baseada em modelos na predição da mortalidade infantil**

Utilizando os valores estimados pelo modelo esférico, procedeu-se à obtenção dos valores preditos pela krigagem clássica para a variável dependente mortalidade infantil. Esses valores foram utilizados no procedimento de comparação entre a krigagem clássica e a baseada em modelos.

Para o procedimento baseado em modelos, utilizaram-se os valores simulados pelo método MCMC e os valores iniciais  $\sigma^2 = 2,598$ ;  $\phi = 400,5$ ;  $\tau^2 = 0,38$  e  $\beta = 1,9721$ . Foram ajustados, pelo método da máxima verossimilhança Monte Carlo, os modelos esférico e Matérn com valores de kappa 0,5; 1,0; 1,5 e 2,0, dos quais o mais bem ajustado foi selecionado pelo critério de Akaike.

Na Tabela 3 são fornecidos os valores de AIC para cada um dos cinco modelos utilizados.

Tabela 3 Valores de AIC para os modelos ajustados aos dados de mortalidade infantil por meio da geoestatística baseada em modelos

MODELOS TEÓRICOS	VALORES DE AIC
Esférico	-88,76
Matérn com kappa 0,5	-88,84
Matérn com kappa 1,0	-88,66
Matérn com kappa 1,5	-88,64
Matérn com kappa 2,0	-88,62

De acordo com os dados da Tabela 3, o modelo que melhor se ajustou aos dados de mortalidade infantil por meio da geoestatística baseada em modelos foi o Matérn com valor de kappa 0,5. Para Christensen, Diggle e Ribeiro Júnior (2003), é interessante utilizar, como função de correlação, um membro da família Matérn, já que o parâmetro kappa ( $\kappa$ ) leva a uma melhor suavização do processo  $S(x)$ .

A partir do ajuste do modelo Matérn com kappa 0,5, obtiveram-se os parâmetros  $\sigma^2 = 0,6595$ ;  $\phi = 109,6954$ ;  $\tau^2 = 0$  e  $\beta = 0,4690$ . Esses parâmetros foram utilizados na predição de valores pelo uso da geoestatística baseada em modelos.

Para a escolha entre a geoestatística clássica e a geoestatística baseada em modelos, como melhor metodologia a ser usada na predição de valores para a mortalidade infantil, utilizou-se o erro absoluto em conjunto com o critério de Akaike (AIC). Os resultados obtidos encontram-se na Tabela 4.

Tabela 4 Valores do erro absoluto e de AIC para a predição de valores usando a geoestatística clássica e a geoestatística baseada em modelos

	ERRO ABSOLUTO	AIC
Geoestatística clássica	4,05	272,4
Geoestatística baseada em modelo	1,55	-88,84

De acordo com os dados da Tabela 4, a geoestatística baseada em modelos apresentou-se como a metodologia mais adequada ao estudo da

mortalidade infantil, visto ter obtido os menores valores, tanto para o erro absoluto quanto para o AIC.

Um fato relevante em estudos utilizando geoestatística baseada em modelos é que essa metodologia considera explicitamente, na composição do modelo, a distribuição que os dados seguem, simula valores a partir deles e, assim, encontra estimativas em locais não amostrados. Isso contribui para estimativas mais precisas, o que não acontece na geoestatística clássica, que usa os dados originais para encontrar valores em locais não amostrados, sem considerar a distribuição seguida pelos dados.

Os valores preditos, obtidos com o uso da geoestatística baseada em modelos, em que se utilizou um modelo linear generalizado geoestatístico com distribuição de Poisson e função de ligação logarítmica, foram utilizados para verificar a existência de correlação entre covariáveis selecionadas e a variável mortalidade infantil.

### 4.3 Covariáveis e o modelo linear generalizado geoestatístico

Na Tabela 5 são mostradas algumas análises descritivas para as covariáveis. As letras de A a G representam as covariáveis descritas no Quadro 1.

Tabela 5 Algumas análises descritivas para as covariáveis representadas pelas letras de A a G descritas na Tabela 1

	Média	Mediana	Mínimo	Máximo	Desvio padrão	CV(%) <sup>*</sup>
A	259,0	272,5	4,0	370,0	71,51	27,6
B	142,6	146,0	1,0	243,0	40,34	28,3
C	2,8	2,0	0,0	10,0	2,59	92,5
D	27,82	25,0	0,0	63,0	13,66	49,1
E	32750	23110	0,0	122000	26278,9	80,2
F	161700	128600	971	537400	105344,8	65,1
G	902,9	940,0	17,0	1396,0	258,64	28,6

\*Coeficiente de variação

De acordo com os dados da Tabela 5, percebe-se, pelo CV, que, C, D, E e F, respectivos representantes das covariáveis “número de mulheres em idade fértil analfabetas”, “número de residências com mais de 6 moradores”, “renda mensal da mulher” e “renda mensal do homem”, apresentam variabilidade bem alta, enquanto A, B e G, que representam, respectivamente as covariáveis “número de mulheres em idade fértil”, “número de mulheres em idade de risco gestacional” e “densidade demográfica do setor censitário”, apresentam CV bem menor, o que indica variabilidade menor para essas três covariáveis.

Para cada uma das covariáveis presentes no Quadro 1 fez-se uma análise geoestatística clássica, em que foram obtidos os valores preditos por meio da krigagem ordinária.

Em seguida, foi utilizado o teste de correlação de Spearman para verificar quais covariáveis apresentaram maior correlação com a variável mortalidade infantil. Para isso, utilizaram-se os valores preditos de cada covariável e os valores preditos da variável mortalidade infantil, sendo estes últimos obtidos por meio da krigagem baseada em modelos.

Na Tabela 6 são apresentados os resultados encontrados utilizando-se a correlação de Spearman para cada covariável e a variável dependente mortalidade infantil.

Observando-se os dados da Tabela 6, pode-se concluir que apenas a covariável “número de residências com mais de 6 moradores” não se correlaciona com a mortalidade infantil.

Observando-se os dados da Tabela 6, pode-se concluir que apenas a covariável “número de residências com mais de 6 moradores” não se correlaciona com a mortalidade infantil.

Tabela 6 Valores do coeficiente de correlação de Spearman com os respectivos p-valores para a relação entre cada covariável e a variável dependente mortalidade infantil

COVARIÁVEIS	VALOR DO COEFICIENTE DE CORRELAÇÃO DE SPEARMAN	P-VALOR
Número de mulheres em idade fértil	0,55	$2,2 \times 10^{-16}$
Número de mulheres em idade de risco gestacional	-0,21	$2,2 \times 10^{-16}$
Número de mulheres em idade fértil analfabetas	0,03	$2,2 \times 10^{-16}$
Número de residências com mais de 6 moradores	-0,07	$5,4 \times 10^{-1}$
Renda mensal da mulher	-0,09	$2,2 \times 10^{-16}$
Renda mensal do homem	-0,12	$2,2 \times 10^{-16}$
Densidade demográfica do setor censitário	-0,08	$2,6 \times 10^{-6}$

Observando-se os dados da Tabela 6, pode-se concluir que apenas a covariável “número de residências com mais de 6 moradores” não se correlaciona com a mortalidade infantil.

Considerando-se o p-valor e o valor do coeficiente de correlação de Spearman, ambos expressos na Tabela 6, as duas covariáveis mais correlacionadas com a mortalidade infantil são “número de mulheres em idade fértil” e “número de mulheres em idade de risco gestacional”. Para estas covariáveis, produziu-se o mapa de predição, que foi confrontado, nas Figuras 8 e 9, com o mapa de predição da variável mortalidade infantil, obtido utilizando-se a geoestatística baseada em modelos.

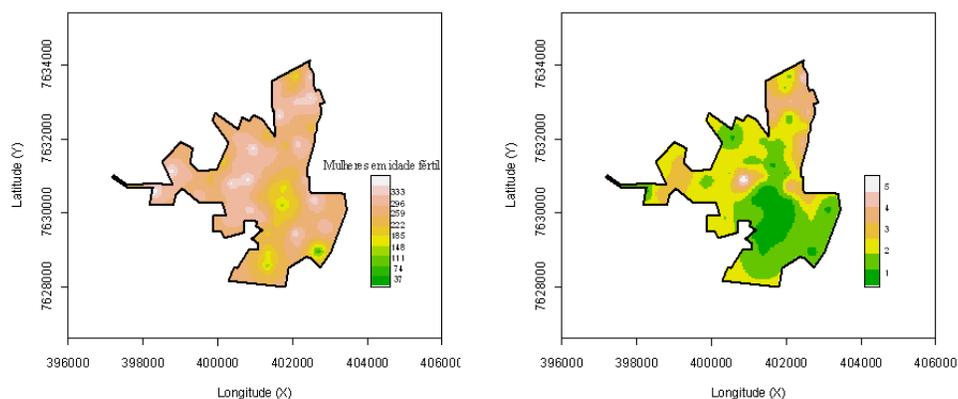


Figura 8 Mapas de predição: covariável “número de mulheres em idade fértil” (à esquerda) e variável “mortalidade infantil” (à direita), para a cidade de Alfenas, MG, entre 2000 e 2004

Percebe-se, visualmente, pela Figura 8 (ambas as legendas mostram valores que crescem da cor verde para a cor branca) que, na maior parte das regiões onde há maior incidência de mulheres em idade fértil, há também maior incidência de mortes infantis.

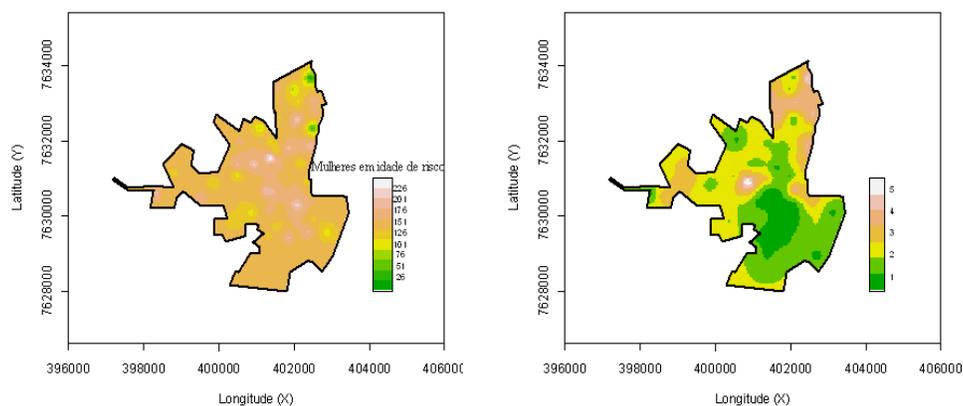


Figura 9 Mapas de predição: covariável “mulheres em idade de risco gestacional” (à esquerda) e variável “mortalidade infantil” (à direita) para a cidade de Alfenas, MG, entre 2000 e 2004

Assim como na Figura 8, percebe-se, visualmente, pela Figura 9 (as legendas também mostram valores que crescem da cor verde para a cor branca) que, na maior parte das regiões onde há maior incidência de casos de mulheres em idade de risco gestacional, o número de mortes infantis também aumenta.

Essas percepções visuais já eram esperadas, visto que o teste de correlação de Spearman demonstrou que realmente há uma relação entre estas duas covariáveis e a variável mortalidade infantil.

## 5 CONCLUSÕES

- a) O uso da modelagem geoestatística baseada em modelos foi mais adequado que a geoestatística clássica para lidar com dados de mortalidade infantil.
- b) A modelagem utilizando geoestatística baseada em modelos possibilitou o ajuste do modelo log-linear com distribuição de Poisson.
- c) O estudo permitiu identificar áreas da cidade de Alfenas, MG, em que há possibilidade maior de incidência de mortes infantis, correlacionadas ao número de mulheres em idade fértil e ao número de mulheres em idade de risco gestacional. Isso poderá servir para gerir formas de intervir nessas regiões, de maneira a minimizar as ocorrências desse evento.
- d) Como futuro trabalho, podem-se incluir as covariáveis diretamente no modelo geoestatístico e propor critérios para a seleção dessas covariáveis.

## REFERÊNCIAS

- ANDRADE, C. L. T. et al. Desigualdades sócio-econômicas do baixo peso aos nascer e da mortalidade perinatal no Município do Rio de Janeiro, 2001. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 20, p. 44-51, 2004. Supl.1.
- ASSUNÇÃO, R. M. **Estatística espacial com aplicações em epidemiologia, economia e sociologia**. São Carlos: Associação Brasileira de Estatística, 2001. Disponível em: <[http://www.est.ufmg.br/~assuncao/cursos\\_eea.htm](http://www.est.ufmg.br/~assuncao/cursos_eea.htm)> Acesso em: 11 dez. 2009.
- BAILEY, T. C.; GATRELL, A. C. **Interactive spatial data analysis**. London: Longman Scientific and Technical, 1995. 413 p.
- BIELE, J.; RESENDE, M. D. V. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. **Revista de Matemática e Estatística**, São Paulo, v. 20, p. 39-65, 2002.
- BIGGAR, J. W.; NIELSEN, D. R.; VIEIRA, S. R. Spatial variability of field-measured infiltration rate. **Soil Science Society American Journal**, Madison, v. 45, p. 1040-1048, 1981.
- CALDWELL, J. C. Education as a factor in mortality decline: an examination of Nigerian data. **Population Studies**, London, v. 33, n. 3, p. 395-413, 1979.
- CARVALHO, B. G.; COSTA, M. C. N. Mortalidade infantil e seus componentes em Salvador, BA, 1980-1991. **Informe Epidemiológico do SUS**, Salvador, v. 7, n. 4, p. 35-41, 1998.
- CARVALHO, M. S.; SANTOS, R. S. Análise de dados espaciais em saúde pública: métodos, problemas, perspectivas. **Caderno de Saúde Pública**, Rio de Janeiro, v. 21, n. 2, p. 361-378, 2005.
- COSTA, G. N. **Mortalidade perinatal, determinantes biológicos, de atenção à saúde materno infantil e socioeconômicos**: uma análise das desigualdades entre os bairros do Recife. 2008. 156 f. Tese (Doutorado em Saúde Pública) - Centro de Pesquisas Aggeu Magalhães/Fundação Oswaldo Cruz, Recife, 2008.
- COSTA, M. C. N. et al. Mortalidade infantil e condições de vida: a reprodução das desigualdades sociais em saúde na década de 90. **Caderno de Saúde Pública**, Rio de Janeiro, v. 17, n. 3, p. 555-567, 2001.

COUTO, E. V.; OLIVEIRA, F. S. C.; VIOLA, D. N. Estimação dos parâmetros para o modelo autologístico espacial com estrutura de vizinhança de primeira ordem. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 18., 2008, São Pedro. **Anais...** São Pedro, 2008.

CRESSIE, N. Fitting variogram models by weighted least squares. **Revista Mathematical Geology**, Netherlands, v. 17, p. 563-586, 1985.

CRESSIE, N. **Statistics for spatial data**. New York: J. Wiley, 1993. 900 p.

CHRISTENSEN, O. F.; DIGGLE, P. J.; RIBEIRO JÚNIOR, P. J. An introduction to model-based geostatistics. In: MØLLER, J. (Ed.). **Spatial statistics and computational methods**. New York: Springer, 2003. p. 43-46.

CHRISTENSEN, O. F.; RIBEIRO JÚNIOR, P. J. GeoRglm: a package for generalised linear spatial models. **R-News**, Auckland, v. 2, n. 2, p. 26-28, 2002.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agrônômica**. Piracicaba: ESALQ, 2001. 113 p.

DIGGLE, P. J.; MOYEED, R. A.; TAWN, J. A. Model-based geostatistics. **Applied Statistics**, London, v. 47, n. 3, p. 299-350, 1998.

DIGGLE, P. J.; RIBEIRO JÚNIOR, P. J. GeoR: A Package for geostatistical analysis. **R-News**, Auckland, v. 1, n. 2, p. 15-18, 2001.

DIGGLE, P. J.; RIBEIRO JÚNIOR, P. J. **Model-based geostatistics**. Caxambu: ABE, 2000. 129 p.

DIGGLE, P. J.; RIBEIRO JÚNIOR, P. J. **Model-based geostatistics**. New York: Springer, 2007. 228 p.

DUARTE, E. C. et al. Expectativa de vida ao nascer e mortalidade no Brasil em 1999: análise exploratória dos diferenciais regionais. **Revista Panamericana de Salud Pública**, Washington, v. 12, n. 6, p. 436-444, 2002.

DRUCK, S. et al. **Análise espacial de dados geográficos**. Brasília: EMBRAPA, 2004. Disponível em: < <http://www.dpi.inpe.br/libros.php> >. Acesso em: 16 nov. 2009.

EMILIANO, P. C. **Fundamentos e aplicações dos critérios de informação:** Akaike e Bayesiano. Lavras. 2009. 92 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2009.

FARAH, M. F. S. Gênero e políticas públicas. **Revista de Estudos Feministas**, Florianópolis, v. 12, n. 1, p. 47-71, 2004.

FUNDO DAS NAÇÕES UNIDAS PARA INFÂNCIA. **Situação mundial da infância:** meninas, educação e desenvolvimento. Brasília, 2004.

GOTLIEB, S. L. D.; LAURENTI, R.; MELLO JORGE, M. H. P. Análise da qualidade das estatísticas vitais brasileiras: a experiência de implantação do SIM e do SINASC. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 12, n. 3, p. 643-654, 2007.

GUIMARÃES, E. C. **Geoestatística básica e aplicada**. Uberlândia: Universidade Federal de Uberlândia, 2004. Apostila.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Projeção da população do Brasil por sexo e idade para op 1980-2050**. Rio de Janeiro, 2008. 93 p. (Série Estudos e Pesquisas, v. 24).

ISAAKS, E. H.; SRIVASTAVA, R. **An introduction to applied geostatistics**. New York: Oxford University, 1989. 561 p.

KEEN, C. S.; RIP, M. R.; WOODS, D. L. Intraurban variations of neonatal and post-neonatal mortality in a developing city. **Social Science and Medicine**, Amsterdam, v. 25, p. 889-894, 1987.

KOZU, K. T. et al. **Mortalidade infantil:** causas e fatores de risco: um estudo bibliográfico. 2001. Disponível em: <<http://www.medstudents.com.br/original/original/mortinf/mortinf.htm>>. Acesso em: 9 mar. 2009.

LANDIM, P. M. B.; STURARO, J. R. **Krigagem indicativa aplicada à elaboração de mapas probabilísticos de riscos**. Rio Claro: UNESP, 2002. Lab. Geomatemática. Texto Didático 6. Disponível em: <<http://www.rlc.fao.org/es/prioridades/transfron/sig/pdf/kindicativa.pdf>>. Acesso em: 2 jan. 2010.

LAURENTI, R. Mortalidade infantil nos Estados Unidos, Suécia e Estado de São Paulo. **Revista de Saúde Pública**, São Paulo, v. 21, n. 3, p. 268-273, 1987.

- LEAL, M. C.; SZWARCOWALD, C. L. Características da mortalidade neonatal no Estado do Rio de Janeiro na década de 80: uma visão espaço-temporal. **Revista de Saúde Pública**, São Paulo, v. 31, n. 5, p. 457-465, 1997.
- LIMA, V. M. C.; SOARES JUNIOR, J. S.; VIOLA, D. N. Modelo autológico espacial com estrutura de vizinhança de terceira ordem: uma avaliação do estimador de pseudo-verossimilhança. In: ESCOLA DE MODELOS DE REGRESSÃO, 11., 2009, Recife. **Anais...** Recife, 2009.
- MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias**. 2. ed. São Paulo: USP, 2006. 428 p.
- MATHERON, G. Principles of geostatistics. **Economic Geology**, Littleton, v. 58, p. 1246-1266, 1963.
- McCULLAGH, P.; NELDER, J.A. **Generalized linear models**. 2. ed. London: Chapman e Hall, 1989. 261 p.
- MELLO, J. M. et al. Ajuste e seleção de modelos espaciais de semivariograma visando à estimativa volumétrica de *Eucalyptus grandis*. **Scientia Forestalis**, Piracicaba, v. 1, p. 25-37, 2005.
- MORAIS NETO, O. L. et al. Diferenças no padrão de ocorrência da mortalidade neonatal e pós-neonatal no Município de Goiânia, Brasil, 1992-1996: análise espacial para identificação das áreas de risco. **Caderno de Saúde Pública**, Rio de Janeiro, v. 17, n. 5, p. 1241-1250, 2001.
- MYUNG, I. J. Tutorial on maximum likelihood estimation. **Journal of Mathematical Psychology**, Orlando, v. 47, p. 90-100, 2003.
- NASCIMENTO, L. F. C. et al. Análise espacial da mortalidade neonatal no Vale do Paraíba, 1999 a 2001. **Revista de Saúde Pública**, São Paulo, v. 41, n. 1, p. 94-100, 2007.
- NOGUEIRA, D. A. **Análise espacial da mortalidade infantil no município de Alfenas, MG**. 2008. 101p. Tese (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2008.
- OLIVEIRA, M. C. N. **Métodos de estimação de parâmetros em modelos geoestatísticos com diferentes estruturas de covariâncias: uma aplicação ao teor de cálcio no solo**. 2003. 153 p. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 2003.

OLIVEIRA, M. C. N. et al. Estimação dos parâmetros de uma função de covariância no estudo da variabilidade espacial do teor de cálcio no solo. **Revista de Matemática e Estatística**, São Paulo, v. 24, n. 4, p. 117-132, 2006.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Viena: R Foundation for Statistical Computing, 2009. Disponível em: <<http://www.R-project.org>>. Acesso em: 15 jul. 2009.

RIBEIRO JÚNIOR, P. J. **Métodos geoestatísticos no estudo da variabilidade espacial de parâmetros do solo**. 1995. 99 p. Dissertação (Mestrado em Agronomia) – Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 1995.

RODRIGUEZ YI, J. L. et al. **Análise espacial da distribuição e dinâmica da violência na cidade de São Paulo entre os anos 1996 e 1999**. São José dos Campos: INPE, 2000.

SHIMAKURA, S. E. et al. Distribuição espacial do risco: modelagem da mortalidade infantil em Porto Alegre, Rio Grande do Sul, Brasil. **Caderno de Saúde Pública**, Rio de Janeiro, v. 17, n. 5, p. 1251-1261, 2001.

SIEGEL, S. **Estatística não paramétrica (para ciências do comportamento)**. São Paulo: McGraw-Hill, 1975. 320 p.

SIMÕES, C. C. S. **Perfis de saúde e de mortalidade no Brasil: uma análise de seus condicionantes em grupos populacionais específicos**. Brasília: Organização Pan-Americana da Saúde, 2002. 141 p.

VIEIRA, S. R. Uso de geoestatística em estudos de variabilidade espacial de propriedades do solo. In: ALVAREZ V. H.; NOVAIS, R. F.; SCHAEFER C. E. G. R. (Org.). **Tópicos em ciência do solo**. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 2000. v.1, p.1-54.

VIOLA, D. N. **Deteção e modelagem de padrão espacial em dados binários e de contagem**. 2007. 118 p. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 2007.