

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Discussão sobre a obtenção de
funções semivariograma a partir de
distribuições de probabilidade

por

Silvânia Ferreira Conceição

Orientador: Prof. Dr. Alan Ricardo da Silva

Maio de 2013

Silvânia Ferreira Conceição

**Discussão sobre a obtenção de
funções Semivariograma a partir de
distribuições de probabilidade**

Dissertação apresentada ao Departamento de
Estatística do Instituto de Ciências Exatas
da Universidade de Brasília como requisito
parcial à obtenção do título de Mestre em
Estatística.

Universidade de Brasília
Brasília, Maio de 2013

Universidade de Brasília-UnB

Instituto de Ciências Exatas

Departamento de Estatística

Coordenador: Prof. Dr. Gustavo L. Gilardoni

Dissertação aprovada como requisito parcial para obtenção do grau de mestre em estatística pelo programa de pós graduação em estatística da Universidade de Brasília e seguinte banca examinadora:

Prof. Dr. Alan Ricardo da Silva (orientador) - Est/UnB

Prof. Dr. Afrânio Márcio Corrêa Vieira - Est/UnB

Prof. Dr. Paulo Justiniano Ribeiro Júnior - Est/UFPR

CIP - Catalogação Internacional na Publicação

Conceição, Silvânia Ferreira

Discussão sobre a obtenção de funções semivariograma a partir de distribuições de probabilidade / Silvânia Ferreira Conceição. Brasília: UnB, 2013.

63p. :il.

Dissertação (Mestrado) - Universidade de Brasília, Brasília, 2013.

1. Semivariograma, 2. Krigagem Ordinária, 3. Família Exponencial.

CDU-519.2

E-mail: cfsilv@yahoo.com.br.

Agradecimentos

Agradeço a Deus, princípio de tudo, pela proteção e cuidado incondicional;

A minha família por entender a minha ausência;

Ao professor Doutor Alan Ricardo da Silva pelas orientações, ensinamentos e demonstração de paciência;

Ao professor Doutor Paulo Justiniano Ribeiro Júnior pela disponibilidade e esclarecimentos;

À professora Doutora Denise Nunes Viola, sempre solícita e dedicada;

À CAPES pelo apoio financeiro;

E, finalmente, aos colegas do mestrado pelo carinho, apoio, pelos bons momentos compartilhados e pela troca de conhecimentos.

Sumário

Agradecimentos	ii
Lista de Figuras	3
Resumo	4
Abstract	5
1 Introdução	6
2 Geoestatística	8
2.1 Introdução	8
2.2 Variáveis Regionalizadas (VR)	9
2.3 Processo espacial gaussiano	10
2.4 As funções covariograma e variograma	11
2.4.1 Modelo Exponencial de Semivariograma	15
2.4.2 Modelo Gaussiano de Semivariograma	16
2.4.3 Modelo Potência de Semivariograma	17
2.4.4 Modelo Esférico de Semivariograma	17
2.4.5 Modelo Cúbico de Semivariograma	17
2.4.6 Modelo Penta-Esférico de Semivariograma	18
2.4.7 Modelo Seno de Semivariograma	18
2.4.8 Modelo Mátern de Semivariograma	19
2.5 Estimação dos parâmetros do semivariograma	19
2.5.1 Método dos Mínimos Quadrados Ponderados	19
2.5.2 Método da máxima verossimilhança	21

2.6	Introdução à krigagem	21
2.6.1	Krigagem ordinária	22
3	Modelos de semivariograma	26
3.1	Limitações de Funções Semivariograma	26
3.1.1	Representação Espectral	27
3.1.2	Gerando modelos de variogramas e covariâncias a partir de uma função de probabilidade	28
3.2	A família exponencial	31
3.3	A família exponencial como modelo de semivariograma	35
3.3.1	Distribuição Normal	36
3.3.2	Distribuição Exponencial	38
3.3.3	Distribuição binomial negativa	38
3.3.4	Distribuição log-normal	39
3.3.5	Distribuição gamma	43
3.3.6	Distribuição de Poisson	47
4	Análise dos Resultados	52
4.1	Efeito dos semivariogramas propostos sobre a Krigagem ordinária . .	52
5	Conclusões	59
5.1	Limitações do Trabalho	60
5.2	Sugestões para trabalhos futuros	60
	Referências Bibliográficas	61

Lista de Figuras

2.1	Parâmetros do variograma	14
2.2	Representação gráfica dos modelos teóricos de variograma	15
3.1	Comportamento do semivariograma.	41
3.2	Comportamento do semivariograma para $a < 1$	44
3.3	Comportamento do semivariograma para diferentes alcances.	45
3.4	Curvas teóricas do modelo de semivariograma Poisson.	49
3.5	Curvas dos modelos clássicos de semivariograma e dos propostos.	51
4.1	Mapa de krigagem com base no semivariograma log-normal com alcance prático.	53
4.2	Mapa de krigagem com base no semivariograma log-normal (pequenas distâncias).	54
4.3	Erro padrão das estimativas de krigagem com base no semivariograma log-normal (pequenas distâncias).	54
4.4	Mapa de krigagem com base no semivariograma gaussiano.	55
4.5	Erro padrão das estimativas com base no semivariograma gaussiano.	56
4.6	Mapa de krigagem com base no semivariograma esférico.	56
4.7	Mapa de krigagem com base no semivariograma exponencial.	57
4.8	Erro padrão das estimativas com base no semivariograma exponencial.	57

Resumo

A distribuição espacial de um conjunto de variáveis fica caracterizada quando a posição geográfica contribui para análise e interpretação dos resultados. A análise exploratória é a primeira etapa a ser realizada em qualquer estudo. No contexto da geoestatística, esta análise é feita por meio de um semivariograma, que é uma função capaz de medir o grau de dependência espacial entre pares de observações separados por uma distância h . O processo de construção de uma função semivariograma não é trivial. Ao especificar uma função semivariograma, deve-se garantir que esta seja definida positiva. A fim de desmistificar eventuais falsas impressões quanto a viabilidade de funções como modelos de semivariograma foram sugeridos candidatos, utilizando como exemplo de funções de probabilidade, escritas sob a forma da família exponencial. As distribuições utilizadas foram gamma, log-normal, Poisson e binomial negativa. O estudo mostrou que não é recomendável utilizar funções escritas sob a forma da família exponencial para construir modelos de semivariograma, pois não há garantias de que o modelo a ser reproduzido seja válido. A construção dos modelos deve ser pautada nas propriedades das funções covariância. Tentativas *ad hoc* são muito específicas e não devem ser utilizadas para obter uma função definida positiva, uma vez que a solução encontrada não pode ser generalizada.

Palavras Chave: *Semivariograma, Krigagem Ordinária, Funções de Probabilidade e Família Exponencial.*

Abstract

The spatial distribution of a set of variables is characterized when the geographic position contributes to the analysis and for the interpretation of the results. The exploratory analysis is the first stage to be done in any study. In the geostatistics context the analysis is done by semivariogram, a function which measures the spatial dependence between pairs of observations separated by a distance h . Construct a semivariogram function is not trivial. The chosen function must be positive definite. For elucidate possible false impressions of viability in function semivariogram models were suggested as candidates probability functions, written in the form of the exponential family. Considering that, four semivariogram models from exponential family were derived: gamma, log-normal, Poisson and negative binomial. The study showed that it is not recommended to use functions written in the exponential family form to construct a semivariogram function, since there is no guarantee that the model will be valid. The construction of models should be based on the properties of the functions covariance. *Ad hoc* attempts are very specific and should not be used to obtain a positive-definite function, since the solution can not be generalized.

key words: *Semivariogram, Ordinary Kriging, Probability Functions and Exponential Family.*

Capítulo 1

Introdução

Os conceitos estatísticos exercem grande influência na maioria dos campos do conhecimento humano. Deste modo, uma variedade de técnicas estatísticas e modelos matemáticos são utilizados com o objetivo de identificar qual o comportamento das variáveis em estudo e como elas se relacionam. As ferramentas mais usuais de análise estatística requerem que as variáveis em estudo sejam independentes. No entanto, há situações em que as variáveis medidas estão correlacionadas no espaço.

Diz-se que um conjunto de observações é espacialmente distribuído quando a sua localização geográfica é importante para a análise e interpretação dos resultados (Bailey e Gatrell, 1995). Neste caso, convém adotar técnicas que permitam acomodar a dependência espacial existente entre os valores observados. Em geostatística a estrutura de correlação espacial inerente aos dados é considerada e a análise parte do princípio de que as observações mais próximas são mais semelhantes do que as mais afastadas. Sendo assim, é fácil intuir que a variabilidade do atributo medido aumentará ao passo que a distância entre os pares de observações aumenta. Daí a importância de conhecer a posição geográfica das unidades amostrais para empregar a geoestatística como ferramenta de análise.

Frequentemente, experimentos na esfera da agronomia, da mineração, das ciências da saúde, entre outras, são conduzidos com a finalidade de estimar valores da variável em locais onde não houve medida. Uma das técnicas geoestatísticas mais utilizadas para a predição de valores não observados é a de interpolação por krigagem, que utiliza a autocorrelação espacial como medida de ponderação. A autocorrelação espacial

é modelada utilizando um semivariograma, ferramenta que descreve quantitativamente a variação no espaço de um fenômeno regionalizado. Desta forma, é razoável pensar que quanto melhor o ajuste da função semivariograma mais precisas serão as estimativas.

Para que um modelo de semivariograma seja considerado apropriado a função covariância que o descreve precisa ser semidefinida positiva. Tal condições não é fácil de garantir e a dificuldade em avaliar a adequabilidade de um candidato a modelo de semivariograma é maior quando se trabalha com grandes dimensões devido a obstáculos matemáticos (Christakos, 1984). Assim, investigar as propriedades de uma função semivariograma possibilita identificar se uma função qualquer pode ser aceita como modelo teórico.

Cada semivariograma possui singularidades, uma vez que sua formulação envolve diferentes restrições no que tange a viabilidade da função que o representa. Este estudo visa explorar situações que podem induzir a equívocos no processo de construção de semivariograma utilizados na krigagem utilizando como exemplo a família exponencial. As análises serão fundamentadas, inicialmente, nos principais modelos existentes e também supondo situações diversas.

Este documento foi organizado em 5 Capítulos. O Capítulo 2 versa sobre os modelos de semivariograma, a utilização dos métodos dos mínimos quadrados ponderados e da máxima verossimilhança para a estimação dos parâmetros do semivariograma teórico e a técnica de krigagem ordinária. O Capítulo 3 discute as limitações de funções semivariograma, apresenta com brevidade a família exponencial, além de fazer uma explanação acerca de algumas particularidades dos principais modelos de semivariograma, bem como levanta questões sobre o processo de construção de modelos fundamentados nas propriedades da família exponencial. O Capítulo 4 é dedicado à exposição e discussão dos resultados alcançados e no Capítulo 5 constam as considerações finais do trabalho.

Capítulo 2

Geoestatística

Este Capítulo tem por finalidade apresentar e discutir os aspectos estruturais das funções de variograma utilizadas na krigagem. Neste tópico serão abordados conceitos como semivariograma método dos mínimos quadrados ponderados, método da máxima verossimilhança e krigagem.

2.1 Introdução

A geoestatística tem como objetivo entender o comportamento dos dados espaciais no que tange a variabilidade inerente ao fenômeno em estudo (Chilès e Delfiner, 1999). A análise espacial utilizando os conceitos da geoestatística ganhou força a partir dos anos 60 quando o engenheiro Daniel Krige e o matemático George Matheron, ao analisarem a correlação espacial entre as variáveis em trabalhos na área de mineração, impulsionaram a expansão da técnica para outros campos de aplicação, tais como agricultura, hidrologia entre outros. Considerando que a geoestatística lida com a incerteza espacial, os dados espaciais são vistos como variáveis aleatórias e a variação do processo é definida admitindo que há uma relação entre os valores das variáveis cujas localizações são próximas umas das outras (Fischer e Getis, 2010).

O estudo geoestatístico inclui um conjunto de procedimentos que podem ser utilizados para analisar e modelar as relações entre determinadas características de fenômenos espaciais. A primeira etapa consiste em uma análise exploratória, realizada por meio de um semivariograma, seguida da modelagem, que visa fornecer

soluções para o problema de estimação por meio da técnica de krigagem (Sen, 2009). Esta metodologia é baseada na teoria das variáveis regionalizadas (VR), cujo conceito será abordado a seguir.

2.2 Variáveis Regionalizadas (VR)

Para Chilès e Delfiner (1999), Sen (2009) dentre outros autores, uma variável é definida como regionalizada quando representa uma característica de algum fenômeno distribuído em um espaço contínuo. Esta variável é denotada por $Z(u)$, em que u identifica a localização geográfica de onde foi realizada a medição. As variáveis regionalizadas reúnem dois aspectos aparentemente contraditórios: ainda que selecionadas ao acaso no espaço, elas apresentam a estrutura de correlação que refletem a natureza do fenômeno. Admitindo que as posições das amostras são selecionadas aleatoriamente, as variáveis regionalizadas podem ser vistas como uma dentre tantas realizações possíveis ao acaso. Isto é, elas podem ser interpretadas como uma realização $z(u)$ de uma função aleatória que, no contexto da geoestatística, pode ser definida como sendo a distribuição conjunta das variáveis aleatórias $Z(u_i)$, $i = 1, 2, \dots, n$.

O modelo probabilístico obtido deve ser capaz de reproduzir tais distribuições ou pelo menos alguns de seus momentos. Os dois primeiros momentos de uma função aleatória são dados pela média, que é o valor esperado $\mu_u = E[Z(u)]$ e pela covariância $C(u_i, u_j) = \{E[Z(u_i) - \mu_{u_i}][Z(u_j) - \mu_{u_j}]\}$. Em geral, na prática, apenas uma realização da variável $Z(u_i)$ é observada. Neste caso, para estimar um valor para a variável em um ponto onde não houve medida, é usual estabelecer restrições a fim de que a VR seja estacionária. A restrição implica que os momentos da variável aleatória $Z(u_i + h)$ devem ser constantes para qualquer vetor de distâncias h .

Neste contexto, segundo Chilès e Delfiner (1999) e Sen (2009) existem duas hipóteses que caracterizam a estacionariedade de uma função aleatória e ao menos uma precisa ser assumida para fins de inferência. As hipóteses serão definidas a seguir:

Hipótese 2.2.1 (Estacionariedade de segunda ordem). *Admite que a média é constante e a covariância entre dois pares quaisquer $Z(u)$ e $Z(u + h)$ existe e depende somente de h .*

Similarmente, tem-se que $E[Z(u)] = \mu$, constante $\forall u \in D$ e $Cov(Z(u_i), Z(u_j)) = C(u_i - u_j) = C(h) \forall u_i, u_j \in D$ e $h = \|u_i - u_j\|$, em que $C(\cdot)$ é a função covariância estacionária ou covariograma.

Hipótese 2.2.2 (Intrínseca). *Estabelece que os incrementos $Z(u) - Z(u+h)$ tenham esperança zero e variância somente em função de h .*

A estacionariedade intrínseca é menos restritiva se comparada com a estacionariedade de segunda ordem. Segundo Cressie (1993), a estacionariedade intrínseca é definida pelas expressões: $E[Z(u+h) - Z(u)] = 0$ e $Var[Z(u+h) - Z(u)] = 2\gamma(h)$, em que a quantidade $\gamma(h)$ é a função semivariograma, principal ferramenta de análise para dados espacialmente distribuídos, conforme será abordado nas próximas seções.

A variável regionalizada é definida por um valor numérico que está associado a localização da amostra. A maneira que as VR se relacionam umas com as outras pode sofrer alterações em função da orientação das amostras, da distância entre os pares de observações ou ambos. Neste sentido, surgem dois novos conceitos: anisotropia e isotropia. O primeiro indica que a relação entre os pares de observações separados por uma distância h depende da direção que esta distância é medida. Já no caso da isotropia, o comportamento das variáveis não muda com a direção e depende apenas da distância h . Todos os conceitos trabalhados aqui são voltados para dados isotrópicos. Modelos com anisotropia são abordados com detalhes em Isaaks e Srivastava (1989), Chilès e Delfiner (1999), Cressie (1993) dentre outros autores.

2.3 Processo espacial gaussiano

Um processo espacial gaussiano, $\{Z(u) : u \in \mathbb{R}^d\}$, é um processo estocástico, tal que um conjunto finito dimensional $D = \{Z(u_1), \dots, Z(u_n)\}$ segue uma distribuição gaussiana multivariada, que é caracterizada pelas funções média e covariância. Geralmente uma variável aleatória não é completamente descrita por alguns parâmetros. Entretanto, a variável aleatória gaussiana pode ser definida completamente por sua média e covariância. Desta forma, um processo gaussiano é dito estacionário se a média for constante para todas as posições u e a covariância depender apenas do vetor de distâncias h (Diggle e Ribeiro Jr., 2007; Isaaks e Srivastava, 1989).

Muitas das técnicas utilizadas para modelar dados geoestatísticos são fundamentadas em processos gaussianos. Segundo Cressie (1993), a importância de utilizar processos gaussianos está na facilidade de trabalhar com tal distribuição. De acordo com Schlather et al. (2012), a suposição de normalidade é fundamental para simulação de dados espaciais, uma vez que a distribuição normal multivariada mantém-se estável sob a soma e o condicionamento.

2.4 As funções covariograma e variograma

Segundo Schlather et al. (2012) a função covariância desempenha um papel importante na modelagem de dados espaciais, uma vez que além de caracterizar um processo gaussiano, também é um elemento chave para realizar interpolação por krigagem e para simular dados com dependência espacial. Uma condição para a utilização dessa ferramenta é a continuidade do processo e que este seja normalmente distribuído. A análise variográfica é a primeira e indispensável etapa a ser realizada, pois permite diagnosticar se existe dependência espacial entre duas observações. Quando duas unidades amostrais são espacialmente dependentes, espera-se obter variações cada vez maiores à medida que a distância entre elas aumenta. Isto é, observações geograficamente mais próximas se comportam de maneira mais similar do que as mais afastadas. Neste sentido, as funções covariograma e variograma podem ser vistas como uma medida da dispersão das diferenças entre os valores de pontos no espaço, separados por uma distância h .

Sob a suposição de estacionariedade a função $C(\cdot)$ mencionada na seção 2.2, também chamada de autocovariância, pode ser representada por :

$$\begin{aligned} Cov[Z(u), Z(u+h)] &= E[Z(u) - \mu][Z(u+h) - \mu] \\ &= E[Z(u)Z(u+h) - \mu^2] \\ &= C(h). \end{aligned} \tag{2.1}$$

A covariância se relaciona com a função correlação, por meio da expressão:

$$\rho(h) = \frac{C(h)}{C(0)}. \tag{2.2}$$

Para que as funções covariância e correlação sejam admitidas como covariograma e correlograma, respectivamente, importa que o valor da covariância no ponto $h = 0$ seja positivo, isto é, $C(0) > 0$ (Cressie, 1993).

Assumindo a hipótese de estacionariedade verifica-se que a covariância é uma função par, conforme mostram os passos a seguir:

$$\begin{aligned} C(h) = Cov[Z(u), Z(u+h)] &= E[Z(u) - \mu][Z(u-h) - \mu] \\ &= E[Z(u)Z(u-h) - \mu^2] \\ &= C(-h). \end{aligned}$$

Sob a hipótese de estacionariedade intrínseca a função variograma de um processo em dois pontos quaisquer é definida como:

$$\begin{aligned} Var[Z(u) - Z(u+h)] &= E[\{Z(u) - Z(u+h)\}^2] \\ &= 2\gamma(h). \end{aligned} \tag{2.3}$$

A função semivariograma se relaciona com o covariograma. Entretanto, essa relação só é estabelecida quando o estudo envolve um processo estacionário de segunda ordem:

$$\begin{aligned} Var[Z(u) - Z(u+h)] &= Var[Z(u)] + Var[Z(u+h)] - 2Cov[Z(u), Z(u+h)] \\ &= C(0) - C(h) \\ &= 2\gamma(h). \end{aligned} \tag{2.4}$$

A equivalência entre as funções covariância e semivariograma é dada pela expressão (2.4). Admitindo apenas a hipótese intrínseca a relação representada pela Equação (2.4) não é válida, uma vez que a função covariância pode não existir (Webster e Oliver, 2007).

O variograma, $2\gamma(h)$, também pode ser expresso em função da autocorrelação. Utilizando a equivalência apresentada pela Equação (2.2) e substituindo em (2.4) obtém-se:

$$\gamma(h) = C(0)[1 - \rho(h)]. \tag{2.5}$$

Existem três tipos de variograma:

- variograma real, sempre desconhecido;
- variograma experimental, ajustado a partir dos valores observados;
- variograma teórico, representado por um modelo matemático.

O variograma real pode ser estimado a partir do variograma experimental. Usualmente a variabilidade espacial é medida pelo semivariograma, que corresponde à metade do variograma. Um estimador clássico, baseado no método dos momentos, proposto por Matheron (1962) é representado por:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(u_i) - Z(u_i + h)]^2, \quad (2.6)$$

em que $\hat{\gamma}(h)$ representa o semivariograma experimental e $N(h)$ o número de pares medidos $Z(u)$ e $Z(u + h)$, separados pelo vetor de distâncias h (Cressie, 1993).

Para determinar o semivariograma experimental a partir do estimador clássico recomenda-se introduzir limites de tolerância para as distâncias, formando classes de distâncias ou *lags*. A semivariância é calculada para cada par de observações, cuja distância esteja compreendida dentro dos limites pré-especificados e esse processo é repetido para cada *lag*. Na presença de anisotropia, além das distâncias, deve-se considerar também as direções.

Segundo Druck et al. (2004) o semivariograma é uma ferramenta básica de suporte às técnicas de krigagem, pois permite representar quantitativamente a variação de um fenômeno regionalizado no espaço. Entretanto, as semivariâncias calculadas a partir do semivariograma experimental não pode ser utilizada como pesos no processo de predição por krigagem, uma vez que este semivariograma pode não satisfazer as condições necessárias para que um modelo seja válido. Assim, ajusta-se ao semivariograma experimental uma função matemática que descreva, continuamente, a variabilidade ou correlação espacial existente nos dados, dando origem ao semivariograma teórico. Essa função deve ser tal que represente o comportamento do semivariograma amostral em relação a uma certa distância h e respeite, por exemplo, a restrição de ser positiva definida.

A interpretação do semivariograma teórico se dá a partir da leitura de seus parâmetros, que podem ser identificados na Figura 2.1.

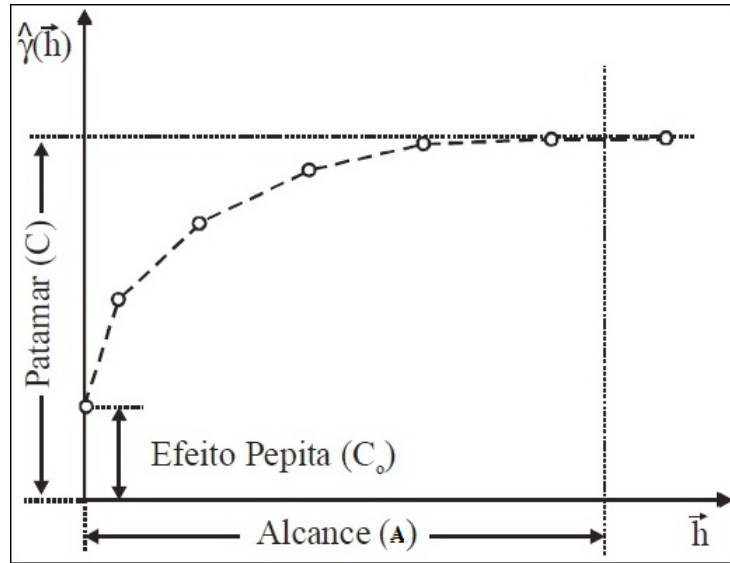


Figura 2.1: Parâmetros do variograma
Fonte: Druck et al. (2004)

- Alcance (A) : é a distância dentro da qual as amostras apresentam-se correlacionadas espacialmente, e que corresponde ao nível em que a variabilidade se estabiliza ($A \geq 0$);
- Patamar (C): é o valor do semivariograma correspondente a seu alcance. Isto é, o nível em que a variância da diferença entre os pares amostrais torna-se aproximadamente constante ($C \geq 0$);
- Efeito Pepita (C_0): é o valor da semivariância para a distância zero e representa uma componente da variabilidade espacial que não pode ser relacionada a uma causa específica. Se $C_0 > 0$ então a contribuição do modelo é definida por $C_1 = C - C_0$, $C_1 \geq 0$ (Druck et al., 2004).

Geralmente o ajuste de modelos teóricos em geoestatística é feito de maneira iterativa, em que a cada passo novos valores para os parâmetros são fornecidos até que o ajuste seja considerado satisfatório. Cressie (1993) aborda diversos métodos analíticos para estimar os parâmetros do semivariograma, tais como o método de máxima verossimilhança, máxima verossimilhança restrita e o de mínimos quadrados.

Neste trabalho serão abordados os métodos dos mínimos quadrados ponderados e máxima verossimilhança. Outras modalidades do método dos mínimos quadrados, bem como abordagem sobre máxima verossimilhança restrita, podem ser vistas com detalhes em Lele (1997); Gotway (1991); Cressie (1993); Zhang et al. (1995); Jian et al. (1996) dentre outros autores.

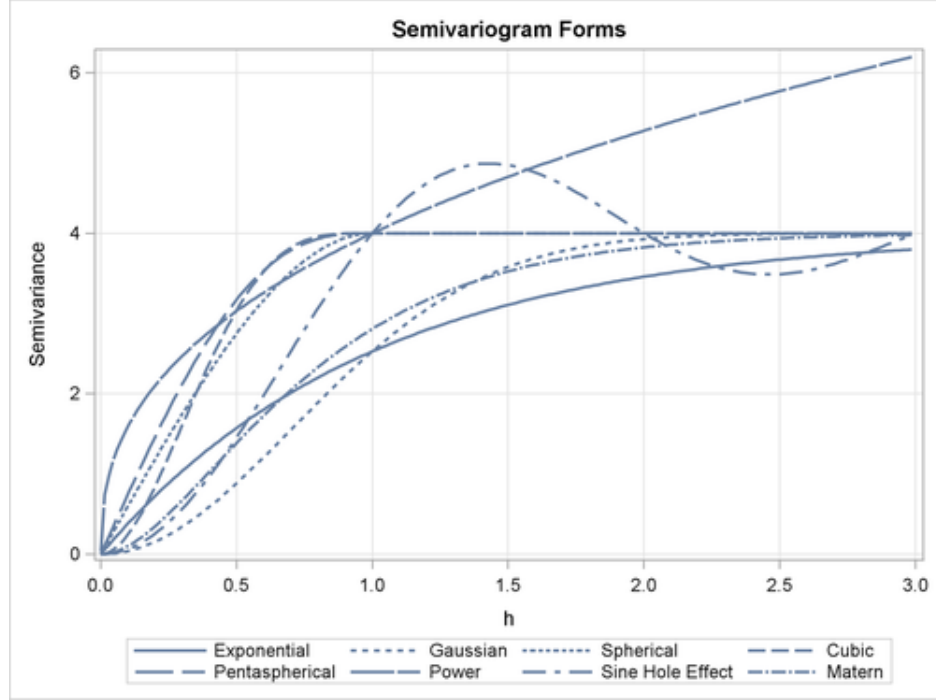


Figura 2.2: Representação gráfica dos modelos teóricos de variograma
Fonte: SAS (2011)

A Figura 2.2 ilustra os principais modelos teóricos de semivariograma isotrópicos. A formulação matemática para cada modelo será descrita na próxima seção.

2.4.1 Modelo Exponencial de Semivariograma

O modelo exponencial de semivariograma é válido em qualquer dimensão $\mathbb{R}^d$, $d \geq 1$ e é descrito pela Equação (2.7), com parâmetro de escala $a > 0$. Este e os demais modelos serão expressos em função de $r = |h|$, que indica a distância euclidiana entre os pares de observações.

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + C_1 \left[1 - \exp\left(-\frac{3r}{a}\right)\right] & \text{se } r \neq 0. \end{cases} \quad (2.7)$$

Equivalentemente, o semivariograma exponencial também pode ser escrito conforme a Equação (2.8)

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + C_1 \left[1 - \exp\left(-\frac{r}{a}\right)\right] & \text{se } r \neq 0. \end{cases} \quad (2.8)$$

Alguns semivariogramas atingem o seu patamar a uma distância r finita, que é denotada como alcance. O semivariograma exponencial atinge o patamar apenas assintoticamente, quando $r \rightarrow \infty$. Então, para efeito de análise, adota-se como alcance prático a distância que corresponde a 95% do patamar (C_1). Sendo assim, o alcance prático para este modelo, observando a Equação (2.8) vale $3a$ (Chilès e Delfiner, 1999; Webster e Oliver, 2007).

$$\begin{aligned} \gamma(r) &= 1 - \exp\left(-\frac{r}{a}\right) \\ 0.95 &= 1 - \exp\left(-\frac{r}{a}\right) \\ 0.05 &= \exp\left(-\frac{r}{a}\right) \\ 3 &\approx \frac{r}{a} \\ r &\approx 3a. \end{aligned} \quad (2.9)$$

A expressão (2.9) mostra o passo a passo realizado até a definição do alcance prático $A = 3a$. O alcance prático é atingido quando a distância assume o valor equivalente a $3a$ se o semivariograma for da forma (2.8). Utilizando a Equação (2.7) o valor encontrado para a já corresponde ao alcance prático, isto é $A = a$.

2.4.2 Modelo Gaussiano de Semivariograma

O modelo gaussiano, tal como o exponencial, também é válido em qualquer dimensão $\mathbb{R}^d$, $d \geq 1$. Além disto, seu patamar é atingido assintoticamente e o alcance prático é $A = \sqrt{3}a$. O semivariograma Gaussiano é definido pela Equação (2.10):

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + C_1 \left[1 - \exp\left(-\frac{r}{a}\right)^2\right] & \text{se } r \neq 0. \end{cases} \quad (2.10)$$

O alcance prático para o modelo Gaussiano é encontrado de maneira análoga ao

do semivariograma exponencial. Um cuidado ao trabalhar com o semivariograma Gaussiano é observar o efeito pepita, pois quando este é nulo o modelo apresenta problemas numéricos (Webster e Oliver, 2007).

2.4.3 Modelo Potência de Semivariograma

O modelo potência é válido em $\mathbb{R}^d$, $d \geq 1$. Sua representação é dada pela Equação (3.4):

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + cr^\alpha & \text{se } r \neq 0. \end{cases} \quad (2.11)$$

em que $c > 0$ é o coeficiente de declividade e $0 < \alpha < 2$ é o expoente.

O modelo potência não tem patamar. Além disso, quando $\alpha = 0$ tem-se o efeito pepita puro.

2.4.4 Modelo Esférico de Semivariograma

O modelo esférico é válido até a dimensão 3. Sua representação matemática é dada pela seguinte equação:

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + C_1 \left[\frac{3}{2} \left(\frac{r}{a} \right) - \frac{1}{2} \left(\frac{r}{a} \right)^3 \right] & \text{se } 0 < r \leq a; \\ C_0 + C_1 & \text{se } r > a. \end{cases} \quad (2.12)$$

O semivariograma esférico exibe um comportamento linear perto da origem e atinge o patamar quando $r = a$, seu parâmetro de escala. Para este modelo $A = a$.

2.4.5 Modelo Cúbico de Semivariograma

O modelo cúbico é assim denominado pelo fato de que o termo irregular de menor grau é o r^3 .

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + C_1 \left[7 \left(\frac{r}{a} \right)^2 - \frac{35}{4} \left(\frac{r}{a} \right)^3 + \frac{7}{2} \left(\frac{r}{a} \right)^5 + \frac{3}{4} \left(\frac{r}{a} \right)^7 \right] & \text{se } 0 < r \leq a; \\ C_0 + C_1 & \text{se } r > a. \end{cases} \quad (2.13)$$

O modelo cúbico é descrito pela Equação (2.13). Uma vantagem deste modelo é que ele é suave na origem. Seu gráfico se assemelha ao do modelo esférico.

2.4.6 Modelo Penta-Esférico de Semivariograma

O modelo penta-esférico é considerado um semivariograma válido em até três dimensões. Isto é, um modelo válido em $\mathbb{R}^d$, $d \leq 3$.

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + C_1 \left[\frac{15}{8} \left(\frac{r}{a} \right) - \frac{5}{4} \left(\frac{r}{a} \right)^3 + \frac{3}{8} \left(\frac{r}{a} \right)^5 \right] & \text{se } 0 < r \leq a; \\ C_0 + C_1 & \text{se } r > a. \end{cases} \quad (2.14)$$

2.4.7 Modelo Seno de Semivariograma

O semivariograma seno exibe flutuações periódicas, que se tornam cada vez menos pronunciadas a medida que o *lag* cresce. Este modelo é considerado válido em $\mathbb{R}^d$, $d \leq 3$ e sua equação é dada por:

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + C_1 \left[1 - \frac{\text{sen}\left(\frac{\pi r}{a}\right)}{\frac{\pi r}{a}} \right] & \text{se } r > 0. \end{cases} \quad (2.15)$$

O modelo seno pode resultar em correlações negativas devido a periodicidade do processo. Em consequência da relação estabelecida pela Equação (2.2), o valor mínimo que a função pode assumir não deve ser menor do que -0.218 , atingido quando $h \simeq 4.5a$.

2.4.8 Modelo Mátern de Semivariograma

A partir da função Mátern é possível derivar outros modelos, por exemplo o modelo exponencial quando $\nu = 0.5$. Sua fórmula é dada por:

$$\gamma(r) = \begin{cases} 0 & \text{se } r = 0; \\ C_0 + C_1 \left[1 - \frac{2}{\Gamma(\nu)} \left(\frac{r\sqrt{\nu}}{a} \right)^\nu K_\nu \left(\frac{2r\sqrt{\nu}}{a} \right) \right] & \text{se } r > 0, \nu > 0. \end{cases} \quad (2.16)$$

em que K é a função Bessel, $\Gamma(\nu)$ é a função Gama e ν é o parâmetro de suavização.

O semivariograma Mátern é válido em $\mathbb{R}^d$, $d \geq 1$ e pode assumir qualquer tipo de comportamento próximo a origem, assumindo a forma $r^{2\nu}$ se ν não for inteiro e $r^{2\nu} \log(r)$ para ν inteiro.

2.5 Estimação dos parâmetros do semivariograma

2.5.1 Método dos Mínimos Quadrados Ponderados

O ajuste do modelo semivariograma por mínimos quadrados ponderados é utilizado para encontrar a solução ótima para o conjunto de parâmetros do semivariograma. Neste contexto, o melhor conjunto de parâmetros é o que minimiza a soma do quadrado das diferenças entre os valores experimentais e os preditos pelo modelo teórico do semivariograma, em que cada diferença é ponderada diretamente pelo inverso da variância do semivariograma experimental. Sendo assim, a Equação (2.17) deve ser minimizada para a obtenção dos parâmetros:

$$\sum_{j=1}^k \frac{N(h_j)}{2} \left[\frac{\hat{\gamma}(\underline{h}(j))}{\gamma(\underline{h}(j); \underline{\eta})} - 1 \right]^2, \quad (2.17)$$

em que j é o *lag* e $\underline{\eta}$ representa o conjunto de parâmetros (Cressie, 1985).

Das et al. (2012) em seu trabalho, ilustra a utilização desta metodologia com aplicação em dados de precipitação, além de estudar as propriedades assintóticas do estimador citado.

As estimativas são obtidas por meio de um processo iterativo, o que requer que os parâmetros sejam inicializados. Os valores iniciais a serem atribuídos aos parâmetros devem ser tais que facilitem a convergência. Na ausência de conhecimento prévio

sobre as flutuações no semivariograma experimental, Jian et al. (1996) sugerem que os seguintes valores sejam utilizados para iniciar o método de iteração:

1. Para todos os modelos descritos anteriormente o valor padrão para efeito pepita é dado por:

$$C_{0_{inicial}} = \text{Max} \left[0, \gamma(h_1) - \frac{h_1}{h_2 - h_1} (\hat{\gamma}(h_2) - \hat{\gamma}(h_1)) \right]$$

2. Para o modelo potência os padrões são:

$$c_{inicial} = \frac{(\hat{\gamma}(h_{k-2}) + \hat{\gamma}(h_{k-1}) + \hat{\gamma}(h_k)) / (3 - C_{0_{inicial}})}{h_k - h_1}$$

$$\alpha_{inicial} = 1, 0.$$

3. Para todos os outros modelos teóricos citados neste documento:

$$a_{inicial} = \frac{h_k}{2}$$

$$C_{1_{inicial}} = \frac{\hat{\gamma}(h_{k-2}) + \hat{\gamma}(h_{k-1}) + \hat{\gamma}(h_k)}{3} - C_{0_{inicial}},$$

em que h_k indica o k -ésimo *lag*. Mais detalhes sobre a técnica mínimos quadrados ponderados podem ser vistos em Diggle e Ribeiro Jr. (2007); Cressie (1993); Das et al. (2012) dentre outros autores.

Após o processo de modelagem é importante medir a qualidade do ajuste realizado. Esta medição é feita por meio do critério de Akaike (AIC), que segundo Cressie (1985), para este contexto é representado por :

$$AIC = k \ln(R) + 2q, \quad (2.18)$$

em que k identifica a classe ou *lag* do semivariograma, q é o número de parâmetros do modelo e R representa o quadrado médio dos resíduos entre os valores experimentais

e os preditos pelo modelo teórico do semivariograma ajustado dado por:

$$R = \sum_{j=1}^k \frac{1}{kC_0^2} [\hat{\gamma}(\underline{h}(j)) - \gamma(\underline{h}(j); \underline{\eta})]^2. \quad (2.19)$$

2.5.2 Método da máxima verossimilhança

O método de máxima verossimilhança é amplamente utilizado como um meio de estimativa de parâmetros. Segundo Schabenberger e Gotway (2005), no contexto de predição espacial, o método parte do princípio de que os dados são observados a partir de uma distribuição gaussiana multivariada com vetor média $\mu = (E[Z(x_1)], \dots, E[Z(x_N)])^T$ e matriz de covariância $\Sigma(\eta)$. Desta forma, a densidade de probabilidade conjunta da variável $Z(x_i), i = 1, 2, \dots, n$ é dada por:

$$f(z) = (2\pi)^{-\frac{n}{2}} |\Sigma(\eta)|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (z - \mu)' \Sigma(\eta)^{-1} (z - \mu) \right).$$

As estimativas dos parâmetros são produzidas a partir da maximização da verossimilhança ou, equivalentemente, pela minimização da log-verossimilhança negativa. Assim, assumindo que $Z \sim \text{gaus}(X\beta, \Sigma(\eta))$, então a matriz de variâncias e covariâncias espacial ($\Sigma(\eta)$) pode ser estimada por meio da Equação (2.20):

$$L(X\beta, \eta) = \frac{1}{2} (n \ln(2\pi) + \ln(|\Sigma(\eta)|) + (Z - X\beta)' \Sigma(\eta)^{-1} (Z - X\beta)). \quad (2.20)$$

A qualidade do ajuste é medida pelo Critério de Akaike, fazendo:

$$AIC = -2 \ln(\text{verossimilhança maximizada}) + 2(\text{número de parâmetros}).$$

2.6 Introdução à krigagem

A krigagem é uma técnica de interpolação voltada para a predição de valores de variáveis contínuas no espaço. Este nome foi atribuído em homenagem ao engenheiro de minas Daniel Krige que na década de 60 introduziu o método, posteriormente aprimorado pelo matemático francês Georges Matheron (Bailey e Gatrell, 1995). O desenvolvimento da krigagem tinha como propósito inicial a solução de problemas da

área de mineração, petróleo, engenharia de controle e redução da poluição e saúde pública, mas devido ao sucesso nos resultados alcançados a técnica se expandiu e passou a ser utilizada fora do ambiente para o qual foi idealizada.

Para Journel (1986) considerar a krigagem simplesmente como um método de interpolação é no mínimo um descuido. A estimação por krigagem está associada a um processo estocástico, permite medir os erros das estimativas e ter uma ideia do quão confiáveis são os resultados. O processo de estimação envolve a atribuição de pesos a cada unidade amostral e esta ponderação é determinada com base no semivariograma (Druck et al., 2004).

A krigagem compreende um conjunto de técnicas de estimação e predição de superfícies baseados na estrutura de correlação espacial tais como: krigagem ordinária (um métodos dos mais utilizados), krigagem simples, krigagem lognormal, krigagem universal, krigagem fatorial, cokrigagem ordinária (que é a extensão da krigagem ordinária para duas ou mais variáveis), krigagem indicatriz, krigagem disjuntiva e krigagem probabilística (Cressie, 1993). Há também a krigagem bayesiana, que inclui medidas de incerteza para a estrutura de covariância.

O desenvolvimento deste trabalho está voltado para a krigagem ordinária e as demais modalidades da técnica não serão abordadas, mas podem ser vistas em Handcock e Stein (1993), Cressie (1993), Diggle e Tawn (1998), Schmidt (2001), Webster e Oliver (2007), Diggle e Ribeiro Jr. (2007).

2.6.1 Krigagem ordinária

A krigagem ordinária é a mais comum e mais utilizada na prática. Esta técnica tem por objetivo estimar o valor de uma variável aleatória, Z em um ou mais pontos não amostrados. Considerando o processo estocástico D , descrito em (1.3), um processo estacionário com função semivariograma $\hat{\gamma}$, então qualquer atributo desconhecido de Z no ponto u_0 pode ser estimado pela combinação linear dos seus vizinhos:

$$\hat{Z}(u_0) = \sum_{i=1}^n \lambda_i Z(u_i), \quad (2.21)$$

em que u identifica uma posição em duas dimensões, λ_i são os pesos determinados de forma a minimizar o erro das estimativas, tal que $\sum_{i=1}^n \lambda_i = 1$ (Journal, 1986).

Se $\hat{Z}(u_i)$ um estimador não tendencioso de $Z(u_0)$, então a diferença média entre o valor real e o estimado deve ser nula (Equação 2.22). Além disso, se este estimador tiver variância mínima, então ele produzirá a melhor estimativa para $Z(u_0)$.

$$E \left[\hat{Z}(u_0) - Z(u_0) \right] = 0. \quad (2.22)$$

A variância é estimada por:

$$Var \left[\hat{Z}(u_0) \right] = E \left[\hat{Z}(u_0) - Z(u_0) \right]^2 = 2 \sum_{i=1}^N \lambda_i \hat{\gamma}(u_i, u_0) - \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \hat{\gamma}(u_i, u_j), \quad (2.23)$$

em que $\gamma(u_i, u_j)$ é a semivariância entre os pontos u_i e u_j e $\gamma(u_i, u_0)$ é a semivariância entre o i -ésimo ponto e o ponto u_0 .

Para cada estimativa existe uma variância de krigagem associada, a qual pode-se denotar por $\sigma^2(u_0)$ e que pode ser definida pela Equação (2.23). Os pesos que minimizam estas variações são obtidos usando o método dos multiplicadores de Lagrange. Este procedimento dá origem a um sistema de equações lineares ou krigagem ordinária (Journal, 1989).

Os pesos λ_i extraídos do sistema de krigagem ordinária são substituídos na Equação (2.21), através da qual se obtém a estimativa da variância:

$$\sigma^2(u_0) = \sum_{i=1}^N \lambda_i \gamma(u_i, u_0) + \Psi(u_0),$$

em que Ψ é o multiplicador de Lagrange.

Em notação matricial as equações de krigagem são representadas por $\mathbf{A}\lambda = \mathbf{b}$, em que $\mathbf{A}$ é a matriz das semivariâncias dos valores amostrados, λ é a matriz coluna que contém os pesos λ_i e o multiplicador de Lagrange e $\mathbf{b}$ é a matriz coluna das semivariâncias entre os valores amostrados e o ponto a ser estimado. Então:

$$\mathbf{A} = \begin{bmatrix} \gamma(u_1, u_1) & \gamma(u_1, u_2) & \dots & \gamma(u_1, u_N) & 1 \\ \gamma(u_2, u_1) & \gamma(u_2, u_2) & \dots & \gamma(u_2, u_N) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(u_N, u_1) & \gamma(u_N, u_2) & \dots & \gamma(u_N, u_N) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix}, \quad \lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \Psi(u_0) \end{bmatrix}$$

$$\text{e } \mathbf{b} = \begin{bmatrix} \gamma(u_1, u_0) \\ \gamma(u_2, u_0) \\ \vdots \\ \gamma(u_N, u_0) \\ 1. \end{bmatrix}$$

A matriz $\mathbf{A}$ é inversível e tem como componentes as semivariâncias calculadas a partir de um modelo de semivariograma válido, de maneira que os pesos podem ser obtidos da seguinte forma:

$$\lambda = \mathbf{A}^{-1}\mathbf{b} \quad (2.24)$$

e a variância é dada por:

$$\hat{\sigma}^2(u_0) = \mathbf{b}'\lambda. \quad (2.25)$$

A variância das estimativas serão maiores ao passo que a distância entre o ponto a ser estimado e os pontos onde houve medição aumenta. Desta forma, uma vez que os pesos são atribuídos em função do semivariograma quanto mais próximo o ponto medido for do alvo, maior será seu peso.

A metodologia acima mencionada pode ser considerada como krigagem pontual, pois as estimativas produzidas nesse processo são obtidas tendo como suporte locais individuais. Quando, ao invés de uma posição individual, uma região é considerada como suporte o procedimento é chamado de krigagem em blocos (Sen, 2009). Muitas vezes o interesse do pesquisador não está em prever o valor para uma variável localizada em um ponto específico, mas sim em um valor médio por áreas. Neste caso, aplica-se a krigagem em blocos ao invés da krigagem pontual e o valor da estimativa obtida representa todo o suporte. As expressões a serem utilizadas são bem parecidas

com a da estimação pontual, sendo suficiente substituir o termo u_0 pelo bloco B . Sendo assim, a matriz coluna b para krigagem em blocos é expressa conforme segue:

$$\mathbf{b} = \begin{bmatrix} \gamma(u_1, B) \\ \gamma(u_2, B) \\ \vdots \\ \gamma(u_N, B) \end{bmatrix},$$

em que $\bar{\gamma}(u_i, B) = \frac{1}{|b|} \int_B \gamma(u_i, u)$.

A variância será representada por:

$$\hat{\sigma}^2(u_0) = \mathbf{b}'\lambda - \bar{\gamma}(B, B).$$

Neste sentido, os elementos de b passam a ser as semivariâncias entre as observações e o bloco B . Cada região é representada pelas coordenadas do seu ponto central e as semivariâncias são calculadas considerando cada variável Z_i e o ponto central do bloco.

Capítulo 3

Modelos de semivariograma

Este Capítulo tem por objetivo explicitar as implicações da formação e utilização de modelos de semivariograma a partir de uma função inadequada. Inicialmente serão apresentadas as limitações das funções semivariograma, em seguida será introduzida a família exponencial, suas propriedades e serão analisados os modelos log-normal, Poisson, binomial negativa, a partir dos quais pretende-se exemplificar situações diversas no processo de estimação.

3.1 Limitações de Funções Semivariograma

A escolha de uma função para o covariograma ou semivariograma não pode ser arbitrária. Conforme foi ilustrado na seção 2.6, essas medidas de variabilidade irão compor as equações do sistema de krigagem ordinária para a obtenção dos ponderadores de cada observação no processo de interpolação. Sendo assim, a matriz das covariâncias deve ser definida positiva para garantir a unicidade na solução do sistema de krigagem ordinária (Isaaks e Srivastava, 1989).

Neste contexto, admitindo a hipótese de estacionariedade de segunda ordem apresentada na seção (2.2), a covariância $C(r)$ deve satisfazer a Equação (3.1), cuja expressão representa a variância dos erros de predição discutido na Seção 2.6.

$$\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(r) \geq 0, \quad (3.1)$$

em que λ_i e λ_j representam os pesos.

De fato, considerando as variâncias das combinações lineares $\sum_{i=1}^n Z(u_i)$ chega-se a Equação (3.1), com $\lambda \in \mathbb{R}$ e $u_i \in \mathbb{R}^d \quad \forall \quad n \in \mathbb{N}$ (Schlather et al., 2012). Então, a restrição de ser definida positiva, imposta a função covariância, é condição necessária e também suficiente (Schlather, 1999). Quando o processo é intrinsecamente estacionário a dependência espacial não pode ser medida por meio da função covariância. Então, alternativamente, a variabilidade espacial é medida pela função semivariograma que deverá satisfazer a condição representada pela expressão (3.2), em que $-\gamma(h)$ é condicionalmente não negativo quando a soma dos pesos é nula:

$$-\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(r) \geq 0. \quad (3.2)$$

Incrementos $Z(r) - Z(0)$ intrinsecamente estacionário tem esperança nula se e somente se a seguinte expressão for satisfeita:

$$\lim_{r \rightarrow \infty} \frac{\gamma(r)}{r^2} = 0.$$

Se o processo é totalmente aleatório, então r^2 deve crescer em um ritmo maior do que o do semivariograma $\gamma(r)$.

3.1.1 Representação Espectral

Teorema 3.1.1 (Teorema de Askey). *Se $C(r)$ é uma função limitada de r , quando $|r| \rightarrow \infty$, é diferenciável para $r > 0$ e tem a propriedade de que o gráfico da função $-C'(r)$, $0 < r < \infty$ é uma curva côncava, então $C(r)$ é uma função de correlação que corresponde a densidade espectral $T(\omega)$, atingindo seu máximo em $\omega = 0$ e é monotonamente decrescente para valores positivos e negativos de ω .*

Segundo o teorema de Bochner (Schlather, 1999; Yaglom, 1987), uma função $C : \mathbb{R} \rightarrow \mathbb{C}$ é contínua e definida positiva se puder ser escrita como uma Transformada de Fourier de uma medida finita e não negativa. A transformada correspondente a covariância $C(r)$ é chamada de função de densidade espectral, aqui denotada por $T(\omega)$. Segundo Yaglom (1962), se a função de covariância é conhecida, então a densidade espectral pode ser obtida por meio da Equação (3.3):

$$T(\omega) = \frac{1}{2\pi} \int_{R^n} \exp(-i\omega r) C(r) dr, \quad (3.3)$$

em que $i = \sqrt{-1}$.

$$C(r) = \frac{1}{2\pi} \int_{R^n} \exp(i\omega r) T(\omega) d\omega. \quad (3.4)$$

Considerando a simetria da função covariância a Equação (3.3) pode ser reescrita como:

$$T(\omega) = \frac{1}{2\pi} \int_{R^n} \cos(\omega r) C(r) dr.$$

Equivalentemente, pode-se calcular uma Transformada de Fourier a partir da função de autocorrelação, $\rho = \frac{C(r)}{C(0)}$, que dará origem a uma função espectral normalizada:

$$t(\omega) = \frac{1}{2\pi} \int_{R^n} \cos(\omega r) \rho(r) dr.$$

Neste caso, $t(r)$ é a Transformada de Fourier de $\rho(r)$ e vice versa (Webster e Oliver, 2007).

$$\rho(r) = \frac{1}{2\pi} \int_{R^n} \cos(\omega r) t(r) dr.$$

Sendo assim, as funções $T(\omega)$ e $t(\omega)$ são medidas finitas e não negativas. Uma observação importante é que a transformada de Fourier de uma função isotrópica preserva essa propriedade.

3.1.2 Gerando modelos de variogramas e covariâncias a partir de uma função de probabilidade

Nesta seção foram apresentadas condições para que uma função represente adequadamente um modelo de variograma ou covariância. A abordagem deste tópico está voltada para a construção de modelos teóricos de covariância a partir de uma distribuição de probabilidade. A ideia é fazer o processo inverso. Ao invés de testar a viabilidade de uma função como modelo de variograma o que se propõe é compor um variograma válido utilizando a função de correlação.

Schoenberg (1938) discute a construção de funções isotrópicas definida positiva.

Teorema 3.1.2 (Schoenberg 1938). *Uma função $C : \mathbb{R} \rightarrow \mathbb{C}$, é real, contínua, isotrópica e definida positiva se puder ser escrita conforme a Equação (3.5):*

$$C(r) = \varphi(r) = \int_{[0, \infty)} \Gamma\left(\frac{n}{2}\right) \left(\frac{2}{r\omega}\right)^p J_p(\omega r) dF(\omega), \quad (3.5)$$

em que $p = (n - 2)/2$, J_p é a função Bessel de primeiro tipo e ordem p e F é uma medida não decrescente limitada por $\omega \geq 0$ chamada de função de distribuição espectral.

Segundo Christakos (1984), é mais fácil construir uma função densidade de probabilidade do que uma covariância não negativa definida. O autor sugere que o resultado acima seja utilizado para gerar funções de correlação a partir de distribuições de probabilidade. Seja $\rho = C_n(r)/C(0)$ a função correlação sob a hipótese de estacionariedade e fazendo a divisão da Equação (3.5) por $C(0)$ em ambos os lados da igualdade, tem-se que o quociente $F(\omega)/C(0)$ é a função espectral e pode ser vista como a função $P_m(m)$ de um vetor aleatório n -dimensional, isto é $P_m(m) = F(\omega)/C(0)$ e isotrópica. Christakos (1984) sugere três passos para gerar funções covariância e variograma válidas:

1. Construir uma função de densidade de probabilidade P_m para um vetor aleatório m ;
2. Substituir P_m em (3.5) para obter a covariância correspondente $C_n(r)$;
3. Substitui $C_n(r)$ em (2.4) para estimar o variograma correspondente $\gamma(r)$.

A título de ilustração o autor supõe que exista um vetor aleatório m distribuído uniformemente sobre uma superfície esférica n -dimensional de raio v . Então a função de densidade de probabilidade de m será:

$$P_m(m = \omega) = \frac{\Gamma(n/2)}{2\pi^{n/2}\omega^{n-1}}\delta(\omega - v).$$

A função espectral $T_n(\omega) = P_m(\omega)C(0)$ é não negativa para todo ω . Substituindo em (3.5) e em seguida usando a Equação (2.4) obtém-se o seguinte modelo de variograma:

$$\gamma_n(r) = C(0)[1 - 2^p\Gamma(n/2)(vr)^{-p}J_p(vr)] \quad (3.6)$$

Considerando $n = 1$, o modelo (3.6) resulta em $\gamma_1(r) = C(0)[1 - \cos(vr)]$, para $n = 2$ tem-se $\gamma_2(r) = C(0)[1 - J_0(vr)]$ e para $n = 3$ $\gamma_3(r) = C(0)[1 - [\sin(vr)/vr]]$

Seguindo o mesmo procedimento, assumindo a P_m como sendo da forma exponencial, o variograma correspondente será:

$$\gamma(r) = C(0) \left[1 - \left(1 + \frac{r^2}{a^2} \right)^{-(n+1)/2} \right] \quad (3.7)$$

Fazendo $P_v = \left(1 + \frac{\omega^2}{a^2} \right)^{-l}$ com $l > n/2$. Então o variograma correspondente é dado por:

$$\gamma(r) = m[1 - mr^v K_v(r/a)]$$

em que $v = (1 - n)/2$ e m é uma constante.

Pensando na dimensão 3 por exemplo, chega-se a seguinte função, que corresponde ao modelo esférico de semivariograma:

$$\gamma_3(r) = \begin{cases} C(0) \left\{ \frac{3}{2}cr - \frac{1}{2}c^3r^3 \right\} & , \quad r \leq 1/c \\ C(0) & , \quad r > 1/c \end{cases} \quad (3.8)$$

As restrições impostas a r são necessárias para que o modelo (3.8) seja positivo. Substituindo (3.9) em (2.4) obtém-se o variograma representado por (3.8). Estes resultados sugerem que a partir da função correlação pode-se gerar mais facilmente modelos de semivariogramas válidos, utilizando a relação representada pela Equação (2.5). Por exemplo, segundo Diggle e Ribeiro Jr. (2007), a família esférica tem função correlação da forma:

$$\rho(r) = \begin{cases} 1 - \frac{3}{2} \left(\frac{r}{a} \right) + \frac{1}{2} \left(\frac{r}{a} \right)^3 & , \quad 0 \leq r \leq a \\ 0 & , \quad r > a \end{cases} \quad (3.9)$$

em que $\phi > 0$ e $a = 1/c$.

Utilizando (2.5) o modelo esférico de variograma correspondente é encontrado. Da mesma forma, sabendo que a família de funções covariância exponencial potência é definida pela função de correlação dada por:

$$\rho(u) = \exp \left\{ - \left(\frac{u}{\phi} \right)^q \right\} \quad (3.10)$$

em que $\phi > 0$ é um parâmetro de escala e q um parâmetro de forma, em que $0 < q \leq 2$, também é possível chegar a um modelo de semivariograma válido por meio da relação apresentada em (2.5), desde que o processo seja estacionário. Fazendo $q = 1$, por exemplo, chega-se ao modelo exponencial de semivariograma.

Segundo Yaglom (1987) a classe de funções correlação normalizada coincide com as funções características das distribuições de probabilidade. Além disso, as funções $\phi = \exp\{-(\vartheta u)^q\}$ são funções características de distribuições estáveis e simétricas e as transformadas de Fourier desta expressão dão as densidades espectrais correspondentes para as funções de correlação e também as densidades de distribuições de probabilidade estável e simétrica. A partir deste raciocínio a seção seguinte irá abordar uma tentativa de construção de funções correlação a partir da família exponencial.

3.2 A família exponencial

Uma família de funções de probabilidade é chamada família exponencial se pode ser escrita da forma:

$$f(y|\theta) = h(y)c(\theta) \exp \left\{ \sum_{i=1}^l \varpi_i(\theta) d_i(y) \right\}, i = 1, \dots, l. \quad (3.11)$$

Os componentes são tais que $h(y) \geq 0$, $d_i(y)$ não depende de θ , $c(\theta) \geq 0$ e $\varpi_1(\theta), \dots, \varpi_l(\theta)$ são funções dos valores possíveis para o vetor de parâmetros θ (Casella e Berger, 2001).

Várias distribuições conhecidas podem ser escritas na forma (3.11), tais como: Poisson, binomial, binomial negativa, normal, gamma, log-normal entre outras. O conceito trabalhado aqui é voltado para a família exponencial uniparamétrica, o que requer a suposição de que um dos parâmetros das distribuições biparamétricas seja conhecido. Para reconhecer se uma família de funções de densidade ou de probabilidade pertence a família exponencial é necessário identificar se o suporte da distribuição independe do parâmetro e verificar se a família pode ser escrita na forma (3.11), conforme será visto nos exemplos que seguem.

Supondo que y segue uma distribuição normal com média μ e variância σ^2 , então

sua função densidade de probabilidade é dada por:

$$f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \text{ para } -\infty < y < \infty \quad (3.12)$$

e pode ser reescrita sob a forma da família exponencial como:

$$f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(\frac{y^2}{\sigma^2} + \frac{\mu y}{\sigma^2}\right), \quad (3.13)$$

em que

$$\begin{aligned} h(y) &= 1 \quad \forall \quad y; \\ c(\theta) = c(\mu, \sigma) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \text{ para } -\infty < \mu < \infty, \quad \sigma > 0; \\ \varpi_1(\mu, \sigma) &= \frac{1}{\sigma^2}, \quad \sigma > 0; \\ \varpi_2(\mu, \sigma) &= \frac{\mu}{\sigma^2}, \quad \sigma > 0; \\ d_1(y) &= -\frac{y^2}{2}; \\ d_2(y) &= y. \end{aligned}$$

Agora supondo que y tem distribuição de Poisson com média μ , sua função de probabilidade é da forma:

$$P(y = s|\mu) = \frac{\mu^y}{y!} \exp\{-\mu\} \quad (3.14)$$

e pode ser escrita sob a forma da família exponencial como:

$$\begin{aligned} P(y = s|\mu) &= \exp\{y \ln \mu - \mu - \ln y!\} \\ &= \frac{1}{y!} \exp\{\mu\} \exp\{y \ln(\mu)\}. \end{aligned}$$

em que

$$\begin{aligned}h(y) &= \frac{1}{y!}; \\c(\mu) &= \exp\{-\mu\}; \\ \varpi(\mu) &= \ln(\mu); \\d(y) &= y.\end{aligned}$$

Se y segue uma distribuição binomial negativa com parâmetros μ e j , sua função de probabilidade é da forma:

$$P(y = s|\mu, j) = \binom{j + y - 1}{y} \left(\frac{j}{\mu + j}\right)^j \left(\frac{\mu}{\mu + j}\right)^y$$

que pode ser representada pela seguinte expressão:

$$P(y = s|\mu, j) = \binom{j + y - 1}{y} \left(\frac{j}{\mu + j}\right)^j \exp\left\{y \ln\left(\frac{\mu}{\mu + j}\right)\right\}$$

em que

$$\begin{aligned}h(y) &= \binom{j + y - 1}{y}; \\c(\mu) &= \left(\frac{j}{\mu + j}\right)^j; \\ \varpi(\mu) &= \ln\left(\frac{\mu}{\mu + j}\right); \\d(y) &= y.\end{aligned}$$

Supondo que y segue uma distribuição gamma com parâmetros α e β , sua função de probabilidade é da forma:

$$f(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\beta y\}, \quad \alpha > 0, \quad \beta > 0 \text{ e } y \geq 0.$$

Quando escrita sob a forma da família exponencial obtém-se:

$$f(y|\alpha, \beta) = \frac{1}{y} \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\{\alpha \ln(y) - \beta y\},$$

em que

$$\begin{aligned} h(y) &= \frac{1}{y}; \\ c(\alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}; \\ \varpi_1(\theta) &= \alpha; \\ \varpi_2(\theta) &= \beta; \\ d_1(y) &= \ln(y); \\ d_2(y) &= -y. \end{aligned}$$

Se $\alpha = \beta$, então:

$$f(y|\alpha, \alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\alpha y\}. \quad (3.15)$$

Supondo que y segue uma distribuição log-normal com parâmetros μ e σ^2 , sua função de probabilidade é da forma:

$$f(y|\mu, \sigma^2) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\ln(y) - \mu)^2\right\} \text{ para } y > 0, \quad (3.16)$$

em que $E(y) = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}$.

Reescrevendo a densidade log-normal sob a forma da família exponencial obtém-se:

$$f(y|\mu, \sigma^2) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\sigma^2}(\ln(y)^2 - 2\mu\ln(y))\right\},$$

$$\begin{aligned}
h(y) &= \frac{1}{y}; \\
c(\mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\}; \\
\varpi_1(\theta) &= \frac{1}{\sigma^2}; \\
\varpi_2(\theta) &= \frac{\mu}{\sigma^2}; \\
d_1(y) &= -\frac{\ln(y)^2}{2}; \\
d_2(y) &= \ln(y).
\end{aligned}$$

Mais referências e maiores detalhes sobre as propriedades da família exponencial podem ser encontradas em Casella e Berger (2001).

3.3 A família exponencial como modelo de semi-variograma

No Capítulo 2 foram apresentadas as condições para que uma função possa ser escrita como um modelo teórico de covariância ou semivariograma. A ideia dessa seção é utilizar os recursos apresentados no Capítulo 2 para mostrar alguns cuidados importantes para a construção de modelos de semivariograma. Entendendo as funções de probabilidade ($f.p$) como medidas positivas, Christakos (1984) utilizou este fato como argumento para construir famílias de modelos isotrópicos de semivariogramas, se amparando nas propriedades que definem uma $f.p$. A metodologia apresentada na seção 3.1 contribuiu para, no mesmo contexto, pensar em construir outros modelos de variograma, mas com aborgagem diferenciada. Esta ilustração será feita de maneira informal e por meio de exemplos, apresentando ideias introdutórias sobre uma possível representação de semivariogramas definidos a partir de distribuições de probabilidade escritas sob a forma da família exponencial. O desenvolvimento do estudo se dará por meio das funções binomial negativa, log-normal, gamma e Poisson, as quais foram reescritas sob a forma da família exponencial. Em seguida será realizada uma discussão pautada no comportamento de tais funções, fazendo uma conexão com as restrições

necessárias para que uma expressão matemática possa representar adequadamente um semivariograma.

Como passo inicial será tomada como exemplo a função de densidade de probabilidade normal, representada pela Equação (3.16), bem como sua forma quando escrita sob a família exponencial, Equação (3.2). Conforme descrito no Capítulo 2, expressão (3.2), correspondente ao semivariograma gaussiano, representa um modelo isotrópico válido em $\mathbb{R}^d \forall d$. Segundo DeGroot (1989), a Equação (3.16) descreve uma função definida não negativa, característica essa que corrobora com a definição de modelos válidos para representar um semivariograma. Sendo assim, no tópico no seguir será feita uma explanação de ideias concatenadas visando obter como resultado o modelo gaussiano de semivariograma, tendo como ponto de partida a distribuição normal. Desenvolvido este raciocínio, nos tópicos seguintes, outras distribuições serão utilizadas como meio factível para a tentativa de criação de novos modelos de semivariograma.

3.3.1 Distribuição Normal

Conforme abordado no Capítulo anterior, uma função é considerada válida para representar um semivariograma se for definida positiva. Uma função pode ser assim classificada se puder ser escrita como uma transformada de Fourier. No contexto de função densidade de probabilidade, a transformada de Fourier é justamente a função característica (Feuerverger e Mureika, 1977). Além disto, segundo o Teorema de Bochner - Khintchine (Magalhães, 2011) uma função contínua $\psi : \mathbb{R} \rightarrow \mathbb{C}$ com $\psi(0) = 1$ é uma função característica se e somente se, for definida positiva.

A função de densidade normal é definida positiva e pode ser escrita como uma transformada de Fourier. Sua função característica é representada por:

$$\phi_y(t) = e^{\left(it\mu - \sigma^2 \frac{t^2}{2}\right)}. \quad (3.17)$$

Sendo assim, é plausível pensar na própria densidade como função de semivariograma. Examinando a positividade, conforme demonstrado em DeGroot (1989), observa-se que a expressão resultante é semelhante ao semivariograma gaussiano a

menos de uma constante:

$$\int_{-\infty}^{\infty} f(y|\mu, \sigma^2) dy = 1. \quad (3.18)$$

Então fazendo $v = y - \mu$ e substituindo em (3.16), obtém-se:

$$\int_{-\infty}^{\infty} f(y|\mu, \sigma^2) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv \quad (3.19)$$

Em seguida fazendo:

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv, \quad (3.20)$$

Então $I = \sigma\sqrt{2\pi}$.

Desta forma, chega-se ao seguinte resultado:

$$\begin{aligned} I^2 = I \times I &= \int_{-\infty}^{\infty} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv \int_{-\infty}^{\infty} \exp\left(-\frac{w^2}{2\sigma^2}\right) dw \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(v^2 + w^2)\right] dw dv \end{aligned}$$

Fazendo transformação para coordenadas polares, em que $v = r \cos \theta$, $w = r \sin \theta$ e $r^2 = v^2 + w^2$, obtém-se:

$$\begin{aligned} I^2 &= \int_0^{2\pi} \int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr d\theta \\ &= 2\pi \int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr. \end{aligned} \quad (3.21)$$

Confrontando a Equação (3.21) com o semivariograma gaussiano, constata-se que ambos diferem-se apenas pela constante 2π . Observa que fazendo $\sigma^2 = \frac{1}{2\pi}$, então o exponencial ficará da forma $\exp(-\pi r^2)$, que corresponde, justamente, a densidade distribuição isotrópica gaussiana multidimensional com variância igual a $\frac{1}{2\pi}$ (Chilès e Delfiner, 1999). Neste sentido, tendo em vista que a função correlação dada pela Equação (3.10), a própria densidade gaussiana, a menos da constante, representa uma função correlação válida para modelar um semivariograma, correspondente a Equação (3.10), quando $q = 2$. Então substituindo em (2.5) obtém-se o semivariograma gaussiano.

A ideia apresentada aqui motivou o interesse em discutir a possibilidade de criação de modelos de semivariograma fundamentados em distribuições de probabilidade es-

critas sob a forma da família exponencial, utilizando-as como função correlação. A seguir será apresentado um exemplo utilizando a densidade exponencial.

3.3.2 Distribuição Exponencial

Da mesma forma que a densidade normal, a função exponencial possui função característica e é definida positiva em seu domínio. A sua densidade é da forma:

$$f(y) = \frac{1}{\mu} \exp\left(-\frac{y}{\mu}\right)$$

Esta função, exceto pela constante $\frac{1}{\mu}$, corresponde a função correlação potência quando $q = 1$ (Equação (3.10)). A proposta aqui é mostrar que a densidade exponencial escrita sob a forma da família exponencial também pode ser utilizada como função correlação e, conseqüentemente, dar origem a um semivariograma válido. A densidade exponencial reescrita sob a forma da família exponencial fica da forma:

$$f(y|\mu) = \exp\left(\ln\left(\frac{1}{\mu}\right) - \mu y\right).$$

Substituindo esta expressão na Equação (2.5), tal como no caso da normal, obtém-se uma expressão correspondente ao semivariograma exponencial, a menos de uma constante, neste caso $\ln\left(\frac{1}{\mu}\right)$:

$$\gamma(r) = C(0) + C_1 \left[1 - \exp\left(\ln\left(\frac{1}{\mu}\right) - \mu y\right)\right].$$

Desenvolvendo um raciocínio sob essa ótica, a família exponencial pode ser vista como uma família de funções com potencial para representar adequadamente uma função correlação para formular um semivariograma. Para ilustrar a metodologia e discutir as implicações de estender essa aplicação para a família exponencial, serão estudadas a seguir as distribuições binomial negativa, Poisson, gamma e log-normal.

3.3.3 Distribuição binomial negativa

A proposta de utilização da função binomial negativa num cenário em que as variáveis mensuradas devem ser contínuas está fundamentada na ideia de que os

eventos, ainda que discretos, podem ocorrer em qualquer ponto do espaço e se observados ao longo do tempo podem descrever um fenômeno pseudo-contínuo. Para a construção deste modelo de semivariograma a função binomial negativa precisou ser reparametrizada em termos de μ . A Equação (3.22) representa a forma final do modelo, que foi montado seguindo os mesmos passos do modelo apresentado na seção anterior. Estudando o comportamento desta função, foi possível observar que a curva gerada é semelhante à do semivariograma exponencial, conforme pode ser visto na Figura 3.5. Neste caso não houve a necessidade de estabelecer restrições por se tratar de uma função crescente e que resulta valores sempre positivos. Sendo assim, a Expressão (3.22) é um modelo de semivariograma válido:

$$\gamma(r) = C(0) + C_1 \left[1 - \exp \left\{ r \ln \left(\frac{a}{a+1} \right) \right\} \right]. \quad (3.22)$$

Embora o procedimento tenha sido bem sucedido nesta aplicação, uma observação importante a se fazer é que existe uma semelhança da curva gerada com a modelo de semivariograma exponencial. Esta função é pode ser vista como uma reparametrização do semivariograma exponencial com alcance igual a $-\frac{1}{\ln(\frac{a}{a+1})}$, quando a tende para infinito. A seguir mais uma tentativa será realizada, utilizando agora, o modelo log-normal.

3.3.4 Distribuição log-normal

Utilizando o raciocínio descrito nas seções anteriores, neste tópico serão apresentados alguns cuidados importantes na tentativa de construir um modelo de semivariograma válido a partir da função log-normal reescrita sob a forma da família exponencial, considerando $\mu = 0$.

O primeiro passo dessa construção foi observar a positividade da função a fim de identificar se a mesma pode representar uma correlação válida para originar um semivariograma. O gráfico do semivariograma proposto a partir da distribuição log-normal sob a forma da família exponencial é sempre positivo para qualquer distância maior do que zero. Para $a > 1$ a curva apresenta a forma típica de uma função correlação, já quando $a < 1$ a curva apresenta intervalos com mudança de comportamento, variando entre crescente e decrescente, sendo que este último estado predomina até a

estabilização da função.

Ao aplicar Equação (2.5) começam a surgir resultados negativos (Figura 3.1 (a)). A partir deste cenário, fica nítida a necessidade de impor restrições a fim de tornar a função positiva e desejável para modelar um semivariograma. Ao passo que as restrições foram sendo adicionadas, o modelo tornou-se cada vez mais limitado. A Equação (3.23) define um modelo que atende somente grandes distâncias, com $a \geq 1$, pois o comportamento decrescente para $0 < a < 1$ desaparece (Figura 3.1 (b)). Sendo assim, surgiu a necessidade de utilizar uma parametrização diferenciada de acordo com o a magnitude do alcance conforme mostra a Equação (3.24). Desta forma, o modelo passou a contemplar as pequenas distâncias ($a < 1$), entretanto, também ficou limitado, pois não consegue trabalhar adequadamente dados alcances superiores a 1 unidade de medida.

Para representar adequadamente um semivariograma, além de ser definida positiva a função deve ser, também, não decrescente. Mais especificamente, o exponencial de um valor negativo sempre retornará um valor entre 0 e 1, além de representar uma função decrescente, condizente com o comportamento da função correlação. Então, subtraindo este exponencial de 1, consegue-se reunir duas condições que tornam a função representativa para um semivariograma. Entretanto, para a densidade log-normal, esta tentativa não correspondeu às expectativas, pois mesmo após reparametrizar a função e inserir restrições não foi possível eliminar o comportamento constante próximo da origem. A mudança de comportamento no intervalo $a \in [0; 1)$ pode estar relacionada com o fato de ser justamente o intervalo em que o logaritmo assume valores negativos.

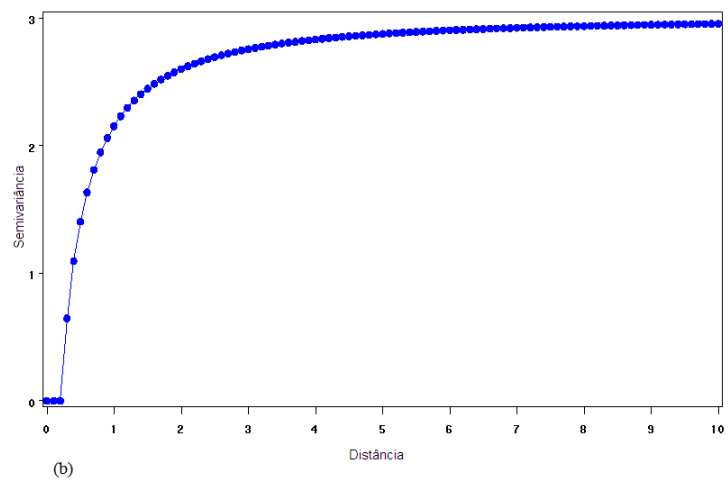
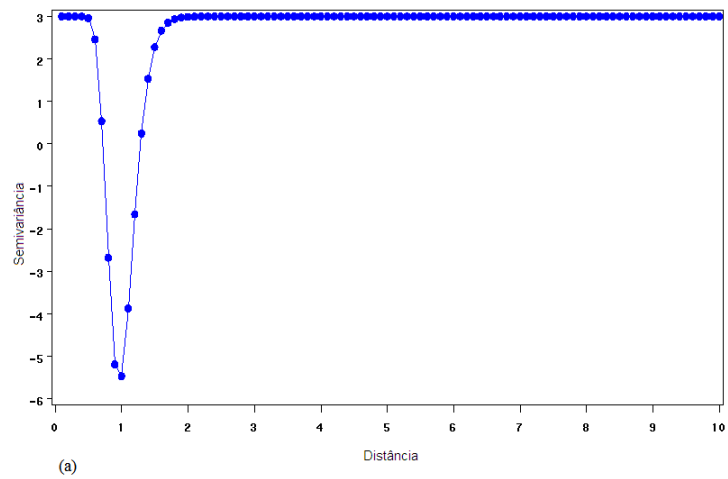


Figura 3.1: Comportamento do semivariograma.

- Distâncias moderadas $a > 1$:

$$\gamma(r) = \begin{cases} 0 \\ C_0 + C_1 \left[1 - \exp \left\{ -\frac{1}{2a^2} \left(\frac{\ln(r)^2}{a^2} + (a^2) \ln(\pi r^2 a^2) \right) \right\} \right] \\ C_0 \end{cases} \quad \begin{aligned} & , \text{ se } r = 0; \\ & , \text{ se } \frac{(\ln(r))^2}{a^2} > -\ln(\pi r^2 a^2) \text{ e } (\ln(r) \geq 0 \text{ ou } (\ln(r)/a)^2 \leq a); \\ & , \text{ caso contrário.} \end{aligned} \quad (3.23)$$

- Pequenas distâncias $a < 1$:

$$\gamma(r) = \begin{cases} 0 \\ C_0 + C_1 \left[1 - \exp \left\{ -\frac{1}{2} \left[(\ln(r))^2 + \ln \left(\frac{\pi r^2}{a^2} \right) \right] \right\} \right] \\ C_0 \end{cases} \quad \begin{aligned} & , \text{ se } r = 0; \\ & , \text{ se } (a \ln(r))^2 > -\ln \left(\frac{\pi r^2}{a^2} \right) \text{ e } (\ln(r) \geq 0 \text{ ou } (a \ln(r))^2 \leq a); \\ & , \text{ caso contrário.} \end{aligned} \quad (3.24)$$

Nota-se que, na verdade, a expressão tomada como função correlação não é como a Equação (3.10), mas sim da forma $f(g(r)) = \exp\left\{-\left(\frac{g(r)}{\phi}\right)^q\right\}$ para a qual não há qualquer garantia de viabilidade para compor um modelo de semivariograma. Tal como os modelos de semivariograma exponencial e gaussiano apresentados no Capítulo 2, o modelo sugerido aqui também apresenta patamar assintótico e sendo assim, faz-se necessário definir o alcance prático. Se tratando de uma composição de funções, chamando de $g(r)$ a quantidade dentro do expoente, (Equação (3.23)), tem-se que o alcance que corresponde a 95% do patamar será o valor de r que satisfaz $g(r) = 6a^2$. Entender a generalidade de um modelo é importante para definir o tipo de fenômeno que ele é capaz de representar. A utilização de um modelo que é aparentemente conveniente, sem estudar as suas particularidades pode comprometer o processo inferencial. As implicações da utilização de modelos de semivariograma inadequados serão abordadas com mais detalhes no próximo Capítulo. No tópico seguinte mais uma construção será apresentada. Desta vez utilizando a densidade gamma.

3.3.5 Distribuição gamma

A função densidade de probabilidade gamma, na qual a tentativa de construção do próximo modelo de semivariograma está fundamentada, é bastante flexível e pode assumir diversas formas a depender da escolha de seus parâmetros. Sendo assim, vários modelos importantes são derivados desta função. Leemis e Mcqueston (2008) mostram, por meio de um diagrama, que o modelo gamma pode ser visto como uma família de distribuições, uma vez que está associado a várias funções. Como o interesse aqui é trabalhar apenas com funções uniparamétricas o modelo foi desenvolvido considerando $\alpha = \beta$. Reescrevendo a função gamma sob a forma da família exponencial, o modelo teórico de semivariograma resultante pode ser representado pelas Expressões (3.25) e (3.26). As restrições impostas à função foram necessárias para garantir resultados sempre positivos. Tal como a densidade log-normal, a densidade gamma, quando escrita em termos da família exponencial também apresentou comportamento desfavorável, não atendendo a todas as distâncias. A primeira parametrização, Equação (3.25), consegue modelar satisfatoriamente apenas peque-

nas distâncias. Já com a segunda parametrização, Equação (3.26), a função passa a ter um comportamento oposto, assumindo forma irregular para pequenos alcances (Figura 3.2).

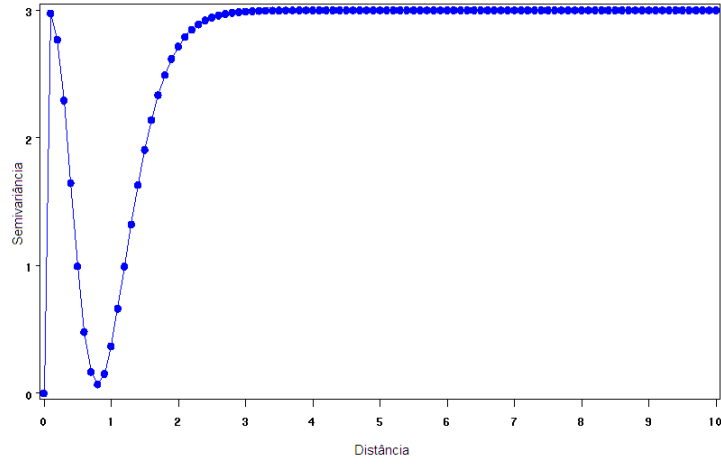


Figura 3.2: Comportamento do semivariograma para $a < 1$.

Com a finalidade de eliminar a mudança de comportamento da curva restrições foram adicionadas ao modelo. Entretanto, as modificações não foram suficientes para evitar que o gráfico fique constante em pontos próximos a origem. Quando menor o valor do alcance, maior o intervalo em que a curva permanece constante em zero, conforme ilustra a Figura (3.3).

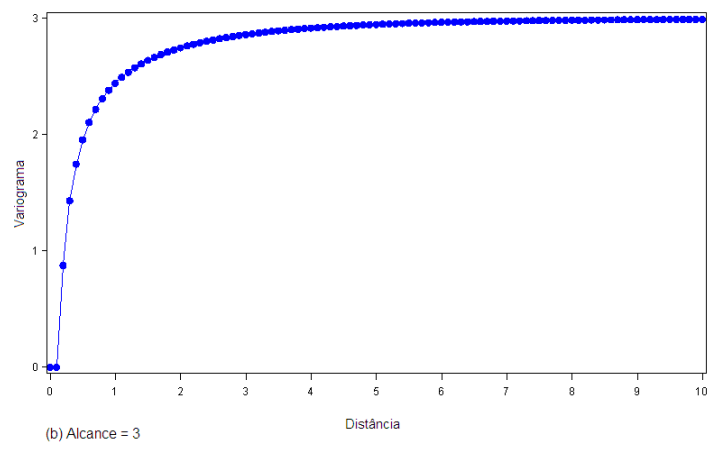
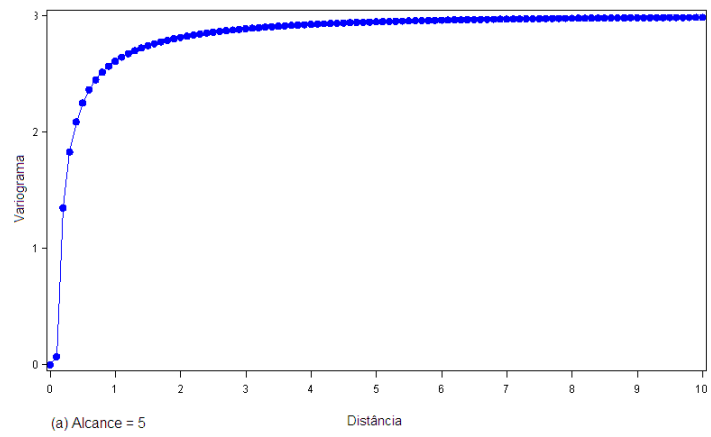


Figura 3.3: Comportamento do semivariograma para diferentes alcances.

1. Pequenas distâncias $a < 1$:

$$\gamma(r) = \begin{cases} 0 & , \text{ se } r = 0; \\ C_0 + C_1 [1 - \exp\{aln(r) - ar - \ln(r) + (a)\ln(a) - \ln(\Gamma(a))\}] & , \text{ se } [\ln(r) - r + \ln(a)]a < \ln(r) + \ln(\Gamma(a)); \\ C_0 & , \text{ caso contrário} \end{cases} \quad (3.25)$$

2. Distâncias moderadas $a \geq 1$:

$$\gamma(r) = \begin{cases} 0 & , \text{ se } r = 0; \\ C_0 + C_1 [1 - \exp\{\frac{1}{a}\ln(r) - \frac{1}{a}r - \ln(r) + \frac{1}{a}\ln(\frac{1}{a}) - \ln(\Gamma(\frac{1}{a}))\}] & , \text{ se } [\ln(r) - r + \ln(\frac{1}{a})]/a < \ln(r) + \ln(\Gamma(\frac{1}{a})); \\ C_0 & , \text{ caso contrário} \end{cases} \quad (3.26)$$

Além de todas as limitações citadas, este modelo apresenta, ainda, uma descon-tinuidade no ponto $r = 0$. O próximo modelo a ser apresentado é baseado na função de probabilidade Poisson, a qual foi desenvolvida para modelar eventos discretos.

3.3.6 Distribuição de Poisson

A função Poisson de probabilidade foi desenvolvida para modelar eventos discre-tos entretanto, num contexto em que sucessivas ocorrências são medidas ao longo do tempo num espaço contínuo, tais eventos podem ser vistos como pseudo-contínuos, uma vez que podem ser medidos em qualquer posição no espaço. Se um a ocorrência de determinado evento segue uma distribuição de Poisson, então sua função de pro-babilidade é dada pela Equação (3.14), que pode ser reescrita sob a forma da família exponencial. Trazendo para o contexto de semivariograma, em que a distância passa a ser a variável medida, experimentalmente o termo fatorial foi adaptado, sendo sub-stituído pela função $\Gamma(\cdot)$, visto que a distância é contínua. A conveniência desta modificação pode ser justificada pelo fato de que para um número s inteiro, $\Gamma(s + 1)$ equivale a $s!$. Com esta medida, além de garantir o cálculo do fatorial a função $\Gamma(\cdot)$ trabalhará a parte contínua, evitando assim, a ocorrência de saltos na curva. Para esta construção, o termo constante μ foi subtraído, uma vez que sua influência sobre o gráfico é o deslocamento da curva para baixo. Com a ausência deste termo a curva surge da origem. Desta forma, as Equações resultantes são (3.27) e (3.28).

Tal qual os modelos de semivariograma construídos a partir das distribuições log-normal e gamma, o modelo fundamentado na distribuição de Poisson apresenta im-pedimentos para a sua aplicabilidade. Dependendo da magnitude das distâncias en-volvidas no estudo o modelo somente será viável para uma ou outra parametrização.

1. Pequenas distâncias :

$$\gamma(r) = \left\{ C_0 + C_1[1 - \exp\{r \ln(a) - \ln(\Gamma(r+1))\}] \right\}, \text{ se } r \ln(a) < \ln(\Gamma(r+1)) \quad (3.27)$$

2. Grandes distâncias:

$$\gamma(r) = \left\{ C_0 + C_1[1 - \exp\{r \ln(\frac{1}{a}) - \ln(\Gamma(r+1))\}] \right\}, \text{ se } r \ln(1/a) < \ln(\Gamma(r+1)) \quad (3.28)$$

Este é um modelo com patamar assintótico, cujo alcance prático é dado pelo valor da distância r que satisfaz a expressão:

$$\exp \left[\frac{\ln(\gamma(r+1)) - 3}{r} \right] = a.$$

A Figura (3.4) ilustra o comportamento da curva gerada pelo modelo construído a partir da distribuição de Poisson. A Figura 3.4 (a) apresenta o comportamento da curva próxima a origem sem restrições, em que nota-se a presença de valores negativos para as semivariâncias. A Figura 3.4 (b) mostra o gráfico do modelo com as restrições, ressaltando o comportamento constante próximo à origem após eliminar os valores negativos.

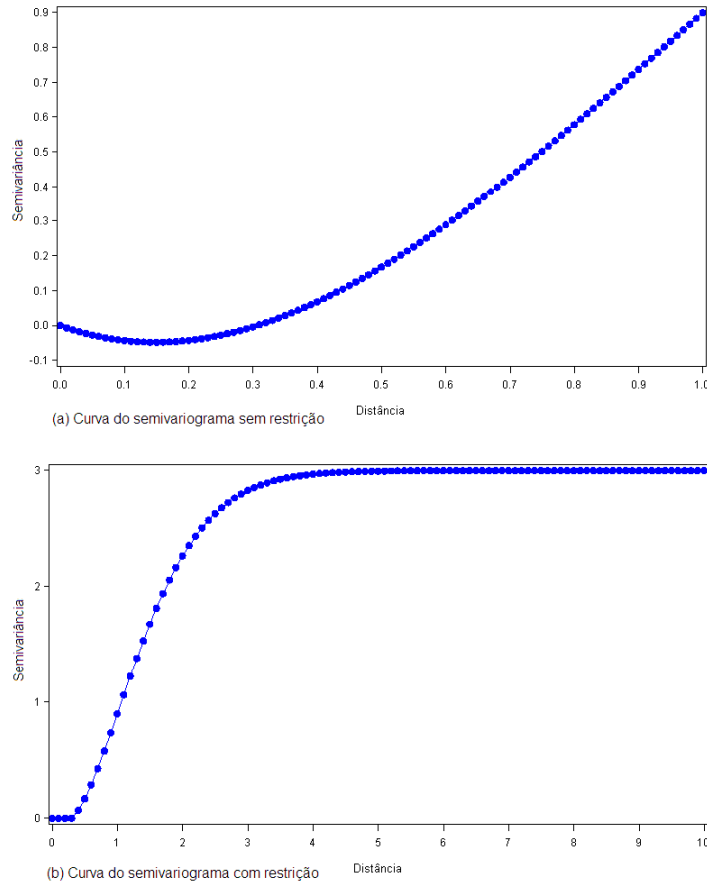


Figura 3.4: Curvas teóricas do modelo de semivariograma Poisson.

As três últimas propostas de construção de modelos de semivariogramas, utilizando funções sob a forma da família exponencial como função correlação falharam,

uma vez que foram necessárias muitas restrições, as quais levaram o modelo a ficar muito limitado em seu domínio. Desta forma, não recomenda-se utilizar quaisquer funções escritas sob a forma da família exponencial como função correlação, pois não há garantias de que o modelo a ser reproduzido será válido.

Os exemplos citados reforçam a ideia de que a ‘forma’ exponencial por si só, não reúne todas as condições necessárias para representar um semivariograma. Observa-se, entretanto, que todas as tentativas apresentadas tem em comum a presença da função logaritmo, o que pode estar produzindo um efeito de desregularização. De fato, a literatura geoestatística (Cressie, 1993; Chilès e Delfiner, 1999; Sen, 2009), faz referência a um modelo de semivariograma chamado semivariograma logaritmo e aponta limitações no domínio desta função. Uma das implicações deste modelo é a variação infinita no ponto $r = 0$. Esta observação é importante, pois conduz ao raciocínio correto de que para um semivariograma ser escrito em função da covariância é preciso checar a existência de $C(0)$.

O modelo logaritmo de semivariograma é representado pela seguinte equação, a qual é baseada na hipótese intrínseca:

$$\gamma(r) = \log(r).$$

A aplicação deste modelo é recomendada em situações específicas quando as variáveis são regularizadas por um suporte de amostragem. Desta forma, a medição não ficará restrita a um ponto (Chilès e Delfiner, 1999).

Segundo Chilès e Delfiner (1999), a família de funções covariância exponencial potência é representada pela seguinte expressão:

$$C(r) = \exp\left(\left(-\frac{r}{a}\right)^q\right),$$

Em que $0 \leq q < 2$. Sendo assim, $C(0)$ existe e é finito. Logo, é possível escrever um semivariograma a partir das Equações (2.4) e (2.5). Já para a composição de funções realizada aqui este resultado não é válido, pois $C(0)$ não está definida. Um outro aspecto a ser observado é o fato de que o valor zero faz parte do domínio das funções que representam os semivariogramas gaussiano e exponencial. O mesmo não

ocorre com as funções propostas. Além disso, o modelo gaussiano de semivariograma equivale a distribuição normal com média zero e ao restringir seu domínio a valores positivos a curva gerada assume uma forma típica da função correlação. Já os modelos apresentados aqui não dispõem dessa característica.

Face a todas as restrições e limitações apresentadas para cada tentativa, não há elementos suficientes para configurar os modelos tentativos como válidos para representar um semivariograma. A Figura 3.5 apresenta as curvas dos modelos apresentados, destacando o comportamento próximo a origem, em análises pontuais, para as distribuições log-normal, gamma e Poisson. Para tornar as curvas comparáveis foram admitidos valores iguais para cada parâmetro em todos os modelos.

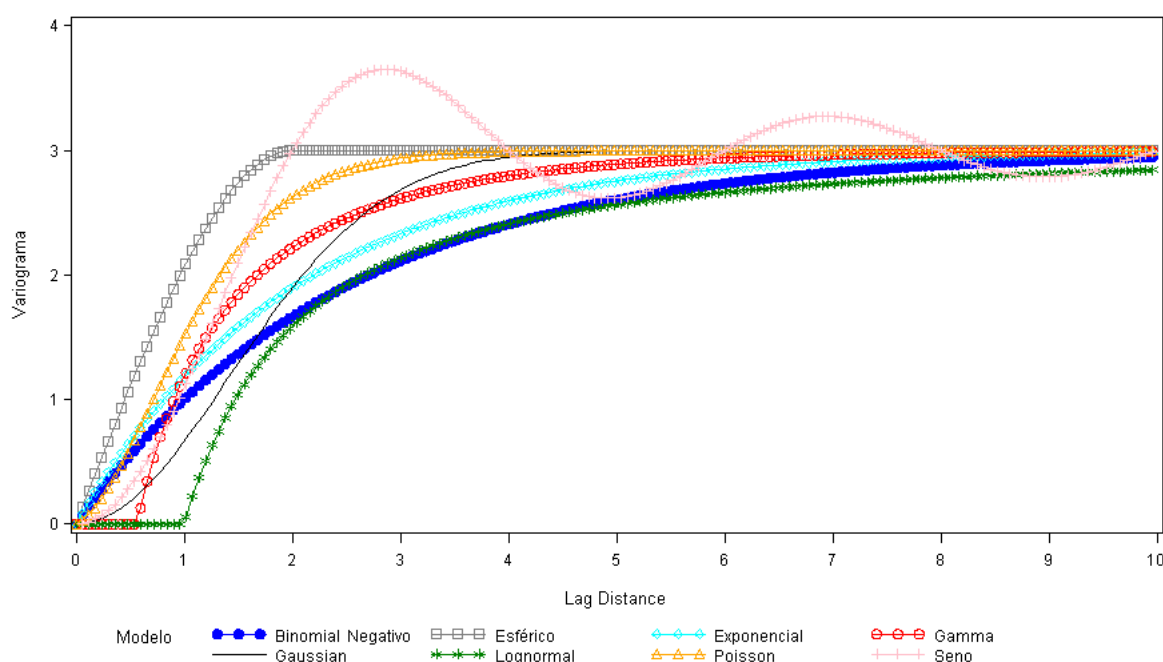


Figura 3.5: Curvas dos modelos clássicos de semivariograma e dos propostos.

A construção dos modelos partindo de premissas equivocadas, resultou em funções inadequadas. No Capítulo seguinte, será feita uma exposição das implicações do uso de tais modelos no sistema de krigagem.

Capítulo 4

Análise dos Resultados

Esse Capítulo apresenta uma análise dos resultados obtidos para os modelos propostos de semivariograma por meio da técnica de krigagem ordinária, ressaltando o efeito dos parâmetros na predição.

4.1 Efeito dos semivariogramas propostos sobre a Krigagem ordinária

Segundo Fischer e Getis (2010) os pesos do sistema de krigagem dependem do semivariograma e são, portanto, sensíveis ao modelo do semivariograma, ao efeito pepita, ao alcance e ao patamar. O conjunto de dados **meuse** (Pebesma, 2004), disponível em R (2010), foi utilizado para ilustrar a krigagem ordinária e avaliar os efeitos dos semivariogramas aqui desenvolvidos sobre processo de predição e a variável analisada foi **ltpb**.

A expectativa nesse tipo de modelagem é que os pontos mais próximos ao alvo tenham maior peso. Entretanto, utilizando o modelo sugerido a partir da distribuição log-normal, verifica-se que os pesos de krigagem são todos iguais e a estimativa produzida nada mais é do que a média dos valores da vizinhança. Ocorre que se as distâncias forem muito grandes o modelo fica constante, uma vez que o logaritmo retornará valores muito parecidos. Isto implica que o modelo não é adequado para trabalhar com grandes distâncias. A Figura (4.1) mostra o mapa de predições de krigagem, utilizando a matriz de semivariâncias obtida por meio do modelo construído

segundo a distribuição log-normal, bem como os erros da estimativas.

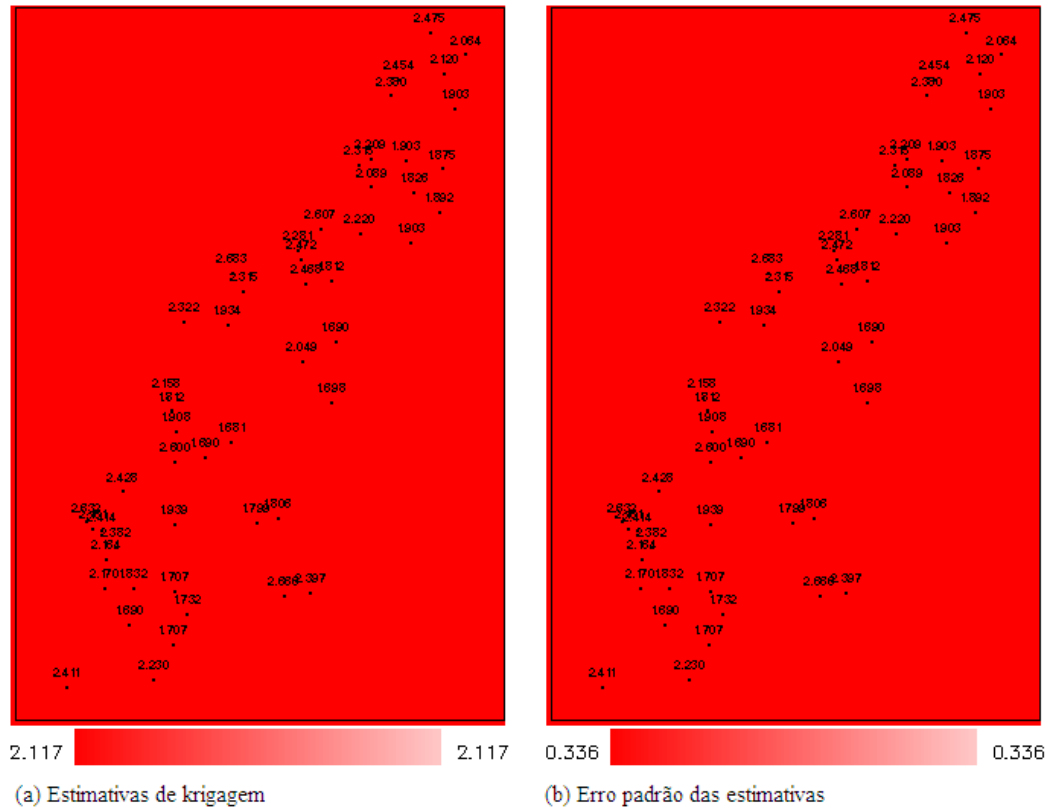


Figura 4.1: Mapa de krigagem com base no semivariograma log-normal com alcance prático.

Para avaliar o efeito do modelo log-normal na krigagem ao trabalhar com valores pequenos para as distâncias, estas foram transformadas de metros para quilômetros, ou seja, as distâncias foram divididas por 1.000. Os mapas de krigagem e os erros das estimativas podem ser visualizados por meio das Figuras 4.2 e 4.3, respectivamente.

Observa-se que o modelo sugerido conseguiu captar a variabilidade do fenômeno após o artifício utilizado para diminuir a magnitude das distâncias (Figuras 4.2 (a) e 4.3 (a)). Entretanto, ao utilizar o alcance prático no processo de predição, o sistema de krigagem ordinária retornou valores negativos para os erros das estimativas. Isto ratifica a inviabilidade do modelo formulado por meio de tentativas *ad hoc*, as quais podem não fornecer solução geral, comprovando que o semivariograma sugerido não é válido para dados de qualquer natureza. De fato, para um semivariograma ser considerado válido, a matriz das semivariâncias originada por essa função deve ser tal que não gere valores negativos para os erros das estimativas.

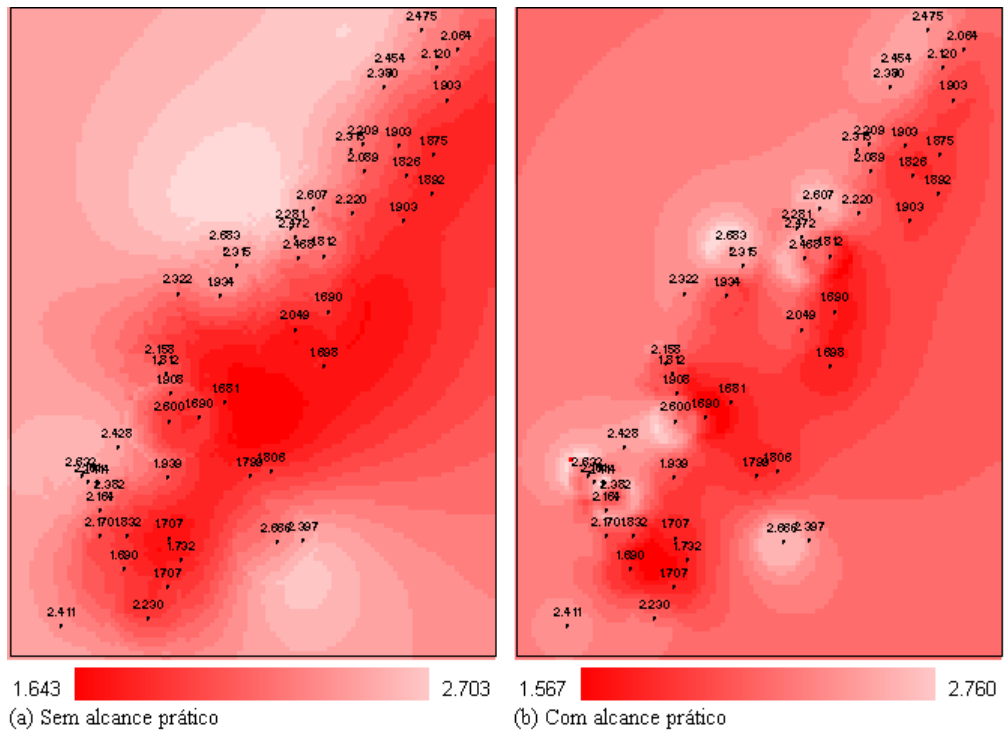


Figura 4.2: Mapa de krigagem com base no semivariograma log-normal (pequenas distâncias).

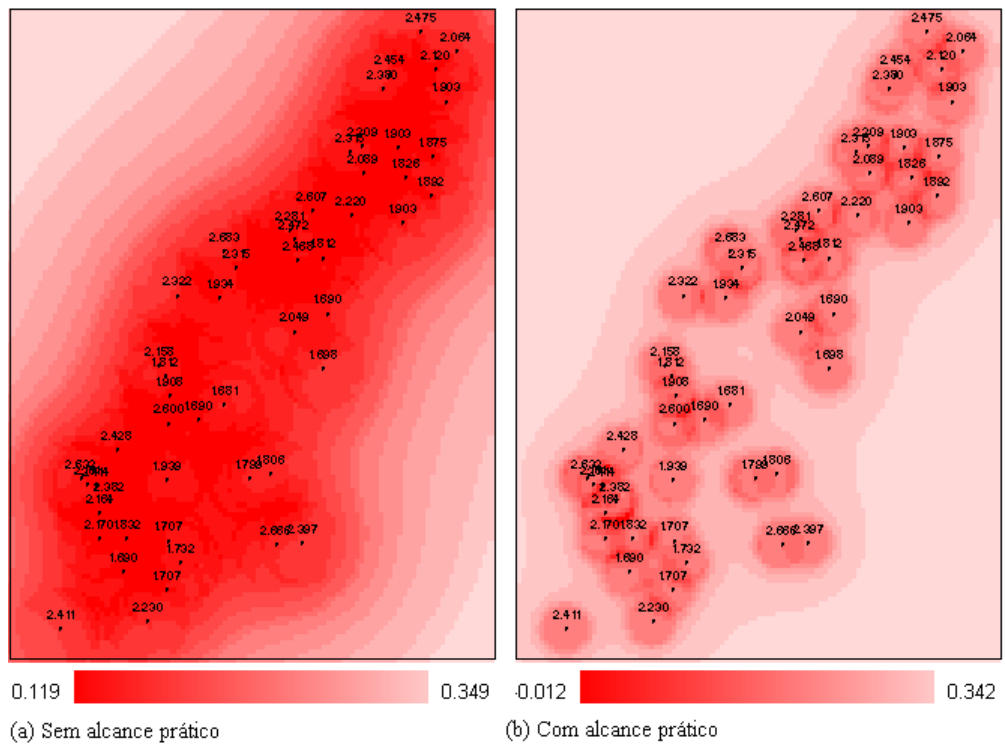


Figura 4.3: Erro padrão das estimativas de krigagem com base no semivariograma log-normal (pequenas distâncias).

Segundo Stein (1988), para obter predições ótimas a função de semivariograma deve estar bem especificada ao longo de distâncias relativamente curtas. Isto é o comportamento do semivariograma próximo a origem precisa ser capturado. Sendo assim, os modelos desenvolvidos não contribuem para estimativas de krigagem eficientes, uma vez que não captam a variabilidade dos dados cujas distâncias estão próximas à origem.

Utilizando a mesma base de dados foram ajustados os modelos gaussiano, esférico e exponencial com a finalidade de comparar os resultados da krigagem. Comparando as Figuras (4.4), (4.6) e (4.7) verifica-se um padrão no comportamento das estimativas, independente do modelo de semivariograma utilizado. O modelo de semivariograma fundamentado na função log-normal para pequenas magnitudes das distâncias (Figura 4.2) conseguiu reproduzir um padrão espacial equivalente ao apresentado quando utilizado os modelos teóricos gaussiano, exponencial e esférico. Observando essa situação específica, o modelo desenvolvido parece satisfatório, exceto pelo fato de produzir erros negativos para as estimativas de krigagem.

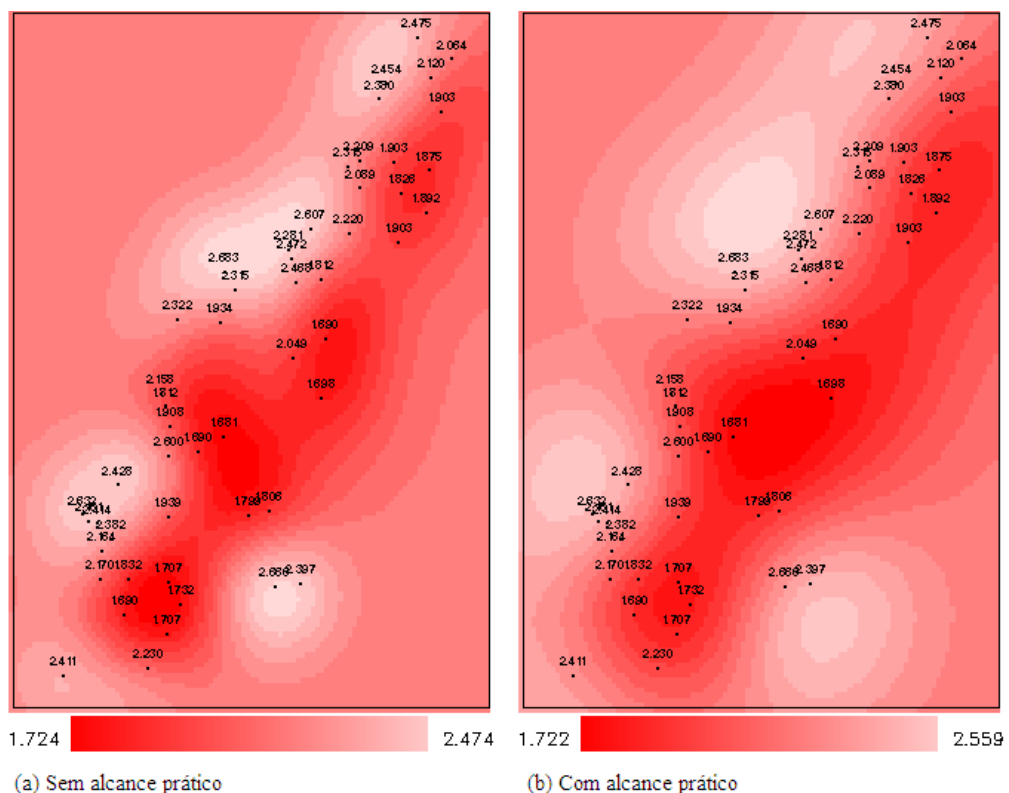


Figura 4.4: Mapa de krigagem com base no semivariograma gaussiano.

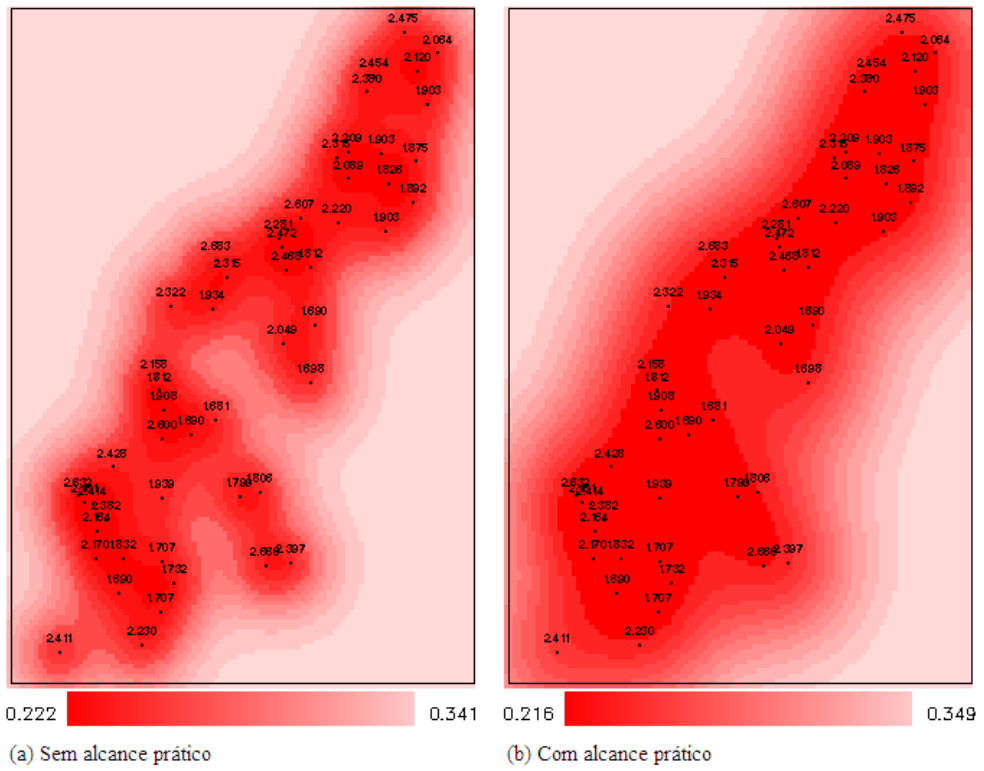


Figura 4.5: Erro padrão das estimativas com base no semivariograma gaussiano.

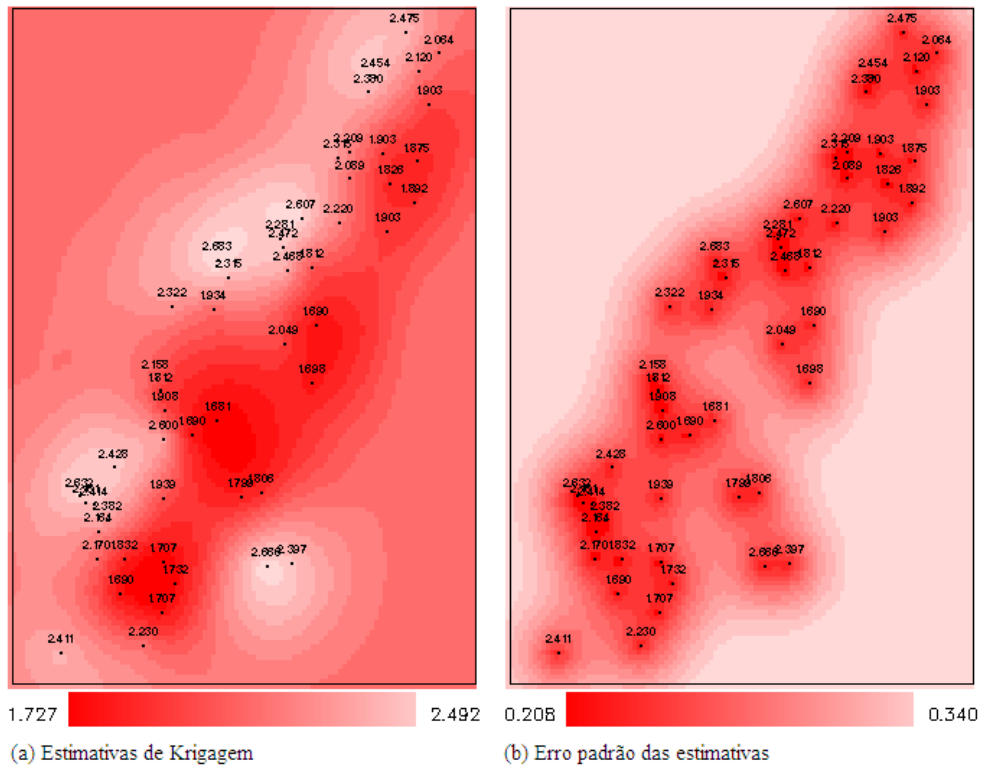


Figura 4.6: Mapa de krigagem com base no semivariograma esférico.

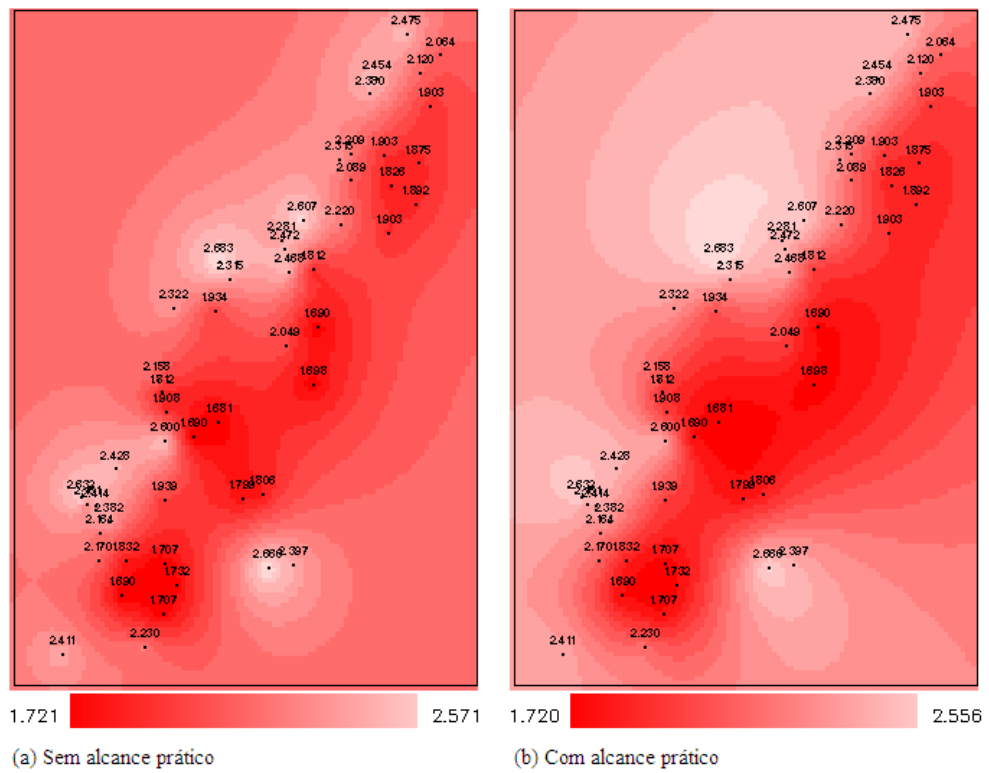


Figura 4.7: Mapa de krigagem com base no semivariograma exponencial.

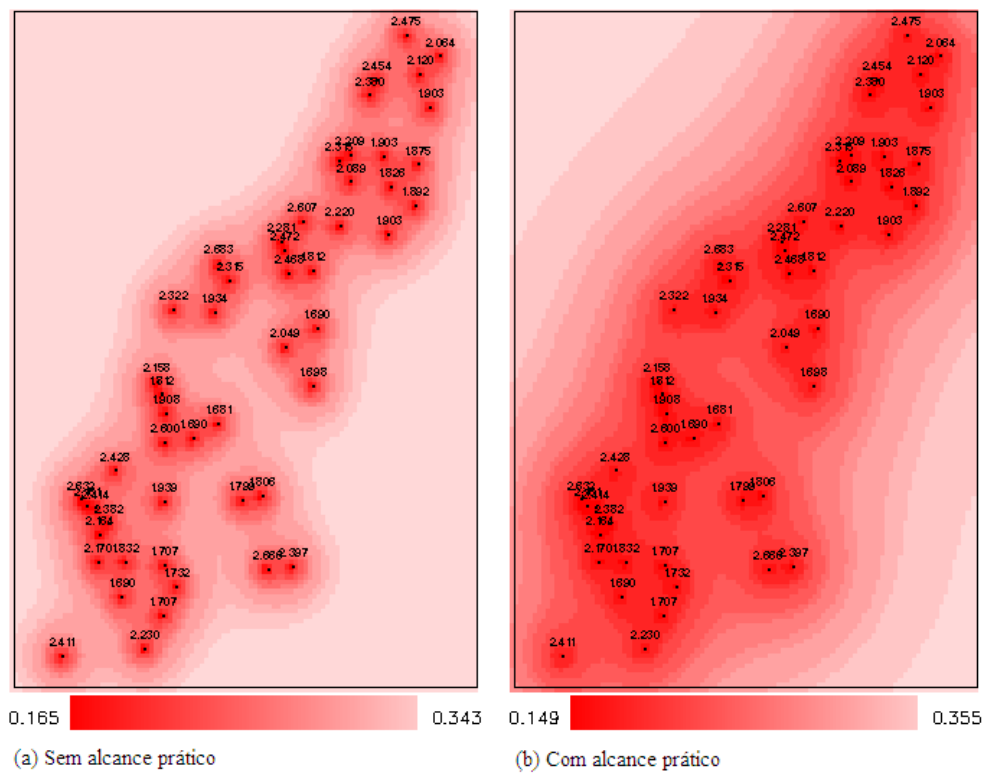


Figura 4.8: Erro padrão das estimativas com base no semivariograma exponencial.

Para avaliar o impacto que o alcance exerce sobre as estimativas de krigagem ordinária, dois modelos clássicos de semivariograma foram utilizados: o gaussiano e o exponencial. Ambos os modelos possuem patamar assintótico e, portanto, alcance prático. As Figuras (4.5) e (4.8), respectivamente, mostram que com o alcance prático o erro mínimo das estimativas de krigagem diminuiu. Segundo Webster e Oliver (2007), isto ocorre porque a medida que o valor do alcance é subestimado os pesos das observações localizadas próximos ao ponto a ser estimado diminuem. Da mesma forma, os pesos para as observações mais distantes sofrem um incremento.

Os três modelos desenvolvidos apresentam características similares quanto a forma gráfica. Todos são sensíveis a grandes distâncias, retornando valores constantes para as semivariâncias. Outra característica comum é o comportamento próximo a origem. Sendo assim, os modelos geram os mesmos impactos sobre o sistema de krigagem ordinária.

Para analisar o efeito da modelagem em situações diversas, foram realizadas simulações de dados com dependência espacial utilizando a função de correlação gaussiana. No entanto, os modelos não se ajustaram aos dados, o que ratifica o fato de que os semivariogramas propostos tem muitas particularidades e só conseguem modelar, eventualmente, em situações muito específicas. Desta forma, entendeu-se não haver necessidade de apresentar esses resultados aqui.

Capítulo 5

Conclusões

O estudo realizado mostrou que a função binomial negativa, apesar de desenvolvida para modelar eventos discretos, pode representar adequadamente um modelo de semivariograma quando escrita sob a forma da família exponencial. Entretanto, ficou comprovado que a expressão resultante apresenta a mesma estrutura do modelo exponencial de semivariograma, diferindo apenas pela parametrização. As distribuições log-normal, gamma e Poisson também foram adaptadas para função semivariograma. Devido a tantas restrições impostas para tais modelos, não há elementos suficientes para considerá-los como funções de semivariogramas válidas. A partir dos resultados alcançados, verificou-se que não é possível estender a utilização da família de funções correlação exponencial potência como uma composição de funções, conservando suas propriedades.

Os modelos de semivariograma propostos não captam adequadamente a variabilidade de qualquer fenômeno e em qualquer unidade de medida das distâncias, comprometendo as estimativas de krigagem. Para grandes distâncias os modelos produzem variabilidade constante, fazendo com que a krigagem ordinária retorne o valor médio da variável analisada para representar os pontos não observados. Outro impacto sobre o processo de predição é quando trabalha-se com pequenas distâncias, pois os modelos desenvolvidos apresentam forma constante para distâncias próximas à origem e mais uma vez a variabilidade real dos dados não é captada. Os resultados ratificaram a importância de utilizar o alcance prático em modelos com patamar assintótico, pois a alteração deste parâmetro interfere nos erros das estimativas de krigagem.

5.1 Limitações do Trabalho

Em qualquer área do conhecimento, o desenvolvimento científico está sujeito a restrições. Neste sentido, algumas limitações identificadas durante o estudo foram:

- A funcionalidade dos modelos foi testada considerando apenas variáveis de ponto;
- As distribuições estudadas apresentaram em sua estrutura o logaritmo da distância;
- Apenas quatro funções foram estudadas;
- O estudo foi conduzido pautado em tentativas *ad hoc*, sendo desejável soluções analíticas.

No entanto, estas limitações não invalidam o trabalho, posto que os resultados alcançados podem ser entendidos como uma possível fonte de pesquisa para outros estudos.

5.2 Sugestões para trabalhos futuros

Face aos conceitos trabalhados aqui recomenda-se para trabalhos futuros:

- Avaliar o comportamento dos modelos utilizando variáveis regularizadas por um suporte de amostragem;
- Utilizar modelos que não apresentem o logaritmo em sua estrutura;

Referências Bibliográficas

- Bailey, T. C. & Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*, (2nd ed.). Prentice Hall.
- Casella, G. & Berger, R. L. (2001). *Statistical Inference*, (2nd ed.). Duxbury.
- Chilès, J. P. & Delfiner, P. (1999). *Geostatistics Modeling Spatial Uncertainty*, (1st ed.). Wiley.
- Christakos, G. (1984). On the problem of permissible covariance and variogram models. *Water Resources Research*, 20:251–265.
- Cressie, N. A. C. (1985). Fitting models by weighted least squares. *Mathematical and Geology*, 17:605–608.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley Interscience.
- Das, S., Rao, T. S., & Boshnakov, G. N. (2012). On the estimation of parameters of variograms of spatial stationary isotropic random processes. Technical report, School of Mathematics. The University of Manchester.
- DeGroot, M. H. (1989). *Probability and Statistics*, (2nd ed.). Addison Wesley.
- Diggle, P. J. & Ribeiro Jr., P. J. (2007). *Model-based Geostatistics*. Springer.
- Diggle, P. J. & Tawn, J. A. (1998). Model-based geostatistics. *Applied Statistics*, 47:299–350.
- Druck, S., Carvalho, M. S., Câmara, G., & Monteiro, A. M. V. (2004). *Análise Espacial de Dados Geográficos*. EMBRAPA.
- Feuerverger, A. & Mureika, R. A. (1977). The empirical function characteristic and its applications. *The Annals of Statistics*, 5:88–97.
- Fischer, M. M. & Getis, A. (2010). *Handbook of Applied Spatial Analysis*. Springer.

- Gotway, C. A. (1991). Fitting semivariogram models by weighted least squares. *Computers and Geosciences*, 17:171–172.
- Handcock, M. S. & Stein, M. L. (1993). A Bayesian analysis of Kriging. *American Statistical Association and American Society for Quality*, 35:403–410.
- Isaaks, E. H. & Srivastava, R. M. (1989). *Applied Geostatistics*. Oxford University Press.
- Jian, X., Olea, R. A., & Yu, Y.-S. (1996). Semivariogram modeling by weighted least squares. *Computers and Geosciences*, 22:381–391.
- Journal, A. G. (1986). Geostatistics: Models and tools for the earth sciences. *Mathematical Geology*, 18:119–140.
- Journal, A. G. (1989). *Fundamentals of Geostatistics in Five Lessons*. Washington: American Geophysical Union.
- Leemis, L. M. & Mcqueston, J. T. (2008). Univariate distribution relationships. *The American Statistician*, 62:47.
- Lele, S. (1997). Estimating functions for semivariogram estimation. *Lecture Notes-Monograph Series, Institute of Mathematical Statistics*, 32:381–396.
- Magalhães, M. N. (2011). *Probabilidade e Variáveis Aleatórias*, (3rd ed.). EDUSP.
- Pebesma, E. J. (2004). geoestatística multivariada em s: O pacote gstat. *Computers & Geosciences*, 30:683–691.
- R (2010). *The R Foundation for Statistical Computing*. <http://CRAN.R-project.org/doc/FAQ/R-FAQ.html>.
- SAS (2011). *SAS on line doc, Versão 9.3*. Cary, NC: SAS Institute Inc.
- Schabenberger, O. & Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC.
- Schlather, M. (1999). Introduction to positive definite functions and to unconditional simulation of random fields. Technical report, Departament of Mathematics and Statistics Faculty of Applied Sciences, Lancaster University, UK.
- Schlather, M., Porcu, E., & Montero, J. M. (2012). *Advances and Challenges in Space-time Modelling of Natural Events*, (1st ed.). Springer.

- Schmidt, A. M. (2001). *Bayesian Spatial Interpolation of Enviromental Stations*. PhD thesis, University of Sheffield.
- Schoenberg, I. J. (1938). Metric spaces and completely monotone functions. *The Annals of Mathematics*, 39:811–841.
- Sen, Z. (2009). *Spatial Modeling Principles in Earth Sciences*. Springer.
- Stein, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics, Institute of Mathematical Statistics*, 16:55–63.
- Webster, R. & Oliver, M. A. (2007). *Geostatistics for Enviromental Scientistis*, (2nd ed.). John Wiley and Sons Ltd.
- Yaglom, A. M. (1962). *An Introduction to the Theory of Stationary Random Functions*. Prentice Hall.
- Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions II, Supplementary Notes and References*. Springer Verlag, New York, Berlin, Heidelberg.
- Zhang, X. F., Eijkeren, J. C. H. V., & Heemink, A. W. (1995). On the weighted least squares method for fitting semivariogram model. *Computers and Geosciences*, 21:605–608.