EDSON ANTONIO ALVES DA SILVA

APLICAÇÃO DE MÉTODOS GEOESTATÍSTICOS MULTIVARIADOS EM PROBLEMAS DE MAPEAMENTO DE VARIÁVEIS DE SOLO E PLANTAS

CURITIBA

MARÇO DE 2007

EDSON ANTONIO ALVES DA SILVA

APLICAÇÃO DE MÉTODOS GEOESTATÍSTICOS MULTIVARIADOS EM PROBLEMAS DE MAPEAMENTO DE VARIÁVEIS DE SOLO E PLANTAS

Projeto de Tese apresentada como requisito parcial à obtenção do grau de Doutor em Ciências, pelo Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná

Orientador: Prof. PhD. Paulo Justiniano Ribeiro Jr.

CURITIBA

MARÇO DE 2007

Termo de Aprovação

EDSON ANTONIO ALVES DA SILVA

APLICAÇÃO DE MÉTODOS GEOESTATÍSTICOS MULTIVARIADOS EM PROBLEMAS DE MAPEAMENTO DE VARIÁVEIS DE SOLO E PLANTAS

Projeto aprovado como requisito parcial para obtenção do grau de Doutor em Ciências, pelo Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná pela seguinte banca examinadora:

Prof. PhD. Paulo Justiniano Ribeiro Jr
Universidade Federal do Paraná

Prof. Dr. Anselmo Chaves Neto
Universidade Federal do Paraná

Prof. Dr. Joel Maurício Corrêa da Rosa
Universidade Federal do Paraná

Prof. Dr. Miguel Angel Uribe-Opazo
Universidade Estadual do Oeste do Paraná

Curitiba, 12 de março de 2007

Sumário

Lista de Figuras

1 INTRODUÇÃO

A grande explosão demográfica que acompanha o desenvolvimento da espécie humana tem exigido cada vez mais um significativo aumento na produção e distribuição de alimentos, pois saciar a fome é uma das suas necessidades mais primárias. A atividade agrícola atual não tem conseguido sucesso em oferecer mais alimentos e simultaneamente preservar o meio ambiente. Os resultados das pesquisas científicas não atingem, em grande escala, a consciência cultural do produtor rural, ávido pelo lucro rápido e sem riscos econômicos.

É incorreto pensarmos que as fronteiras agrícolas se estabelecem nos limites de cada propriedade rural. O ecossistema é um sistema altamente correlacionado onde os recursos disponíveis em um local decorrem das transformações ao longo de milhares de anos de evolução e desenvolvimento do globo terrestre. Uma propriedade rural não representa um sistema fechado. Os insumos aplicados tendem a se distribuírem além de seus limites geográficos. Os recursos naturais ali demandados em um dado momento, sem controle ou critério, podem levar posteriormente à sua falta ou mesmo um esgotamento definitivo, não só naquela propriedade, como também em toda uma região. Se considerarmos os recursos naturais compartilhados, tais como os recursos hídricos, então um manejo isolado em uma propriedade poderá produzir conseqüências danosas às outras ou mesmo ao meio-ambiente local.

Tomemos como exemplo o Estado do Paraná – que tem sido historicamente um dos maiores produtores de grãos do país, com seu potencial econômico agrícola e uma localização privilegiada em relação ao Mercado Comum do Sul - MERCOSUL. Sua região Oeste é responsável por aproximadamente um terço da produção de grãos do Estado, tendo sua economia baseada principalmente na produção de soja e trigo, através de muitas propriedades disputando

os recursos naturais da região. Outro exemplo é na região Nordeste do Estado de Santa Catarina, particularmente nos pequenos municípios de Rio Negrinho e Doutor Pedrinho onde juntos, dispõem de 232 indústrias ligadas ao setor madereiro, abastecidas por grandes áreas de reflorestamento de pinus e eucalipto, interferindo com a economia e o meio-ambiente dessas duas cidades.

Por outro lado, a globalização da economia mundial e a grande demanda por mais alimentos exigem que a agricultura brasileira desenvolva tecnologias que possibilitem a competição de nossos produtos no mercado mundial e o aumento da produtividade para atender o crescimento populacional. O aumento da produtividade é normalmente acompanhado pelo aumento do uso dos insumos agrícolas. Estes insumos compreendem os insumos biológicos, insumos mecânicos, água e insumos químicos. O uso de insumos químicos tem sido identificado como o principal fator de contaminação da água e do solo (BAKHSH et al., 1997). Deduz-se, portanto, que estes insumos, ao mesmo tempo em que auxiliam no aumento da produtividade agrícola, apresentam grande perigo para o solo e mananciais de água. Visando aumentos progressivos de produtividade, os agricultores utilizam o máximo de fertilizantes e corretivos, considerando como uniforme o solo de cada área de cultivo. Entretanto, cada talhão pode ter considerável variação em seus atributos. Com o aumento da área do talhão, a diferença entre as necessidades da cultura e a taxa de aplicação empregada em função da média tendem a ser maiores e a otimização das aplicações de insumos pode ajudar a proteger o meio ambiente. Para esse autor, a aplicação da Agricultura de precisão (AP) nas propriedades agrícolas requer o uso de tecnologias emergentes que possuam o potencial de discriminar, à uma resolução refinada, a variabilidade espacial dos diversos fatores associados à produção e direcionar o sistema mecanizado a aplicar os insumos otimizados com o auxílio de aparelhos com Sistema de Posicionamento Global, popularmente conhecido por GPS (acrónimo do inglês *Global Positioning System*) e tecnologia SIG, acrónimo de Sistema de Informações Geográficas que lidam com informação geográfica na forma de dados geográficos. Essas tecnologias geram uma grande quantidade de dados, normalmente expressos na forma de mapas temáticos, gerenciados por softwares especialistas de apoio a decisão no manejo agrícola. Visando tais aumentos progressi-

vos de produtividade os agricultores utilizam o máximo de fertilizantes e corretivos. Entretanto, cada talhão pode ter considerável variação em seus atributos.

Uma prática da agricultura de precisão está fundamentada basicamente na existência da variabilidade espacial dos fatores produtivos e, portanto, da própria quantidade produzida pela cultura, constituindo a sua representação gráfica uma das mais importantes ferramentas destinadas a sua análise (BALASTREIRE; ELIAS; AMARAL, 1997). Molin (1997) considerou que a AP será o próximo desafio a ser vencido pelo agricultor brasileiro.

Muitos autores empregaram em seus trabalhos sensores para detectar a energia electromagnética proveniente do campo e registrar em filmes ou na forma digital. São instrumentos básicos do Sensoriamento Remoto, que também é uma importante ferramenta de aquisição de dados para a AP. Estes instrumentos não invasivos fornecem dados e possibilitam a detecção de fenômenos e o acompanhamento de determinados alvos a longas distâncias, como por exemplo o diagnóstico de déficit de nitrogênio pela emissão de uma cor característica em um espectro de luz. O acompanhamento do desenvolvimento de uma cultura em tempo real e a correção dos fatores deficientes no instante que é diagnosticado é uma das metas mais importantes e ousadas da AP (CAPELLI, 1999).

A geoestatística se insere nesse contexto como um método que utiliza procedimentos estatísticos aplicados a problemas cujos dados provêm de fonômenos naturais e que são espacialmente distribuídos e autocorrelacionados, ou seja, consideram não só o valor obtido para uma determinada variável, mas também sua posição, expressa por um sistema de coordenadas. Assim, o comportamento do evento estatístico pode ser descrito pelas diferenças entre as informações obtidas em função da distância que as separa. O valor em uma determinada posição poderá ser estimado pelas informações de posições vizinhas. Atualmente a Geoestatística é popular em muitas áreas das ciências e da indústria para avaliar dados correlacionados no espaço ou no tempo.

Tanto em experimentos baseados nos conceitos de Agricultura de Precisão quanto experimentos de outras áreas que envolvem a estatística espacial, particularmente a geoestatística,

usam procedimentos univariados para a representação do comportamento de suas variáveis em áreas de manejo. Entretanto, em problemas reais, os fenômenos frequentemente ocorrem sob circunstâncias multivariadas e espacialmente correlacionados.

Existe disponível na literatura, muitos trabalhos envolvendo métodos geoestatísticos multivariados mas ainda cabe investigações para se determinar as condições em que uma análise multivariada para os problemas representam um ganho efetivo na qualidade dos resultados, no confiabilidade do processo, na eficiência, sobretudo na predição. Cabe espaço também para se avaliar as características dos diferentes modelos propostos, ou seja, tanto aqueles baseados em variogramas e na estrutura da matiz de correlação como aqueles baseados em modelos de regressão. Em decorrência dessas avaliações, poderá surgir novas proposições ou recomendações de estratégias de modelagem que levem a uma viabilidade computacional, – grande limitação nos métodos atuais, ou ainda, ampliar a interpretabilidade dos resultados.

Este trabalho pretende estudar o rendimento das cultivares soja e pinus, correlacionando-as com variáveis agrícolas de solo para um eficiente manejo localizado de nutrientes. Empregaremos os métodos geoestatísticos multivariados devido ao grande conjunto de variáveis preditoras disponíveis para o resultado agrícola.

2 PROCESSOS GEOESTATÍSTICOS

2.1 MODELO GAUSSIANO UNIVARIADO

2.1.1 Geometria do espaço geoestatístico

Neste trabalho serão considerados dados espaciais as informações observadas de um fenômeno aleatório ocorrido em um sistema solo-planta, distribuído em uma região de um espaço bidimensional. Não serão abordados dados que representem polígonos de uma região (sub-área) e nem dados que representem processos pontuais, como a ocorrência positiva ou negativa de um atributo. Estaremos aqui interessados somente em dados vinculados a um processo aleatório gaussiano de variação contínua e mensurável.

O formato básico para dados geoestatísticos univariados que empregaremos será aquele adotado por Diggle e Ribeiro Jr (2007), ou seja:

$$\{(x_i;y_i) : x_i \in \mathbb{R}, y_i \in \mathbb{R}, i : 1,2,\ldots,n\}$$

onde:

x_i : indica a localização espacial da i-ésima coordenada em uma região do espaço bi-dimensional ($\mathbb{R}^2$).

y_i : indica uma medida escalar da variável aleatória contínua $Y = (y_1, y_2, \ldots y_n)$, tomada na x_i-ésima localização.

Um particular resultado y da variável Y pode ocorrer em qualquer localização x de uma

região contínua. Assumimos aqui que as localizações $x_i : i = 1,2,\ldots,n$ formam uma malha fixa ou estocasticamente independente, onde serão obtidas as medidas de y_i.

Um processo gaussiano é definido como um conjunto de n variáveis aleatórias onde a distribuição finito-dimensional de qualquer subconjunto de variáveis tomadas desse conjunto, terá distribuição gaussiana multivariada com o número de variáveis do subconjunto. Assim, o conjunto $\{S(x_i) : x_i \in \mathbb{R}^2; i : 1,2,\ldots,n\}$, será o processo estocástico gaussiano que descreverá, de maneira teórica o comportamento de um fenômeno em uma área, onde supomos que esse processo tenha uma distribuição contínua e que o evento Y ocorra devido a sua lei de probabilidades. O modelo geoestatístico apropriado que adotaremos será então baseado em um processo estocástico espacial $S(\mathbf{x})$, gaussiano, contínuo, que irá representar nosso fenômeno de interesse em uma área de um espaço bidimensional ou, eventualmente, em uma reta de um espaço unidimensional. Entendemos aqui o processo estocástico gaussiano univariado como sendo um modelo probabilístico definido por um conjunto de variáveis aleatórias gaussianas $\{S(\mathbf{x}) : x \in \mathbb{R}^2\}$ em que os $S(x_i)$ são medidas de mesma natureza, que ocorrem em diferentes locais do espaço (WALLER; GOTWAY, 1965). Assim, $Y = \{y_1, y_2, \ldots, y_n\}$ será um vetor aleatório de dimensão n contendo as medidas da realização do evento, onde cada y_i terá função densidade de probabilidade gaussiana dada, segundo Mood, Graybill e Boes (1974), por:

$$f_Y(y_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \; exp\left\{-\frac{1}{2}\left(\frac{y_k - \mu}{\sigma_k}\right)^2\right\} \quad k = 1,2,\ldots,n \tag{2.1}$$

Uma realização do evento Y corresponde a um conjunto de observações em n localizações distintas e fixas, onde cada resultado é, em si, o resultado de uma variável aleatória $Y_k = Y(x_k) = y_k$, $k = 1,2,\ldots,n$. Uma realização de Y é então a ocorrência de n variáveis aleatórias gaussianas y_k com distribuição de probabilidades dadas por 2.1, cada uma com uma única observação e que pode ser modelada como:

$$y(x_i) = \mu(x_i) + S(x_i) + \delta_i; \quad i = 1,\ldots,n \tag{2.2}$$

onde:

- $y(x_i)$ será uma variável aleatória contínua com distribuição normal de média $E\left[y_i|S(x_i)\right] = \mu(x_i) + S(x_i)$ e variância condicional $Var(y_i|S(x_i)) = \tau^2$;
- $\mu(x_i) = \beta_0 + \beta_1 d_1(x_i) + \beta_2 d_2(x_i) + \ldots + \beta_p d_p(x_i)$ que pode ser representado matricialmente como $\mathbf{D}\beta$ é efeito espacial externo associado a t variáveis $d(x_i)$, diferentes de $y(x_i)$ mas que irão depender da localização x_i. Os coeficientes β são constantes a serem determinadas. Esse componente torna o modelo não estacionário.
- $\left\{S(x_i) : x_i \in \mathbb{R}^2\right\}$ é um processo gaussiano multivariado com média zero e variância σ^2 e função de correlação $\rho(u_{ij}) = Corr\{S(x_i), S(x_j)\}$ onde $u_{ij} = \|x_i - x_j\|$ é a distância euclidiana que separa duas coordenadas quaisquer x_i e x_j ;
- δ_i são erros aleatórios mutuamente independentes com distribuição normal de média zero e variância τ^2, ou seja, $\delta_i \sim \mathrm{N}(0; \tau^2)$.

A distribuição de probabilidade da variável aleatória Y será então:

$$\mathbf{Y} \sim N\left(\mathbf{D}\beta, \sigma^2\mathbf{R} + \tau^2\mathbf{I}\right) \tag{2.3}$$

onde:

- σ^2 é a variância (constante) e $\mathbf{R}$ é uma matriz de tamanho $n \times n$ cujos elementos representa as correlações entre variáveis observadas em diferentes localizações;
- τ^2 representa a variância do erro δ_i e $\mathbf{I}$ a matriz identidade de tamanho $n \times n$.

2.1.2 Tendência devido a Estacionariedade e Isotropia

Segundo Waller e Gotway (1965), dois conceitos devem ser estabelecidos antes de se modelar um processo espacial: estacionariedade e isotropia. Matematicamente um processo será estacionário quando for invariante às translações em um espaço multidimensional, ou seja, a relação entre dois eventos em um processo estacionário dependerá somente de suas posições relativas. Será isotrópico quando for invariante às rotações em torno da origem de um sistema

de referência, ou seja, não deverá depender da orientação do eixo que liga suas posições no espaço.

O conceito de estacionariedade ocorre quando existe uma variação natural, – própria da área, que interfere no comportamento do processo, como por exemplo: a declividade sistemática de um solo que interfere nas características de fertilidade, umidade e compactação, importantes para se avaliar a variação da produtividade de uma área. Para estudar esse efeito, pesquisadores costumam modelar a média μ como uma função das localizações x, destacando os efeitos de tendência por modelos de regressão polinomial e utilizando o resíduo, para então prosseguir com a análise. Modelos assim não são cientificamente explicados pois as correlações com direções definidas não dão informações sobre o processo causador do efeito.

Esse modelo que afeta a média do processo $Y(x)$ pode ser feita relativamente às covariáveis $d(x)$. É o equivalente aos fatores em uma análise estatística tradicional. Muitas pesquisas são feitas em áreas onde existem sub-áreas de características próprias que afetam o processo $S(x)$ em estudo. Neste caso o conceito é semelhante ao delineamento de experimentos em blocos, que retira do resíduo uma fonte de variação conhecida. As informações $d(x)$ são normalmente tomadas nas mesmas coordenadas do processo principal $S(x)$ e não são tratadas como um segundo processo $S^*(x)$, mantendo assim o aspecto univariado da análise.

De maneira um pouco mais formal, dizemos que o processo é estacionário na média se $\mu(x_i) = \mu,\ \forall\, x_i$ e estacionário na variância se as covariâncias para cada par de coordenadas forem função somente da distância euclidiana u_{ij} e para $\rho(u_{ij} = 0) = \sigma^2$.

Um outro aspecto importante sobre o modelo gaussiano é quando há uma certa falta de estacionariedade na sua estrutura de correlação. Um pressuposto razoável é supor que seu valor decai a medida que a distância entre as localizações aumenta. Se supormos que a taxa de variação independe do ângulo do eixo formado entre essas localizações, dizemos que o processo é isotrópico, senão dizemos ser anisotrópico.

Essa forma direcional de se avaliar o comportamento das correlações é chamado de efeito direcional que na sua forma mais simples, — e talvez mais comum, é chamado aniso-

tropia geométrica. Este tipo de anisotropia ocorre quando a estrutura de covariância apresenta alongamentos e rotações em relação aos eixos das coordenadas. Desta forma podemos caracterizar esse efeito através de dois parâmetros: o ângulo de anisotropia ψ_A que dá a direção do efeito e a razão de anisotropia $\psi_R > 1$ que dá a relação entre o eixo maior e o eixo menor da elipse formada.

Na prática ψ_A e ψ_R são informações desconhecidas que podem ser convenientemente incorporadas ao modelo geoestatístico para serem estimadas. Uma vez conhecida a tendência devido à anisotropia, poderemos, para efeito de análise, transformar as coordenadas. Se (a,b) é a coordenada de um ponto x no plano cartesiano $\mathbb{R}^2$, representando um vetor, poderemos contrair/extender e/ou rotacionar esse vetor aplicando transformações lineares conforme dada na equação 2.1.2 (KOLMAN, 1997):

$$(a',b') = (a,b) \begin{pmatrix} cos(\psi_A) & -sen(\psi_A) \\ sen(\psi_A) & cos(\psi_A) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \psi_R^{-1} \end{pmatrix}$$

Tanto o efeito de tendência direcional quanto o regional como o efeito de anisotropia têm papel fundamental na análise do processo $S(x)$ pois permitem melhorar o conhecimento subjetivo do fenômeno em estudo, entretanto, devem ser modelados e eliminados.

Segundo Matheron (1973), um tipo de modelo não-estacionário é o modelo intrínseco. Ele considera um caminho aleatório $S(x) = S(x-1) + Z(x)$ com $Z \sim N(0,1)$, ou seja, uma função aleatória intrínseca é um processo estocástico $S(x)$ com incrementos estacionários. Assim, o processo $D_u(x) = S(x) - S(x-u)$ será dito estacionário para todo $u \in \mathbb{R}^2$.

A principal diferença entre uma predição obtida com modelo intrínseco e modelo estacionário, é que se for usado o primeiro, a predição em uma localização x será influenciada pelo ambiente local dos dados, ou seja, por observações medidas em locais próximos de x. Considerar uma hipótese intrínseca para os dados significa supor que as diferenças entre os valores apresentam fraco incremento, ou seja, as diferenças serão localmente estacionárias. Já com o

emprego de modelos estacionários, as predições serão afetadas pelo ambiente global dos dados.

2.1.3 Covariância e Variograma

Consideremos o modelo dado pela equação 2.2 (supondo estacionariedade) como sendo aquele que descreve o conjunto Y das variáveis observadas de um determinado pelo processo $S(x)$. Sejam assim, y_i e y_j observações tomadas em quaisquer duas localizações separadas por uma distância $u_{i;j}$, então, $Var(y_i - y_j)$ registra a variação da diferença dos valores medidos separados por essa distância. Assim, fixando $\mu = 0$:

$$\begin{aligned} Var(y_i - y_j) &= Var(y_i) + Var(y_j) - 2\,Cov(y_i; y_j) \\ Var(y_i - y_j) &= Var(S(x_i) + \delta_i) + Var(S(x_j) + \delta_j) - 2\,Cov(y_i; y_j) \end{aligned} \tag{2.4}$$

Como $S(x_i)$ e δ_i são processos diferentes e independentes, então:

$$\begin{aligned} Var(y_i) &= Var(S(x_i) + \delta_i) = Var(S(x_i)) + Var(\delta_i) \\ Var(y_j) &= Var(S(x_j) + \delta_j) = Var(S(x_j)) + Var(\delta_j) \end{aligned}$$

$$Var(y_i) = Var(y_j) = \sigma^2 + \tau^2 \tag{2.5}$$

Em estatística o coeficiente de correlação de Pearson (ρ) mede o grau e a direção (positiva ou negativa) da correlação entre duas variáveis (MONTGOMERY; PECK, 1955). Se aplicada no contexto da geoestatística temos:

$$\rho(u_{ij}) = \frac{Cov(y_i; y_j)}{\sqrt{Var(y_i)Var(y_j)}} = \frac{Cov(y_i; y_j)}{\sqrt{\sigma^2\sigma^2}} = \frac{Cov(y_i; y_j)}{\sigma^2}$$

então:

$$Cov(y_i; y_j) = \sigma^2 \rho(u_{ij}) \tag{2.6}$$

Notar que, pela equação 2.6, caso aceita a hipótese de estacionariedade, a correlação

entre dois valores medidos de Y, irá depender somente da distância que os separa. Esta função será monótona decrescente, restrita a $\rho(0) = 1$ e $\lim_{u\to\infty} \rho(u) = 0$ para $u \geq 0$

Assim, substituindo os resultados das equações 2.5 e 2.6 na equação 2.4 obtemos:

$$\begin{aligned} Var(y_i - y_j) &= (\sigma^2 + \tau^2) + (\sigma^2 + \tau^2) - 2\,\sigma^2 \rho(u_{ij}) \\ Var(y_i - y_j) &= 2\,\tau^2 + 2\,\sigma^2(1 - \rho(u_{ij})) \\ Var(y_i - y_j) &= 2\left(\tau^2 + \sigma^2(1 - \rho(u_{ij}))\right) \end{aligned}$$

Definimos então $\frac{1}{2}Var(Y_i - Y_j)$ como sendo a semivariância teórica, denotada por $\gamma(u_{ij})$ e escrevemos:

$$\gamma(u_{ij}) = \tau^2 + \sigma^2(1 - \rho(u_{ij})) \tag{2.7}$$

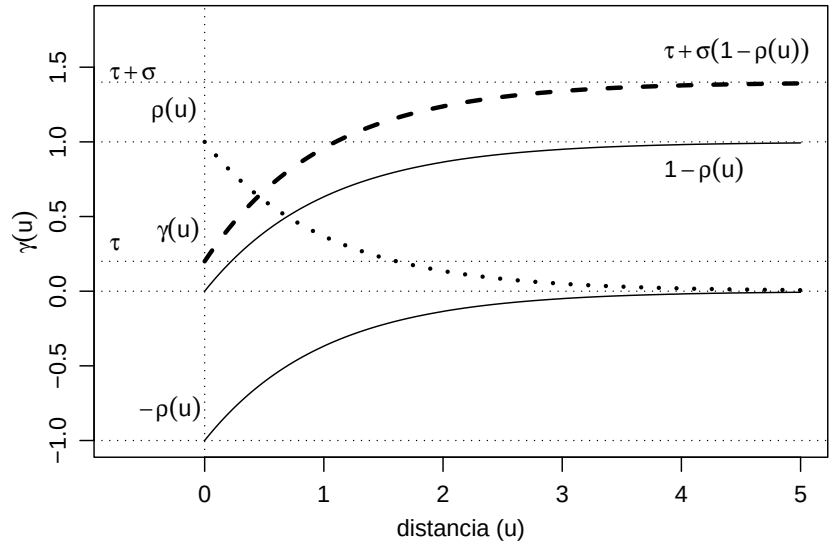

Figura 2.1: Etapas da transformação da função de correlação (linha pontilhada) para a função semivariograma (linha tracejada)

A figura 2.1 mostra o comportamento gráfico da função semivariância onde podemos notar o papel funtamental da função de correlação pois é ela que representa a propriedade desejada para o modelo. Segundo Diggle e Ribeiro Jr (2007), sendo o processo, estacionário, a semivariância é o equivalente teórico para a função covariância, com a vantagem de ser uma excelente ferramenta de análise de dados, especialmente em condições de um experimento conduzido em malha regular.

2.1.4 Transformação da Variável Resposta

Existem muitas razões importantes amplamente discutidas na literatura para se transformar dados estatísticos buscando obter uma forma de distribuição próxima da distribuição normal de probabilidades. Eventos que têm evolução não linear, representados por forte assimetria na distribuição de frequências de seus dados, requerem tranformações logarítmicas convertendo o problema em uma escala de evolução mais aditiva, levando a distribuição em direção a um comportamento mais simétrico, próximo de uma distribuição gaussiana. Já transformações do tipo raiz quadrada ou arco-seno tendem a estabilizar a variância para amostras de uma distribuição de Poisson e Binomial, respectivamente. Esse tipo de transformação torna os dados mais homocedásticos.

Box e Cox (1964) apresentam um método de transformação que basicamente consiste da adequação a uma família paramétrica numa generalização empírica do modelo gaussiano, na qual a escolha da transformação mais adequada corresponde a estimar um parâmetro λ. Uma vez escolhido, procede-se com a seguinte operação nos dados observados:

$$Y^* = \begin{cases} \left(\dfrac{Y^\lambda - 1}{\lambda}\right) & \text{se } \lambda \neq 0 \\ \log Y & \text{se } \lambda = 0 \end{cases} \tag{2.8}$$

2.2 FUNÇÕES DE CORRELAÇÃO

2.2.1 Continuidade e Diferenciabilidade da função de correlação

A estrutura de correlação pode desempenhar um papel decisivo na escolha do modelo geoestatístico pois esta escolha irá afetar diretamente a suavidade da superfície gerada. É ela que estabelece o comportamento de uma característica pontual em sua vizinhança. Medidas matemáticas aceitas para se avaliar essa suavidade são a continuidade e a diferenciabilidade da função associada ao processo. Bartlett (1955) afirma que um processo estocástico estacionário

com função de correlação $\rho(u)$ será k-vezes diferenciável se e somente se $\rho(u)$ for 2 k-vezes diferenciável na origem.

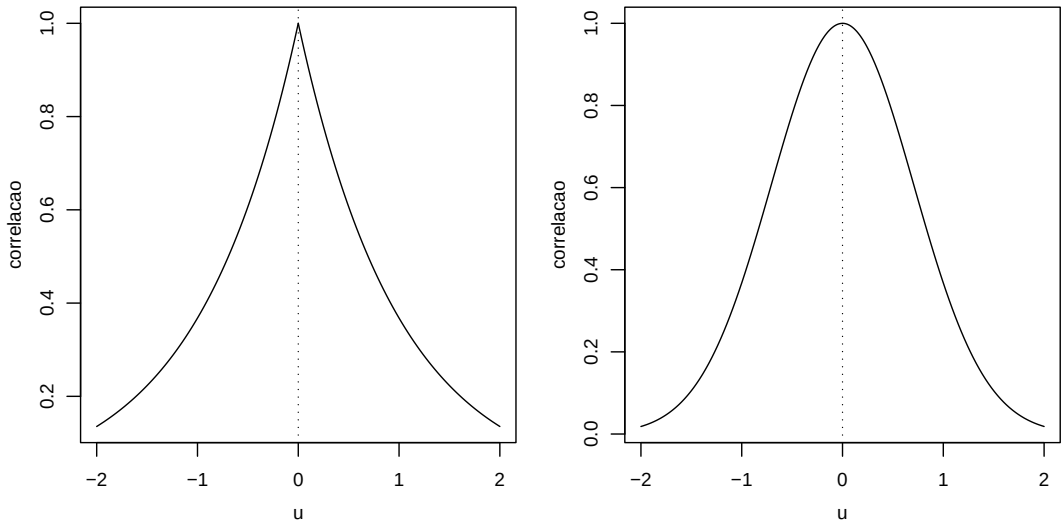

Figura 2.2: A figura de esquerda corresponde ao comportamento da função de correlação exponencial de ordem 1 ($exp(-u)$) onde a reta tangente à função no ponto $u = 0$ é vertical (não diferenciável). A figura da direita corresponde a mesma função de correlação exponencial "poder" com potência igual a 2 ($exp(-u^2)$), com reta tangente igual a zero em $u = 0$ (diferenciável).

Na figura 2.2 temos o comportamento básico de função de correlação que é diferenciável e de função que não o é. Ambas figuras ilustram o caso de funções contínuas em todo o domínio das distâncias u, o que é mais frequentemente adotado, embora possa ocorrer descontinuidades na origem. A figura da esquerda apresenta um ponto "problema" que é o ponto $u = 0$ onde a função não é diferenciável. Já a figura da direita mostra uma função contínua e derivável em todos os seus pontos.

O processo $S(x)$ é desconhecido e tipicamente não diretamente observável, assim, a experiência do pesquisador com o fenômeno estudado deve ser usada para uma boa escolha do modelo de correlação espacial. Se o evento em questão tem variações mais abruptas, modelos com números menores de derivadas deverão ser preferidos e se tem variações mais suaves, utiliza-se números maiores.

Ilustramos um exemplo desse efeito na figura 2.3 onde gerou-se simulações do processo $S(x)$. Ele foi gerado simulando 200 resultados de um processo estocástico estacionário, isotrópico, com taxas de decaimento equivalentes. Nela foram empregados duas situações:

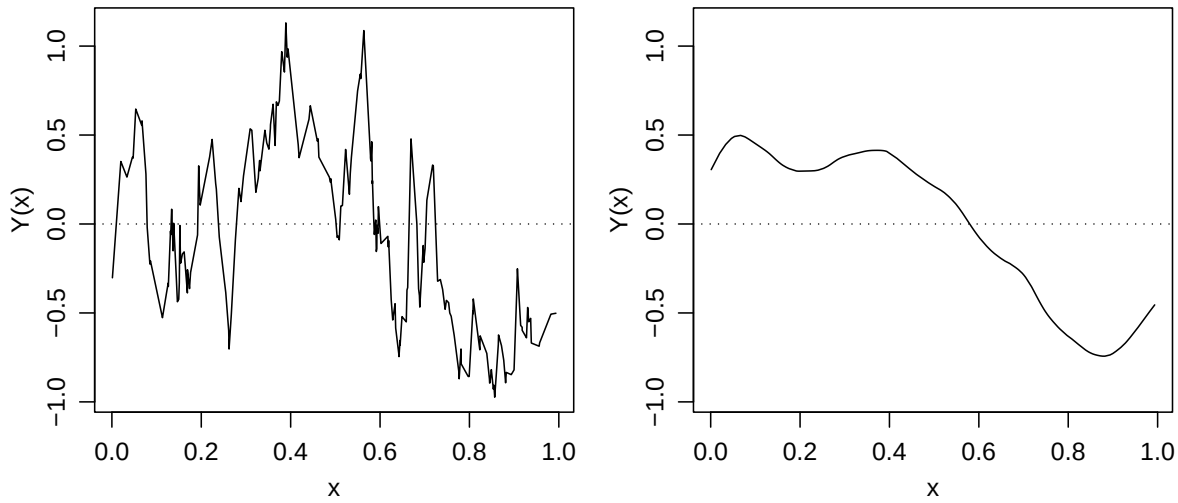

Figura 2.3: A figura da esquerda representa um processo de variações abruptas ao longo de uma transecção unidimensional, associada a uma função de correlação não-diferenciavel. A figura da direita mostra um processo com variações mais suaves ao longo da mesma transecção, mas associada a uma função de correlação duas vezes diferenciável.

função contínua não diferenciável (esquerda) onde notamos variações bruscas da superfície gerada pelo processo e função contínua diferenciável (direita) onde as variações são mais suaves. Cabe aqui salientar que o processo é o mesmo (exponencial), diferindo apenas na diferenciabilidade da função de correlação.

Devemos lembrar que correlações com variações muito suaves perto da origem podem produzir efeitos de *quasi*-multicolinearidade, levando a dificuldades computacionais na solução numérica da álgebra envolvida no processo. Uma vez que se supõe diminuir a similaridade regional a longas distâncias, sendo no máximo nula, então é rezoável escolher o conjunto de funções de correlações que sejam definidas positiva. Esta condição impõe restrições. Assim, para um conjunto de localizações x_i e uma constante real a_i, a combinação linear $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Y_i;Y_j) \geq 0 \ \forall \ i;j$ implicando que somente algumas famílias paramétrica específicas de função de correlação, como as dadas a seguir, terão uso prático.

2.2.2 Função de correlação de Matérn

Matérn (1986) apresenta uma classe de funções de correlação que é considerada uma das mais completas por englobar outras funções de correlação, pela simples escolha de um parâmetro de diferenciabilidade. Ela é dada por:

$$\rho(u,\phi,\kappa) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{u}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{u}{\phi}\right) \tag{2.9}$$

onde $K_{\kappa}(\delta)$, $\delta = \frac{u}{\phi}$ é a função modificada de Bessel de terceiro tipo (ABRAMOWITZ; STEGUN, 1965) dada por:

$$K_{\kappa}(\delta) = \begin{cases} \left(\frac{\pi}{2\sin\pi\delta}\right)\{I_{-\kappa}(\delta) - I_{\kappa}(\delta)\} & \kappa \neq 0,1,2,\ldots \\ \\ \lim\limits_{p\to\kappa}\left(\frac{\pi}{2\sin\pi p}\right)\{I_{-\kappa}(\delta) - I_{\kappa}(\delta)\} & \kappa = 0,1,2,\ldots \end{cases}$$

sendo que:

$I_{\kappa}(\delta) = \sum\limits_{j=0}^{\infty} \frac{(\delta/2)^{\kappa+2j}}{j!\Gamma(\kappa+j+1)}$ para $\kappa = 0,1,2,\ldots$ e

$\Gamma(\kappa) = \int_0^{\infty} t^{\kappa-1}e^{-t}dt \quad \kappa > 0$ é a função Gamma.

O parâmetro $\phi > 0$ dá a taxa na qual a função de correlação cai a zero com o aumento da distância u. O parâmetro $\kappa > 0$ é chamado de ordem do modelo de Matérn e determina a suavidade com que o sinal $S(x)$ cai a zero. O comportamento dessa função pode ser vista na figura 2.4.

2.2.3 Função de correlação da Família Esférica

A função de correlação dessa família é definida como:

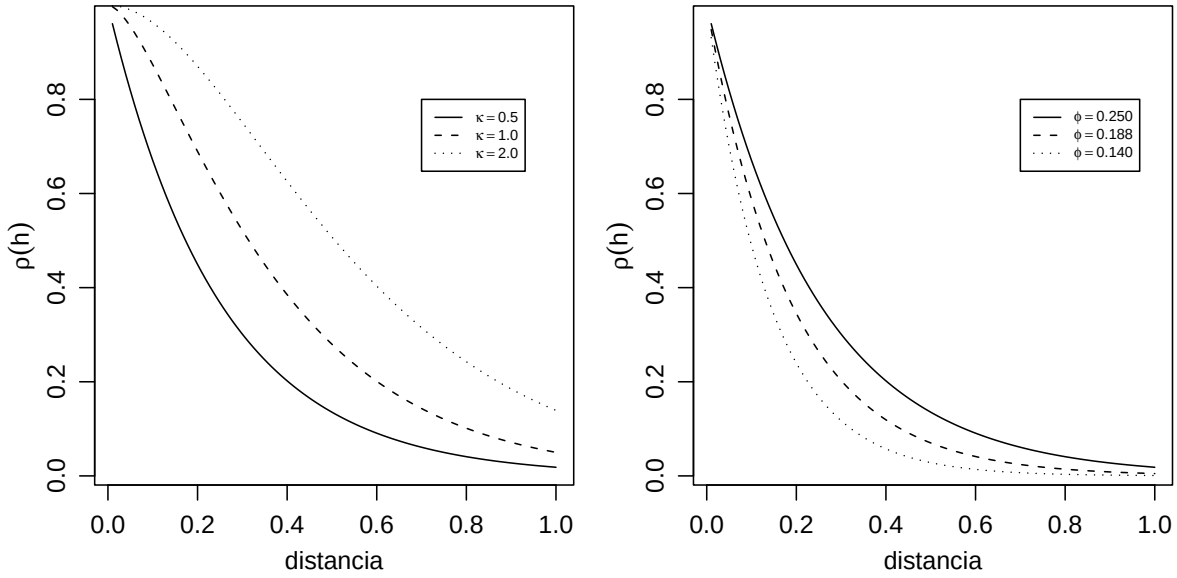

Figura 2.4: À esquerda a figura ilustra o comportamento da função de correlação de Matérn com o parâmetro $\phi = 0,25$ fixo e diferentes valores para o parâmetro de diferenciabilidade κ. Na figura da direita, para um mesmo valor de $\kappa = 0.5$, variou-se o parâmetro ϕ que controla a taxa de decaimento da função.

$$\rho(u;\phi) = \begin{cases} 1 - \frac{3}{2}\left(\frac{u}{\phi}\right) + \frac{1}{2}\left(\frac{u}{\phi}\right)^3 & 0 \leq \phi \\ 0 & u > \phi \end{cases} \tag{2.10}$$

O nome desta função se deve ao fato de que $\rho(u;\phi)$ tem uma interpretação geométrica como sendo o volume de interseção de duas esferas cujos centros estejam separadas de uma distância u (DIGGLE; Ribeiro Jr, 2007). Essa função de correlação tem alcance finito e depende somente do parâmetro de escala ϕ. O comportamento gráfico dessa função pode ser vista na figura 2.5 à esquerda.

2.2.4 Função de correlação da Família Exponencial "Poder" de ordem κ

A função de correlação dessa família é definida como:

$$\rho(u;\phi;\kappa) = e^{-\left(\frac{u}{\phi}\right)^{\kappa}} \quad \text{para } \phi > 0 \text{ e } 0 < \kappa \leq 2 \tag{2.11}$$

Nesta função, se $\kappa < 2$, o processo $S(x)$ é contínuo mas não é diferenciável e se se $\kappa \geq 2$ pode ser infinitamente diferenciável. Existem dois casos particulares para essa ela. No caso de $\kappa = 1$ e função será chamada exponencial, já para $\kappa = 2$ a função será chamada de gaussiana (figura 2.5 à direita).

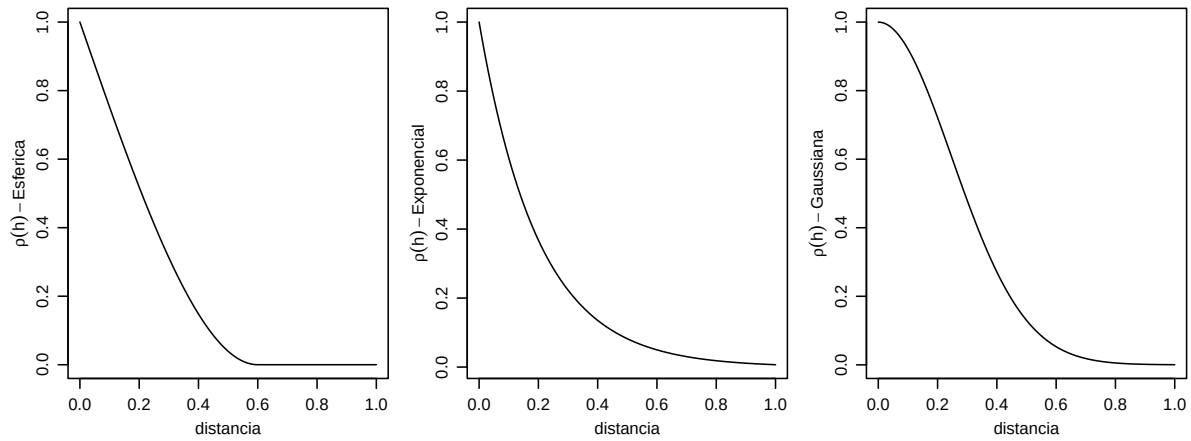

Figura 2.5: O gráfico da esquerda mostra uma função de correlação esférica com o parâmetro $\phi = 0,6$. O gráfico do centro ilustra o comportamento de uma função de correlação exponencial de ordem K ($\kappa = 1$) e $\phi = 0,2$, correspondendo a uma função denominada Exponencial.O gráfico da direita ilustra também o comportamento de uma função de correlação exponencial de ordem K mas com $\kappa = 2$ e $\phi = 0,35$, correspondendo a uma função denominada Gaussiana.

Toda a metodologia geoestatística está baseada na correlação existente entre as medidas tomadas em duas coordenadas distintas. As formas das funções apresentadas atendem ao pressuposto de que as observações mais próximas são, provavelmente, mais similares entre si do que aquelas muito afastadas. Isso dá o caráter regionalizado de um atributo ou uma propriedade em áreas agrícolas.

Existe na literatura muitas outras propostas de funções de correlação que atendem a fenômenos específicos. Das funções apresentadas, a mais empregada é a de Matérn pois ela permite maior flexibilidade na variação dos parâmetros por descreverem a diferenciabilidade do processo e a extensão da dependência espacial. Ela será a nossa escolha no desenvolvimento deste trabalho.

2.3 ESTIMAÇÃO DE PARÂMETROS

2.3.1 Modelagem e estimação de parâmetros de tendência não-estacionária

O modelo geoestatístico completo idealizado para o processo mensurável Y é dado pela equação 2.2 como:

$$y(x_i) = \mu(x_i) + S(x_i) + \delta_i \quad i = 1, 2, \ldots, n$$

Vamos aqui assumir uma estrutura de dependência espacial para a média $\mu(x_i)$ como:

$$\mu(x_i) = \beta_0 + \beta_1 d_1(x_i) + \ldots + \beta_k d_k(x_i) = \beta_0 + \sum_{j=1}^{k} \beta_j d_{j_i} \tag{2.12}$$

ou, na sua forma matricial como:

$$\mu = \mathbf{D}\,\beta \tag{2.13}$$

onde: $\mu = \begin{pmatrix} \mu_1 & \mu_2 & \ldots & \mu_n \end{pmatrix}', \beta = \begin{pmatrix} \beta_1 & \beta_2 & \ldots & \beta_k \end{pmatrix}', \varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_n \end{pmatrix}'$

$$\mathbf{D}(\mathrm{x}) = \begin{pmatrix} 1 & d_{11} & d_{12} & \ldots & d_{1k} \\ 1 & d_{21} & d_{22} & \ldots & d_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d_{n1} & d_{n2} & \ldots & d_{nk} \end{pmatrix};$$

sendo a matriz **D** uma matriz de posto completo, ou seja, $n \geq k$. Os coeficiente podem facilmente serem obtidos empregando-se o método dos mínimos quadrados (MONTGOMERY; PECK, 1955). Sob a hipótese de independência entre as observações, a função de mínimos quadrados para o problema pode ser escrita como:

$$MSQ(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(\mu_i - \beta_0 - \sum_{j=1}^{k} \beta_j d_j \right)^2 \tag{2.14}$$

A solução que minimiza a equação 2.14 em termos de β, segundo Montgomery e Peck (1955) é aquela que satisfaz:

$$\frac{\partial}{\partial \beta_0} MSQ(\beta) = -2 \sum_{i=1}^{n} \left(\mu_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j d_{ij} \right) = 0$$

$$\frac{\partial}{\partial \beta_j} MSQ(\beta) = -2 \sum_{i=1}^{n} \left(\mu_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j d_{ij} \right) d_{ij} = 0 \text{ para } j = 1, 2, \ldots, k \text{ e}$$

$i = 1, 2, \ldots, n$

Expandindo o somatório e simplificando obtemos o seguinte sistema de equações normais de mínimos quadrados:

$$\begin{cases} n\hat{\beta}_0 & +\hat{\beta}_1 \sum_{i=1}^{n} d_{i1} & +\hat{\beta}_2 \sum_{i=1}^{n} d_{i2} & +\ldots & +\hat{\beta}_k \sum_{i=1}^{n} d_{ik} & = \sum_{i=1}^{n} \mu_i \\ \hat{\beta}_0 \sum_{i=1}^{n} d_{i1} & +\hat{\beta}_1 \sum_{i=1}^{n} d_{i1}^2 & +\hat{\beta}_2 \sum_{i=1}^{n} d_{i1} d_{i2} & +\ldots & +\hat{\beta}_k \sum_{i=1}^{n} d_{i1} d_{ik} & = \sum_{i=1}^{n} d_{i1} \mu_i \\ \hat{\beta}_0 \sum_{i=1}^{n} d_{i2} & +\hat{\beta}_1 \sum_{i=1}^{n} d_{i1} d_{i2} & +\hat{\beta}_2 \sum_{i=1}^{n} d_{i2}^2 & +\ldots & +\hat{\beta}_k \sum_{i=1}^{n} d_{i2} d_{ik} & = \sum_{i=1}^{n} d_{i2} \mu_i \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\beta}_0 \sum_{i=1}^{n} d_{ik} & +\hat{\beta}_1 \sum_{i=1}^{n} d_{ik} d_{i1} & +\hat{\beta}_2 \sum_{i=1}^{n} d_{ik} d_{i2} & +\ldots & +\hat{\beta}_k \sum_{i=1}^{n} d_{ik}^2 & = \sum_{i=1}^{n} d_{ik} \mu_i \end{cases}$$

A solução dessas equações normais dá os estimadores de mínimos quadrados para β. Usando uma notação da álgebra linear, a função de mínimos quadrados para a equação 2.13 será dada por:

$$\begin{aligned} MSQ(\beta) &= \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (\mu - \mathbf{D}\beta)'(\mu - \mathbf{D}\beta) \\ &= \mu'\mu - \beta'\mathbf{D}'\mu - \mu'\mathbf{D}\beta + \beta'\mathbf{D}'\mathbf{D}\beta \\ &= \mu'\mu - 2\beta'\mathbf{D}'\mu + \beta'\mathbf{D}'\mathbf{D}\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \beta} MSQ(\beta) &= \frac{\partial}{\partial \beta} \left(\mu'\mu - 2\beta'\mathbf{D}'\mu + \beta'\mathbf{D}'\mathbf{D}\beta \right) = \\ &= -2\mathbf{D}'\mu + 2\mathbf{D}'\mathbf{D}\hat{\beta} = 0 \end{aligned}$$

assim, $\mathbf{D}'\mathbf{D}\hat{\beta} = \mathbf{D}'\mu$ e portanto $\hat{\beta}$ pode ser estimado como:

$$\hat{\beta} = \left(\mathbf{D}'\mathbf{D} \right)^{-1} \mathbf{D}'\mu$$

Considerando que μ_i representa a média de uma única observação na localização x_i, então esse valor coincide com o valor observado y_i e podemos assim escrever o estimador dos

coeficientes do modelo de tendência como sendo:

$$\hat{\beta} = \left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{Y}$$

Se os dados não são independentes e conhecermos a matriz de covariância associada Σ do modelo (que é nosso caso), então o método será denominado mínimos quadrados generalizados. O modelo será inflacionado na quantidade de parâmetros a serem estimados. Essa estimativa será dada por:

$$\hat{\beta} = \left(\mathbf{D}'\Sigma^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\Sigma^{-1}\mathbf{Y} \tag{2.15}$$

Assumindo-se que $\mathbf{Y}$ tem distribuição normal multivariada, então $\hat{\beta}$ é o estimador de mínimos quadrados para β, com suas importantes propriedades, coincidindo com o estimador de máxima verossimilhança.

Uma vez identificada e modelada a tendência, esta deve ser eliminada do conjunto observado, subtraindo-a deles de seguinte forma:

$$\mathbf{Y}^{*} = \mathbf{Y} - \mathbf{D}\hat{\beta} \tag{2.16}$$

2.3.2 Ajuste de modelo ao semivariograma por mínimos quadrados

Definimos a função semivariância teórica $\gamma(u)$ para o processo gaussiano idealizado pela equação 2.2 como sendo aquela dada pela equação 2.7, ou seja, $\gamma(u) = \tau^2 + \sigma^2(1-\rho(u))$. Já o semivariograma teórico trata-se do gráfico da função semivariância *versus* a distância *u* que separa duas posições.

Para Journel e Huijbregts (1978) o semivariograma é uma ferramenta muito utilizada para representar o mecanismo de dependência espacial. A sua forma padrão pode ser visualizada na figura 2.6.

Nesse gráfico a função semivariância, que é uma função monótona não decrescente,

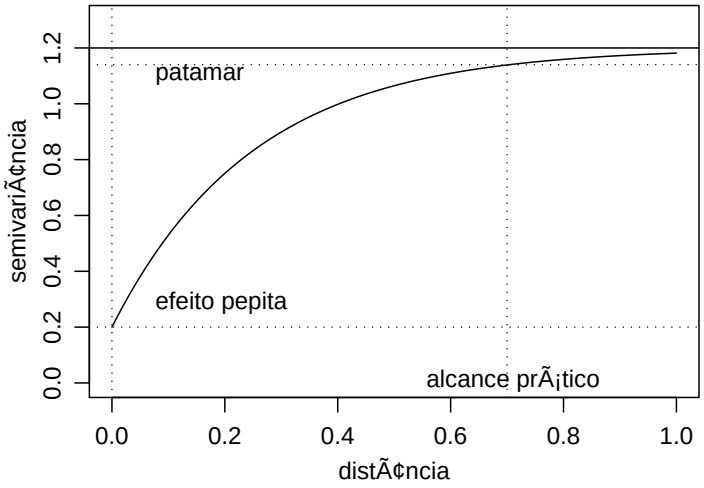

Figura 2.6: Comportamento padrão da função semivariância. O elementos principais que a compõem são: o alcance prático ϕ associado à função de correlação, a variância de pequena escala ou efeito pepita, que corresponde a τ^2 e a contribuição σ^2, ambos presentes na equação 2.7

.

depende somente do comportamento da função de correlação $\rho(u)$. O efeito pepita (*nugget*) representa a variância de pequena escala τ^2. O patamar (*sill*) dado por $\tau^2 + \sigma^2$ representa a variância total do processo Y e o alcance prático de dependência espacial (*range*) é determinado por um parâmetro ϕ que controla o decaimento da função de correlação. Como a função de correlação é assintoticamente decrescente, sua variação será muito pequena para grandes valores de u, podendo ser considerada estável para efeitos práticos. Segundo Diggle e Ribeiro Jr (2007) uma convenção adotada por este modelo é considerar atingido o patamar quando, para um dado u_0, a correlação fica $\rho(u_0) \simeq 0,05$. Não há uma razão científica para se adotar esse valor de corte, pode ser considerada uma quantidade razoável para a estabilização da função de correlação e, consequentemente, da função semivariância. Esse valor u_0 é denominado de alcance prático. Em termos da função semivariância, seu valor é obtido com o valor de u_0 tal que $\gamma(u_0) = \tau^2 + 0,95\ \sigma^2$.

Para a modelagem de um processo gaussiano isotrópico estacionário, o problema se reduz a definir a função de correlação mais apropriada ao fenômeno e estimar os parâmetros μ, τ^2, σ^2 e ϕ, na situação mais simples.

A estimativa de Matheron (MATHERON, 1963) para a semivariância teórica envol-

vendo duas medias do processo Y será: $v_{ij} = \frac{1}{2}(y_i - y_j)^2$, denominado de semivariância experimental ou empírica. Uma área contendo n coordenadas amostrais fornecerá $\binom{n}{2}$ pares do tipo (u_{ij}, v_{ij}). Este será, dependendo do número de coordenadas amostrais, um conjunto muito grande de pares. O seu gráfico é denominado semivariograma experimental e alguns autores o chamam de nuvem variográfica. Seu aspecto é dado pela figura 2.7.

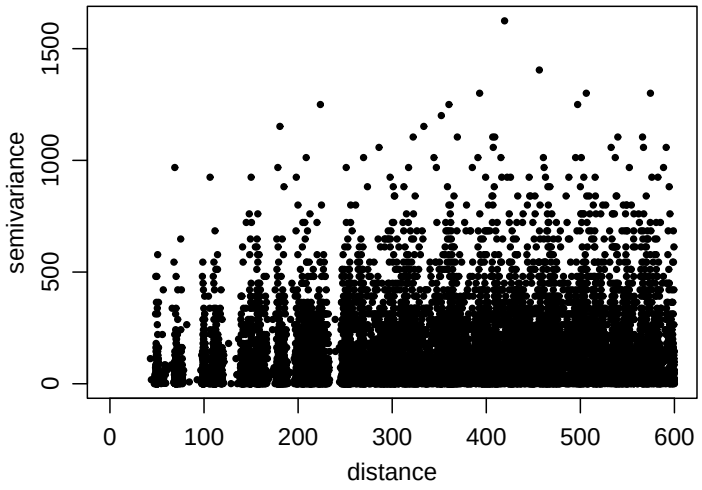

Figura 2.7: Variograma empírico de dados de concentração de cálcio em uma área com 178 pontos amostrais, coletados por pesquisadores do PESAGRO e EMBRAPA-Solos, Rio de Janeiro-RJ (OLIVEIRA, 2003)

Devido ao grande número de pontos no gráfico do semivariograma empírico, bem como a forte dispersão dos pontos à grandes distâncias, ele se torna uma figura de difícil interpretação, no sentido de se tornar difícil aderir visualmente um bom modelo variográfico por seus pontos. Diggle e Ribeiro Jr (2007) dizem que esse comportamento errático se deve ao fato de que a distribuição amostral marginal de cada ordenada v_{ij} ser proporcional a uma distribuição qui-quadrado com 1 grau de liberdade, sendo portanto, fortemente assimétrica e com alto coeficiente de variação.

Visando facilitar o aspecto computacional do processo e ter uma interpretação gráfica plausível, Pannatier (1996) sugeriu dividir em poucos intervalos a variação das distâncias u e representar, no ponto médio de cada intervalo, o valor médio do grupo das semivariâncias relativas a esse intervalo. O semivariograma se reduz a uns poucos pontos, permitindo facilmente o ajuste de um modelo variográfico teórico usando como critério de ajuste, métodos baseados em minimizar o erro médio quadrático, dado pela diferença entre o valor médio de v para uma

distância u_0 representante do intervalo e o valor teórico nessa mesma distância, ou seja, um erro do tipo $(\gamma(u_0) - v(u_0))^2$. O gráfico típico resultante desse procedimento é mostrado na figura 2.8.

O estimador pelo método dos momentos mais utilizado para a semivariância é aquele proposto por Matheron (1962) e definido como:

$$\hat{\gamma}(u) = \frac{1}{|2N(u)|} \sum_{N(u)} \big(y(x_i) - y(x_j)\big)^2 \tag{2.17}$$

onde $N(u) = \big\{(x_i, x_j) : x_i - x_j = u; i, j = 1, 2, \ldots, n\big\}$ é o conjunto das diferentes distâncias u que separam as coordenadas x. Para Braga (1990), se Y for uma função aleatória estacionária, então esse estimador, sob a hipótese intrínseca, é não-tendencioso e não-viciado para para a média mas muito afetado por observações atípicas (outliers).

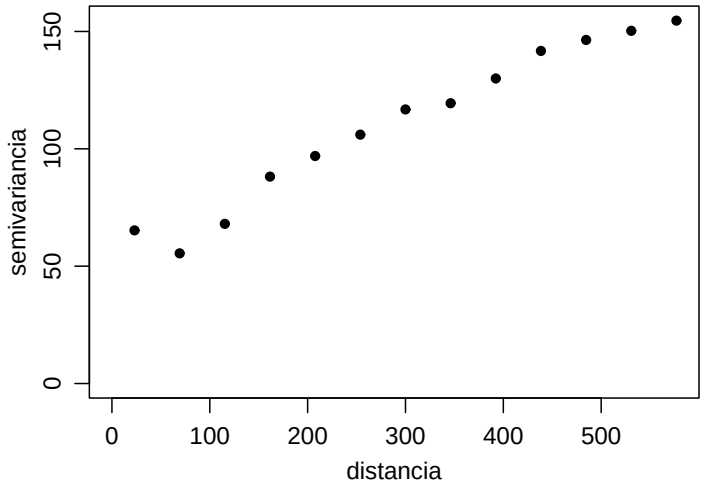

Figura 2.8: Variograma empírico agrupado em classes (“binado”) de dados de concentração de cálcio em uma área com 178 pontos amostrais, coletados por pesquisadores do PESAGRO e EMBRAPA-Solos, Rio de Janeiro-RJ (OLIVEIRA, 2003)

Essa abordagem vem sendo atotada por diversos autores em estudos que envolvem aplicações agrícolas. Reichardt, Vieira e Libardi (1986) estudaram 50 dados de pH de solo, de amostras coletadas com espaçamento de 1 m, em transecção de um área de Latossolo Vermelho-escuro orto localizado em Araras-SP, cultivada com cultura de cana-de-açúcar. A técnica de autocorrelação que empregaram nos dados mostrou que observações de pH eram correlacionadas espacialmente até uma distância de 5 m. Observaram ainda que, para as amostras serem

consideradas independentes e completamente casualizadas, deveriam ser espaçadas de, pelo menos, 10 m. Com seu trabalho, os autores concluíram que a variabilidade espacial do solo pode ser definida corretamente e que a geoestatística era a alternativa certa às metodologias tradicionais.

Prevedello (1987) estudou a magnitude da variabilidade espacial de 47 parâmetros (físicos e químicos) de um solo com Terra Roxa Estruturada, em uma área de 4810 m^2, em Piracicaba-SP, onde foi aplicado o manejo de uma cultura de arroz de sequeiro. O autor utilizou em seu experimento uma estrutura regular de 4x13, totalizando 52 pontos amostrais, separados 10 m entre si. Avaliou e discutiu a dependência espacial pela análise do autocorrelograma e do semivariograma, usando o estimador clássico de Matheron. Assim, com o emprego da teoria das variáveis regionalizadas, estabeleceu subunidades de amostragem ou de manejo individualizado, considerando-as independentes. Concluiu ainda que a área total não se mostrou homogênea para nenhum dos 47 parâmetros estudados, contrariando o que havia inicialmente suposto.

Mohamed, Evans e Shiel (1996) usaram a geoestatística para examinar a variabilidade geográfica em uma área de terra e descobrir, pela distribuição espacial a melhor densidade amostral, no sentido de obterem as propriedades de colheita e distribuição das características do solo com poucas amostras. Com o emprego do semivariograma experimental determinado pelo estimador clássico de Matheron, detectaram uma estrutura de variabilidade no solo. Com isso puderam utilizar seus parâmetros para efetuarem a interpolação de dados para produção de mapas de contornos.

Yang et al. (1998) estudaram a influência da topografia no rendimento da colheita, pela variabilidade de cinco campos em declive, da região de Palouse, em Washington-USA. Os autores desenvolveram um sistema de informações geográficas (GIS) para o manejo e análise do rendimento de trigo, juntamente com informações georreferenciadas sobre a variabilidade da topografia. Identificaram também o padrão de variabilidade do rendimento do trigo dentro de cada região plantada, para cada uma das cinco regiões estudadas e avaliaram a relação entre

rendimento e atributos de topografia. Descreveram o padrão de variabilidade espacial pelo semivariograma, que mostraram claramente uma estrutura de dependência espacial justificando o emprego do manejo localizado.

2.3.3 Ajuste de modelos e estimação dos parâmetros por máxima verossimilhança

Considerando o caso estacionário do modelo geoestatístico univariado dado pela equação 2.2, onde o processo $S(x_i)$ pode ser escrito como um conjunto de observações Y com distribuição de probabilidades de acordo com a equação 2.3, os parâmetros gerais do modelo a serem estimados são: $\Theta = (\beta, \sigma^2, \phi, \tau^2)$ onde, como já foi dito, ϕ é um parâmetro da função de correlação.

A variável aleatória $\mathbf{Y} = \{Y(x_1), Y(x_2), \ldots, Y(x_n)\}$, que representa um conjunto de realizações em n coordenadas, forma um processo gaussiano multivariado, ou seja, $\mathbf{Y} \sim NMV(\mu; \Sigma)$ onde μ é um vetor de números reais, todos iguais a μ e Σ é a matriz de variâncias e covariâncias de tamanho $n \times n$, com as propriedades de ser simétrica e definida positiva. Então, a distribuição conjunta de $\mathbf{Y}$, segundo (DUDEWICZ; MISHRA, 1988) será:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}}\, e^{-\frac{1}{2}(\mathbf{y}-\mu)'\blacksquare^{-1}(\mathbf{y}-\mu)}$$

para todo vetor $\mathbf{y}$ de números reais.

Sendo Y um processo gaussiano correlacionado, sua função de verossimilhança será composta pela sua distribuição conjunta de probabilidades dada por:

$$L(\theta) = f(\theta; \mathbf{y}) = \frac{1}{(2\pi)^{n/2}(|\sigma^2\mathbf{R} + \tau^2\mathbf{I}|)^{1/2}}\, exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{D}\beta)'(\sigma^2\mathbf{R} + \tau^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{D}\beta)\right\},$$

A função de log-verossimilhança será:

$$l(\theta) = -\frac{1}{2}\log(2\pi)^n - \frac{1}{2}\log(|\sigma^2\mathbf{R} + \tau^2\mathbf{I}|) - \frac{1}{2}(\mathbf{y} - \mathbf{D}\beta)'(\sigma^2\mathbf{R} + \tau^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{D}\beta)$$

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\sigma^2\mathbf{R} + \tau^2\mathbf{I}|) + \right.$$
$$\left. + (\mathbf{y} - \mathbf{D}\beta)'(\sigma^2\mathbf{R} + \tau^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{D}\beta)\right] \quad (2.18)$$

Fazendo $\frac{\tau^2}{\sigma^2} = \nu^2$ então $Var(\mathbf{Y}) = \Sigma = \sigma^2\mathbf{R} + \tau^2\mathbf{I} = \sigma^2\left(\mathbf{R} + \frac{\tau^2}{\sigma^2}\mathbf{I}\right) = \sigma^2\mathbf{V}$

Substituindo $\sigma^2\mathbf{R} + \tau^2\mathbf{I}$ por $\sigma^2\mathbf{R}$ na equação(2.18), vem:

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\sigma^2\mathbf{V}|) + (\mathbf{y} - \mathbf{D}\beta)'(\sigma^2\mathbf{V})^{-1}(\mathbf{y} - \mathbf{D}\beta)\right] \quad (2.19)$$

Agora substituindo $\sigma^2\mathbf{R} + \tau^2\mathbf{I}$ por Σ na mesma equação (2.18), temos:

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\Sigma|) + (\mathbf{y} - \mathbf{D}\beta)'(\Sigma)^{-1}(\mathbf{y} - \mathbf{D}\beta)\right] \quad (2.20)$$

Desenvolvendo os produtos matriciais da equação 2.20 resulta em:

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\Sigma|) + \mathbf{y}'\Sigma^{-1}\mathbf{y} - 2\mathbf{y}'\Sigma^{-1}\mathbf{D}\beta + \beta'\mathbf{D}'\Sigma^{-1}\mathbf{D}\beta\right] \quad (2.21)$$

onde $\mathbf{y}'\Sigma^{-1}\mathbf{D}\beta$ é um escalar pois $\mathbf{y}_{1\times n}$, $\Sigma_{n\times n}$, $\mathbf{D}_{n\times n}$ e $\beta_{n\times 1}$.

Segundo (KOLMAN, 1997), se $\mathbf{A}$ é uma matriz quadrada simétrica definida positiva e $\beta = [\beta_1, \beta_2, ..., \beta_n]'$ um vetor, então:

a) $\frac{\partial \mathbf{Ax}}{\partial \beta} = \mathbf{A}'$ (transposta)

b) $\frac{\partial \mathbf{x}'\mathbf{Ax}}{\partial \mathbf{x}} = 2\mathbf{Ax}$ (forma quadrática)

Desses resultados obtemos a derivada parcial da função de log-verossimilhança de θ com relação a β

$$\frac{\partial l(\theta)}{\partial \beta} = -\frac{1}{2}(-2(\mathbf{y}'\Sigma^{-1}\mathbf{D})' + 2(\mathbf{D}'\Sigma^{-1}\mathbf{D})\beta) = \mathbf{D}'\Sigma^{-1}\mathbf{y} - \mathbf{D}'\Sigma^{-1}\mathbf{D}\beta$$

Se $\frac{\partial l(\theta)}{\partial \beta} = \mathbf{0}$ então teremos que $\mathbf{D}'\Sigma^{-1}\mathbf{y} - \mathbf{D}'\Sigma^{-1}\mathbf{D}\hat{\beta} = \mathbf{0}$ e assim obtemos o estimador

para o parâmetro β dado por:

$$\hat{\beta} = (\mathbf{D}'\Sigma^{-1}\mathbf{D})^{-1}\mathbf{D}'\Sigma^{-1}\mathbf{y} \tag{2.22}$$

Considerando também que $\Sigma = \sigma^2\mathbf{V}$ e que $|\Sigma| = (\sigma^2)^n|\mathbf{V}|$, então a equação 2.19 fica:

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\sigma^2\mathbf{V}|) + \mathbf{y}'(\sigma^2\mathbf{V})^{-1}\mathbf{y} - \beta'\mathbf{D}'(\sigma^2\mathbf{V})^{-1}\mathbf{D}\beta\right]$$

$$\begin{aligned} l(\theta) &= -\frac{1}{2}\left[n\log(2\pi) + \log[(\sigma^2)^n|\mathbf{V}|] + \frac{\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}}{\sigma^2} - 2\frac{\mathbf{y}'\mathbf{V}^{-1}\mathbf{D}\beta}{\sigma^2} + \frac{\beta'\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\beta}{\sigma^2}\right] \\ &= -\frac{1}{2}\left[n\log(2\pi) + n\log(\sigma^2) + \log|\mathbf{V}| + \frac{(\mathbf{y}-\mathbf{D}\beta)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{D}\beta)}{\sigma^2}\right] \end{aligned}$$

onde $[(\mathbf{y}-\mathbf{D}\beta)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{D}\beta)]/\sigma^2$ é uma soma de quadrados ponderada pela matriz de covariâncias.

Calculando a derivada de $l(\theta)$ com relação a σ^2 obtemos:

$$\frac{\partial l(\theta)}{\partial\sigma^2} = -\frac{1}{2}\left[\frac{n}{\sigma^2} - \frac{(\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{D}\hat{\beta})}{(\sigma^2)^2}\right]$$

Se $\frac{\partial l(\theta)}{\partial\sigma^2} = 0$ e considerando o vetor de parâmetros $(\beta, \sigma^2, \phi, \nu^2)'$, tem-se:

$$\begin{aligned} -\frac{n}{\hat{\sigma^2}} + \frac{(\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{D}\hat{\beta})}{(\sigma^2)^2} &= 0 \\ \frac{(\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{D}\hat{\beta})}{(\hat{\sigma}^2)^2} &= n \end{aligned}$$

$$\hat{\sigma}^2_{\phi,\nu^2} = \frac{(\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}_{\phi,\nu^2}(\mathbf{y}-\mathbf{D}\hat{\beta})}{n} \tag{2.23}$$

Retomando a equação (2.22) e substituindo Σ por $\sigma^2\mathbf{V}$ teremos:

$$\begin{aligned}\hat{\beta} &= \left(\frac{\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}}{\sigma^2}\right)^{-1}\frac{\mathbf{D}\mathbf{V}^{-1}\mathbf{y}}{\sigma^2} \\ &= (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}\sigma^2\frac{\mathbf{D}\mathbf{V}^{-1}\mathbf{y}}{\sigma^2} \\ &= (\mathbf{D}^{-1}\mathbf{V}^{-1}_{\phi,\nu^2}\mathbf{D})^{-1}\mathbf{D}\mathbf{V}^{-1}_{\phi,\nu^2}\mathbf{y} \qquad (2.24)\end{aligned}$$

que depende somente dos parâmetros ϕ e ν^2. Neste caso, a matriz de correlação será dada por:

$$\mathbf{V} = \begin{pmatrix} 1+\nu^2 & \rho(u_{12}) & \dots & \rho_(u_{1n}) \\ \rho(u_{21}) & 1+\nu^2 & \dots & \rho(u_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(u_{n1}) & \rho(u_{n2}) & \dots & 1+\nu^2 \end{pmatrix}$$

A função log-verossimilhança concentrada será então dada por:

$$\begin{aligned}l(\phi,\nu^2) &= -\frac{1}{2}\left[n\log(2\pi)+n\log\left(\frac{(\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}_{\phi,\nu^2}(\mathbf{y}-\mathbf{D}\hat{\beta})}{n}\right)+\log|\mathbf{V}|+\right. \\ &\quad \left.+\frac{(\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{D}\hat{\beta})}{\left(\frac{\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{D}\hat{\beta}}{n}\right)}\right] \\ l(\phi,\nu^2) &= -\frac{1}{2}\left[n\log(2\pi)+n\log\left(\frac{(\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}_{\phi,\nu^2}(\mathbf{y}-\mathbf{D}\hat{\beta})}{n}\right)+\log|\mathbf{V}|+n\right] \\ l(\phi,\nu^2) &= -\frac{1}{2}\left[n\log(2\pi)+n\log\left((\mathbf{y}-\mathbf{D}\hat{\beta})'\mathbf{V}^{-1}_{\phi,\nu^2}(\mathbf{y}-\mathbf{D}\hat{\beta})\right)-n\log n+\right. \\ &\quad \left.+\log|\mathbf{V}|+n\right] \qquad (2.25)\end{aligned}$$

Para um modelo estacionário, a menos das constantes, a função $l(\phi,\nu^2)$ fica:

$$l(\phi,\nu^2) \propto -\frac{n}{2}\left((\mathbf{y}-\mu)'\mathbf{V}^{-1}_{\phi,\nu^2}(\mathbf{y}-\mu)\right)-\frac{\log|\mathbf{V}|}{2} \qquad (2.26)$$

Esta função recebe como argumentos, o vetor das observações do processo $\mathbf{Y}$ e a matriz

das distâncias de cada coordenada com as demais, qua permitirá obter **V** pela escolha conveniente de uma função de correlação $\rho(u_{ij})$. A maximização dessa função segundo os parâmetros envolvidos, fornecerá o modelo de correlação espacial com a estimativa de seus parâmetros.

Funções côncavas são aquelas cujo gráfico está sempre acima ou sobre qualquer corda traçada numa região entre seus pontos, ou, equivalentemente, seu gráfico está abaixo da reta tangente ao seu ponto de máximo. Neste sentido, tanto a função de verossimilhança quanto a função log-verossimilhança são funções côncavas, garantindo assim a existência de um ponto de máximo local.

Para obtermos a melhor estimativa para os parâmetros, devemos encontrar simultaneamente o valor dos parâmetros que irão maximizar essa função. Muitos programas computacionais, incluindo o geoR (Ribeiro Jr; DIGGLE, 2001), possuem algoritmos eficientes para estimar esses parâmetros. A questão importante a se destacar aqui é que esse método, usado para aderir um modelo teórico com a melhor estimativa de seus parâmetros, envolve todas as observações amostrais, sem a necessidade dos agrupamentos feito nos ajustes através de variogramas, evitando os erros decorrentes.

No caso de um processo gaussiano a função log-verossimilhança é facilmente obtida, mas nem sempre será tão simples. O efeito da transformação de variáveis proposta por Box e Cox (1964) pela equação 2.8 contorna o problema para as distribuições assimétricas e/ou com a presença de valores discrepantes, mas para distribuições leptocúrticas, como é o caso da distribuição *t-Student*, poderemos ter dificuldades na sua construção. O desejável seria desenvolver o método para outras famílias de distribuições, permitindo assim maior flexibilidade na definição da distribuição de probabilidades envolvida no processo, mas não seguiremos essa linha neste trabalho.

Outra restrição no uso do método da otimização da função log-verossimilhança está relacionada à forma suave de variação de certas funções de correlação, ou seja, aquelas funções que são diferenciáveis um número grande de vezes. Nestes casos, a matriz de correlação poderá apresentar colunas muito parecidas numericamente, impossibilitando numericamente sua

inversão. Muitos pesquisadores atualmente envolvem em seus trabalhos, a escolha de modelo de correlação e ajuste dos parâmetros por este método.

2.4 PREDIÇÃO LINEAR ESPACIAL

Um aspecto importante da modelagem estatística é a utilização do modelo obtido para efetuar predições. Empregamos aqui o termo predição como sendo uma conjectura ou suposição sobre um resultado de Y desconhecido que poderá ou não acontecer. A meta é realizar boas estimativas de quantidades que variam continuamente no espaço, em função de um conjunto discreto de observações obtidas dispersamente em uma área. Esse procedimento, sob certas circunstâncias, é chamado krigagem, termo este criado por G. Matheron em reconhecimento ao trabalho do engenheiro de minas D. G. Krige (KRIGE, 1951), sendo a krigagem ordinária a mais utilizada. O método estima um valor em um ponto arbitrário de uma região fechada onde a função de correlação do processo é conhecida, empregando o conjunto de pontos amostrais conhecidos, distribuídos pela área.

Isaaks e Srivastava (1989) citam vários métodos de estimação pontual como: método poligonal de desagrupamento, método da triangulação, método do inverso do quadrado das distâncias, método dos vizinhos mais próximos. Mas para eles, a krigagem ordinária é tida como um método BLUE, acrónimo do inglês *Best Linear Unbiased Estimator* – melhor estimador não viciado e de variância mínima. Segundo eles, o método é linear porque seus estimadores são feitas a partir de combinações lineares sobre as observações amostrais disponíveis, é não viciado pois o erro médio residual é zero e “melhor” porque dentre outros estimadores é o que leva à menor variância do erro.

Segundo esse autor, em uma coordenada arbitrária, digamos x_0, as estimativas serão dadas por:

$$\hat{y}(x_0) = \sum_{i=1}^{n} \omega_i y(x_i)$$

onde os x_0 são as coordenadas onde se deseja efetuar uma estimativa, ω_i é o peso associado

à i-ésima observação $y(x_i)$, sujeito à restrição $\Sigma\omega_i = 1$, que garante a não tendenciosidade do preditor.

Journel e Huijbregts (1978) também salientaram que, no caso de processos não-estacionários, serão necessárias algumas condições de ausência de viés. Para eles a limitação à classe de estimadores lineares é natural, uma vez que são necessários somente os momentos de segunda ordem da função de covariância.

Schabenberger e Gotway (2005) fazem distinção estre estimação e predição, por serem expressões muitas vezes tidas como equivalentes. Em um modelo básico de regressão linear simples $Y(x_i) = \beta_0 + \beta_1 S(x_i) + \varepsilon_i$, os erros ε_i não são correlacionados (são independentes) e os coeficientes β_0 e β_1 são estimados (por métodos de mínimos quadrados, conforme equação 2.15) e se prediz o valor $\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 S(x_0)$. Não fica claro se $\hat{Y}(x_0)$ é um preditor de $Y(x_0)$ como uma "resposta" em x_0 ou é um estimador de $E[Y(X_0)]$. Apesar de que estimar uma quantidade fixa ou predizer uma quantidade aleatória ser uma questão menor, sua importância fica clara ao se considerar uma incerteza associada a essas quantidades. No caso da geoestatística, apesar do total desconhecimento do processo $S(x)$, a aplicações com predição são frequentemente mais empregadas do que aquelas que buscam a estimação de uma média.

O modelo de predição linear, — sinônimo de krigagem, dependendo se a média do processo é ou não conhecida, proposto por esses autores é dado por:

$$\hat{Y}(x_0) = \hat{\mu} + \mathbf{r}'\Sigma^{-1}(Y(x) - \hat{\mu})$$

onde $\mathbf{r} = Cov(Y(x), Y(x_0))$ e Σ é a matriz de variâncias e covariâncias das variáveis observadas.

A variância da predição, segundo eles, será:

$$Var(\hat{Y}(x_0)) = \sigma^2 - \mathbf{r}'\Sigma^{-1}\mathbf{r} + \frac{(1 - \mathbf{1}'\Sigma^{-1}\mathbf{r})^2}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}$$

Segundo Goovaerts (1997), o estimador de krigagem é um estimador de regressão linear $\hat{S}(x)$ definido como:

$$\begin{aligned}\hat{S}(x) &= \mu(x) + \sum_{i=1}^{n} \lambda_i(x)\,[Y_i - \mu(x)] \\ &= \mu(x) + \sum_{i=1}^{n} \lambda_i(x) Y_i - \sum_{i=1}^{n} \lambda_i(x)\mu(x) \\ &= \left(1 - \sum_{i=1}^{n} \lambda_i(x)\right)\mu(x) + \sum_{i=1}^{n} \lambda_i(x) Y_i\end{aligned}$$

onde $\mu(x)$ é a função média, $\mathbf{Y}$ o vetor de observações e $\lambda(x)$ a função peso.

Tomando novamente $S(x)$ um processo estacionário e $\mathbf{Y}$ um vetor de variáveis aleatórias cujos valores são observáveis e T outra variável aleatória, cujo valor desejamos estimar, $\mathbf{Y}$ terá distribuição normal multivariada com média constante $\mu\mathbf{1}$ e variância $\sigma^2\mathbf{R} + \tau^2\mathbf{I}$. $T = T(S)$ é a meta de predição. Se $T = S(x_0)$ então a distribuição conjunta de $\mathbf{T}$ e $\mathbf{Y}$ será normal multivariada e a distribuição condicional de $\mathbf{T}$ dado $\mathbf{Y=y}$ será normal com média $\mu_T + \rho_{TY}\left(\frac{\sigma_T}{\sigma_Y}\right)(y - \mu_Y)$ e variância $\sigma_T^2(1 - \rho_{TY}^2)$. Em notação matricial podemos escrever:

$$(T,Y) \sim MVN\left(\mu\mathbf{1}, \begin{bmatrix} \sigma^2 & \sigma^2\mathbf{r}' \\ \sigma^2\mathbf{r} & \sigma^2\mathbf{R} + \tau^2\mathbf{I} \end{bmatrix}\right) \text{ e } (T|Y) \sim MVN\,[E(T|Y); Var(T|Y)]$$

Para Diggle e Ribeiro Jr (2007) o estimador pontual $\hat{T} = E[T|Y]$ será o valor que minimiza o erro médio quadrático $MSE(\hat{T}) = E(\hat{T} - T)^2$ e assim eles escrevem:

- $\hat{T}(x_0) = E(T|Y) = \mu + \mathbf{r}'\mathbf{V}^{-1}(\mathbf{Y} - \mu\mathbf{1})$
- $Var(\hat{T}(x_0)) = Var(T|Y) = \sigma^2\left(1 - \mathbf{r}'\mathbf{V}^{-1}\mathbf{r}\right)$

onde $\mathbf{V} = \sigma^2\mathbf{R} + \tau^2\mathbf{I}$ e $\mathbf{r}$ é o vetor de correlação entre a posição dos valores observados e a posição do valor y_0 a ser predito.

No caso do valor de μ ser desconhecido, então ele poderá ser estimado por:

$$\hat{\mu} = (1'\mathbf{V}^{-1}1)^{-1}1'\mathbf{V}^{-1}\mathbf{Y}$$

2.5 PROCESOS ESTOCÁSTICOS ESPACIAIS MULTIVARIADOS

Para Ver Hoef e Cressie (1993), em ciências da terra é frequente o interesse em predizer conjuntamente uma grande quantidade de variáveis. Normalmente se prediz uma variável por vez, usando dados de um mesmo tipo (krigagem) ou utilizando informações adicionais de outra variável tomada nas mesmas coordenadas (krigagem com co-variável). O modelo bivariado mostra que a predição de uma variável com base em uma outra variável correlacionada, mas em locais diferentes (cokrigagem) resulta em predições menos precisas. Predições espaciais multivariadas permitem construir regiões de predição multivariada. Esses autores relacionam e comparam predições baseadas no variograma cruzado, predições espaciais multivariadas e estimação de parâmetros por mínimos quadrados generalizados.

Os modelos geoestatísticos multivariados dizem respeito a um conjunto de variáveis aleatórias gaussianas dadas por:

$$\left\{Y_1(\mathbf{x}), Y_2(\mathbf{x}), \ldots, Y_p(\mathbf{x}) : Y_k(\mathbf{x}) \in S_k(\mathbf{x}); x_i \in \mathbb{R}^2; i = 1, 2, \ldots, n\right\} \tag{2.27}$$

Essas varíaveis são georreferenciadas, tomadas em uma mesma região geográfica, todas com igual interesse científico e com distribuição conjunta de probabilidades. É uma situação pouco realística pois esta descrição não leva a uma interpretação física no sentido prático, entretanto o será se descrevermos a distribuição condicional de uma das variáveis, eleita de interesse primário, condicionada a uma ou mais variáveis espacialmente localizadas. Neste caso exige-se que todas as variáveis sejam tomadas nas mesmas posições geográficas e que haja uma certa correlação entre elas. Outra situação prática ocorre quando a variável primária for de difícil aquisição, então, poderemos formar um conjunto das variáveis restantes, – supostamente de fácil observação, como o conjunto de variáveis preditoras que, modeladas adequadamente, permitirão fazer estimativas da variável primária em locais onde foram obtidas as demais variáveis. Neste caso, as variáveis podem ser em quantidades, tipos e localizações diferentes. Pretendemos aqui abordar ambos os casos e ainda utilizar o suporte da análise multivariada de componentes principais (ACP) para a redução do número de variáveis envolvidas no problema. Inicialmente

apresentaremos o problema geoestatístico multivariado envolvendo duas variáveis, sendo uma a principal e a outra, secundária. Estenderemos, a seguir, o caso de uma variável primária e um conjunto de variáveis secundárias.

2.5.1 Modelos geoestatístico bivariado

Consideremos o seguinte processo gaussiano estacionário bivariado $\{\mathbf{S}(\mathbf{x}) = S_1(x_i), S_2(x_j) : x_i, x_j \in \mathbb{R}; i = 1,2,\ldots,r; j = 1,2,\ldots,s\}$, com $E(S_1(x_i)) = 0$; $E(S_2(x_i)) = 0$ e $Var(S_1(x_i)) = \sigma_1^2$ e $Var(S_2(x_i)) = \sigma_2^2$. A matriz de covariância será dada por:

$$\Sigma = \left(\begin{array}{c|c} Cov(S_1;S_1) & Cov(S_1;S_2) \\ \hline Cov(S_2;S_1) & Cov(S_2;S_2) \end{array} \right)$$

Expandindo essa matriz temos:

$$\Sigma = \left(\begin{array}{cccc|cccc} \sigma_{1,1}^{(1,1)} & \sigma_{1,2}^{(1,1)} & \ldots & \sigma_{1,r}^{(1,1)} & \sigma_{1,1}^{(1,2)} & \sigma_{1,2}^{(1,2)} & \ldots & \sigma_{1,s}^{(1,2)} \\ \sigma_{2,1}^{(1,1)} & \sigma_{2,2}^{(1,1)} & \ldots & \sigma_{2,r}^{(1,1)} & \sigma_{2,1}^{(1,2)} & \sigma_{2,2}^{(1,2)} & \ldots & \sigma_{2,s}^{(1,2)} \\ \ldots & \ldots & & \ldots & \ldots & \ldots & & \ldots \\ \sigma_{r,1}^{(1,1)} & \sigma_{r,2}^{(1,1)} & \ldots & \sigma_{r,r}^{(1,1)} & \sigma_{r,1}^{(1,2)} & \sigma_{r,2}^{(1,2)} & \ldots & \sigma_{r,s}^{(1,2)} \\ \hline \sigma_{1,1}^{(2,1)} & \sigma_{1,2}^{(2,1)} & \ldots & \sigma_{1,r}^{(2,1)} & \sigma_{1,1}^{(2,2)} & \sigma_{1,2}^{(2,2)} & \ldots & \sigma_{1,s}^{(2,2)} \\ \sigma_{2,1}^{(2,1)} & \sigma_{2,2}^{(2,1)} & \ldots & \sigma_{2,r}^{(2,1)} & \sigma_{2,1}^{(2,2)} & \sigma_{2,2}^{(2,2)} & \ldots & \sigma_{2,s}^{(2,2)} \\ \ldots & \ldots & & \ldots & \ldots & \ldots & & \ldots \\ \sigma_{s,1}^{(2,1)} & \sigma_{s,2}^{(2,1)} & \ldots & \sigma_{s,r}^{(2,1)} & \sigma_{s,1}^{(2,2)} & \sigma_{s,2}^{(2,2)} & \ldots & \sigma_{s,s}^{(2,2)} \end{array} \right)$$

Nesta matriz, o bloco superior esquerdo representa as autocovariancias da variável Y_1 e o bloco inferior direito as autocovariâncias das variável Y_2. Os blocos superior direito e inferior esquerdo representam as covariâncias entre as variáveis Y_1 e Y_2. Os índices à acima dos elementos da matriz representam as variáveis envolvidas e os índices à abaixo correspondem às

localizações. Assim, o elemento $\sigma_{4,3}^{(2,1)}$ representa a covariância entre a variável Y_2 medida na localização x_4 e Y_1 medida na localização x_3. De uma forma geral, essa matriz de covariâncias não estabelece que as coordenadas devam ser totalmente ou parciamente coincidentes. Iremos considerar, doravante, a notação x_i para a i-ésima coordenada da variável Y_1 e x'_j a j-ésima coordenada da variável Y_2.

Modelos bivariados podem ser escritos como uma junção de modelos univariados.

$$\begin{cases} Y_{1,i} = \mu_1 + S_1(x_i) + \tau_1 Z & i = 1,2,\ldots,m \\ Y_{2,j} = \mu_2 + S_2(x_j) + \tau_2 Z & j = 1,2,\ldots,n \end{cases}$$

Consideraremos aqui que as variáveis Y_1 e Y_2 não precisarão ser, necessariamente, co-localizados e nem tomadas o mesmo número de vezes, ou seja, podem ou não serem coincidentes na área (ver figura 2.9). Z representa erro gaussiano aleatório de média zero e variância unitária. $S(x)$ é um processo gaussiano multivariado com vetor média $\mu\mathbf{1}$ e variância $\sigma^2\mathbf{R}$. Esse fica então:

$$\begin{cases} Y_{1,i} = \mu_1 + \sigma_1 R_1(x_i) + \tau_1 Z_i & i = 1,2,\ldots,m \\ Y_{2,j} = \mu_2 + \sigma_1 R_2(x_j) + \tau_2 Z_j & j = 1,2,\ldots,n \end{cases}$$

Devemos aqui considerar 4 possibilidades distintas para modelos assim especificados, considerando as características de seus elementos, assumindo que Y_1 e Y_2 ocorrem simultaneamente em uma mesma área de um espaço bidimensional:

a) Sendo $\tau_1 = \tau_2 = 0$ e $R_1(x)$ independente de $R_2(x)$, então Y_1 será independente de Y_2, ou seja, não serão correlacionados. Um problema escrito desta maneira, exigirá a estimação de 3 parâmetros: σ_1, ϕ_1 em $R_1(x)$ e ϕ_2 em $R_2(x)$.

b) Sendo $\tau_1 = \tau_2 = 0$ e $R_1(x)$ idêntico a $R_2(x)$, então Y_1 será perfeitamente correlacionado com Y_2. Um problema escrito desta maneira, exigirá a estimação de 2 parâmetros: σ e ϕ em $R(x)$.

c) Sendo $\tau_1 \neq \tau_2$ e $R_1(x)$ idêntico a $R_2(x)$ então Y_1 será parcialmente correlacionado com Y_2,

provocando uma dispersão difusa, dependendo da variância σ^2. Um problema modelado desta maneira, exigirá a estimação de 4 parâmetros: τ_1, τ_1, σ, ϕ em $R(x)$.

d) Numa situação mais mais realística, os modelos poderiam ser escritos como:

$$\begin{cases} Y_{1,j} = \mu_1 + \sigma_{0,1}R_0(x_i) + \sigma_1 R_1(x_i) \quad i = 1,2,\ldots,m \\ Y_{2,j} = \mu_2 + \sigma_{0,2}R_0(x_j) + \sigma_2 R_2(x_i) \quad j = 1,2,\ldots,n \end{cases} \tag{2.28}$$

Aqui teremos Y_1 correlacionado com Y_2 devido a presença de uma mesma matriz de correlação $R_0(x)$ em ambos os modelos. Um problema escrito desta maneira, exigirá a estimação de 9 parâmetros: μ_1, μ_2, $\sigma_{0,1}$, $\sigma_{0,2}$, σ_1, σ_2, ϕ_1 (em $R_0(x)$), ϕ_2 (em $R_1(x)$) e ϕ_3 (em $R_2(x)$). Supomos aqui que $\tau_1 = \tau_2 = 0$ então, neste caso, o semivariograma inicia no zero, mas esses parâmetros poderiam estar presentes no modelo, levando à necessidade de estimar um total de 11 parâmetros.

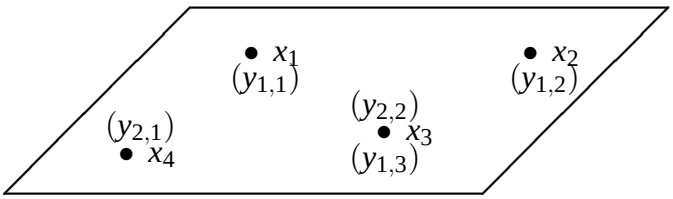

Figura 2.9: Representação ilustrativa de uma área típica com processos geoestatísticos bivariados contendo quatro localizações amostrais, onde as variáveis não são co-localizadas e nem oferecem o mesmo número de observações

Mood, Graybill e Boes (1974) definem a covariância entre duas variáveis aleatórias, digamos Y_1 e Y_2 como sendo:

$$Cov[Y_1(x);Y_2(x)] = E[(Y_1(x) - \mu_{Y_1})(Y_2(x) - \mu_{Y_2})]$$

onde $\mu_{Y_1} = E(Y_1(x))$ e $\mu_{Y_2} = E(Y_2(x))$ e define o coeficiente de correlação entre elas como sendo:

$$\rho_{Y_1;Y_2} = \frac{Cov[Y_1(x);Y_2(x)]}{\sigma_{Y_1}\sigma_{Y_2}}$$

onde $\sigma^2_{Y_1} = Var(Y_1(x))$ e $\sigma^2_{Y_2} = Var(Y_2(x))$.

Vamos aqui determinar, numa notação compatível com Goovaerts (1997), uma função

que estimará a correlação entre as variáveis Y_1 e Y_2 (nesta ordem), quando elas estiverem separadas por uma mesma distância, digamos, $h_k; k = 1,2,\ldots,s$ sendo s a quantidade de pares que suportem essa distância. Idealizamos então um vetor $\mathbf{h} = (h_1, h_2, \ldots, h_s)'$ contendo todas as distâncias possíveis para o cenário contento ambas as variáveis. A figura 2.10 ilustra essa intenção.

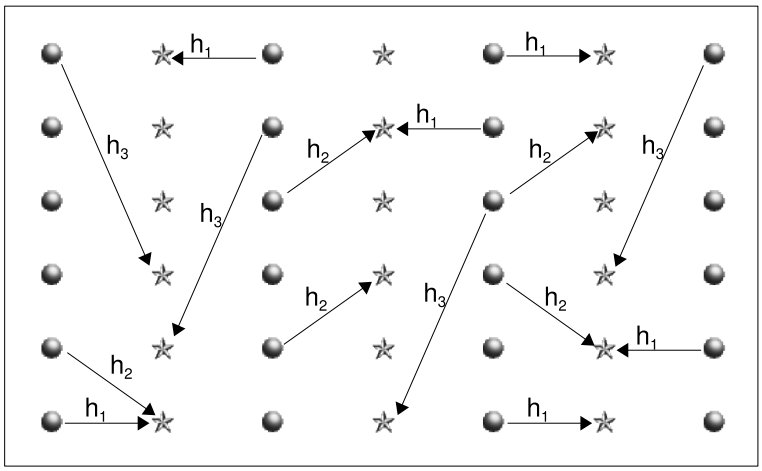

Figura 2.10: Grid regular com locação amostral de duas variáveis sendo os círculos a primeiro e as estrelas a segunda. As setas estabelecem a direção das correlações e os h, através de seus índices indicam o grupo de correlações entre variáveis separadas por uma mesma distância.

Ele define a função covariância como:

$$C_{1;2}(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{k=1}^{N(\mathbf{h})} y_1(x_k) y_2(x'_k) - \mu_1 \mu_2$$

onde $\hat{\mu}_1 = \frac{1}{N(\mathbf{h})} \sum_{k=1}^{N(\mathbf{h})} y_1(x_k)$ e $\hat{\mu}_2 = \frac{1}{N(\mathbf{h})} \sum_{k=1}^{N(\mathbf{h})} y_2(x'_k)$ e $N(\mathbf{h})$ é o número de pares pertencentes à mesma classe de distâncias e direção, $\hat{\mu}_1$ e $\hat{\mu}_2$ são respectivamente as médias de Y_1 e Y_2 nas suas respectivas coordenadas do conjunto formado pelas distâncias $\mathbf{h}$. Se tomarmos como exemplo a figura 2.10 para a distância h_1, $\hat{\mu}_1$ seria a média das observações $y(x_i)$ (círculos) do conjunto dessas distâncias e $\hat{\mu}_2$ a média das observações $y(x'_i)$ (estrelas) do mesmo conjunto.

A covariância obtida para essas diferentes distâncias é chamada de função covariância cruzada experimental. De maneira geral $C_{(1;2)}(\mathbf{h}) \neq C_{(1;2)}(-\mathbf{h})$.

A estimativa do correlograma cruzado será dada por:

$$\rho_{1;2}(\mathbf{h}) = \frac{C_{(1;2)}(\mathbf{h})}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

onde $\sigma_1^2 = \frac{1}{N(\mathbf{h})} \sum_{k=1}^{N(\mathbf{h})} (y_1(x_k) - \hat{\mu}_1)^2$ e $\sigma_2^2 = \frac{1}{N(\mathbf{h})} \sum_{k=1}^{N(\mathbf{h})} (y_2(x'_k) - \hat{\mu}_2)^2$ sendo que σ_1^2 e σ_2^2 são as variâncias de Y_1 e Y_2 nas suas respectivas coordenadas do conjunto formado pelas distâncias $\mathbf{h}$.

2.5.2 Semivariograma Cruzado

Segundo Isaaks e Srivastava (1989), o coeficiente de correlação, utilizado para descrever o comportamento espacial de uma variável isolada, pode ser empregado também para descrever a continuidade espacial entre duas variáveis distintas, medidas simultaneamente em cada coordenada amostral. Para isso, definiremos um processo espacial p-dimensional como uma coleção de variáveis $\mathbf{Y}(x) = \{\mathbf{Y}_1(x), \mathbf{Y}_2(x), \ldots, \mathbf{Y}_p(x)\}$ onde $x \in \mathbb{R}^2$ são as coordenadas regionais em comum e cada variável é um processo estocástico em si. Neste caso, a função covariância de $\mathbf{Y}(x)$ será uma matriz simétrica p-dimendional $\Gamma(x, x')$ onde seu $j;k$-ésimo elemento será:

$$\gamma_{jk}(x; x') = Cov\left\{\mathbf{Y}_j(x); \mathbf{Y}_k(x')\right\}$$

Quando $\mathbf{Y}(x)$ for estacionário, $\gamma_{jj}(x; x') = Var\left[\mathbf{Y}_j(x)\right] = \sigma^2 \mathbf{R}_j + \tau^2 \mathbf{I}_n$ conforme vimos na equação 2.3, representanto o autovariograma de $\mathbf{Y}_i(x)$. Analogamente, considerando $\mathbf{Y}_i(x)$ e $\mathbf{Y}_j(x)$ processos diferentes e seguindo o mesmo raciocínio apresentado na equação 2.4, podemos escrever o semivariograma cruzado para essas duas variáveis como (CRESSIE; WIKLE, 1998; DIGGLE; Ribeiro Jr, 2007):

$$\gamma_{ij}(x; x') = \frac{1}{2}(\sigma_i^2 + \sigma_j^2) + \frac{1}{2}(\tau_i^2 + \tau_j^2) - \sigma_i \sigma_j \rho(u_{ij}) \tag{2.29}$$

Devemos estar atentos aqui para o fato de que os índices i e j dizem respeito às

variáveis $\mathbf{Y}_i(x)$ e $\mathbf{Y}_j(x)$ e u_{ij} representa a distância entre a coordenada x da variável $\mathbf{Y}_i(x)$ e a coordenada x' da variável $\mathbf{Y}_j(x)$ e que a função variograma irá depender também somente da distância euclidiana u_{ij} entre essas coordenadas através da função de correlação $\rho(u_{ij})$. Essa função de correlação, chamada de função de correlação cruzada.

O estimador pelo método dos momentos, para a equação 2.29, segundo Wakernagel (2003) é:

$$\hat{\gamma}_{rs}(u) = \frac{1}{|2N(u)|} \sum_{N(h)} \big(Y_r(x_i) - Y_r(x_j)\big)\big(Y_s(x_i) - Y_s(x_j)\big) \tag{2.30}$$

onde $N(u) = \big\{(x_i, x_j) : x_i - x_j = u; i, j = 1, 2, \ldots, n\big\}$ é o conjunto das diferentes distâncias que separam as coordenadas x. Y_r e Y_s dois processos distintos ocorrendo simultaneamente na mesma área.

Para Mata (1997), com o semivariograma cruzado é possível verificar o relacionamento entre duas variáveis espacialmente medidas, mostrando se a variabilidade de uma é acompanhada pela variabilidade da outra variável. A avaliação da estrutura de dependência espacial pode ser feita através do gráfico estimado de acordo com a equação 2.30, relacionando a variável compatível com a produção com as demais variáveis consideradas preditoras do processo.

2.5.3 Cokrigagem Convencional

Isaaks e Srivastava (1989) apresentam a cokrigagem como um método de estimação, envolvendo a correlação cruzada entre variáveis secundárias e uma variável primária. A grande utilidade do método, alegada pelo autor, é que as variáveis secundárias podem apresentar características favoráveis à sua obtenção, como baixo custo, fácil acesso, dentre outras, que podem ser utilizadas para estimar variáveis primárias sujeitas a subamostragem.

Tomemos novamente dois processos estocásticos $\mathbf{Y}_1$ e $\mathbf{Y}_2$ distintos, mas ocorrendo simultaneamente em uma região, onde supomos, por conveniência de notação, $\mathbf{Y}_1$ a variável primária. No caso de uma única variável, para alguma coordenada onde não tenhamos um

valor medido, poderemos estimá-lo por krigagem usando uma combinação linear de pesos w associados a valores conhecidos como: $\hat{y}_0 = \sum_{i=1}^{n} w_i y_i$ onde y_i é o valor medido na i-ésima coordenada x. No caso de duas variáveis, a estimativa por cokrigagem, com um modelo linear de corregionalização, será obtida por uma combinação linear das duas variáveis, como:

$$\hat{y}_1(x_0) = \sum_{i=1}^{n} a_i y_1(x_i) + \sum_{j=1}^{n} b_j y_2(x_j) \tag{2.31}$$

onde $\hat{y}_1(x_0)$ é a estimativa da variável primária em uma particular localização x_0 não amostrada, $\mathbf{y}_1(x_i) = (y_1(x_1), y_1(x_2), \ldots y_1(x_n))$ são os dados da variável primária observados em n localizações da área A, $\mathbf{y}_2(x_i) = (y_2(x_1), y_2(x_2), \ldots y_2(x_m))$ são os dados da variável secundária observados em m localizações da mesma área, que podem ser parcialmente ou totalmente coincidentes ou isoladas com com relação às localizações da variável primária, $a_1, a_2, \ldots, a_n$ e $b_1, b_2, \ldots, b_m$ são, respectivamente, os pesos de krigagem a serem determinados, associados às observações $y_1(x_i)$ e $y_2(x_j)$.

Sendo $(\hat{y}_1(x_0) - y_1(x_0))$ o erro de predição na coordenada x_0 então:

$$Var(\hat{y}_1(x_0) - y_1(x_0)) = \mathbf{w}'\mathbf{C}\mathbf{w} \tag{2.32}$$

onde:

$\mathbf{w}' = (a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m, -1)$

$\mathbf{Y}^* = (Y_1(x_1), Y_1(x_2), \ldots, Y_1(x_n), Y_2(x_1), Y_2(x_2), \ldots, Y_2(x_m))$

$\mathbf{C}$ é a matriz de covariância de $\mathbf{Y}^*$.

Resolvendo o lado direito da equação 2.31 obtemos:

$$\begin{aligned} Var(\hat{y}_1(x_0) - y_1(x_0)) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Y_1(x_i); Y_1(x_j)) + \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j Cov(Y_2(x_i); Y_2(x_j)) + \\ &+ 2\sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(Y_1(x_i); Y_2(x_j)) - \\ &- 2\sum_{i=1}^{n} a_i Cov(Y_1(x_i); Y_1(x_0)) - \\ &- 2\sum_{j=1}^{m} b_j Cov(Y_2(x_j); Y_1(x_0)) + \\ &+ Cov(Y_1(x_0); Y_1(x_0)) \end{aligned} \tag{2.33}$$

As condições a que os pesos de cokrigagem devem satisfazer são de que:

a) Devem levar a uma estimativa não viciada, o que ocorrerá se $\sum_{i=1}^{n} a_i = 1$ e $\sum_{j=1}^{m} b_j = 0$, o que pode ser comprovado só aplicando a definição de estimador não-viciado dada por Mood, Graybill e Boes (1974). De fato:

$$E(\hat{y}_1(x_0)) = E\left(\sum_{i=1}^{n} a_i y_1(x_i) + \sum_{j=1}^{m} b_j y_2(x_j)\right) = \mu_1 \sum_{i=1}^{n} a_i + \mu_2 \sum_{j=1}^{m} b_j = \mu_1$$

b) A variância do erro dado pela equação 2.33 deverá ser a menor possível, para escolhas convenientes dos pesos.

Introduzindo os multiplicadores de Lagrange ϑ_1 e ϑ_2 na equação 2.32 obtemos:

$$Var(\hat{y}_1(x_0) - y_1(x_0)) = \mathbf{w}'\mathbf{C}\mathbf{w} + 2\vartheta_1\left(\sum_{i=1}^{n} a_i - 1\right) + 2\vartheta_2\left(\sum_{j=1}^{n} b_i\right) \tag{2.34}$$

Sob a condição dada pelo item (a) acima, a expressão 2.34 não muda. Ela poderá ser minimizada então, derivando-se a equação em relação a cada um dos pesos, inclusive os multiplicadores de Lagrange e igualando-se a zero, o que resulta em:

$$\begin{aligned} \frac{\partial}{\partial a_k}(Var(\hat{y}_1(x_0) - y_1(x_0))) &= 2\sum_{i=1}^{n} a_i Cov(Y_1(x_i); Y_1(x_k)) + \\ &+ 2\sum_{i=1}^{n} b_i Cov(Y_1(x_i); Y_2(x_k)) - \\ &- 2Cov(Y_1(x_0); Y_1(x_k)) + 2\vartheta_1 = 0 \text{ para } k = 1, 2, \ldots, n \end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial b_k}\left(Var(\hat{y}_1(x_0)-y_1(x_0))\right) &= 2\sum_{i=1}^{n} a_i Cov(Y_1(x_i);Y_2(x_k))+ \\ &+ 2\sum_{i=1}^{n} b_i Cov(Y_1(x_i);Y_2(x_k))- \\ &- 2Cov(Y_1(x_0);Y_2(x_k))+2\vartheta_2 = 0 \text{ para } k=1,2,\ldots,m\end{aligned}$$

$$\frac{\partial}{\partial \vartheta_1}\left(Var(\hat{y}_1(x_0)-y_1(x_0))\right) = 2\sum_{i=1}^{n} a_i - 1 = 0$$

$$\frac{\partial}{\partial \vartheta_2}\left(Var(\hat{y}_1(x_0)-y_1(x_0))\right) = 2\sum_{i=1}^{n} b_i = 0$$

A variância do erro fica:

$$\begin{aligned}Var(\hat{y}_1(x_0)-y_1(x_0)) &= Cov(Y_1(x_0);Y_1(x_0))+\vartheta_1-\sum_{i=1}^{n} a_i Cov(Y_1(x_i);Y_1(x_0)) \\ &\quad -\sum_{j=1}^{m} b_j Cov(Y_2(x_j);Y_1(x_0))\end{aligned}$$

O método de cokrigagem poderá ser escrito em termos de semivariogramas desde que as covariâncias cruzadas seja simétricas. A continuidade espacial será modelada utilizando semivariogramas posteriormente convertidos para as covariâncias equivalentes pela transformação:

$$C_{\mathbf{Y}_1\mathbf{Y}_2}(u) = \gamma_{\mathbf{Y}_1\mathbf{Y}_2}(\infty) - \gamma_{\mathbf{Y}_1\mathbf{Y}_2}(u)$$

e empregados na matriz de krigagem dada por **Cw** = **D** onde:

$$\mathbf{C} = \left(\begin{array}{c|c|c} C(\mathbf{Y}_1;\mathbf{Y}_1) & C(\mathbf{Y}_1;\mathbf{Y}_2) & \mathbf{1} \\ \hline C(\mathbf{Y}_2;\mathbf{Y}_1) & C(\mathbf{Y}_2;\mathbf{Y}_2) & \mathbf{1} \\ \hline \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \end{array}\right) \quad \mathbf{w} = \left(\begin{array}{c} \mathbf{a} \\ \hline \mathbf{b} \\ \hline -\vartheta_1 \\ -\vartheta_2 \end{array}\right) \quad \mathbf{D} = \left(\begin{array}{c} C(\mathbf{Y}_1;\mathbf{Y}_0) \\ \hline C(\mathbf{Y}_2;\mathbf{Y}_0) \\ \hline 1 \\ 0 \end{array}\right)$$

Este sistema de equações para a cokrigagem é válida somente para estimação pontual. Para estimativas da média, dado uma região, poderá ser estimado um número suficientemente grande de pontos em coordenadas de um grid regular e então se obter a média das estimativas. Isaaks e Srivastava (1989) alertam que, para que a solução das equações existam e sejam únicas, o conjunto das autocorrelações e das correlações cruzadas devem formar matrizes que sejam definidas positivas. Dizem ainda que, se as variáveis forem obtidas nas mesmas coordenadas,

as estimativas por cokrigagem e krigagem ordinária serão idêntivas.

A condição necessária que garantirá que a matriz de correlação seja definida positiva será dada por:

$$\omega' \mathbf{C} \omega = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j C(i;j) > 0$$

onde $\omega = (\omega_1, \omega_2, \ldots, \omega_n)'$ é o vetor de pesos de krigagem, onde pelo menos um de seus elementos deve ser diferente de zero.

Essa condição garante que a variância de qualquer variável aleatória formada pela combinação linear ponderada pelos pesos ω de outras variáveis aleatórias será positiva, ou seja, iremos garantir que a variância do erro de estimação dado por $\hat{Y}(x_0) - Y(x_0)$ será positivo.

Ver Hoef e Barry (1998) usam o termo cokrigagem para se referir a uma predição de uma variável primária em uma específica localização x_0 a partir de um conjunto multivariado de dados e o termo predição espacial quando se deseja predizer um vetor de variáveis aleatórias (de diferentes tipos) também uma específica localização x_0.

Eles destacam três os problemas com a aplicação da cokrigagem tradicional. O primeiro surge quando se pretende minimizar o erro médio quadrático de predição usando o semivariograma cruzado, na forma porposta por Journel e Huijbregts (1978). O procedimento será viável, segundo eles, quando a função de covariância cruzada for uma função par e de reflexão simétrica, ou seja, $C_{(i;j)}(\mathbf{h}) = \mathbf{C}_{(\mathbf{i};\mathbf{j})}(-\mathbf{h})$. A condição de simetria é muito restritiva e pode tornar questionável o uso da função tradicional do semivariograma cruzado.

O segundo problema será o de estimar o semivariograma cruzado quando os dados de ambas as variáveis envolvidas forem tomados nas mesmas coordenadas. Os autores propõem uma adaptação do que chamaram pseudo-variograma cruzado, dado por:

$$2\gamma_{(k;m)}(x_i;x_j) \equiv Var\left(Y_k(x_i) - Y_m(x_j)\right)$$

o que elimina a necessidade das variáveis estarem localizadas nas mesmas coordenadas.

O terceiro problema é que é difícil produzir modelos de semivariogramas cruzados

válidos que sejam consistentes com os modelos de semivariogramas conhecidos. Por modelos válidos eles se referem àqueles cuja variância de predição se mantém positiva.

2.5.4 Modelos geoestatísticos multivariados

Tomemos o conjunto dado pela expressão 2.27 como uma coleção p-dimensional de variáveis aleatórias. A matriz Σ de covariâncias desse conjunto será dada por:

$$\Sigma = \begin{pmatrix} Cov(\mathbf{Y_1};\mathbf{Y_1}) & Cov(\mathbf{Y_1};\mathbf{Y_2}) & \dots & Cov(\mathbf{Y_1};\mathbf{Y_p}) \\ Cov(\mathbf{Y_2};\mathbf{Y_1}) & Cov(\mathbf{Y_2};\mathbf{Y_2}) & \dots & Cov(\mathbf{Y_2};\mathbf{Y_p}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\mathbf{Y_p};\mathbf{Y_1}) & Cov(\mathbf{Y_p};\mathbf{Y_2}) & \dots & Cov(\mathbf{Y_p};\mathbf{Y_p}) \end{pmatrix}$$

sendo a diagonal da matriz a matriz de autocorrelação de cada variável $Y_k : k = 1,2,\dots,p$ do conjunto de variáveis aleatórias escolhido. Os elementos fora da diagonal represemtam a matriz correlação cruzada para cada combinação de pares de variáveis. Essa matriz é uma expansão daquela matriz para o caso bivariado anteriormente apresentada. Ela deve ser uma matriz quadrada, simétrica, definida positiva e passível de decomposição para se obter sua inversa.

Se tomarmos qualquer par de variáveis, digamos $(Y_c(a);Y_d(b))$ então o elemento $Cov(Y_c(a);Y_d(b)) = \left[\sigma_{ab}^{cd}\right]$. Neste caso estamos afirmando que a covariância (e a correlação) se estabelece entre a variável $Y_c(\mathbf{x})$ tomada na coordenada a e a variável $Y_d(\mathbf{x})$ tomada na coordenada b. Uma propriedade imediata, é a sua natureza simétrica, ou seja, $\left[\sigma_{ab}^{cd}\right] = \left[\sigma_{ba}^{dc}\right]$. Se o processo associado à variável Y_k for estacionário, então $\left[\sigma_{ab}^{cd}\right] = Var(Y_k(x_j)) = \sigma_k^2$ e $\left[\sigma_{ab}^{cd}\right]$ para $c \neq d$ irá depender somente das distância u.

A matriz de correlação será dada por $\mathbf{R}(\mathbf{u})$, cujos elementos serão dados por:

$$\left[\rho_{ab}^{cd}\right] = \frac{\sigma_{ab}^{cd}}{\sqrt{\sigma_a^2\sigma_b^2}} = \frac{\sigma_{ab}^{cd}}{\sigma_a\sigma_b}$$

Quando $a = b$, a função $\rho^{aa}(u) = \rho^{bb}(u)$ corresponderá à função de correlação do

processo univariado $Y_a(\mathbf{x})$ e $\rho^{aa}(-u) = \rho^{aa}(u)$. Se $a \neq b$ a função $\rho^{ab}(u)$ será chamada função de correlação cruzada de $Y_a(\mathbf{x})$ e $Y_b(\mathbf{x})$, mas não será necessariamente simétrica na matriz $\mathbf{R}(\mathbf{u})$, mas ainda assim satisfará a condição de que $\rho^{ab}(u) = \rho^{ba}(-u)$ (DIGGLE; Ribeiro Jr, 2007).

Pebesma e Wesseling (1998) apresentam em seu artigo um modelo de predição multivariada envolvendo variáveis cruzadas correlacionadas. O modelo utilizado para cada variável $Y_k; k = 1, 2, \ldots, p$ é aquele definido pela equação 2.2, portanto, um processo não estacionário. O modelo multivariado, neste caso de envolvimento de todas as variáveis do conjunto, é dado por:

$$\mathbf{Y} = \mathbf{D}\beta + \mathbf{S}(\mathbf{x})$$

onde $\mathbf{D}\beta$ corresponde à tendência externa do modelo aplicada às respectivas variáveis. As matrizes envolvidas são:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D_1} & 0 & \ldots & 0 \\ 0 & \mathbf{D_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{D_p} \end{pmatrix} \text{ onde } \mathbf{D_k} = \begin{pmatrix} 1 & d_{11}^{(k)} & d_{12}^{(k)} & \ldots & d_{1q}^{(k)} \\ 1 & d_{21}^{(k)} & d_{22}^{(k)} & \ldots & d_{2q}^{(k)} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & d_{n_1 1}^{(k)} & d_{n_2 2}^{(k)} & \ldots & d_{n_k q}^{(k)} \end{pmatrix}$$

para $n_{()}$ representando o número de coordenadas relativas à k-ésima variável externa d e q é o número de variáveis externas associada a um particular processo Y_k. $\mathbf{S}(x) = \{S_1(x), S_2(x), \ldots, S_p(x)\}$.

O melhor estimador linear não viciado será:

$$\hat{\mathbf{Y}}(x_0) = \mathbf{d}(x_0)\hat{\beta} + \mathbf{r}'\mathbf{V}^{-1}(\mathbf{Y}(x) - \mathbf{D}\hat{\beta})$$

onde

$$\mathbf{d}(x_0) = \begin{pmatrix} \mathbf{d_1}(x(0) & 0 & \ldots & 0 \\ 0 & \mathbf{d_2}(x(0) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{d_p}(x(0) \end{pmatrix}$$

e $d_k(x_0)$ correspondendo à linha da matriz **D** que contém o valor correspondente de $Y_(x_0)$, **r** a matriz de correlações de cada variável com o ponto x_0 a ser estimado e **V** a matriz de correlações obtidas à partir da matriz de covariâncias multivariadas dada pela equação 2.5.4.

A variância do erro de predição será dado por:

$$Var(\hat{\mathbf{Y}}(x_0) = \hat{\beta} - \mathbf{r}'\mathbf{V}^{-1}\mathbf{r} + (d(x_0) - \mathbf{r}'\mathbf{V}^{-1}\mathbf{D})(\mathbf{D}^{-1}\mathbf{V}^{-1}\mathbf{D})^{-1}(d(x_0) - \mathbf{r}'\mathbf{V}^{-1}\mathbf{r})^{-1}$$

Para Ver Hoef e Cressie (1993) este modelo não impõe restrições ao número de variáveis e cada variável pode ter um número diferente de localizações.

Couto e Cunha (2002) afirmam que o pantanal matogrossense apresenta muitas unidades de pedopaisagens com áreas periodicamente inundáveis, onde a amostragem é difícil devido a elevada variabilidade espacial inter e intra estratos. Para sua pesquisa coletaram e analisaram cento e onze amostras sistemáticas 5×10 com cinco atributos físicos e quinze atributos químicos em três ecossistemas. Efetuaram uma análise de componentes principais e fatorial. Das amostras restaram quatro componentes que explicaram 77% da variância total e dois fatores que mostraram a melhor separação entre as pedopaisagens. Utilizaram, nas estimativas dos semivariogramas para os componentes principais os softwares GS+ produzido e comercializado pela empresa *Gamma Design Software* (www.gammadesign.com) e o software Surfer produzido também pela empresa estadosunidense *Golden Software, Inc.* (www.goldensoftware.com). Relatam os aspectos da análise multivariada como apoio às aplicações geoestatísticas mas não explicitam o emprego da geoestatística multivariada.

Filzmoser e Reimann (2002) discutem e comparam métodos e propriedades da análise de componentes principais e da análise fatorial. Eles expõem as vantagens em se aplicar métodos multivariados robustos em geoestatística. Ilustram porém, com aplicações a um conjunto de dados geoquímicos, aplicações da geoestatística univariada.

2.5.5 Redução de variáveis por componentes principais

A análise de componentes principais é amplamente utilizada em pesquisas e mais recentemente, vem sendo aplicada a conjuntos de variáveis com dados autocorrelacionados em modelos geoestatísticos.

Este tipo de análise estatística de dados tem a finalidade de transformar linearmente variáveis correlacionadas em seus componentes principais não correlacionados e organizar esses componentes em ordem decrescente de suas variâncias. A idéia é reduzir a quantidade de dados aos componentes que retêm a maior parte da variância total do conjunto de variáveis. Deve-se aqui atentar para o fato de que, na presença de valores discrepantes (*outliers*) a variabilidade dos dados poderá ser comprometida, altarando o papel da variável portadora desses valores no processo de análise dos componentes do conjunto. Para processos gaussianos, os componentes escolhidos podem ser tidos como fatores.

Seja $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$ um processo estocástico p-dimensional onde cada variável Y_k ($k = 1, 2, \ldots, p$) segue o modelo definido pela equação 2.3. Queremos explicar a estrutura de covariância desse processo para a redução de seu número de variáveis devido a redundâncias ou de uma interpretação correlacional (JOHNSON; WICHERN, 1992). Esse tipo de análise de dados é tido como um processo intermediário para investigações mais amplas como regressão múltipla ou análise de agrupamentos.

Tomemos então o vetor $\mathbf{Y}$ e a partir dele, construímos a matriz de covariâncias $\Sigma = E\left[(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)'\right]$ sendo $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ o vetor das médias relativas a cada variável do vetor $\mathbf{Y}$. Já os elementos da matriz de covariâncias amostral são:

$$[s_{kk'}] = \frac{1}{n-1}\sum_{i=i}^{n}(y_{ik} - \bar{y_k})(y_{ik'} - \bar{y_{k'}}) \qquad k, k' = 1, 2, \ldots, p \tag{2.35}$$

ou, em forma matricial, $\mathbf{S} = (n-1)^{-1}(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})'$

Decompondo a matriz de covariâncias obtemos os p pares de autovalores e autovetores associados $(\lambda_k; \mathbf{e}_k)$, tais que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Para Kolman (1997) sendo **Y** um conjunto de vetores em um mesmo espaço vetorial, então um outro vetor **Cp**, nesse mesmo espaço vetorial será uma combinação linear dos vetores de Y se existirem números reais $a)1, a_1, \ldots, a_p$ tais que **Cp** = $a_1Y_1, a_2Y_2, \ldots a_pY_p$. Assim, segundo Reis (1997), podemos assim escrever o vetor **Y** como uma combinação de seus elementos como:

$$\begin{aligned} Cp_1 &= a_{11}Y_1 + a_{12}Y_2 + \ldots + a_{1p}Y_p \\ Cp_2 &= a_{21}Y_1 + a_{22}Y_2 + \ldots + a_{2p}Y_p \\ &\vdots \\ Cp_p &= a_{p1}Y_1 + a_{p2}Y_2 + \ldots + a_{pp}Y_p \end{aligned}$$

sendo Cp_k a k-ésima componente principal (não correlacionada) aquela cuja variância seja a maior possível, ou seja:

- $Var(Cp_k) = \mathbf{e}_k'\Sigma\mathbf{e}_k = \lambda_k$
- $Cov(Cp_j; Cp_k) = \mathbf{e}_j'\Sigma\mathbf{e}_k = 0$
- $\sum_{k=1}^{p} Var(Y_i) = \sum_{k=1}^{p} Var(Cp_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$

Determinamos a porcentagem de contribuição de cada componente, como:

$$\%CCp_k = \lambda_k \left(\sum_{j=1}^{p} \lambda_j \right)^{-1} \tag{2.36}$$

asssim, aquelas primeiras m variáveis Y_k que acumularem maior porcentagem, poderão ser substituídas pelas m componentes principais, reduzindo assim, o número de variáveis sem grande perda na variabilidade do processo.

Segundo Johnson e Wichern (1992) o coeficiente de correlação entre as componentes e as variáveis primárias Y_k é dada por:

$$\rho(Cp_i; Y_k) = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{VarY_k}} = \frac{e_{ik}sqrt\lambda_i}{\sigma_{ii}} \qquad \text{onde } i; k = 1, 2, \ldots, p$$

Apesar da correlação entre as p variáveis e seus componentes principais ajudar a interpretar o papel dos componentes, ela mede somente a contribuição univariada de um particular Y_k para formar a componente Cp_k e não a sua importância na presença das demais. Quando as variáveis Y_k forem processos medidos em escalas diferentes, recomenda-se utilizar a sua padronização para que possam ser comparáveis. O processo de seleção de componentes principais a partir de variáveis padronizadas Z_k se dá a partir da matriz de correlações **R** obtida como:

$$\mathbf{R} = \left(\left(V^{-0,5}\right)^{-1}\right)\sigma\left(\left(V^{-0.5}\right)^{-1}\right) \tag{2.37}$$

onde Σ é a matriz de covariâncias, $V^{-0,5} = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_{22}} & \dots & 0 \\ \vdots & \vdots & & 0 \\ 0 & 0 & \dots & \sqrt{\sigma_{pp}} \end{pmatrix}$

Decompondo a matriz de correlações **R** obteremos também os p pares de autovalores e autovetores ordenados $(\lambda_k; \mathbf{e}_k)$. A k-ésima componente principal padronizada será:

$$Cp_k = \sum_{k=1}^{p} e_k Z_k \qquad \text{onde } Z_k = \frac{Y_k - \overline{Y_k}}{\sqrt{s_{kk}}} \tag{2.38}$$

A porcentagem de variação explicada por cada componente será dada pela equação 2.36 e a correlação entre componente e variável padronizada será dada por:

$$\rho(Cp_k, Z_l) = e_{kl}\sqrt{\lambda_k}; \qquad k;l : 1,2,\dots,p \tag{2.39}$$

Wackernagel (1998) diz que é de vital importância tal tipo de análise para verificar se os dados são intrinsecamente correlacionados, senão o método geoestatístico multivariado poderá gerar resultados viesados.

Para se detectar uma correlação intrínseca em dados autocorrelacionados no espaço, há a necessidade de se verificar se os dados seguem um modelo de correlação intrínseca. Nesse modelo, todo autovariograma e todo variograma cruzado de duas variáveis Y_i, Y_j serão propor-

cionais a um variograma geral $\gamma(u)$, ou seja,

$$\gamma_{ij} = b_{ij}\gamma(u) \qquad \text{para } i, j = 1, 2, \ldots, n$$

onde os b_{ij} são coeficientes.

Uma corregionalização (um conjunto de variáveis espacialmente correlacionadas) é intrinsecamente correlacionada quando o quociente:

$$\frac{\gamma_{ij}(u)}{\sqrt{\gamma_{ii}\gamma_{jj}(u)}} = \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} = r_{ij}$$

é constante para qualquer distância u. Notar que a correlação entre duas variáveis não depende de u, diferentemente da autocorrelação de cada uma das variáveis separadamente.

A correlação intrínseca pode ser avaliada determinando suas componentes principais para a seguir determinar o variograma cruzado entre os primeiros componentes principais. No caso de existência de correlação intrínseca, o variograma cruzado resultante será nulo, caso contrário, as componentes serão correlacionadas espacialmente em alguma região do espaço e então o modelo deverá ser preterido a favor de outros modelos de corregionalização.

3 OBJETIVOS

O objetivo geral deste trabalho é contribuir para a ampliação nas aplicações geoestatísticas em experimentos agrícolas desenvolvidos na região Oeste do Paraná, incorporando métodos multivariados que permitam a elaboração de mapas temáticos melhores com uma possível redução no número de amostras de variáveis de interesse principal pela sua correlação com outras variáveis agrícolas que sejam mais facilmente disponíveis.

Os objetivos específicos deste projeto de tese são:

- Ampliar a revisão bibliográfica com pesquisas do estado da arte;
- Elaborar modelo de correlação espacial multivariada com base nas diferentes combinações da matriz do covariância cruzada e autocovariancia;
- Elaborar a função de verossimilhança para modelos geoestatísticos multivariados;
- Construir uma algorítimo para de otimização para os componentes de modelos gaussianos multivariados.
- Avaliar, pela análise de componentes principais e análise fatorial, a viabilidade de redução do número de processos;
- Analisar um conjuntos de dados experimentais de processos estocásticos gaussianos multivariados;
- Analisar um conjuntos de dados simulados de processos estocásticos gaussianos multivariados.

4 METODOLOGIA

Para este trabalho serão utilizados dados de pesquisa realizada na Universidade Estadual do Oeste do Paraná – Unioeste, em área de Latossolo Roxo, com declividade média de 0,19%, em área de 1,33 ha, localizado no Centro de Pesquisa Eloy Gomes, da Cooperativa Central Agropecuária de Desenvolvimento Tecnológico e Econômico Ltda. (COODETEC), situada na BR 467, km 98, em Cascavel-PR. Nessa área, no final do ano de 1997, cultivou-se soja em sistema de semeadura direta. Em abril de 1998, após serem demarcadas 256 parcelas de 7,20 $\times$ 7,20 m, a produção de cada parcela foi colhida e pesada. Simultaneamente foram tomadas, em cada parcela, amostras do solo para a análise química.

Para modelar a variabilidade espacial e correlacioná-la com a cultura implantada na área, serão utilizados os atributos químicos: pH, Matéria Orgânica (%), Potássio ($Cmol_c \times dm^{-3}$), Fósforo ($mg \times dm^{-3}$) e Índice de Saturação de Bases (%).

As amostras foram obtidas com 7 cm de diâmetro e 15 cm de profundidade dentro de cada uma das 256 parcelas, estruturadas em um grid de 7,20 $\times$ 7,20 m, com carreador de 2,4 m em uma das direções, usando-se o sistema desalinhado, sistemático estratificado de Wollenhaupt e Wolkowski (1994). Para a Produtividade, foram colhidas e identificadas as parcelas de 5,0 $\times$ 5,0 m, excluídos bordaduras e carreador (ver Figura 4.1).

Será utilizado também um segundo conjunto de dados provenientes de banco de dados cartográficos gerado em padrões e formatos Arc Gis, disponíveis no setor de informações geográficas (SIG) da Universidade Federal do Paraná - UFPr e da empresa Modo Battistella Reflorestamento S/A – MOBASA. O estudo foi desenvolvido em parcelas de inventários florestais contínuos com plantio de Pinus da espécie *P. Taeda L.* em fazendas situadas no município de

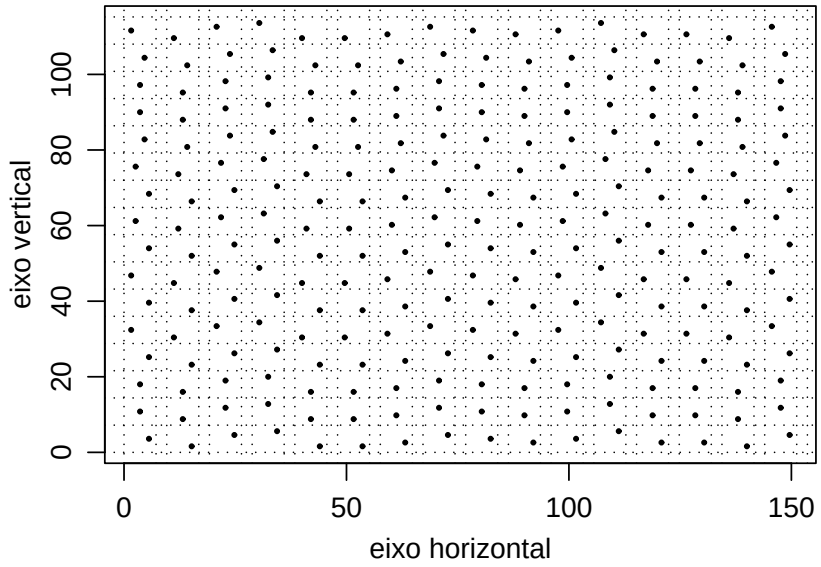

Figura 4.1: Grid amostral com locação das parcelas e pontos amostrais em sistema desalinhado, sistemático estratificado (WOLLENHAUPT; WOLKOWSKI, 1994).

Rio Negrinho no Estado de Santa Catarina sob domínio das bacias e coberturas sedimentares na região do patamar oriental da Bacia do Paraná e na unidade do patamar de Mafra-SC. Trata-se de uma área de 2.252,11 ha localizado no Norte do Estado onde o relevo é quase plano, com cotas altimétricas diminuindo de Leste para Oeste, atingindo valores entre 650 a 740 metros. A geologia é representada pelo grupo Itararé compreendendo todo o pacote de sedimentos de origem glacial e periglacial relacionado ao Carbonífero Superior e Permiano Inferior.

Foram efetuados levantamentos pedológicos com prospecção por tradagem e em perfis em barrancos de estrada, acompanhada de coleta de amostras para análises químicas de: $pH(CaCl_2)$, Fósforo disponível (P), Potássio disponível (K), Al^{3+}, Carbono orgânico, $H + Al^{3+}$, Soma de bases (SB), capacidade de troca catiônica (CTC) e Saturação por bases ($V\%$) e análises granulométricas de: areia, silte e argila.

Foram analisadas nas parcelas, árvores com idades que variavam de 11 a 15 anos, onde foram medidas o diâmetro (cm) a 1,3 m de altura, a altura média (m) das árvores da parcela, o número de árvores por hectare, a altura dominante (m) das 10 maiores árvores, a área basal (m^3), o volume médio (m^3/ha) e incremento médio anual (m^3). Essas foram consideradas as variáveis principais por estarem relacionadas com algum interesse econômico.

O delineamento geoestatístico foi feito em *grid* irregular (figura 4.2) registrando-se os pontos amostrais em coordenadas ortogonais UTM (*Universal Transverse Mercator*)com auxílio de aparelho de posicionamento por satélite (GPS) e anotando-se, para cada localização a análise do material geológico, as medições das árvores, a profundidade efetiva do perfil do solo (horizontes A + B), altura estimada do lençol freático, a posição na encosta, o percentual de declividade e a altitude.

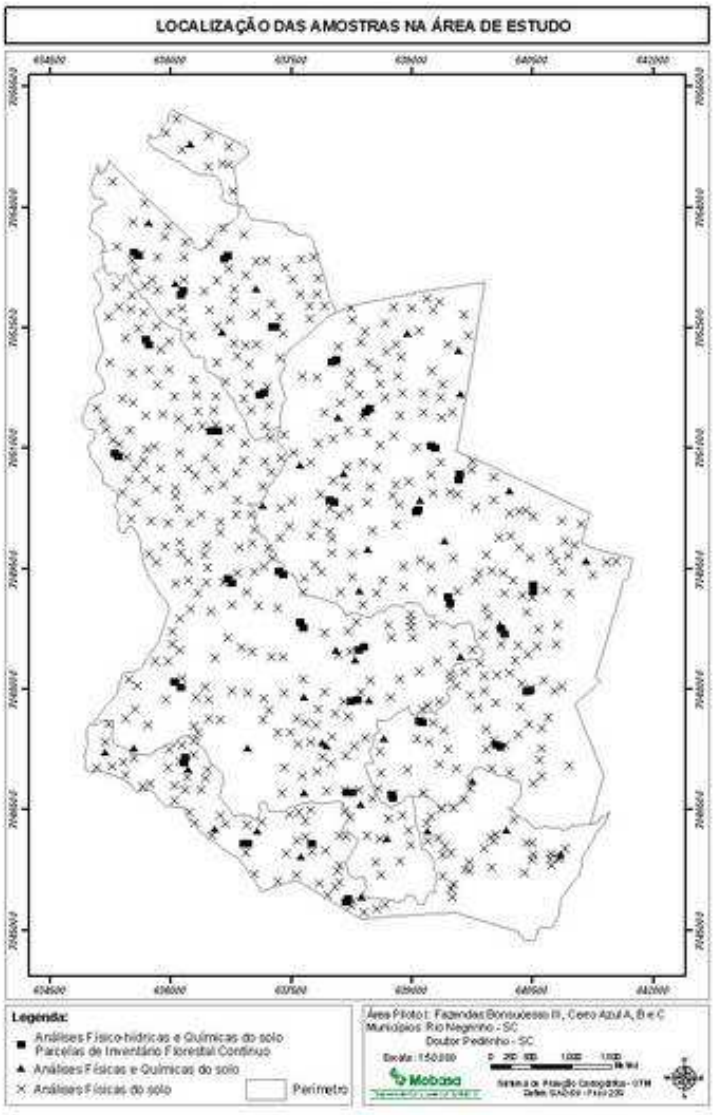

Figura 4.2: Grid amostral com locação das parcelas e pontos amostrais na fazenda MOBASA. Os 35 pontos retangulares representam as coordenadas de análises Fisico-Hídricas e Químicas, os 18 pontos triangulares representam as coordenadas de análises Físicas e Químicas e os 555 pontos em cruz representam as análises Físicas.

4.1 ANÁLISE GEOESTATÍSTICA

Adotaremos ao longo das análises estatísticas, recursos computacionais baseados em programas livres, que atendam a licença GPL (*General Public Licence*), dentre eles, o ambiente operacional GNU/Linux, o pacote estatístico R (R Development Core Team, 2006) e os pacotes geoestatísticos geoR (Ribeiro Jr; DIGGLE, 2001) e Gstat (PEBESMA; WESSELING, 1998). Com os pacotes geoestatísticos será possível calcular os semivariogramas amostrais (direcionais e cruzados), ajustar modelos válidos aos semivariogramas, ajustar modelos lineares de corregionalização, efetuar estimação linear e simulações por krigagem e cokrigagem e produzir gráficos e mapas temáticos. O pacote estatístico, além do suporte às funções dos pacotes geoestatísticos, permitirá a análise convencional dos dados bem como a análise multivariada de componentes principais. Os *scripts* desevolvidos serão apresentados no final como anexo.

4.1.1 Estatística descritiva

Será estudado inicialmente o enfoque estatístico tradicional para o conjunto de variáveis aleatórias em cada um dos problemas. Será empregada uma análise descritiva, visando inicialmente identificar e avaliar parâmetros como homogeneidade, normalidade, pontos discrepantes e tendência direcional. Essas análises preliminares servirão para obter indicativos de atendimento aos pressupostos do modelo geoestatístico e para referências exploratórias no ajuste de parâmetros.

Para a avaliação exploratória da variabilidade espacial, serão elaborados os semivariogramas experimentais calculados pela função semivariância $\hat{\gamma}(u)$ dada pela equação 2.17. Devido ao fato dos semivariogramas empíricos serem de difícil interpretação (ver figura 2.7) os resultados serão divididos em poucos intervalos de variação das distâncias u, representando no ponto médio de cada intervalo de classe, o valor médio das semivariâncias relativas a esse intervalo (ver figura 2.8). Entretanto, para a estimação de parâmetros será utilizado o conjunto de dados original e não o semivariograma.

4.1.2 Ajuste de um modelo teórico ao semivariograma experimental

O método de ajuste de um modelo teórico aos semivariogramas empíricos obtidos que adotaremos será o da máxima verossimilhança, ou seja, através da maximização da função log-verossimilhança dada pela equação 2.23. Neste processo, escolheremos o modelo de correlação que levar ao maior valor da função, assim seremos capazes de fixar os parâmetros do modelo que melhor expliquem o resultado experimental.

Inicialmente buscaremos modelar, conforme equação 2.13, algum efeito direcional ou tendência não estacionária que porventura esteja presente na variabilidade espacial do conjunto de variáveis, em cada problema. Empregando o método dos mínimos quadrados ordinários (equação 2.14) obteremos uma estimativa dos parâmetros do modelo (equação 2.15) para assim remover tais efeitos, conforme a equação 2.16, ou inserindo esses parâmetros no modelo de estimação por máxima verossimilhança (equação 2.23).

Outra verificação importante nessa fase da análise será constatação da gaussianiedade do processo estocástico, representado pelas observações Y. Empregaremos o processo de transformação de Box & Cox (BOX; COX, 1964) conforme a equação 2.8. Calcularemos o perfil da função log-verossimilhança para o parâmetro λ de transformação em uma região de 95% de confiança em torno do valor máximo da função e assim, teremos um intervalo de valores prováveis para a escolha do tipo de transformação, incluindo a possibilidade de não se transformar.

Seguindo a análise, buscaremos estimar, para cada variável, tanto o “melhor” modelo geoestatístico quanto seus parâmetros, utilizando o método da máxima verossimilhança segundo os vários modelos de correlação das famílias já apresentadas. O principal critério de escolha será o próprio valor de máximo da função.

Os mesmos procedimentos serão aplicados ao semivariograma cruzado, embora sua interpretação seja mais direcionada aos aspectos de corregionalização.

4.1.3 Seleção de variáveis

O método geoestatístico visa produzir um mapa temático presumido de uma variável principal (variável resposta) em função de uma ou mais variáveis preditoras. Quando se trata de uma única variável preditora, então o modelo se torna um processo bivariado, mas nem sempre é este o caso. Na maioria dos experimentos agrícolas dispomos de um conjunto de variáveis preditoras relacionadas às características do solo e da região geográfica. Uma seleção de variáveis pode ser um procedimento razoável para diminuir a quantidade de preditoras na variabilidade total do processo, podendo assim serem construídos mapas temáticos que exijam menores recursos computacionais.

Faremos uma análise de componentes principais padronizados e produziremos produziremos o mapa de produtividade de soja do primeiro conjunto de dados e o rendimento de Pinus no segundo, utilizando tanto as informações principais disponíveis da variável resposta quanto os componentes principais que juntas representem a maior parte da variabilidade total. Para isto, padronizaremos cada variável resposta, determinaremos a matriz de correlações amostrais (equação 2.37), decomporemos essa matriz para extrair os autovalores (ordenados) e autovetores e dessa forma, conforme equação 2.38, obtendo as componentes principais padronizadas.

A porcentagem explicada por cada componente será obtida pela equação 2.36 e assim, escolheremos as primeiras componentes que totalizem uma alta porcentagem de explicação. Examinaremos também as correlações existentes entre as variáveis padronizadas e as componentes principais, utilizando a equação 2.39.

4.1.4 Método de predição linear

A predição linear espacial que adotaremos para predizer o valor de uma variável em uma coordenada geográfica onde não foi efetuada nenhuma medida, será aquela feita pelo método da krigagem. Sobrepondo-se a uma área uma malha de predição com espaçamento suficientemente pequeno, o conjunto de valores preditos nas coordenadas dessa malha, esca-

lonados pelos equivalentes numéricos de uma variação de cor (tons de cinza, por exemplo), permitirá representar como um mapa temático a variação espacial da variável.

Diferentes situações se apresentam com as aplicações dessa técnica. Podemos representar a variação espacial de uma variável isoladamente (krigagem), podemos representar a variação auxiliada pela influência de outra variável, que se denomina krigagem com covariáveis, podemos ainda representar a variação espacial de uma variável primária com o auxílio de outra variável com ela correlacionada, co-localizada ou não. Estaremos aqui analisando dados experimentais e simulados que contemplem esse cenário.

Para os dados do experimento de Agricultura de Precisão, coletados junto à Coodetec, adotaremos como variável principal a produtividade de soja e como variáveis preditivas de interesse secundário, os atributos químicos. Faremos um mapa para representar a variação da produtividade na área por krigagem na sua forma tradicional. Em seguida determinaremos a correlação cruzada da cada atributo químico com a produtividade, estabelecendo assim a existência de alguma estrutura de corregionalização. Essa avaliação será complementada por análise de componentes principais sobre os atributos químicos visando a redução no número de variáveis de caráter preditivo. Em seguida, estaremos elaborando novo mapa de produtividade com a contribuição co-localizada da primeira componente e outro com a contribuição de todos os atributos químicos, ambos por cokrigagem. Como avaliação dos diferentes mapas, iremos comparar as variâncias dos erros de predição.

Para os dados de levantamento de inventários florestais na fazenda Mobasa, pela natureza desses dados iremos elaborar um mapa temático para cada variável de rendimento. Determinaremos também a correlação cruzada de cada variável de rendimento com cada atributo preditor medido na área. Faremos também uma análise de componentes principais com os atributos preditores para escolher as componentes que juntas oferecerem maior contribuição para a variabilidade do conjunto. Faremos também mapas temáticos de cada uma das variáveis de rendimento com as respectivas componentes principais por cokrigagem. Iremos também expandir a malha de coordenadas das variáveis de rendimento com a contribuição da variável Argila,

disponível em número maior que as demais e localizadas em diferentes coordenadas, fazendo novos mapas da variável de rendimento, também por cokrigagem.

Visando consolidar ou mesmo avaliar a robustez dos métodos aplicados, estaremos simulando dados de uma variável primária em *grid* aleatório e dados de outras variáveis secundárias, também em grid aleatório, mas em coordenadas diferentes. Da mesma maneira como procederemos com os dados experimentais, iremos elaborar mapa da variável primária simulada com krigagem e mapa da variável primária simulada com complemento de informação das variáveis secundárias por cokrigagem. Iremos também comparar as variâncias dos erros de predição.

5 CRONOGRAMA

12-03-07	12-03-07	Qualificar o projeto junto ao PPGMNE
12-03-07	30-04-07	Completar a revisão da literatura
		Elaborar metodologia de análise dos dados
30-04-07	20-06-07	Aplicar metodologia aos conjuntos de dados
20-06-07	30-06-07	Simular processos geoestatísticos multivariados
01-07-07	30-09-07	Elaborar relatório preliminar com resultados e discussão
01-10-07	31-12-07	Elaborar versão preliminar do relatório final
01-01-08	28-02-08	Fazer revisão ortográfica e gramatical do relatório final
15-01-08	15-01-08	Marcar data de defesa do relatório final
15-01-08	15-01-08	Definir os membros da banca
01-03-08	01-03-08	Encaminhar relatório final aos membros da banca
01-03-08	30-03-08	Defender a tese

Referências Bibliográficas

ABRAMOWITZ, M.; STEGUN, I. *Handbook of Mathematical Functions*. ninth ed. New York: Dover, 1965.

BAKHSH, A. et al. Spatial distribution of soil attributes affecting crop yield. *ASAE*, p. 1032, 1997.

BALASTREIRE, L. A.; ELIAS, A. I.; AMARAL, J. R. Agricultura de precisão: Mapeamento da produtividade da cultura de milho. *Revista de Engenharia Rural*, p. 97–111, 1997.

BARTLETT, M. *Stochastic Process*. USA: Cambridge University Press, 1955.

BOX, G.; COX, D. An analysis of transformation. *JRSSB*, p. 211–252, 1964.

BRAGA, L. P. V. Geoestatística e aplicações. *Anais do 9º Simpósio Brasileiro de Probabilidade e Estatística do IME/USP - São Paulo*, p. 36p, 1990.

CAPELLI, N. L. Agricultura de precisão: Novas tecnologias para o processo produtivo. *LIE/DMAQAG/ FEAGRI/UNICAMP*, 1999.

COUTO, E. G.; CUNHA, C. N. Application of multivariate geostatistics to identify soil landscapes in the Pantanal of Mato Grosso - Brazil. *Revista Agricultura Tropical*, v. 6, p. 48–65, 2002.

CRESSIE, N.; WIKLE, C. K. The variance-based cross-variogram: You can add apples and oranges. *Mathematical Geology*, v. 30, p. 789–799, 1998.

DIGGLE, P. J.; Ribeiro Jr, P. J. *Model-based Geostatistics*. USA: Springer Series in Statistics, 2007.

DUDEWICZ, E.; MISHRA, S. *Modern Mathematical Statistics*. Singapore: Wiley-Sons, 1988.

FILZMOSER, P.; REIMANN, C. Robust multivariate methods in geostatistics. *W. Gaul and G. Ritter, editors, Classification, Automation, and New Media*, p. 429–436, 2002.

GOOVAERTS, P. *Geostatistics for Natural Resources Evaluation*. Oxford: Oxford University Press, 1997.

ISAAKS, E. H.; SRIVASTAVA, R. M. *Applied GEostatistics*. New York: Oxford University, 1989.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. third ed. USA: Prentice-Hall, 1992.

JOURNEL, A. G.; HUIJBREGTS, C. J. *Mining Geostatistics*. London: Academic Press, 1978.

KOLMAN, B. *Introductory Linear Algebra with Applications*. USA: Prentice Hall, 1997.

KRIGE, D. G. *A Statistical Approach to Some Mine Valuations and Allied Problems at Witwatersrand*. Tese (Doutorado) — University of Witwatersrand, 1951.

MATA, J. D. V. *Variabilidade espacial de indicadores da compactação de Terra Roxa Estruturada, sob dois sistema de preparo, cultivada com feijão (Phaseolus vulgaris L.) irrigado*. 73 p. Tese (Doutorado) — Universidade de São Paulo/ESALQ, 1997.

MATHERON, G. Traite de geostatistique appliquee. *Bureau de Recherches Geologiques et Minieres*, v. 14, p. 1246–1266, 1962.

MATHERON, G. Principles of geostatistics. *Economic Geology*, v. 58, p. 1246–1266, 1963.

MATHERON, G. The intrinsic random function and their application. *Advances in Applied Probability*, p. 508–541, 1973.

MATéRN, B. *Spatial Variation Analysis*. 2nd. ed. Berlin: Springer Verlag, 1986.

MOHAMED, S. B.; EVANS, E. J.; SHIEL, R. S. Mapping techniques and intensity of soil sampling for precision farming. *Proceedings of the 3rd International Conference on Precision Agriculture*, p. 217–226, 1996.

MOLIN, J. P. Agricultura de precisão: Mais um desafio para o agricultor brasileiro. *Plantio Direto*, p. 26–27, 1997.

MONTGOMERY, D. C.; PECK, E. A. *Introduction to Linear Regression Analysis*. USA: Cambridge University Press, 1955.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. *Introduction to the Theory of Statistics*. third ed. Singapore: McGraw-Hill, 1974.

OLIVEIRA, M. C. N. *Métodos de estimação de parâmetros em modelos geoestatísticos com diferentes estruturas de covariâncias: uma aplicação ao teor de cálcio no solo.* Tese (Doutorado) — Universidade de São Paulo/ESALQ, 2003.

PANNATIER, Y. *Variowin 2.2: Software for Epatial Data Analysis in 2D*. New York: Springer, 1996.

PEBESMA, E. J.; WESSELING, C. G. Gstat: a program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, v. 1, p. 17–31, 1998.

PREVEDELLO, B. M. S. *Variabilidade espacial de parâmetros de solo e planta*. 166 p. Tese (Doutorado) — Universidade de São Paulo, 1987.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2006. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

REICHARDT, K.; VIEIRA, S. R.; LIBARDI, P. L. Variabilidade espacial de solos e experimentação de campo. *Revista Brasileira de Ciência do Solo*, v. 10, p. 1–6, 1986.

REIS, E. *Estatística Multivariada Aplicada*. Lisboa: ED. Silabo, 1997.

Ribeiro Jr, P. J.; DIGGLE, P. J. geoR: A package for geostatistical analysis. *R-NEWS*, v. 01, 2001. Http://cran.r-project.org/doc/Rnews.

SCHABENBERGER, O.; GOTWAY, C. A. *Statistical Methods for Spatial Data Analysis*. New York: Chapman-Hall, 2005.

Ver Hoef, J. M.; BARRY, R. P. Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, v. 69, p. 275–294, 1998.

Ver Hoef, J. M.; CRESSIE, N. Multivariate spatial prediction. *Mathematical Geology*, v. 25, p. 219–240, 1993.

WACKERNAGEL, H. Principal component analysis for autocorrelated data: a geoestatistical perspective. *Techinical Report 22/98-G - Centre de Géostatistique - Ecole des Mines de Paris*, 1998. Http://cg.ensmp.fr.

WAKERNAGEL, H. *Multivariate geostatistics: an introduction with applications*. third ed. Germany: Springer, 2003.

WALLER, L. A.; GOTWAY, C. A. *Applied spatial statistics for public health data*. USA: Wiley series in probability and statistics, 1965.

WOLLENHAUPT, N. C.; WOLKOWSKI, R. P. Grid soil sampling. better crops with plant food. *NORCROSS*, v. 78, p. 6–9, 1994.

YANG, C. et al. Spatial variability of field topography and wheat yield in the palouse region of the pacific northwest. *American Society of Agricultural Engineers*, v. 41, p. 17–27, 1998.