UNIVERSIDADE FEDERAL DO PARANÁ

ANA BEATRIZ TOZZO MARTINS

MÉTODOS GEOESTATÍSTICOS APLICADOS A DADOS COMPOSICIONAIS PARA CLASSIFICAÇÃO DE SOLOS

CURITIBA

2008

ANA BEATRIZ TOZZO MARTINS

MÉTODOS GEOESTATÍSTICOS APLICADOS A DADOS COMPOSICIONAIS PARA CLASSIFICAÇÃO DE SOLOS

Projeto de Tese apresentado ao Curso de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, Setor de Ciências Exatas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutora em Métodos Numéricos em Engenharia.

Orientador: Prof. PhD. Paulo Justiniano Ribeiro Junior

CURITIBA

2008

SUMÁRIO

Página

LISTA DE FIGURAS **5**

1 Introdução **8**

2 Revisão da Literatura **11**

2.1 Modelo Geoestatístico Gaussiano 11

2.1.1 Definição do modelo 11

2.1.2 Componentes do modelo 13

2.1.3 Estimação de parâmetros do modelo 18

2.2 Predição linear espacial 22

2.2.1 Conceitos de predição 22

2.2.2 Krigagem 27

2.2.3 Inferência Bayesiana para predição espacial 29

2.3 Matriz de Covariância, Matriz de Correlação Cruzada e Variograma Cruzado 31

2.4 Modelo Multivariado 34

2.5 Dados Composicionais 37

2.5.1 Composição Regionalizada 39

2.5.2 Base Regionalizada 41

2.5.3 Subcomposição Regionalizada 42

2.5.4 Amalgamação e Partição Regionalizada 43

2.5.5 Transformação 44

2.5.6 Perturbação e Potência 46

2.5.7 Estatísticas Descritivas e Domínio de Confiança Para Dados Composicionais 47

2.5.8 Representação Gráfica 48

2.5.9 Estacionariedade 49

2.5.10 Estrutura de Covariância Espacial 50

2.5.11 Estrutura de Covariância Espacial Intrínsica 52

3 Material e Métodos **54**

3.1 Material 54

3.2 Método 55

4 Resultados Esperados **68**

5 Cronograma **69**

LISTA DE FIGURAS

Página

1 Triângulo de Feret. 9

2 Diagrama triangular simplificado, utilizado pela EMBRAPA, para a classificação textural do solo. 10

3 Situações relacionando estacionariedade e isotropia. 14

4 (a)Composições de 3 partes como raios partindo da origem em $\mathbb{R}^3_+$; (b) O simplex $\mathbb{S}^2$. Adaptado de Barceló-Vidal, Martín-Fernández e Pawlowsky-Glahn (2001) 42

5 (a)Interpretação geométrica da formação da subcomposição $\underset{\sim}{\mathrm{W}}_{12}$ da composição $\underset{\sim}{\mathrm{W}}$: (a) em $\mathbb{R}^3_+$; (b) em $\mathbb{S}^2$. Adaptado de Barceló-Vidal, Martín-Fernández e Pawlowsky-Glahn (2001). 43

6 Diagrama ternário para dados do Lago Ártico incluindo o centro da distribuição e região 2-sigma de confiança. 48

7 Gráfico de círculos do logit da porcentagem de areia, silte e argila. 61

8 Localizações (topo à esquerda), logit(porcentagem) vs coordenadas (topo à direita e baixo à esquerda), e histograma (baixo à direita) da areia. . . . 62

9 Boxplot do logit da porcentagem de areia. 63

10 Perfil do log da verossimilhança para o parâmetro λ de transformação de Box-Cox para areia 63

11 Localizações (topo à esquerda), logit(porcentagem) vs coordenadas (topo à direita e baixo à esquerda), e histograma (baixo à direita) do silte. 64

12 Boxplot do logit da porcentagem de silte. 65

13 Perfil do log da verossimilhança para o parâmetro λ de transformação de Box-Cox para silte. 65

14 Localizações (topo à esquerda), logit(porcentagem) vs coordenadas (topo à direita e baixo à esquerda), e histograma (baixo à direita) da argila. . . . 65

15 Boxplot do logit da porcentagem de argila 66

16 Perfil do log da verossimilhança para o parâmetro λ de transformação de Box-Cox para argila. 66

17 Mapas da porcentagem de areia (à esquerda), silte (centro) e argila (à direita). 66

18 Diagrama ternário para areia, silte e argila. 67

19 Diagrama Ternário para areia, silte e argila incluindo o centro da distribuição e regiões de confiança de 2 e 4 sigma. 67

20 Diagrama de dispersão para areia vs silte, areia vs argila e silte vs argila. . 67

LISTA DE SIGLAS

ABNT	-	Associação Brasileira de Normas Técnicas
ALR	-	Transformação razão log-aditiva
AGL	-	Transformação logística generalizada aditiva
CLR	-	Transformação razão log-centrada
EM	-	Erro de medida
EMBRAPA	-	Centro Nacional de Pesquisa de Solos
EQM	-	Erro de Predição quadrático médio
ESALQ	-	Escola Superior de Agricultura "Luiz de Queiroz"
LR	-	Log-razão
MCMC	-	Cadeias de Markov de Monte Carlo
MMQ	-	Método dos mínimos quadrados
NBR	-	Norma Brasilerira Registrada
SQR	-	Soma dos quadrados dos resíduos
UEM	-	Universidade Estadual de Maringá
USP	-	Universidade de São Paulo
VME	-	Variação de pequena escala ou micro escala

1 Introdução

Desde a muito tempo pesquisadores atuam na área de estudos denominada Estatística Espacial e importantes trabalhos têm sido desenvolvidos, por exemplo, por Matheron (1963), Cressie (1993) e muitos outros. Mais Recentemente, em particular na área de geoestatística, surgiram trabalhos como os de Diggle, Tawn e Moyeed (1998), Schabenberger e Pierce (2001) e Diggle e Ribeiro Jr. (2007) com uma perspectiva baseada em modelos e com isto métodos clássicos de inferência baseados em verossimilhança foram aplicados para produzir estimativas mais eficientes dos parâmetros desconhecidos e avaliar a incerteza em predições espaciais.

Trabalhos realizados por Aitchison (1986) em análise de dados composicionais apresentam uma metodologia adequada para analisar dados caracterizados por se apresentarem em forma de proporções e por somarem 1. O autor aponta que métodos estatísticos tradicionais, como por exemplo, o cálculo de correlação entre razões cujos numeradores e denominadores contém partes comuns, são incorretamente aplicáveis a dados desta natureza. Mais recentemente, este tipo de dado vem sendo analisado considerando-se a espacialização das respectivas variáveis (PAWLOWSKY-GLAHN; OLEA, 2004). Seguindo a linha de pesquisa da autora, fica clara a possibilidade de maiores investigações na análise geoestatística de dados composicionais.

Neste trabalho os dados composicionais considerados são de solo. O tamanho das partículas fragmentadas (granulometria) diferencia a areia do silte e da argila. No Brasil, a granulometria segundo a Norma Brasileira Regulamentada-NBR 6502/95 da Associação Brasileira de Normas técnicas-ABNT, tem-se a seguinte classificação:

Classificação	Diâmetro dos Grãos (mm)
Argila	$(0;\ 0,002]$
Silte	$(0,002;\ 0,06]$
Areia	$(2,0;\ 0,06]$
Pedregulho	$(60,0;\ 2,0]$

A proporção relativa da areia, silte e argila no solo, designada pelo termo Textura, pode caracterizar o solo. A versão do triângulo de Feret, Figura 1, é um triângulo

textural formado por estes três elementos, que determinam um ponto no interior do triângulo, com a soma das três frações igual a 100%. Este triângulo é dividido em áreas e conforme a localização do ponto nestas áreas tem-se uma classificação para o solo conforme pode ser visto na Figura 2. A Classificação é importante no combate à erosão, na escolha da cultura a ser plantada, na preservação do meio ambiente, na análise química do solo para adequada adubação, para o manejo e uso do solo. Oliveira (2008?) apresenta mais detalhes sobre o emprego agrícola e não agrícola da classificação que neste trabalho será adotada conforme EMBRAPA (1999).

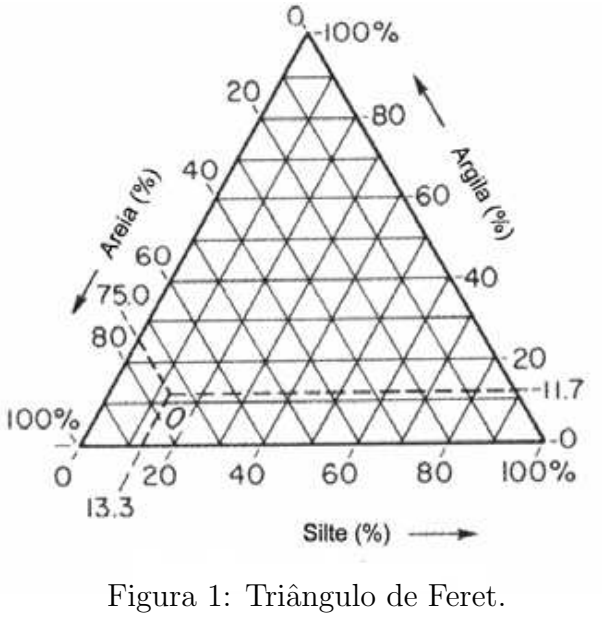

Figura 1: Triângulo de Feret.

O objetivo geral deste trabalho é obter um modelo geoestatístico para dados composicionais de solo que sejam espacializados e multivariados e que compatibilize a estrutura de covariância induzida pelo modelo multivariado com a estrutura de covariância induzida pelo modelo de dados composicionais representando o resultado através de mapas de classificação espacial do solo.

Os objetivos específicos são:

- Investigar formas alternativas à proposta de Pawlowsky-Glahn e Olea (2004) sob o enfoque da geoestatística baseada na declaração explícita de modelos;
- Construir um modelo em que as dependências espacial e entre variáveis sejam con-

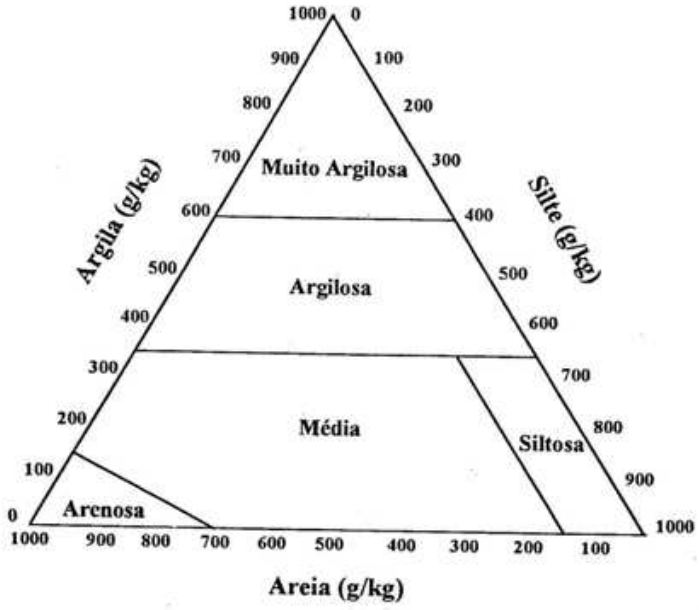

Figura 2: Diagrama triangular simplificado, utilizado pela EMBRAPA, para a classificação textural do solo.

sideradas na obtenção de uma função de covariância válida;

- Derivar métodos de inferência baseados em verossimilhança para a estimação dos parâmetros desconhecidos do modelo;
- Adequar métodos bayesianos aos parâmetros do modelo;
- Desenvolver recursos computacionais livres para análise de dados composicionais;
- Elaborar mapas temáticos de modelos composicionais de solo em estudo de caso.

2 Revisão da Literatura

2.1 Modelo Geoestatístico Gaussiano

2.1.1 Definição do modelo

Williams (2002) define um processo estocástico como a coleção de variáveis aleatórias $\{S(\underline{x}) : \underline{x} \in \mathbb{R}^d\}$ em que d é o número de entradas do vetor de localização $\underline{x}$, e é especificado de maneira consistente pela distribuição conjunta de probabilidade de todo subconjunto finito de variáveis $S(\underline{x}_1), S(\underline{x}_2), ..., S(\underline{x}_n)$. Em particular, para $d = 2$, um processo espacial gaussiano $S = (S(\underline{x}) : \underline{x} \in \mathbb{R}^2)$, é um processo estocástico com a propriedade de que, para qualquer coleção de localizações $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$, com $\underline{x}_i \in \mathbb{R}^2$, $S = (S(\underline{x}_1), S(\underline{x}_2), ..., S(\underline{x}_n))$ tem uma distribuição conjunta gaussiana multivariada e fica completamente especificado pela função média, e pela matriz de covariância cujos elementos correspondem a função $\gamma(\underline{x}_i, \underline{x}_j) = Cov\big(S(\underline{x}_i), S(\underline{x}_j)\big)$ (DIGGLE; RIBEIRO JR., 2007).

Schabenberger e Pierce (2001) destacam que a função de distribuição acumulada é aquela de uma distribuição gaussiana n-variada

$$P\big(S(\underline{x}_1) < s_1,\ ...,\ S(\underline{x}_n) < s_n\big).$$

A geoestatística segundo Diggle, Ribeiro Jr e Christensen (2003) é um ramo da estatística espacial na qual os dados consistem de mensurações $y_1, y_2, ..., y_n$ obtidas nas localizações $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ amostradas em uma região $A \subset \mathbb{R}^2$ espacialmente contínua, relacionados a um fenômeno espacial também contínuo que pode ser tratado como a realização de um processo estocástico $S(\underline{x})$; $\underline{x} \in \mathbb{R}^2$, denominado sinal que em geral não é observável. O valor observado y_i é uma realização de $Y_i = Y(\underline{x}_i)$ que é uma versão ruído de $S(\underline{x}_i)$. O delineamento amostral será de forma que $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ sejam fixos ou estocasticamente independentes de $Y(\underline{x}_1), Y(\underline{x}_2), ..., Y(\underline{x}_n)$.

De acordo com Matérn (1960), a teoria de amostragem espacial mostra que, sob suposições típicas de modelagem, as propriedades são mais eficientemente estimadas por

um delineamento regular do que por um delineamento completamente aleatório. Ainda, se não existir relação entre a escolha de $\underline{x}$ e y, que será o caso deste trabalho, a amostragem será denominada não preferencial, caso contrário, será denominada preferencial (DIGGLE; LEITE; SU, 2007).

A terminologia geoestatística baseada em modelos foi introduzida por Diggle, Tawn e Moyeed (1998) e se caracteriza pela utilização de métodos de inferência estatística baseadas na verossimilhança aplicada a problemas geoestatísticos. Assim, um modelo geoestatístico é a especificação da distribuição conjunta de $S(\underline{x})$ e $Y(\underline{x})$, fatorada como

$$[S(\underline{x}), Y(\underline{x})] = [S(\underline{x})][Y(\underline{x})|S(\underline{x})].$$

Em particular, esse modelo não especifica a distribuição do delineamento amostral $\{\underline{x}_1, \underline{x}_2, ..., \underline{x}_n\}$ o qual será assumido independente de $S(\underline{x})$ e de $Y(\underline{x})$.

De acordo com Diggle, Ribeiro Jr e Christensen (2003), no modelo geoestatístico gaussiano $S(\underline{x}) \sim N(\mu; \sigma^2)$ é um processo gaussiano estacionário, com função de correlação $\rho(u_{ij}) = Corr(S(\underline{x}_i), S(\underline{x}_j))$ em que $u_{ij} = \|\underline{x}_i - \underline{x}_j\|$, $i, j = 1, ..., n$ e Y_i o valor observado na localização $\underline{x}_i$. Essa função de correlação é definida como:

$$\rho(u) = \frac{\gamma(u)}{\sigma^2},$$

que é simétrica em u, ou seja, $\rho(u) = \rho(-u)$.

Os autores definem o modelo geoestatístico plausível como aquele em que a distribuição de Y_i, $i = 1, 2, ..., n$ condicionada a distribuição de S, $S(\cdot)$, é gaussiana com média $S(\underline{x}_i)$ e variância τ^2 e Y_i são mutuamente independentes condicionados em $S(\cdot)$, ou seja:

$$Y_i = S(\underline{x}_i) + Z_i \qquad i = 1, ..., n \tag{1}$$

em que $Z_i \sim N(0; \tau^2)$ são erros aleatórios independentes de $S(\underline{x})$. Esse modelo pode

também ser representado como:

$$\underline{Y} \sim N_n(\mu\underline{1}; \sigma^2\mathbf{R} + \tau^2\mathbf{I}) \tag{2}$$

em que $\underline{1}$ é um vetor com n elementos iguais a 1, $\mathbf{R}$ é uma matriz de ordem $n \times n$ cujos elementos são as correlações $\rho(u_{ij})$ e $\mathbf{I}$ é a matriz identidade de ordem $n \times n$.

Com a reparametrização $\nu^2 = \dfrac{\tau^2}{\sigma^2}$, segue que:

$$\begin{aligned} Var\big(\underline{Y}\big) &= \sigma^2\mathbf{R} + \tau^2\mathbf{I} \\ \frac{1}{\sigma^2}Var\big(\underline{Y}\big) &= \mathbf{R} + \frac{\tau^2}{\sigma^2}\mathbf{I} \\ \frac{1}{\sigma^2}Var\big(\underline{Y}\big) &= \mathbf{V} \end{aligned}$$

onde

$$\mathbf{V} = \mathbf{R} + \nu^2\mathbf{I}, \tag{3}$$

e o modelo da Equação 2 pode ser reescrito como

$$\underline{Y} \sim N_n(\mu\underline{1};\ \sigma^2\mathbf{V}).$$

2.1.2 Componentes do modelo

Segundo Schabenberger e Gotway (2005) um processo estocástico é estacionário se a distribuição espacial de $S(\underline{x})$ é invariante sob translação das coordenadas. Schabenberger e Pierce (2001) afirmam que geometricamente isto implica que a distribuição espacial é invariante sob rotação e estiramento do sistema de localização das amostras. Por outro lado, Bailey e Gatrell (1995) afirmam que isto acontece se as propriedades estatísticas são independentes da localização absoluta em A. Isto implica que a média e a variância são constantes em A e não dependem da localização $\underline{x}$. Implica também que a covariância $Cov\big(S(\underline{x}_i), S(\underline{x}_j)\big)$, $i \neq j$, depende somente das localizações relativas destes dois pontos, da distância u que as separa e da direção entre elas, e não de sua localização absoluta em A. Desta forma, este processo pode ser pensado como o equivalente espa-

cial de uma amostra aleatória em estatística clássica que dá origem a variáveis aleatórias independentes com a mesma média e dispersão (SCHABENBERGER; PIERCE, 2001). Ainda segundo estes autores, a função covariância será denominada isotrópica na ausência de dependência da direção, ou melhor, quando a função de covariância depender somente da distância absoluta (que neste trabalho será considerada a distância euclidiana) entre os pares de pontos. O processo espacial é isotrópico se, em acréscimo à estacionariedade, a covariância depender somente da distância entre dois pontos, e não da direção nos quais estão separados (BAILEY; GATRELL, 1995).

A Figura 3 ilustra uma situação onde os segmentos $\underline{x}_1\underline{x}_2$ e $\underline{x}_3\underline{x}_4$ são estacionários porque estão separados pela mesma distância, têm a mesma direção, estão em locais diferentes em A, e as covariâncias são iguais. Pode-se observar que os segmentos $\underline{x}_1\underline{x}_2$ e $\underline{x}_5\underline{x}_6$ tem a mesma distância mas estão em direções diferentes. Neste caso, ao considerar covariâncias diferentes para os pares $(\underline{x}_1\underline{x}_2)$ e $(\underline{x}_5\underline{x}_6)$, ter-se-ia estacionariedade se, ao rotacionar o par $(\underline{x}_5\underline{x}_6)$ colocando-o na mesma direção de $\underline{x}_1\underline{x}_2$, as covariâncias passarem a ser iguais. Por outro lado, se as covariâncias forem iguais ter-se-ia isotropia pois, em acréscimo a estacionariedade, não houve a necessidade de rotacionar para que tivessem a mesma covariância.

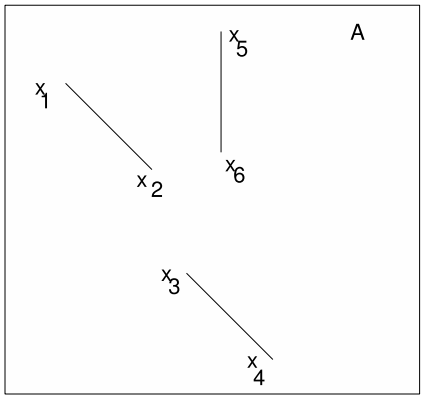

Figura 3: Situações relacionando estacionariedade e isotropia.

Matematicamente, estacionariedade quer dizer que $E\big(S(\underline{x})\big) = \mu$ para todo $\underline{x}$, $Var\big(S(\underline{x})\big) = \sigma^2$ e $\gamma(\underline{x}_i, \underline{x}_j) = \gamma(u)$. Também, um processo estacionário é isotrópico se $\gamma(u) = \gamma(\|u\|)$. Observa-se que o termo estacionário será utilizado para significar um

processo estacionário e isotrópico.

De acordo com Diggle, Ribeiro Jr e Christensen (2003), a especificação da função de correlação, $\rho(u)$, determina a suavidade do processo $S(\underline{x})$. A descrição matemática formal da suavidade de uma superfície espacial é dada por seu grau de diferenciabilidade. $S(\underline{x})$ é quadrado-médio contínuo se $\lim_{u\to 0} E\big(\{S(\underline{x}_i) - S(\underline{x}_j)\}^2\big) = 0$ para todo $\underline{x}$. Da mesma forma, é quadrado-médio diferenciável se existe um processo $S'(\underline{x})$ tal que

$$\lim_{u\to 0} E\left(\left\{\frac{S(\underline{x}_i) - S(\underline{x}_j)}{u} - S'(\underline{x}_i)\right\}^2\right) = 0$$

Então, a diferenciabilidade quadrado-médio de $S(\underline{x})$ está diretamente relacionada com a diferenciabilidade de sua função de covariância através do resultado que diz que se $S(\underline{x})$ é um processo gaussiano estacionário com função de correlação $\rho(u) : u \in \mathbb{R}$, então $S(\underline{x})$ é quadrado-médio contínuo se, e somente se, $\rho(u)$ é contínuo em $u = 0$ e é k vezes quadrado-médio diferenciável se, e somente se, $\rho(u)$ é ao menos $2k$ vezes diferenciável em $u = 0$. A demonstração deste resultado pode ser encontrada no capítulo 2.4 em Stein (1999).

Como em Kitanidis (1997) e em Isaaks e Srisvastava (1989), a matriz de covariância deve ser definida positiva como uma garantia de que a variância de uma variável aleatória formada pela combinação linear ponderada de $S(\underline{x})$, em um número de pontos e que pode ser expressa em termos de funções de covariância, será positiva.

Na estrutura do modelo proposto, a família Matérn de funções de correlação apresentada em Diggle e Ribeiro Jr. (2007) é uma importante função paramétrica de correlação dada por

$$\rho(u, k, \phi) = \frac{1}{2^{k-1}\Gamma(k)} \left(\frac{u}{\phi}\right)^k K_k \left(\frac{u}{\phi}\right)$$

em que $K_k(\cdot)$ denota a função de Bessel modificada de ordem k, $\phi > 0$ é um parâmetro de escala, associado ao alcance e $k > 0$ é um parâmetro de forma que determina a suavidade analítica do processo $S(\underline{x})$, interpretado como uma medida da diferenciabilidade do processo. Especificamente, $S(\underline{x})$ é $\lceil k - 1$ vezes quadrado médio diferenciável, onde o símbolo $\lceil$, denominado ceiling, significa o menor inteiro maior ou igual a k. Devido

a dificuldade de identificação de todos os parâmetros do modelo, os valores de k são escolhidos dentre o conjunto de valores $\{0,5;\ 1,5;\ 2,5\}$ correspondendo respectivamente à não diferenciabilidade ou a um processo estocástico uma ou duas vezes diferenciável na origem. A partir do valor de k pode-se estabelecer a amplitude prática do modelo que é a distância u_0 no qual $\rho(u_0) = \alpha$, onde α é um valor tão pequeno quanto o pesquisador determinar. Para o conjunto estabelecido acima, os valores de u_0 serão aproximadamente $\{3\phi;\ 4,75\phi;\ 5,92\phi\}$, respectivamente. Nesta família, fazendo $k = 0,5$ obtém-se a função de correlação exponencial $\rho(u) = \exp\left(-\frac{u}{\phi}\right)$ e $\lim_{k\to\infty} \rho(u) = \exp\left\{-\left(\frac{u}{\phi}\right)^2\right\}$ obtendo-se a função de correlação gaussiana para a qual $u_0 \simeq \sqrt{3}\phi$.

Em Isaaks e Srisvastava (1989), Kitanidis (1997), Goovaerts (1997), autores da linha tradicional da geoestatística e Schabenberger e Pierce (2001), Diggle e Ribeiro Jr. (2007), autores da geoestatística baseada em modelos, pode-se encontrar outras funções de correlação. Neste trabalho será adotada a função de correlação da família Matérn.

Quando não existe estacionariedade na média a situação mais comum é que $\mu(\underline{x})$, denominada superfície de tendência, seja escrita como um modelo de regressão polinomial usando potências e produtos cruzados das coordenadas cartesianas de $\underline{x}$ como variáveis exploratórias. Diggle e Ribeiro Jr. (2007) afirmam que superfícies de tendência linear e quadrática podem fornecer descritores empíricos úteis da tendência espacial não explicada, mas superfícies de ordem maior devem ser evitadas porque tendências mais complexas podem ser melhor descritas através do componente estocástico do modelo.

A especificação de $\mu(\underline{x})$ pode também ser feita em função de outras variáveis explanatórias e neste caso a média se associa a uma tendência externa.

Neste trabalho, assume-se uma superfície de tendência linear dada por

$$\mu(\underline{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j d_j(\underline{x})$$

em que $d_j(\underline{x})$ são variáveis explanatórias espaciais, dependentes ou não das coordenadas.

Outro componente implícito do modelo é o efeito pepita que é um termo usado para representar a variância τ^2 da variável Z_i. Esta variância pode ser dividida em dois

componentes como

$$\tau^2 = \text{EM} + \text{VME} \tag{4}$$

em que EM significa erro de medida e VME variação de pequena ou micro escala, que representa uma variação não capturada pelo processo $S(\underline{x})$ e que ocorre em uma distância menor do que a menor distância entre duas localizações. Em muitas situações, a distinção dos dois componentes não é possível, mas conforme Equação 4 pode-se determinar o valor de VME se o valor de EM for conhecido. No entanto, isto só será possível se houver mais de uma observação de Y na mesma localização. Nesse caso, a distância será nula e o valor de EM será o resultado da média aritmética das quantidades $v_{ij} = \frac{1}{2}(Y_i - Y_j)^2$, $i = 1, ..., n$, $j > i$.

Quando os dados observados Y são contínuos mas não seguem uma distribuição gaussiana ou em problemas de não estacionariedade da variância, o valor do erro quadrático médio mínimo do preditor é afetado de forma a se obter aproximações ruins. Em muitos casos, através de transformações, é possível que os dados passem a seguir uma distribuição gaussiana.

Box e Cox (1964) propõe a transformação

$$Y^* = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda} & , \quad \lambda \neq 0 \\ \log Y & , \quad \lambda = 0 \end{cases}$$

em que o parâmetro λ introduz flexibilidade ao modelo. Valores interpretáveis para esse parâmetro são:

$$\begin{aligned} \lambda = \quad & 1,0 \quad : \text{ Sem transformação} \\ \lambda = \quad & 0,5 \quad : \text{ Transformação raiz quadrada} \\ \lambda = \quad & 1,5 \quad : \text{ Transformação logarítmica} \\ \lambda = \quad & -1,0 \quad : \text{ Transformação recíproco.} \end{aligned}$$

2.1.3 Estimação de parâmetros do modelo

Tem-se definido um modelo gaussiano linear quando a estrutura de média e de covariância são possíveis de serem estimadas. Os métodos mais comumente utilizados para a estimação de parâmetros são o método dos mínimos quadrados (MMQ) e o método da máxima verossimilhança (MMV).

Na estimação da tendência pelo Método dos Mínimos Quadrados, a média do modelo, $E\big(Y\big)$, é dada como

$$\mu(\underline{x}) = \beta_0 + \sum_{j=1}^{n} \beta_j d_j(\underline{x}),$$

em que $d_j,\ j = 1, ..., n$ são variáveis explanatórias espaciais e $\beta_j,\ j = 0, ..., n$ os parâmetros da regressão linear.

Usando o método dos mínimos quadrados ordinários, as estimativas $\tilde{\beta}_j^*$ são aquelas que minimizam a soma dos quadrados dos resíduos (SQR), e que sob o modelo da Equação 1 é dada por

$$\text{SQR}(\beta) = \sum_{i=1}^{n} Z_i^2(\underline{x}) = \sum_{i=1}^{n} (Y_i - \mu_i - S(\underline{x}_i))^2.$$

Para estimar corretamente os parâmetros da média, os parâmetros da correlação deveriam ser conhecidos mas geralmente não o são. Por outro lado, para estimar os da correlação, seriam necessários os da média que tampouco são conhecidos. O que se faz então é considerar, inicialmente, o modelo $Y_i = \mu_i + Z_i$ que ignora a correlação e cujo SQR é dado por

$$\text{SQR}(\beta) = \sum_{i=1}^{n} Z_i^2(x) = \sum_{i=1}^{n} (Y_i - \mu_i). \tag{5}$$

Matricialmente, a equação de regressão linear múltipla estimada é

$$\underline{\hat{Y}} = \mathbf{D}\underline{\tilde{\beta}}$$

sendo $\mathbf{D}$ uma matriz $n \times p$ de covariáveis, $\underline{\tilde{\beta}}$ o vetor dos parâmetros da regressão e a aplicação do MMQ resultará no estimador consistente

$$\underline{\tilde{\beta}} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\underline{Y}.$$

Assim, com as estimativas $\tilde{\beta}_j$ pode-se calcular o vetor dos resíduos

$$\underline{Z} = \underline{Y} - \mathbf{D}\underline{\tilde{\beta}} \tag{6}$$

com elementos Z_i, $i = 1, ..., n$ que substituídos na Equação 1 permitirão a obtenção das estimativas dos parâmetros de $S(\underline{x})$, e por sua vez a obtenção do estimador $\hat{\mathbf{V}}$ para $\mathbf{V}$ apresentado na Equação 3. Com este estimador é possível se obter uma estimativa de $\underline{\beta}$ mais eficiente dada pela estimativa de mínimos quadrados generalizados conforme a equação

$$\underline{\hat{\beta}} = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{V}^{-1}\underline{Y}, \tag{7}$$

substituindo-se nesta $\mathbf{V}$ por $\hat{\mathbf{V}}$. Este procedimento é repetido até a obtenção da convergência.

Estimação de parâmetros do modelo pelo Método da Função de Verossimilhança:

Por outro lado, utilizando o método da máxima verossimilhança, pode-se obter todas as estimativas ao mesmo tempo. No caso de $\underline{Y}$ apresentar distribuição gaussiana, o $\underline{\hat{\beta}}$ dado pela Equação 7 coincide com a estimativa de máxima verossimilhança.

Admitindo uma superfície de tendência polinomial para $\mu(\underline{x})$, tem-se

$$\underline{Y} \sim N_n(\mathbf{D}\underline{\beta};\, \sigma^2\mathbf{R} + \tau^2\mathbf{I}) \tag{8}$$

onde $\mathbf{D}$ é uma matriz $n \times p$ de covariáveis; $\underline{\beta}$ é o vetor de parâmetros da regressão correspondente e a matriz do correlação $\mathbf{R}$, a matriz de correlação, depende de ϕ e k.

A função de verossimilhança de $\underline{Y}$ será dada por:

$$\begin{aligned} L(\underline{Y}) &= \frac{(2\pi)^{-n/2}}{|\sigma^2\mathbf{R}+\tau^2\mathbf{I}|^{1/2}}\cdot \\ &\quad \exp\{-\frac{1}{2}\left(\underline{Y}-\mathbf{D}\underline{\beta}\right)'\left(\sigma^2\mathbf{R}+\tau^2\mathbf{I}\right)^{-1}\left(\underline{Y}-\mathbf{D}\underline{\beta}\right)\} \end{aligned}$$

Logo, a função de log-verossimilhança é:

$$\begin{aligned} l(\underline{\beta},\tau^2,\sigma^2,\phi) &= log[(2\pi)^{-n/2}] - log(|\sigma^2\mathbf{R}+\tau^2\mathbf{I}|^{1/2}) \\ &\quad -\frac{1}{2}(\underline{Y}-\mathbf{D}\underline{\beta})'(\sigma^2\mathbf{R}+\tau^2\mathbf{I})^{-1}(\underline{Y}-\mathbf{D}\underline{\beta}) \\ &= -0,5\left\{n\,log(2\pi)+log(|\sigma^2\mathbf{R}+\tau^2\mathbf{I}|)\right. \\ &\quad \left.+\,(\underline{Y}-\mathbf{D}\underline{\beta})'(\sigma^2\mathbf{R}+\tau^2\mathbf{I})^{-1}(\underline{Y}-\mathbf{D}\underline{\beta})\right\} \end{aligned} \tag{9}$$

Para proceder a maximização da Equação 9, considera-se $\underline{Y} \sim N_n(\mathbf{D}\underline{\beta};\sigma^2\mathbf{V})$ e os seguintes resultados da álgebra matricial:

$$\frac{\partial}{\partial \underline{X}}(\mathbf{A}\underline{X}) = \mathbf{A}' \tag{10}$$

$$\frac{\partial}{\partial \underline{X}}(\underline{X}'\mathbf{A}\underline{X}) = 2\mathbf{A}\underline{X}, \tag{11}$$

em que $\mathbf{A}$ é uma matriz quadrada de ordem $n \times n$ e $\underline{X}$ um vetor de ordem $n \times 1$. A Equação 9 pode ser reescrita como

$$l(\underline{\beta},\sigma^2) = -\frac{1}{2}\{n\,log(2\pi)+log(|\sigma^2\mathbf{V}|^{1/2})+(\underline{Y}-\mathbf{D}\underline{\beta})'(\sigma^2\mathbf{V})^{-1}(\underline{Y}-\mathbf{D}\underline{\beta})\}. \tag{12}$$

Nessa equação,

$$\begin{aligned} (\underline{Y}-\mathbf{D}\underline{\beta})'(\sigma^2\mathbf{V})^{-1}(\underline{Y}-\mathbf{D}\underline{\beta}) &= \underline{Y}'(\sigma^2\mathbf{V})^{-1}\underline{Y}-\underline{Y}'(\sigma^2\mathbf{V})^{-1}\mathbf{D}\underline{\beta}-\underline{\beta}'\mathbf{D}'(\sigma^2\mathbf{V})^{-1}\underline{Y}+ \\ &\quad \underline{\beta}'\mathbf{D}'(\sigma^2\mathbf{V})^{-1}\mathbf{D}\underline{\beta} \end{aligned}$$

de modo que permita expressar a Equação 11 como:

$$\begin{aligned} l(\underline{\beta}, \sigma^2) &= -\frac{1}{2}\{n\,log(2\pi) + log(|\sigma^2\mathbf{V}|) + \\ &\quad \frac{1}{\sigma^2}[\underline{Y}'\mathbf{V}^{-1}\underline{Y} - 2\underline{Y}'\mathbf{V}^{-1}\mathbf{D}\underline{\beta} + \underline{\beta}'(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})\underline{\beta}]\}. \end{aligned} \tag{13}$$

Aplicando os resultados (11) e (11) na Equação 13 tem-se:

$$\frac{\partial}{\partial \underline{\beta}} l(\underline{\beta}, \sigma^2) = -\frac{1}{\sigma^2}\left[-(\underline{Y}'\mathbf{V}^{-1}\mathbf{D})' + (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})\underline{\beta}\right].$$

Fazendo

$$\frac{1}{\sigma^2}\left(\mathbf{D}'\mathbf{V}^{-1}\underline{Y} - \mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\hat{\underline{\beta}}\right) = 0, \quad \text{vem:}$$

$$\hat{\underline{\beta}} = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{V}^{-1}\underline{Y}. \tag{14}$$

Nesta equação, $\hat{\underline{\beta}}$ é a estimativa de mínimos quadrados generalizada e depende somente de ϕ e ν^2. Reescrevendo a Equação 12 e usando o fato de que $|Var(\underline{Y})| = |\sigma^2\mathbf{V}| = (\sigma^2)^n|\mathbf{V}|$ tem-se que

$$l(\underline{\beta}, \sigma^2) = -\frac{1}{2}\left\{n\,log(2\pi) + n\,log(\sigma^2) + log|\mathbf{V}| + \frac{(\underline{Y} - \mathbf{D}\underline{\beta})'\mathbf{V}^{-1}(\underline{Y} - \mathbf{D}\underline{\beta})}{\sigma^2}\right\} \tag{15}$$

Derivando a Equação 15 em relação a σ^2 vem

$$\frac{\partial}{\partial \sigma^2} l(\underline{\beta}, \sigma^2) = -\frac{1}{2}\left\{\frac{n}{\sigma^2} + (\underline{Y} - \mathbf{D}\underline{\beta})'\mathbf{V}^{-1}(\underline{Y} - \mathbf{D}\underline{\beta}) - \frac{1}{(\sigma^2)^2}\right\}$$

Fazendo $-\frac{1}{2}\left[\frac{n}{\hat{\sigma}^2} - \frac{(\underline{Y} - \mathbf{D}\hat{\underline{\beta}})'\mathbf{V}^{-1}(\underline{Y} - \mathbf{D}\hat{\underline{\beta}})}{(\hat{\sigma}^2)^2}\right] = 0$, tem-se:

$$\hat{\sigma}^2 = n^{-1}\left[(\underline{Y} - \mathbf{D}\hat{\underline{\beta}})'\mathbf{V}^{-1}(\underline{Y} - \mathbf{D}\hat{\underline{\beta}})\right] \tag{16}$$

No ponto de máximo, $\underline{\beta} = \hat{\underline{\beta}}$ e $\sigma^2 = \hat{\sigma}^2$ de forma que substituindo em (15) obtém-se

a log-verossimilhança concentrada

$$l_0(\nu^2, \phi, k) = -\frac{1}{2}\left\{n\,log(2\pi) + n\,log(\hat{\sigma}^2) + log|\mathbf{V}| + n\right\} \tag{17}$$

que recebe as constantes $\underline{Y}$, a matriz $\mathbf{D}$ cujos elementos são os valores das covariáveis e a matriz de distâncias calculadas $\mathbf{V}$. Para obter a estimativa dos parâmetros a Equação 17 deve ser otimizada numericamente com relação a ϕ e ν. Com o valor obtido para $\hat{\phi}$ obtém-se a matriz $\hat{\mathbf{V}}$, substituindo-a na Equação 15 obtém-se $\hat{\underline{\beta}}$ e, consequentemente, $\hat{\sigma}^2$. Desta forma, substituindo-se $\hat{\nu}$ e $\hat{\sigma}^2$ em $\nu^2 = \tau^2/\sigma^2$ obtém-se $\hat{\tau}^2$. O Método Delta (DEGROOT; SCERVISH, 2002) pode então ser aplicado para a obtenção da variância dos estimadores.

Quando os dados $\underline{Y}$ sofrem uma transformação como na seção 2.1.2, aplicando o jacobiano da transformação obtém-se a log-verossimilhança:

$$\begin{aligned} l(\underline{\beta}, \sigma^2, \phi, \nu^2, \lambda) \;\; = \;\; & (\lambda - 1)\sum_{i=1}^{n} log(y_i) - 0,5\left\{n\,log(2\pi) + log|\sigma^2\mathbf{V}(\phi, \nu^2)| + \right. \\ & \left. (\underline{Y}^* - \mathbf{D}\underline{\beta})'[\sigma^2\mathbf{V}(\phi, \nu^2)]^{-1}(\underline{Y}^* - \mathbf{D}\underline{\beta})\right\}. \end{aligned}$$

que é otimizável numericamente.

2.2 Predição linear espacial

2.2.1 Conceitos de predição

Considere o vetor $\underline{\mathrm{S}} = (S(\underline{x}_1), S(\underline{x}_2), ..., S(\underline{x}_n))'$ com distribuição multivariada

$$\underline{\mathrm{S}} \sim N_n(\mu\underline{1};\, \sigma^2\mathbf{R}),$$

em que $\mathbf{R}$ é uma matriz de ordem $n \times n$ com elementos $r_{ij} = \rho(\|\underline{x}_i - \underline{x}_j\|)$ e

$$\underline{Y} \sim N_n(\mu\underline{1};\, \sigma^2\mathbf{V}).$$

O interesse está na predição do processo estacionário $S(\underline{x})$ em uma localização

onde $\underset{\sim}{Y}$ não foi observado. O vetor $\underset{\sim}{Y}$ tem como elementos variáveis aleatórias cujos valores são observados e considera-se T uma variável aleatória cujo valor será predito a partir do valor de $\underset{\sim}{Y}$.

Um preditor pontual de T é uma função qualquer de $\underset{\sim}{Y}$ representada por

$$\hat{T} = t(\underset{\sim}{Y}),$$

que tem erro de predição quadrático médio (EQM) definido como

$$\text{EQM}(\hat{T}) = E\big((T - \hat{T})^2\big).$$

O EQM($\hat{T}$) assume o valor mínimo quando $\hat{T} = E\big(T|\underset{\sim}{Y}\big)$ pois

$$\begin{aligned} E\Big((T - \hat{T})^2\Big) &= E_{\underset{\sim}{Y}}\Big(E_T\big((T - \hat{T})^2|\underset{\sim}{Y}\big)\Big) \\ &= E_{\underset{\sim}{Y}}\Big(Var_T\big((T - \hat{T})|\underset{\sim}{Y}\big) + \{E_T\big((T - \hat{T})|\underset{\sim}{Y}\big)\}^2\Big) \end{aligned} \tag{18}$$

em que os subscritos nos dois operadores da esperança indicam que as esperanças são calculadas com relação a $\underset{\sim}{Y}$ e S, respectivamente.

Dado que $Var_T\big(\hat{T}|\underset{\sim}{Y}\big)$ e $Cov\big(T|\underset{\sim}{Y}, \hat{T}|\underset{\sim}{Y}\big)$ são zero já que condicionado em $\underset{\sim}{Y}$, $\hat{T}$ que é função de $\underset{\sim}{Y}$ é constante, tem-se

$$\begin{aligned} Var_T\big((T - \hat{T}|\underset{\sim}{Y})\big) &= Var_T\big(T|\underset{\sim}{Y}\big) + Var_S\big(\hat{T}|\underset{\sim}{Y}\big) - 2Cov_S\big(T|\underset{\sim}{Y}, \hat{T}|\underset{\sim}{Y}\big) \\ &= Var_T\big(T|\underset{\sim}{Y}\big) \end{aligned}$$

e

$$\begin{aligned} E_T\big((T - \hat{T})|\underset{\sim}{Y}\big) &= E\big(T|\underset{\sim}{Y}\big) - E\big(\hat{T}|\underset{\sim}{Y}\big) \\ &= E\big(T|\underset{\sim}{Y}\big) - \hat{T} \end{aligned}$$

que substituídas na Equação 18 fornecem

$$E\big((T - \hat{T})^2\big) = E_{\underset{\sim}{Y}}\big(Var_T(T|\underset{\sim}{Y}) + \{E\big(T|\underset{\sim}{Y}\big) - \hat{T}\}^2\big) \tag{19}$$

Da equação 19 obtém-se o erro quadrático médio de $\hat{T}$

$$E\big((T-\hat{T})^2\big) \;=\; E_{\underline{Y}}\big(Var_T(T|\underline{Y})\big), \tag{20}$$

quando $\hat{T} = E\big(T|\underline{Y}\big)$.

Nota-se também que

$$\begin{aligned} E\big((T-\hat{T})^2\big) &= Var\big(T-\hat{T}\big) + \{E\big(T-\hat{T}\big)\}^2 \\ &= Var\big(T\big) + Var\big(\hat{T}\big) - 2Cov\big(T,\hat{T}\big) + \{E\big(T\big) - \hat{T}\}^2 \\ &= Var\big(T\big) - 2Cov\big(T,\hat{T}\big). \end{aligned}$$

Então,

$$Var\big(T\big) = E\big((T-\hat{T})^2\big) + 2Cov\big(T,\hat{T}\big)$$

e, consequentemente,

$$E\big((T-\hat{T})^2\big) \leq Var\big(T\big)$$

se T e $\underline{Y}$ são independentes.

Se $S(\underline{x})$ for um processo gaussiano estacionário, os dados $\underline{Y}$ são gerados por um modelo gaussiano estacionário e se T for igual a $S(\underline{x})$, $\big(T,\, \underline{Y}\big) = \big(S(\underline{x}),\, \underline{Y}\big)$ é gaussiana multivariada e a distribuição condicional de T dado $\underline{Y}$ (MOOD; GRAYBILL; BOES, 1974) é também gaussiana com média

$$\mu_{T|\underline{Y}} = \mu_T + \frac{\Sigma_{T\underline{Y}}}{\Sigma_{\underline{Y}\underline{Y}}}(\underline{Y} - \mu_{\underline{Y}})$$

e variância

$$\Sigma_{T|\underline{Y}} = \Sigma_{TT} - \frac{\Sigma_{T\underline{Y}}}{\Sigma_{\underline{Y}\underline{Y}}}\Sigma_{\underline{Y}T}$$

Logo, $(T, \underline{Y})$ é gaussiana multivariada com média $\mu\underline{1}$ e matriz de covariância

$$\begin{bmatrix} \sigma^2 & \sigma^2\underline{r}' \\ \sigma^2\underline{r} & \sigma^2\mathbf{V} \end{bmatrix}.$$

em que $\underline{r}$ é um vetor com elementos $r_i = \rho(\| \underline{x} - \underline{x}_i \|)$.

Desta forma, o preditor do erro quadrado médio mínimo para S é:

$$\begin{aligned} \hat{S} &= \mu + \sigma^2 \underline{r}' \frac{\mathbf{V}^{-1}}{\sigma^2}(\underline{Y} - \mu\underline{1}) \\ &= \mu + \underline{r}'\mathbf{V}^{-1}(\underline{Y} - \mu\underline{1}) \end{aligned} \tag{21}$$

com variância de predição

$$\begin{aligned} Var\big(S|\underline{Y}\big) &= \sigma^2 - \sigma^2 \underline{r}' \frac{\mathbf{V}^{-1}}{\sigma^2}\sigma^2 \underline{r} \\ &= \sigma^2 + (1 - \underline{r}'\mathbf{V}^{-1}\underline{r}). \end{aligned} \tag{22}$$

Como a variância de predição não depende de $\underline{Y}$, da Equação 20 tem-se $E\big((S - \hat{S})^2\big) = Var\big(S|\underline{Y}\big)$.

Ao escrever o preditor de S em termos de $\hat{S}(\underline{x}_0)$ onde $\underline{x}_0$ é a localização de predição, pode-se observar que $\underline{r}'\mathbf{V}^{-1}$ nada mais é do que uma combinação linear da média μ e de Y_i de modo que

$$\begin{aligned} \hat{S}(\underline{x}_0) &= \mu + \sum_{i=1}^{n} a_i(\underline{x}_0)(Y_i - \mu) \\ &= \mu + \sum_{i=1}^{n} a_i(\underline{x}_0)Y_i - \sum_{i=1}^{n} a_i(\underline{x}_0)\mu \\ &= \{1 - \sum_{i=1}^{n} a_i(\underline{x}_0)\}\mu + \sum_{i=1}^{n} a_i(\underline{x}_0)Y_i \end{aligned} \tag{23}$$

onde $a_1(\underline{x}_0), a_2(\underline{x}_0), ..., a_n(\underline{x}_0)$, são denominados pesos de predição. A configuração dos pontos (localizações) é importante na análise. Na distribuição dos pesos, não só a distância entre os pontos é importante mas a posição relativa também influi. Geralmente, para pontos próximos do local de predição são dados pesos mais altos e à medida que a distância aumenta os pesos devem diminuir. Considerando-se n localizações, se não houver dependência espacial os pesos serão iguais a $1/n$. Agora, quanto maior a dependência espacial maior a penalidade para pontos amostrais próximos pois isto acarretaria, entre outros fatores, em aumento de custo nas observações e a informação obtida poderia ser

insignificantemte maior à que seria obtida se fossem substituídos por apenas um ponto, por exemplo. Ainda nesta situação, destaca-se que os pesos atribuídos a pontos colineares ao ponto de predição são distribuídos de forma que o mais próximo recebe peso mais alto e o mais afastado pode até mesmo receber um peso negativo.

Há que se destacar ainda que sendo o modelo $Y(\underline{x}_i) = S(\underline{x}_i) + Z_i$, ao considerar $\tau^2 = 0$ tem-se que $\hat{S}(\underline{x}_i) = y_i$, para todo $i = 1, ..., n$, significando que a superfície $\hat{S}(\underline{x})$ interpola os dados. Mas à medida que o valor de τ^2 aumenta, vai ocorrendo uma suavização da superfície de modo que no limite a superfície torna-se a média do processo.

Segundo Diggle, Ribeiro Jr e Christensen (2003), em muitas aplicações, o foco inferencial não está em $S(\underline{x}_0)$, mas em uma média ou valor máximo, por exemplo. Primeiramente, os autores consideram T qualquer funcional linear de $S(\underline{x})$, ou seja,

$$T = \int_A a(\underline{x}) S(\underline{x})\, d\underline{x}$$

para alguma função peso $a(\underline{x})$. Como já visto, sob o modelo gaussiano, $[T, \underline{Y}]$ é gaussiana multivariada e $[T|\underline{Y} = \underline{y}]$ é gaussiana univariada. A média é dada por

$$E\big(T|\underline{Y}\big) = \int_A a(\underline{x}) E\big(S(\underline{x})|\underline{Y}\big)\, d\underline{x},$$

que resulta

$$\hat{T} = \int_A a(\underline{x}) \hat{S}(\underline{x})\, d\underline{x}.$$

A variância de $T|\underline{Y}$ será

$$Var\big(T|\underline{Y}\big) = \int_A \int_A a(\underline{x}) a(\underline{x}') Cov\big(S(\underline{x}), S(\underline{x}')\big)\, d\underline{x} d\underline{x}'.$$

Em outras palavras, os autores afirmam que dada a superfície predita $\hat{S}(\underline{x})$, é razoável calcular qualquer propriedade linear desta superfície e usar o resultado como o preditor para a propriedade linear correspondente da superfície verdadeira $S(\underline{x})$. Isto não será válido para propriedades não lineares.

2.2.2 Krigagem

Banerjee, Carlin e Gelfand (2004) colocam que o problema é de predição espacial ótima: dados as observações de um processo estocástico $\underline{Y} = (Y(\underline{x}_1), Y(\underline{x}_2), ..., Y(\underline{x}_n))'$ deseja-se predizer a variável Y em uma localização não observada. Em outras palavras, deseja-se encontrar o melhor preditor do valor de $Y(\underline{x}_0)$ baseado nas observações y de Y.

Como visto na seção 2.3.1, o melhor preditor é o que apresenta o menor erro quadrático médio para $T = S(\underline{x}_0)$ e é dado por

$$\hat{T} \;=\; \mu + \underline{r}'\mathbf{V}^{-1}(\underline{Y} - \mu\underline{1}) \tag{24}$$

com variância de predição

$$Var\big(T|\underline{Y}\big) \;=\; \sigma^2(1 - \underline{r}'\mathbf{V}^{-1}\underline{r}) \tag{25}$$

Observa-se que devido os parâmetros do modelo serem quantidades desconhecidas, as estimativas obtidas a partir das Equações 14 e 16 são substituídas nas Equações 24 e 25. O preditor $\hat{T}$ é então linear nos dados e este método é conhecido como krigagem simples.

No método conhecido como krigagem ordinária, o parâmetro média é tratado como desconhecido e os da covariância são conhecidos. Assim, o preditor é escrito como a combinação linear

$$\hat{T} = \hat{S}(\underline{x}) = \sum_{i=1}^{n} a_i(\underline{x})Y_i$$

onde $a_i(\underline{x})$, os pesos de krigagem satisfazem $\sum_{i=1}^{n} a_i(\underline{x}) = 1$ para qualquer localização de predição. Em forma equivalente a Equação 24, substitui-se a média μ pelo estimador de mínimos quadrados generalizados

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

de onde segue que

$$\hat{T} = (\underline{1}'\mathbf{V}^{-1}\underline{1})^{-1}\underline{1}'\mathbf{V}^{-1}\underline{Y} + \underline{r}'\mathbf{V}^{-1}\left[\underline{Y} - (\underline{1}'\mathbf{V}^{-1}\underline{1})^{-1}\underline{1}'\mathbf{V}^{-1}\underline{Y}\right].$$

Quando uma transformação nos dados originais é feita com o objetivo de que estes passem a seguir uma distribuição gaussiana, em geral os dados transformados são escritos como $\underline{Y}^* = h(\underline{Y})$ onde $h(\cdot)$ é uma função, por exemplo, a função logarítmica. Neste caso, $h(\cdot) = log(\cdot)$ implica $h^{-1}(\cdot) = \exp(\cdot)$.

Supondo, então, $T(\underline{x}) = \exp\{\mu + S(\underline{x})\}$, pode-se escrever

$$T(\underline{x}) = \exp\{\mu\} + \exp\{S(\underline{x})\} = \exp\{\mu\} + T_0(\underline{x}).$$

A distribuição de $S(\underline{x})$ dado $\underline{Y}^*$ é gaussiana univariada com média $\hat{S}(\underline{x})$ e variância $v(\underline{x})$ dadas pelas Equações 24 e 25, respectivamente, substituindo-se $\underline{Y}$ por $\underline{Y}^*$. Então, a função geratriz de momentos de $S(\underline{x})$ é,

$$\psi_{S(\underline{x})}(a) = E\Big(\mathrm{e}^{aS(\underline{x})}\Big) = \mathrm{e}^{a\hat{S}(\underline{x})+\frac{1}{2}a^2 v(\underline{x})} \qquad a \in \mathbb{R},$$

a qual para $a = 1$, resulta em

$$\hat{T}_0(\underline{x}) = E\big(T_0(\underline{x})\big) = \mathrm{e}^{\hat{S}(\underline{x})+\frac{v(\underline{x})}{2}},$$

e para $a = 2$ em

$$E\big((T_0(\underline{x}))^2\big) = \mathrm{e}^{2\hat{S}(\underline{x})+2v(\underline{x})},$$

de onde se obtém a variância de predição:

$$Var\big(T_0(\underline{x})|\underline{Y}^*\big) = \mathrm{e}^{2\hat{S}(\underline{x})+v(\underline{x})}[\mathrm{e}^{v(\underline{x})} - 1].$$

2.2.3 Inferência Bayesiana para predição espacial

Na inferência bayesiana se considera os parâmetros desconhecidos do modelo como variáveis aleatórias. Dessa forma, é possível levar em conta as incertezas envolvidas na estimação dos parâmetros. Para isto, considere o vetor de parâmetros $\underline{\theta} = (\underline{\beta}, \sigma^2, \phi, \tau^2)'$ e o vetor aleatório $\underline{Y}$ com distribuição de probabilidade dada pela função $P(\underline{Y}|\underline{\theta})$.

Para $\underline{Y}$ dado como na Equação 8:

$$\underline{Y} \sim NM(\mathbf{D}\underline{\beta};\, \sigma^2\mathbf{R} + \tau^2\mathbf{I})$$

a função de verossimilhança é

$$L(\underline{\theta}|\underline{Y}) \;\propto\; |\sigma^2\mathbf{R} + \tau^2\mathbf{I}|^{-\frac{1}{2}} \cdot \exp\{-\frac{1}{2}\left(\underline{Y} - \mathbf{D}\underline{\beta}\right)'\left(\sigma^2\mathbf{R} + \tau^2\mathbf{I}\right)^{-1}\left(\underline{Y} - \mathbf{D}\underline{\beta}\right)\}. \quad (26)$$

e sendo $\underline{Y}$ e $\underline{\theta}$ vetores aleatórios, $P(\underline{Y}, \underline{\theta}) = P(\underline{Y}|\underline{\theta})\, P(\underline{\theta})$ será a distribuição conjunta. As informações sobre os parâmetros do modelo e que são externas aos dados está refletida na distribuição *priori* $P(\underline{\theta})$. O Teorema de Bayes, combina a distribuição *priori* e a verossimilhança de tal forma que o conhecimento à *priori* sobre os parâmetros é atualizado, após a coleta de dados, usando a relação:

$$P(\underline{\theta}|\underline{Y}) \;\propto\; P(\underline{\theta})\, P(\underline{Y}|\underline{\theta}) \quad (27)$$

A distribuição $P(\underline{\theta}|\underline{Y})$ é denominada distribuição a *posteriori* e é a base da inferência bayesiana sobre os parâmetros do modelo. A distribuição *posteriori* para o modelo $\underline{Y}$ é:

$$\begin{aligned} P(\underline{\beta}, \sigma^2, \phi, \tau^2|\underline{Y}) \;\propto\; & P(\underline{\beta}, \sigma^2, \phi, \tau^2)|\sigma^2\mathbf{R} + \tau^2\mathbf{I}|^{-\frac{1}{2}} \cdot \\ & \exp\{-\frac{1}{2}\left(\underline{Y} - \mathbf{D}\underline{\beta}\right)'\left(\sigma^2\mathbf{R} + \tau^2\mathbf{I}\right)^{-1}\left(\underline{Y} - \mathbf{D}\underline{\beta}\right)\} \end{aligned} \quad (28)$$

Quanto às *prioris*, a escolha é uma questão delicada em inferência bayesiana. *Prioris* que levam a uma *posteriori* da mesma família de distribuições são chamadas *prioris* conjugadas. Essas *prioris* podem ser computacionalmente convenientes mas não

deveriam ser escolhidas somente por isso. Dois casos extremos para a escolha da *priori* são: quando os parâmetros são perfeitamente conhecidos as *prioris* podem ser vistas como distribuições degeneradas nos valores dos parâmetros; quando o conhecimento da *priori* sobre os parâmetros é vaga podem ser adotadas *prioris* não informativas.

Observa-se que não sendo possível derivar analiticamente uma distribuição *posteriori* de modo que esta se apresente como uma distribuição conhecida, o procedimento seria amostrar desta distribuição obtendo-se então as estimativas desejadas. Caso contrário, se não for possível derivá-la analiticamente, utiliza-se de métodos computacionalmente intensivos.

A metodologia da teoria de decisão (Berger (1985), *apud* Ribeiro Jr e Diggle (1999)) leva a escolhas ótimas de estimativas pontuais. As funções de perda definem a qualidade dos estimadores. O erro quadrático médio, por exemplo, corresponde a função de perda quadrática.

Por outro lado, a base da predição bayesiana é a distribuição preditiva $P(\underline{Y}_0|\underline{Y})$. A distribuição preditiva leva em consideração a incerteza sobre os parâmetros calculando a média da distribuição condicional $P(\underline{Y}_0|\underline{Y},\underline{\theta})$, sobre o espaço dos parâmetros, com pesos dados pela distribuição *posteriori* dos parâmetros do modelo $P(\underline{\theta}|\underline{Y})$:

$$\begin{aligned} P(\underline{Y}_0,\underline{Y}) &= \int P(\underline{Y}_0,\underline{\theta}|\underline{Y})\,d\underline{\theta} \\ &= \int p(\underline{Y}_0|\underline{Y},\underline{\theta})p(\underline{\theta}|\underline{Y})\,d\underline{\theta} \end{aligned} \tag{29}$$

A distribuição preditiva pode ainda ser escrita como

$$P(\underline{Y}_0,\underline{Y}) = \int \frac{P(\underline{Y}_0,\underline{Y}|\underline{\theta})P(\underline{\theta}|\underline{Y})}{\int P(\underline{Y}|\underline{\theta})P(\underline{\theta})\,d\underline{\theta}}\,d\underline{\theta}.$$

ssiana

Nota-se que nos métodos geoestatísticos convencionais, os parâmetros do modelo são estimados e são usados na predição, como no item 2.3.3, onde o preditor de krigagem

é baseado na distribuição

$$\underline{Y}_0|\underline{Y} \sim P(\underline{Y}_0|\underline{Y}, \hat{\underline{\theta}}). \tag{30}$$

Assim, a predição bayesiana pode ser interpretada como uma média ponderada das predições utilizando as estimativas dos parâmetros como se fossem os verdadeiros valores.

2.3 Matriz de Covariância, Matriz de Correlação Cruzada e Variograma Cruzado

Ao realizar uma pesquisa os dados coletados, frequentemente por amostragem em localizações espaciais, são multivariados, ou seja, mais de uma variável é mensurada em cada localização. As técnicas multivariadas encontradas em Johnson e Wichern (1998) e Reis (1997), por exemplo, são usadas segundo Bailey e Gatrell (1995) para fins de redução dos dados e exploração do espaço do atributo multidimensional, com o objetivo de identificar um número pequeno de sub-dimensões de interesse dado por combinações dos atributos, que podem então ser examinados de uma perspectiva espacial, explorando padrões espaciais e relacionamentos, ou para uso em classificação e discriminação espacial.

Da perspectiva de Diggle e Ribeiro Jr. (2007), procura-se descrever a distribuição espacial conjunta das variáveis; a distribuição condicional de uma variável resposta de interesse dado uma ou mais covariáveis referenciadas espacialmente, ou ainda, a distribuição conjunta de duas variáveis onde uma é a resposta de difícil obtenção ou de custo elevado e a outra é com esta correlacionada mas de fácil obtenção. Neste caso, combina-se poucas mensurações da difícil com um número maior de mensurações da outra.

Em se tratando de mais de uma variável resposta, considera-se o processo estocástico $\underline{Y} = (Y_1(\underline{x}), ..., Y_d(\underline{x}))'$ de dimensão d, $\underline{x} \in \mathbb{R}^2$, sendo o valor observado da variável resposta associada a localização $\underline{x}_i$ da forma $y_i = (y_{i1}, ..., y_{id})$. Desta maneira, a

função de covariância é a matriz simétrica

$$\Gamma(\underline{x},\underline{x}') = \begin{bmatrix} Cov\big(Y_1(\underline{x}),Y_1(\underline{x}')\big) & Cov\big(Y_1(\underline{x}),Y_2(\underline{x}')\big) & \cdots & Cov\big(Y_1(\underline{x}),Y_d(\underline{x}')\big) \\ Cov\big(Y_1(\underline{x}),Y_2(\underline{x}')\big) & Cov\big(Y_2(\underline{x}),Y_2(\underline{x}')\big) & \cdots & Cov\big(Y_2(\underline{x}),Y_d(\underline{x}')\big) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\big(Y_1(\underline{x}),Y_d(\underline{x}')\big) & Cov\big(Y_2(\underline{x}),Y_d(\underline{x}')\big) & \cdots & Cov\big(Y_d(\underline{x}),Y_d(\underline{x}')\big) \end{bmatrix},$$

denominada matriz de covariância cruzada em que $\gamma_{jk}(\underline{x},\underline{x}') = \gamma_{kj}(\underline{x},\underline{x}')$, $k,j = 1,...,d$.

Agora, se $\underline{Y}(\underline{x})$ for um processo estacionário, a matriz de covariância torna-se

$$\Gamma(\underline{x},\underline{x}') = \begin{bmatrix} \sigma_1^2 & Cov\big(Y_1(\underline{x}),Y_2(\underline{x}')\big) & \cdots & Cov\big(Y_1(\underline{x}),Y_d(\underline{x}')\big) \\ Cov\big(Y_2(\underline{x}),Y_1(\underline{x}')\big) & \sigma_2^2 & \cdots & Cov\big(Y_2(\underline{x}),Y_d(\underline{x}')\big) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\big(Y_d(\underline{x}),Y_1(\underline{x}')\big) & Cov\big(Y_d(\underline{x}),Y_2(\underline{x}')\big) & \cdots & \sigma_d^2 \end{bmatrix},$$

que não é necessariamente simétrica. Consequentemente, a matriz de correlação de $\underline{Y}(\underline{x})$ será

$$\mathbf{R}(u) = \begin{bmatrix} \rho_1(u) & \frac{\gamma_{12}(u)}{\sigma_1\sigma_2} & \cdots & \frac{\gamma_{1d}(u)}{\sigma_1\sigma_d} \\ \frac{\gamma_{21}(u)}{\sigma_2\sigma_1} & \rho_2(u) & \cdots & \frac{\gamma_{2d}(u)}{\sigma_2\sigma_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\gamma_{d1}(u)}{\sigma_d\sigma_1} & \frac{\gamma_{d2}(u)}{\sigma_d\sigma_2} & \cdots & \rho_d(u) \end{bmatrix},$$

onde os $\rho_{jk}(u)$ satisfazem $\rho_{jk}(u) = \rho_{kj}(-u)$ e, ainda, para $j = k$ são as funções de correlação do processo univariado e para $j \neq k$, funções de correlação cruzada.

Ainda de acordo com Diggle e Ribeiro Jr. (2007), existem pelo menos duas maneiras de se definir o variograma cruzado. Na primeira, as variáveis devem ser medidas

nas mesmas localizações e na segunda, não necessariamente. São elas:

$$\begin{aligned}
V^*_{jk}(u) &= \frac{1}{2}Cov\Big([Y_j(\underline{x}) - Y_j(\underline{x}-u)], [Y_k(\underline{x}) - Y_k(\underline{x}-u)]\Big) \\
&= \frac{1}{2}\Big\{Cov\big(Y_j(\underline{x}), Y_k(\underline{x})\big) - Cov\big(Y_j(\underline{x}), Y_k(\underline{x}-u)\big) - Cov\big(Y_j(\underline{x}-u), Y_k(\underline{x})\big) + \\
&\quad Cov\big(Y_j(\underline{x}-u), Y_k(\underline{x}-u)\big)\Big\} \\
&= \frac{1}{2}\gamma_{jk}(\underline{x},\underline{x}) + \frac{1}{2}\gamma_{jk}(\underline{x}-u,\underline{x}-u) - \frac{1}{2}\gamma_{jk}(\underline{x},\underline{x}-u) - \frac{1}{2}\gamma_{jk}(\underline{x}-u,\underline{x}) \\
&= \frac{1}{2}\gamma_{jk}(0) + \frac{1}{2}\gamma_{jk}(0) - \frac{1}{2}\gamma_{jk}(u) - \frac{1}{2}\gamma_{jk}(-u) \\
&= \sigma_j\sigma_k\rho(0) - \frac{1}{2}\{\sigma_j\sigma_k\rho_{jk}(u) + \sigma_j\sigma_k\rho_{jk}(-u)\} \\
&= \sigma_j\sigma_k\{1 - \frac{1}{2}[\rho_{jk}(u) + \rho_{jk}(-u)]\}, \qquad (31)
\end{aligned}$$

e

$$\begin{aligned}
V_{jk}(u) &= \frac{1}{2}Var\big(Y_j(\underline{x}) - Y_k(\underline{x}-u)\big) \\
&= \frac{1}{2}\Big\{Var\big(Y_j(\underline{x})\big) + Var\big(Y_k(\underline{x}-u)\big) - 2Cov\big(Y_j(\underline{x}), Y_k(\underline{x}-u)\big)\Big\} \\
&= \frac{1}{2}\{\sigma_j^2 + \sigma_k^2 - 2\sigma_j\sigma_k\rho_{jk}(u)\} \\
&= \frac{1}{2}(\sigma_j^2 + \sigma_k^2) - \sigma_j\sigma_k\rho_{jk}(u). \qquad (32)
\end{aligned}$$

Trabalhando-se com variáveis padronizadas, mostra-se que

$$\begin{aligned}
V^*_{jk}(u) &= 1 - \frac{1}{2}\big[\rho_{jk}(u) + \rho_{jk}(-u)\big] \\
&= \frac{1}{2}\big[2 - \rho_{jk}(u) - \rho_{jk}(-u)\big] \\
&= \frac{1}{2}\{[1 - \rho_{jk}(u)] + [1 - \rho_{jk}(u)]\} \\
&= 1 - \rho_{jk}(u) \\
&= V_{jk}.
\end{aligned}$$

2.4 Modelo Multivariado

O grande problema encontrado na modelagem é garantir que a matriz de covariância $\Gamma(\underline{x}, \underline{x}')$ seja definida positiva. Uma maneira seria a construção de combinações lineares de componentes independentes cuja descrição podem ser encontradas em Diggle e Ribeiro Jr. (2007), Schmidt e Gelfand (2003), Banerjee, Carlin e Gelfand (2004), Schmidt e Sansó (2006). Procedimentos baseados em separabilidade, corregionalização, médias móveis e convolução são descritos em Banerjee, Carlin e Gelfand (2004).

Considere que $\underline{Y}_1, \underline{Y}_2, ..., \underline{Y}_n$ sigam uma determinada distribuição. Muitas vezes a distribuição de probabilidade de $\underline{Y} = (\underline{Y}_1, \underline{Y}_2, ..., \underline{Y}_n)'$ não é conhecida, mas ao escrever

$$[\underline{Y}, \underline{S}] \quad = \quad [\underline{S}][\underline{Y}|\underline{S}] \tag{33}$$

observa-se que para um conjunto finito de pontos, $\underline{S}$ será gaussiana multivariada e $\underline{Y}|\underline{S}$ será um produto de densidades gaussianas univariadas, já que $\underline{Y}_1$ e $\underline{Y}_2$ que inicialmente são dependentes, dado o conhecimento de S_1 e S_2, tornam-se independentes. Consequentemente, $[\underline{Y}, \underline{S}]$ será uma distribuição gaussiana. Logo, com um processo multivariado gaussiano latente $\underline{S}(\underline{x})$ e uma independência condicional será possível construir o modelo multivariado para $\underline{Y}$ sem a necessidade de que este tenha distribuição gaussiana bastando, para isso, integrar a Equação 33 com relação a $\underline{S}$ obtendo a marginal de $\underline{Y}$

$$[\underline{Y}] = \int [\underline{Y}, \underline{S}]\, d\underline{S} = \int [\underline{S}][\underline{Y}|\underline{S}]\, d\underline{S}.$$

Portanto, este produto integrado em relação a $\underline{S}$ resulta numa distribuição de probabilidade que é o objetivo da modelagem.

No caso bivariado, por exemplo, o modelo será dado por:

$$\begin{cases} Y_{i1} \quad = \quad \mu_1(\underline{x}_i) + S_1(\underline{x}_i) + Z_1(\underline{x}_i) \\ Y_{i'2} \quad = \quad \mu_2(\underline{x}_{i'}) + S_2(\underline{x}_{i'}) + Z_2(\underline{x}_{i'}) \end{cases}$$

em que as respostas

$$\underline{Y} = (\underline{Y}_1, \underline{Y}_2)' = (Y_{11}, Y_{21}, ..., Y_{n_1 1}, Y_{12}, Y_{22}, ..., Y_{n_2 2})'$$

são medidas nas localizações $\underline{x}_{ij},\ \underline{x}_{i'j},\ i, i' = 1, ..., n_j, j = 1, 2,\ \underline{S}(\underline{x}) = \big(S_1(\underline{x}), S_2(\underline{x})\big)$ será considerado um processo gaussiano estacionário com média $\underline{\mu} = (0,\ 0)'$, variâncias $\sigma_j^2 = Var\big(S_j(\underline{x})\big)$ e estrutura de correlação dada por

$$\begin{bmatrix} Corr\big(S_1(\underline{x}), S_1(\underline{x}-u)\big) & Corr\big(S_1(\underline{x}), S_2(\underline{x}-u)\big) \\ Corr\big(S_1(\underline{x}), S_2(\underline{x}-u)\big) & Corr\big(S_2(\underline{x}), S_2(\underline{x}-u)\big) \end{bmatrix},$$

e $Z_j \sim N(0;\ \tau^2)$.

Para Y_{i1} observa-se que $\mu_1(\underline{x}_i)$ será o efeito fixo e o efeito aleatório $S_1(\underline{x}_i)$ será decomposto em dois outros efeitos aleatórios: $S_{01}^*(\underline{x}_i)$ que será comum a Y_{i1} e Y_{i2}, e $S_1^*(\underline{x}_i)$ que será específico a Y_{i1}. O mesmo ocorrerá para Y_{i2}. Mas, como em geral Y_{ij} são medidas em escalas diferentes, a padronização se fará necessária e será dada por $\sigma_{0j}U_0(\underline{x}_i)$ e $\sigma_j U_j(\underline{x}_i)$ sendo $U_0(\underline{x}_i)$ e $U_j(\underline{x}_i)$ efeitos aleatórios padronizados que têm correlação espacial, logo admensionais com as unidades preservadas nas constantes padronizadoras σ_{0j} e σ_j. Estes efeitos apresentarão distribuição gaussiana com vetor de médias iguais a zero e matriz de covariância com variâncias unitárias na diagonal principal e covariâncias cruzadas dadas pela função de correlação adotada. Desta forma, o modelo bivariado se tornará

$$\begin{cases} Y_{i1} &= \mu_1(\underline{x}_i) + \sigma_{01}U_0(\underline{x}_i) + \sigma_1 U_1(\underline{x}_i) + Z_1(\underline{x}_i) \\ Y_{i'2} &= \mu_2(\underline{x}_{i'}) + \sigma_{02}U_0(\underline{x}_{i'}) + \sigma_2 U_2(\underline{x}_{i'}) + Z_2(\underline{x}_{i'}) \end{cases} \tag{34}$$

Será possível calcular a correlação entre Y_1 em uma localização $\underline{x}_i$ e Y_1 em uma outra localização $\underline{x}_{i'},\ i, i' : 1, ..., n_1$ através do modelo espacial dado pela Equação 34, primeira linha, em que $U_0(\underline{x}_i)$ e $U_1(\underline{x}_i)$ contribuirão para este cálculo. Da mesma forma será possível calcular a correlação considerando-se Y_2, segunda linha do modelo (34). Também, a correlação de Y_1 em uma localização $\underline{x}_i$ e Y_2 em uma localização $\underline{x}_{i'}$ será decorrente do modelo (34) e, neste caso, apenas $U_0(\underline{x}_i)$ e $U_0(\underline{x}_{i'})$ contribuirão para o cálculo já que os outros termos são independentes.

Considerando-se apenas duas localizações ($\underline{x}_1$ e $\underline{x}_2$) a matriz $\mathbf{\Sigma}$ será dada por:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{01}^2 + \sigma_1^2 & \sigma_{01}^2\rho_0 + \sigma_1^2\rho_1 & \sigma_{01}\sigma_{02} & \sigma_{01}\sigma_{02}\rho_0 \\ \sigma_{01}^2\rho_0 + \sigma_1^2\rho_1 & \sigma_{01}^2 + \sigma_1^2 & \sigma_{01}\sigma_{02}\rho_0 & \sigma_{01}\sigma_{02} \\ \sigma_{01}\sigma_{02} & \sigma_{01}\sigma_{02}\rho_0 & \sigma_{02}^2 + \sigma_2^2 & \sigma_{02}^2\rho_0 + \sigma_2^2\rho_2 \\ \sigma_{01}\sigma_{02}\rho_0 & \sigma_{01}\sigma_{02} & \sigma_{02}^2\rho_0 + \sigma_2^2\rho_2 & \sigma_{02}^2 + \sigma_2^2 \end{bmatrix}.$$

Por outro lado, o conceito de corregionalização deriva da teoria de variáveis regionalizadas desenvolvidas por Matheron (1965). Esta teoria forma o fundamento para a análise de estrutura espacial, avaliação e estimação de dados multivariados distribuídos espacialmente (PAWLOWSKY-GLAHN; OLEA, 2004).

No modelo probabilístico, uma variável regionalizada $\mathtt{w}(\underline{x})$ será considerada uma realização de uma função aleatória $\mathtt{W}(\underline{x})$, isto é, uma família infinita de variáveis aleatórias construídas em todos os pontos $\underline{x}$ de uma região (WACKERNAGEL, 1998).

Schmidt e Sansó (2006) apontam que o modelo mais básico de corregionalização linear, introduzido por Matheron (1982) é da forma

$$\underline{Y}(\underline{x}) = \mathbf{D}\underline{\mathtt{w}}(\underline{x}) \tag{35}$$

em que $\mathbf{D}$ é uma matriz de dimensão $p \times p$ de posto completo e as componentes de $\underline{\mathtt{w}}(\underline{x})$, $\mathtt{w}_j(\underline{x})$, $j = 1, ..., p$, são processos espaciais independentes e identicamente distribuídos. Isto é, os $\mathtt{w}_j(\underline{x})$ são independentes se $j \neq j'$ mas $Cov\big(\mathtt{w}_j(\underline{x}), \mathtt{w}_j(\underline{x}')\big) = \rho(\underline{x} - \underline{x}')$ independente do valor de j. Fazendo $\mathbf{DD}' = \mathbf{T}$ tem-se o modelo de corregionalização linear intrínsico. A matriz $\mathbf{D}$ pode ser assumida triangular inferior e se $\underline{Y} = (\underline{Y}(\underline{x}_1), \underline{Y}(\underline{x}_2), ..., \underline{Y}(\underline{n}))'$, a matriz de covariâncias será dada por

$$\mathbf{\Sigma} = \mathbf{R} \otimes \mathbf{T} \tag{36}$$

em que $R_{ij} = \rho(\underline{x}_i, \underline{x}_j)$ e $\otimes$ denota o produto de Kronecker. Esta matriz de covariâncias é válida pois resulta do produto de matrizes positiva definidas.

Se na Equação 35, for assumido que $\mathtt{w}_j(\underline{x})$ são independentes mas não identi-

camente distribuídos, pode ser obtido um modelo de corregionalização mais geral. Considerando $\mathtt{w}_j(\underline{x})$ um processo gaussiano com média 0, variância 1, função de correlação estacionária $\rho_j(u), E\big(\underline{\mathrm{Y}}(\underline{x})\big) = 0$ e, a matriz de covariância cruzada de $\underline{\mathrm{Y}}(\underline{x})$ será dada por

$$\mathbf{\Sigma} = \sum_{j=1}^{p} (\mathbf{R}_j \otimes \mathbf{T}_j) \tag{37}$$

em que $\mathbf{T}_j = \underline{\mathrm{d}}_j \underline{\mathrm{d}}_j^{'}$ com $\underline{\mathrm{d}}_j$ sendo a j-ésima coluna de $\mathbf{D}$. Observa-se que $\mathbf{T}_j$ tem posto 1 e $\sum_{j=1} \mathbf{T}_j = \mathbf{T}$. Também, a transformação linear mantém a estacionariedade do processo espacial conjunto.

No caso bivariado em que $p = 2$ e, considerando-se apenas duas localizações ($\underline{x}_1$ e $\underline{x}_2$) a matriz $\mathbf{D}$ será:

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 \\ d_{21} & d_{22} \end{bmatrix}$$

de forma que

$$\mathbf{T}_1 = \begin{bmatrix} d_{11}^2 & d_{11}d_{21} \\ d_{11}d_{21} & d_{21}^2 \end{bmatrix}, \qquad \mathbf{T}_2 = \begin{bmatrix} 0 & 0 \\ 0 & d_{22}^2 \end{bmatrix}$$

e

$$\mathbf{\Sigma} = (\mathbf{R}_1 \otimes \mathbf{T}_1) + (\mathbf{R}_2 \otimes \mathbf{T}_2), \tag{38}$$

ou

$$\mathbf{\Sigma} = \begin{bmatrix} d_{11}^2 & d_{11}d_{21} & d_{11}^2\rho_1 & d_{11}d_{21}\rho_1 \\ d_{11}d_{21} & d_{21}^2 + d_{22}^2 & d_{11}d_{21}\rho_1 & d_{21}^2\rho_1 + d_{22}^2\rho_2 \\ d_{11}^2\rho_1 & d_{11}d_{21}\rho_1 & d_{11}^2 & d_{11}d_{21} \\ d_{11}^2d_{21}\rho_1 & d_{21}^2\rho_1 + d_{22}^2\rho_2 & d_{11}d_{21} & d_{21}^2 + d_{22}^2 \end{bmatrix}$$

2.5 Dados Composicionais

Os dados são ditos composicionais na medida em que registram informação sobre frequências relativas associadas com diferentes componentes de um sistema, por exemplo, proporções associadas com diferentes nutrientes (BUTLER; GLASBEY, 2008).

A análise de dados composicionais foi introduzida nos anos 80 por Aitchison (1982). Tais dados consistem de vetores $\underline{Y}$ de proporções de algum "todo" e apresentam variabilidade de vetor para vetor. Cada vetor é denominado uma composição e os componentes de qualquer composição de B partes $(Y_1, Y_2, ..., Y_B)$ devem satisfazer as exigências de que cada componente é não negativo:

$$Y_1 \geqslant 0, \ldots, Y_B \geqslant 0, \tag{39}$$

e que a soma de todos os componentes é 1:

$$Y_1 + Y_2 + \ldots + Y_B = 1. \tag{40}$$

Aplicações com dados composicionais são frequentemente encontrados em áreas como geologia, agricultura, biologia, literatura, ambiental, estatística médica, economia, etc. As características principais de um conjunto de dados composicionais segundo Reyment e Savazzi (1999) *apud* Labus (2005) são:

- podem ser representados na forma de uma matriz;
- cada linha da matriz corresponde a uma única unidade observacional ou experimental e soma 1 no caso de proporções, ou respectivamente, 100(%) no caso de porcentagem. Algumas vezes outra constante pode ser encontrada devido à alguma manipulação por parte do pesquisador.;
- cada coluna da matriz representa uma única parte (variável);
- cada elemento da matriz é não negativo;
- os coeficientes de correlação mudam se uma das variáveis (partes) é excluída da matriz de dados e as linhas somam 1 ou 100 novamente. O mesmo ocorre se um novo componente é adicionado.

Esta última propriedade significa que alterar uma ou mais variáveis do (ao) conjunto de dados pode ter um efeito numericamente significativo nas correlações entre as variáveis restantes.

Pawlowsky-Glahn e Olea (2004) e Tolosana-Delgado, Otero e Pawlowsky-Glahn (2005) apontam que os dados composicionais apresentam um efeito de correlação espúria o que significa que a aplicação dos métodos estatísticos padrão podem levar a resultados inconsistentes. Isto significa que as covariâncias estão sujeitas a controles não estocásticos, isto é, sofrem distorções devido à restrição da soma totalizar 1 levando à interpretação errônea da estrutura de covariância espacial. Um outro problema, levantado por Labus (2005), é que é uma consequência da Equação 40 o fato de que os componentes não são independentes e isto, segundo Pawlowsky-Glahn e Olea (2004), implica em singularidade da matriz de covariância de uma composição, excluindo por exemplo, o uso de técnicas de estimação como cockrigagem de todos os componentes.

Pawlowsky-Glahn e Olea (2004) relatam que os problemas com correlação espacial espúria e singularidade da matriz de covariância são relacionados à suposição básica de que o espaço amostral é irrestrito, uma suposição implícita na análise estatística de corregionalizações e, a suposição de que a distribuição do erro de estimação em cada ponto da região amostral é gaussiana.

Neste trabalho, o interesse está na teoria de dados composicionais no contexto espacial onde existe dependência espacial entre os locais de observação.

2.5.1 Composição Regionalizada

Aitchison (1986) apresenta duas maneiras de determinar uma composição. Na primeira, a composição fica completamente especificada pelos componentes de um subvetor de b-partes $(Y_1, Y_2, ..., Y_b)$, onde $b = B - 1$,

$$Y_B = 1 - Y_1 - Y_2 - \ldots - Y_b,$$

o que significa que uma composição de B-partes é um vetor b-dimensional podendo ser representado em algum conjunto b-dimensional. Na segunda, especificando b razões r_i dadas por:

$$r_i = Y_i/Y_B \qquad\qquad i = 1, ..., b.$$

A composição é determinada por

$$\begin{aligned} Y_i &= r_i/(r_1 + \cdots + r_b + 1) \qquad i = 1, ..., b. \\ Y_B &= 1/(r_1 + \cdots + r_b + 1). \end{aligned} \quad .$$

Aitchison e Greenacre (2002) apresentam três maneiras equivalentes de se considerar razões dentro de uma composição:

(a) as $\frac{1}{2}B(B-1)$ razões Y_i/Y_j entre os pares de componentes, assumindo $i < j$ ao selecionar o par;

(b) as $B-1$ razões Y_i/Y_B entre os primeiros $B-1$ componentes;

(c) as B razões $Y_i/g(\underline{Y})$ entre os componentes e sua média geométrica $g(\underline{Y}) = \sqrt[B]{Y_1 Y_2 \ldots Y_B}$.

De acordo com Pawlowsky-Glahn e Olea (2004), de qualquer vetor aleatório $\underline{W} = [W_1, W_2, ..., W_B]'$ não negativo que não seja uma composição pode-se obter uma, dividindo-se os componentes individuais pela soma dos componentes, ou seja:

$$Y_i = \frac{W_i}{W_1 + W_2 + \cdots + W_B}.$$

A composição resultante é $\underline{Y} = (Y_1, Y_2, ..., Y_B)'$ e a relação $\frac{W_i}{W_j} = \frac{Y_i}{Y_j}$ é válida para quaisquer índices $i, j = 1, 2, ..., B$, para W_j e $Y_j \neq 0$, enfatizando-se que o dado composicional contém informação somente sobre magnitudes relativas e não absolutas.

Então, considerando $\Omega \subset \mathbb{R}^n$ um domínio espacial e $\mathbb{R}^n$ um espaço real de dimensional n, um vetor função aleatória $\underline{Y}(\underline{x})$ em cada localização $\underline{x} \in \Omega$ onde todos os componentes são positivos e a soma destes é igual a 1, será denominado composição regionalizada.

O espaço amostral natural para $\underline{Y}(\underline{x})$ é o simplex-B unitário, embutido no espaço real B-dimensional $\mathbb{R}^B$, dado por:

$$\mathbb{S}^B = \{\underline{Y}(\underline{x}) \in \mathbb{R}^B;\ Y_i(\underline{x}) > 0,\ i = 1, ..., B;\ \underline{j}'\underline{Y}(\underline{x}) = 1\},$$

sendo $\underline{j}'$ um vetor com elementos iguais a 1. McBratney, De Gruijter e Brus (1992) *apud* Odeh, Tood e Triantafilis (2003) definem um simplex como uma representação geométrica do espaço de atributos, onde uma composição de B partes é representada por um número mínimo de vértices em um espaço de um dado número de dimensões.

2.5.2 Base Regionalizada

Uma base de B partes é um vetor $B \times 1$ de componentes positivos $(W_1, W_2, ..., W_B)$ todos registrados na mesma escala de medida Aitchison (1986). Segue então, que uma base regionalizada (Pawlowsky-Glahn e Olea (2004)) $\underline{W}(\underline{x})$, $\underline{x} \in \Omega \subset \mathbb{R}^n$ é um vetor função aleatória cujos componentes são todos positivos e medidos na mesma escala. Isto implica que o espaço amostral de uma base de B-partes é dado por:

$$\mathbb{R}^B_+ = \{\underline{W}(\underline{x}) \in \mathbb{R}^B;\ W_i(\underline{x}) > 0,\ i = 1, ..., B\}.$$

A transformação de uma base regionalizada em uma composição regionalizada se dá através da função:

$$\begin{array}{rccl} \mathcal{C}\ : & \mathbb{R}^B_+ & \longrightarrow & \mathbb{S}^B \\ & \underline{W}(\underline{x}) & \longrightarrow & \mathcal{C}\big(\underline{W}(\underline{x})\big) = \dfrac{\underline{W}(\underline{x})}{\underline{j}'\underline{W}(\underline{x})}, \end{array}$$

onde $\mathcal{C}$ é denominado operador fechamento. O operador fechamento garante que o vetor resultante seja uma composição.

Barceló-Vidal, Martín-Fernández e Pawlowsky-Glahn (2001) apresentam uma interpretação geométrica para $\mathcal{C}$, desconsiderando a regionalização, que é a intersecção do raio partindo da origem através de $\underline{W}$ e o hiperplano de $\mathbb{R}^B$ definido pela equação $W_1 + W_2 + \ldots + W_B = 1$, como na Figura 5(a). O conjunto de todos estes pontos é o simplex regular de dimensão $b = B - 1$, dado por:

$$\mathbb{S}^b = \{(W_1, W_2, ..., W_B)';\ W_1 > 0, ..., W_B > 0;\ W_1 + \ldots + W_B = 1\}.$$

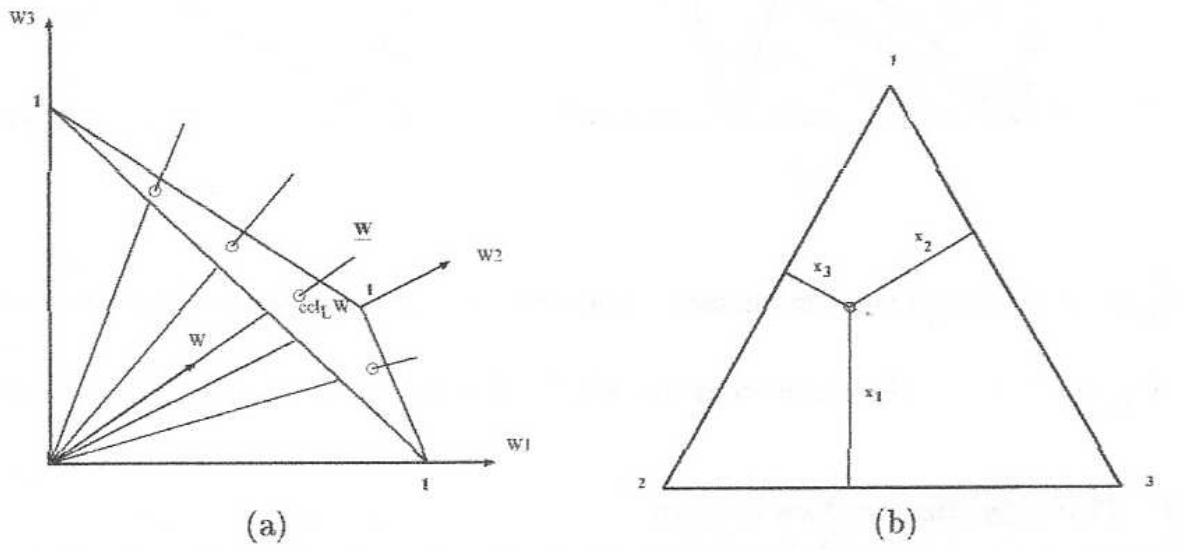

Figura 4: (a)Composições de 3 partes como raios partindo da origem em $\mathbb{R}^3_+$; (b) O simplex $\mathbb{S}^2$. Adaptado de Barceló-Vidal, Martín-Fernández e Pawlowsky-Glahn (2001) Nota: Em particular, nesta figura, $\underline{\mathrm{W}} = \{kW : k \in \mathbb{R}^+\}$.

O simplex $\mathbb{S}^2$ corresponde ao diagrama ternário (Figura 5(b)), um triângulo equilátero de altura unitária. Para qualquer ponto P no triângulo 123 as perpendiculares W_1, W_2 e W_3 de P aos lados opostos $\overline{23}$, $\overline{13}$ e $\overline{12}$, respectivamente, satisfazem $W_1 + W_2 + \ldots + W_B = 1$.

2.5.3 Subcomposição Regionalizada

Uma subcomposição regionalizada é um subconjunto de uma composição regionalizada. Assim, considerando-se $\underline{\mathrm{Y}}(\underline{\mathrm{x}})$ uma composição regionalizada de B partes e $s \subset \{1, 2, ..., B\}$ de forma a tornar $\underline{\mathrm{Y}}_s(\underline{\mathrm{x}})$ um subvetor cujos elementos são os componentes de $\underline{\mathrm{Y}}(\underline{\mathrm{u}})$ correspondentes às partes em s, então uma subcomposição regionalizada será:

$$\mathcal{C}\big(\underline{\mathrm{Y}}_s(\underline{\mathrm{x}})\big) = \frac{\underline{\mathrm{Y}}_s(\underline{\mathrm{x}})}{\underline{\mathrm{j}}'\underline{\mathrm{Y}}_s(\underline{\mathrm{x}})}.$$

Segundo Aitchison (1986), Aitchison e Greenacre (2002), Aitchison e Egozcue (2005) o conceito de subcomposição é importante no que se refere a coerência subcomposicional. Isto significa que, se um pesquisador, tendo acesso à uma composição, também fizer inferência a partir de uma subcomposição desta e, se um outro, com acesso somente a uma subcomposição mas com partes comuns à do primeiro pesquisador fizer inferência

a partir desta, os dois resultados deverão coincidir.

As correlações momento-produto e análise de componentes principais baseadas em covariâncias calculadas no dado composicional em linha, não tem coerência subcomposicional, mas a característica importante de uma subcomposição é que esta preserva relações de razões. Daí, se $\underline{\mathrm{s}}(\underline{\mathrm{x}}) = \mathcal{C}\big(\underline{\mathrm{Y}}_s(\underline{\mathrm{x}})\big)$, então $\dfrac{s_i(\underline{\mathrm{x}})}{s_j(\underline{\mathrm{x}})} = \dfrac{Y_i(\underline{\mathrm{x}})}{Y_j(\underline{\mathrm{x}})}$, $i, j \in s$. Observa-se que as razões são formadas com os componentes do vetor de dado composicional (linha) através das colunas da matriz de dados.

Da mesma forma que na seção anterior Barceló-Vidal, Martín-Fernández e Pawlowsky-Glahn (2001) apresentam uma interpretação geométrica para uma subcomposição. Dada uma composição de B-partes, a formação de uma subcomposição corresponde à projeção ortogonal do raio associado a $\underline{\mathrm{W}}$ em $\mathbb{R}^+$ em um subespaço que é gerado pelos eixos coordenados das partes selecionadas para formar a subcomposição.

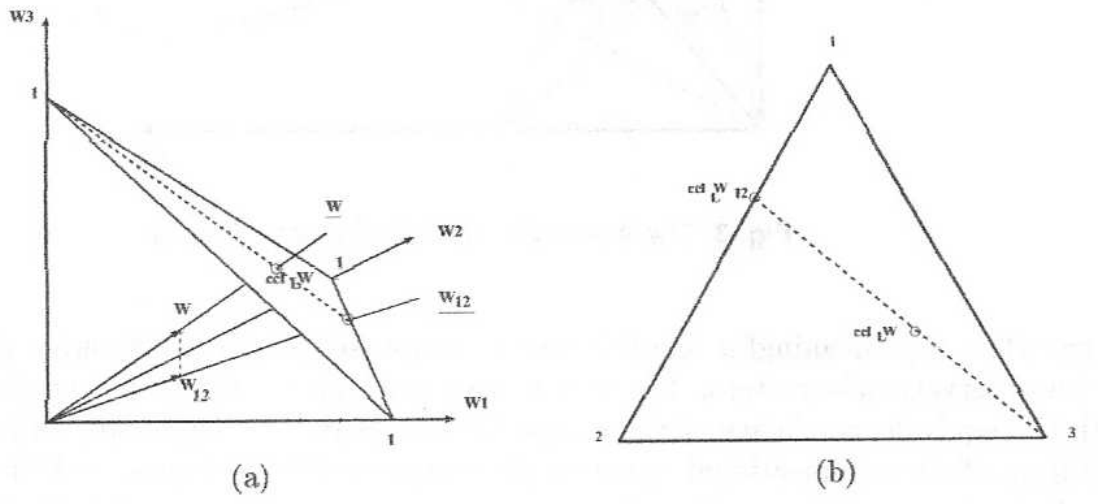

Figura 5: (a)Interpretação geométrica da formação da subcomposição $\underline{\mathrm{W}}_{12}$ da composição $\underline{\mathrm{W}}$: (a) em $\mathbb{R}^3_+$; (b) em $\mathbb{S}^2$. Adaptado de Barceló-Vidal, Martín-Fernández e Pawlowsky-Glahn (2001).
Nota: Em particular, nesta figura, $\underline{\mathrm{W}} = \{kW : k \in \mathbb{R}^+\}$.

2.5.4 Amalgamação e Partição Regionalizada

Seja $\underline{\mathrm{Y}}(\underline{\mathrm{x}}) = [Y_1(\underline{\mathrm{x}}), Y_2(\underline{\mathrm{x}}), ..., Y_B(\underline{\mathrm{x}})]'$ uma composição regionalizada de B partes que é dividida em C ($C \leq B$) subconjuntos mutuamente exclusivos. Seja $\underline{\mathrm{Y}}_i(\underline{\mathrm{x}})$, $i =$

$1, ..., C$, o vetor cujos componentes são os elementos do i-ésimo subconjunto. Pawlowsky-Glahn e Olea (2004) definem amalgamação regionalizada como a composição regionalizada com C componentes em que os componentes de cada um dos C subconjuntos são somados:

$$\underline{\mathrm{A}}(\underline{\mathrm{x}}) = [A_1(\underline{\mathrm{x}}), A_2(\underline{\mathrm{x}}), ..., A_C(\underline{\mathrm{x}})]',$$

com $A_i(\underline{\mathrm{x}}) = \underline{\mathrm{j}}'\underline{\mathrm{Y}}_i(\underline{\mathrm{x}})$.

Ao considerar a amalgamação regionalizada $\underline{\mathrm{A}}(\underline{\mathrm{x}})$ juntamente com as subcomposições regionalizadas, $\underline{\mathrm{s}}_i(\underline{\mathrm{x}}) = \mathcal{C}\big(\underline{\mathrm{Y}}_i(\underline{\mathrm{x}})\big),\ i = 1, ..., C$:

$$P_C\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big) = \big(\underline{\mathrm{s}}_1(\underline{\mathrm{x}}), \underline{\mathrm{s}}_2(\underline{\mathrm{x}}), ...\underline{\mathrm{s}}_C(\underline{\mathrm{x}}); \underline{\mathrm{A}}(\underline{\mathrm{x}})\big)$$

tem-se definida a partição regionalizada de ordem C de $\underline{\mathrm{Y}}(\underline{\mathrm{x}})$. Assim, as informações contidas nos subvetores são preservadas em uma partição.

2.5.5 Transformação

Graf (2006) expõe de forma clara que a restrição dada pela Equação 40 implica que existe, necessariamente, uma correlação negativa entre os componentes e isto faz com que as correlações não sejam diretamente interpretáveis. Para relaxar e/ou evitar esta restrição Aitchison (1986) propôs transformações que generalizam a transformação logística $\ln\left(\frac{Y}{1-Y}\right)$ para um vetor composicional de 2 partes.

Aitchison (1999) afirma que, por fornecer uma estrutura de dependência correta e interpretável para descrever os padrões reais de variabilidade composicional, que permite a investigação coerente da variabilidade subcomposicional, a transformação log-razão é viável para investigar a importância ou irrelevância de componentes individuais. A natureza essencial de uma composição é que as magnitudes relativas dos componentes são as unidades relevantes sob estudo. Estas magnitudes relativas ou razões implicam em tratabilidade e interpretação estatística.

Algumas transformações logísticas como a aditiva, a aditiva modificada, a multiplicativa e a híbrida podem ser encontradas em Aitchison (1982), Aitchison et al. (2000) e

Odeh, Tood e Triantafilis (2003), por exemplo.

No sentido de composições regionalizadas, Pawlowsky-Glahn, Olea e Davis (1995), Pawlowsky-Glahn e Olea (2004) definem a transformação razão log-aditiva (ALR) como a função

$$\begin{aligned} \text{ALR} \ : \ \mathbb{S}^B &\longrightarrow \mathbb{R}^{B-1} \\ \underline{\mathrm{Y}}(\underline{\mathrm{x}}) &\longrightarrow \text{ALR}\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big) = \left(\ln\left(\frac{Y_1(\underline{\mathrm{x}})}{Y_B(\underline{\mathrm{x}})}\right), \ldots, \ln\left(\frac{Y_{B-1}(\underline{\mathrm{x}})}{Y_B(\underline{\mathrm{x}})}\right)\right)'. \end{aligned}$$

Pode-se observar que para $B = 2$, a transformação ALR torna-se a transformação logística. Nota-se também que se a composição possui três componentes, após a transformação a composição passa a ter 2 componentes seguindo uma distribuição normal bivariada.

Por outro lado, a transformação inversa denominada transformação logística generalizada aditiva (AGL) é dada por

$$\begin{aligned} \text{AGL}: \quad \mathbb{R}^{B-1} &\longrightarrow \mathbb{S}^B \\ \text{ALR}\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big) &\longrightarrow \text{AGL}\{\text{ALR}\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)\} = \underline{\mathrm{Y}}(\underline{\mathrm{x}}) = \left(\exp\left\{\ln\left(\frac{Y_1(\underline{\mathrm{x}})}{Y_B(\underline{\mathrm{x}})}\right)\right\}, \ldots, \exp\{0\}\right)'. \end{aligned}$$

Se a restrição de soma constante é $c \neq 0$ a transformação de volta é, então,

$$\underline{\mathrm{Y}}(\underline{\mathrm{x}}) = c \cdot \text{AGL}\{\text{ALR}\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)\}.$$

Uma segunda transformação é a transformação razão log centrada (CLR) definida como

$$\begin{aligned} \text{CLR} \ : \ \mathbb{R}_+^B &\longrightarrow \mathbb{R}^B \\ \underline{\mathrm{W}}(\underline{\mathrm{x}}) &\longrightarrow \text{CLR}\big(\underline{\mathrm{W}}(\underline{\mathrm{x}})\big) = \ln\left(\frac{\underline{\mathrm{W}}(\underline{\mathrm{x}})}{g\big(\underline{\mathrm{W}}(\underline{\mathrm{x}})\big)}\right) \end{aligned}$$

ou

$$\begin{aligned} \text{CLR} \ : \ \mathbb{S}^B &\longrightarrow \mathbb{R}^B \\ \underline{\mathrm{Y}}(\underline{\mathrm{x}}) &\longrightarrow \text{CLR}\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big) = \ln\left(\frac{\underline{\mathrm{Y}}(\underline{\mathrm{x}})}{g\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)}\right) \end{aligned}$$

onde $g\big(\underline{\mathrm{W}}(\underline{\mathrm{x}})\big) = \sqrt[B]{\prod_{i=1}^{B} W_i(\underline{\mathrm{x}})}$ é a média geométrica dos componentes da base regi-

onalizada $\underline{\mathrm{W}}(\underline{\mathrm{x}})$ e $g\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big) = \sqrt[B]{\prod_{i=1}^{B} Y_i(\underline{\mathrm{x}})}$ é a média geométrica dos componentes da composição regionalizada $\underline{\mathrm{Y}}(\underline{\mathrm{x}})$.

2.5.6 Perturbação e Potência

Perturbação no simplex para Eynatten, Barceló-Vidal e Pawlowsky-Glahn (2003) é uma operação que pode ser usada para descrever numericamente mudanças em uma composição e a combinação de perturbação e transformação potência fornece um método para a análise de processos lineares composicionais no simplex. Para Aitchison (1986), Tolosana-Delgado, Otero e Pawlowsky-Glahn (2005), Aitchison e Egozcue (2005), por exemplo, estas operações definem uma estrutura de espaço vetorial, de dimensão $B-1$ no simplex, com a perturbação como uma operação comutativa e a potência como um produto externo. A perturbação corresponde a multiplicar composições componente a componente e dividir cada um pela soma de todos para se obter soma igual a 1, ou seja,

$$\underline{\mathrm{Y}}_1 \oplus \underline{\mathrm{Y}}_2 = (Y_{11}, Y_{12}, ..., Y_{1B}) \oplus (Y_{21}, Y_{22}, ..., Y_{2B}) = \mathcal{C}(Y_{11}Y_{21}, Y_{12}Y_{22}, ..., Y_{1B}Y_{2B}).$$

É a operação análoga a translação no espaço real. A potência, por sua vez, é a análoga a multiplicação por um escalar no espaço real:

$$\alpha \odot (Y_{11}, Y_{12}, ..., Y_{1B}) = \mathcal{C}(Y_{11}^{\alpha}, Y_{12}^{\alpha}, ..., Y_{1B}^{\alpha}).$$

Consequentemente, tem-se o vetor de diferenças composicionais,

$$\underline{\mathrm{Y}}_1 \ominus \underline{\mathrm{Y}}_2 = \underline{\mathrm{Y}}_1 \oplus (-1 \odot \underline{\mathrm{Y}}_2).$$

Acrescentando às operações um produto interno,

$$< \underline{\mathrm{Y}}_1, \underline{\mathrm{Y}}_2 >= \sum_{i=1}^{B} \ln\left(\frac{Y_{1i}}{g(\underline{\mathrm{Y}}_1)}\right) \ln\left(\frac{Y_{2i}}{g(\underline{\mathrm{Y}}_2)}\right)$$

tem-se uma estrutura de espaço Euclidiano real para o simplex onde $g(\underline{\mathrm{Y}}_1) = \sqrt[B]{\prod_{j=1}^{B} Y_{1j}}$ é a média geométrica. Este produto interno induz uma distância (uma medida que pode

ser entendida, por exemplo, como grau de alteração) no simplex, denominada distância de Aitchison, usada para calcular a distância ou diferença entre duas composições e útil para entender a variabilidade dentro de um conjunto de dados:

$$d(\underset{\sim}{Y}_1, \underset{\sim}{Y}_2) = \sqrt{\sum_{i=1}^{B} \left(\ln \left(\frac{Y_{1i}}{g(\underset{\sim}{Y}_1)} \right) - \ln \left(\frac{Y_{2i}}{g(\underset{\sim}{Y}_2)} \right)^2 \right)}$$

Com isto, tem-se a geometria de Aitchison do simplex.

2.5.7 Estatísticas Descritivas e Domínio de Confiança Para Dados Composicionais

A média aritmética não é uma medida representativa em se tratando de dados composicionais por apresentarem distribuição lognormal. Aitchison (1997) *apud* Pawlowsky-Glahn e Olea (2004) sugeriu como medida de tendência central ou centro da distribuição o fechamento da média geométrica dado por:

$$\text{cen}(\underset{\sim}{Y}) = \frac{1}{g_s} \left[g(Y_1) \quad g(Y_2) \quad \ldots \quad g(Y_B) \right]'$$

em que $g(Y_i)$ é a média geométrica do i-ésimo componente e

$$g_s = g(Y_1) + g(Y_2) + \ldots + g(Y_B).$$

Como uma medida de dispersão ou medida de variabilidade total o autor define

$$\text{Totvar}(\underset{\sim}{Y}) = \frac{1}{B} \sum_{i<j} Var \left(\ln \left(\frac{Y_i}{Y_j} \right) \right).$$

O domínio de confiança para $\underset{\sim}{Y} = (Y_1, Y_2, ..., Y_B)$ é um subconjunto do simplex S^B dado por:

$$B_{1-\alpha}(\underset{\sim}{Y}) = \left\{ \underset{\sim}{Y} \in S^B | \left(\ln \left(\frac{\underset{\sim}{Y}_{-B}}{\underset{\sim}{Y}_B} \right) - \mu \right)' \Sigma^{-1} \left(\ln \left(\frac{\underset{\sim}{Y}_{-B}}{\underset{\sim}{Y}_B} \right) - \mu \right) \leq \chi^2_{b;1-\alpha} \right\}$$

em que $\chi^2_{b;1-\alpha}$ é o quantil $(1-\alpha)$ da distribuição qui-quadrado com b graus de liberdade.

2.5.8 Representação Gráfica

A inspeção visual auxilia a compreensão dos dados. Aitchison e Egozcue (2005) relaciona algumas técnicas gráficas como Diagramas de Variação, Diagramas Ternários, Diagramas de Dispersão de Razão e Razão Log, Gráfico Bivariado Composicional e Diagrama de Dispersão das Coordenadas.

O diagrama ternário, segundo Butler e Glasbey (2008) é um triângulo equilátero cujos vértices representam os três componentes da composição. Como exemplo, considere os dados apresentados em Aitchison (1986), de 39 amostras de sedimentos de areia, silte e argila obtidos em um lago Ártico, uma parte do qual pode ser visto na Tabela 1. O diagrama ternário correspondente a estes dados pode ser visualizado na Figura 6.

n	Areia	Silte	Argila
1	77.5	19.5	3
2	71.9	24.9	3.2
3	50.7	36.1	13.2
4	52.2	40.9	6.6
⋮	⋮	⋮	⋮
39	2	47.8	50.2

Tabela 1: Composições (areia, silte, argila) de 39 amostras de sedimentos em um lago Ártico.

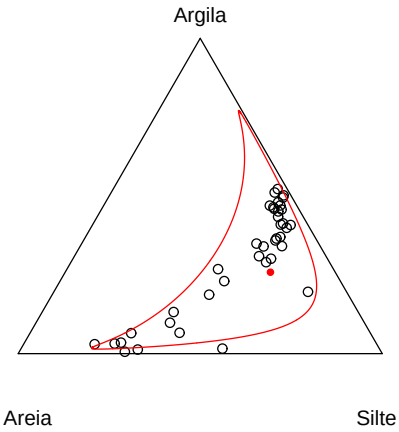

Figura 6: Diagrama ternário para dados do Lago Ártico incluindo o centro da distribuição e região 2-sigma de confiança.

Pontos localizados próximo a um vértice tem altas proporções do componente representado por aquele vértice, enquanto pontos localizados no centro do triângulo tem

proporções iguais para todos os três componentes.

Para localizar uma composição em um diagrama ternário efetua-se três regras de três, uma para cada componente como a seguir:

Comprimento do lado do triângulo	$\rightarrow$	100
Comprimento a ser determinado	$\rightarrow$	Porcentagem referente ao componente desejado,

obtendo-se os comprimentos para cada um dos componentes. Considerando-se agora o vértice V_1, o comprimento relacionado ao componente representado por este vértice será marcado nos lados $\overline{V_1V_2}$ e $\overline{V_1V_3}$ partindo-se dos vértices V_2 e V_3. Trace um segmento de reta unindo estes dois pontos. O mesmo procedimento é feito para os outros dois vértices e no cruzamento destes três segmentos de reta tem-se o ponto que representa a composição.

Segundo Boogaart (2005) todos os pontos localizados sobre uma linha paralela ao eixo oposto a um vértice têm a mesma proporção daquele componente correpondente a este vértice. Esta proporção é igual a distância relativa da linha ao eixo sobre a distância do vértice ao eixo. Por outro lado, todos os pontos em uma linha reta partindo de um dos vértices têm proporções relativas iguais dos outros dois componentes e são representadas pelo ponto onde a linha cruza o eixo oposto.

Além disso, a extensão da variabilidade da razão entre dois componentes se dá avaliando a extensão da intersecção das linhas retas partindo de cada um dos vértices e passando por todos os pontos com o eixo oposto (AITCHISON, 1986).

Por fim, relembrando, para qualquer ponto no diagrama as perpendiculares aos eixos satisfazem a restrição de que a soma deve ser igual a 1.

2.5.9 Estacionariedade

Seja $\underline{\mathrm{Y}}(\underline{\mathrm{x}})$ um vetor função aleatória, $\underline{\mathrm{x}} \in \Omega \subset \mathbb{R}^n$, e f uma função (ex: logaritmo) de modo que $f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)$ também é um vetor função aleatória. De acordo com Pawlowsky-Glahn e Olea (2004), $\underline{\mathrm{Y}}(\underline{\mathrm{x}})$ é uma função estacionária de 2^a ordem se $f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)$ é estacionária de 2^a ordem, isto é, se $f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)$ satisfaz as condições:

a) o vetor de valores esperados $E\Big(f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)\Big) = \mu$ existe e não depende de $\underline{\mathrm{x}}$;

b) a matriz função de covariância

$$Cov\Big(f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}}_1)\big),\, f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}}_2)\big)\Big) = \mathbf{\Sigma}(\underline{\mathrm{x}}_2 - \underline{\mathrm{x}}_1)$$

existe e não depende de $\underline{\mathrm{x}}_1,\ \underline{\mathrm{x}}_2$, mas somente da diferença $\underline{\mathrm{h}} = \underline{\mathrm{x}}_2 - \underline{\mathrm{x}}_1$.

Além disso, $\underline{\mathrm{Y}}(\underline{\mathrm{x}})$ é uma função intrínsica se $f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)$ satisfaz:

a) o vetor de valores esperados $E\Big(f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)\Big) = \mu$ existe e não depende de $\underline{\mathrm{x}}$;

b) a matriz função de covariância

$$Cov\Big(f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}}_2)\big) - \ f\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}}_1)\big)\Big) = \mathbf{\Gamma}(\underline{\mathrm{x}}_2 - \underline{\mathrm{x}}_1)$$

existe e não depende de $\underline{\mathrm{x}}_1,\ \underline{\mathrm{x}}_2$, mas somente da diferença $\underline{\mathrm{h}} = \underline{\mathrm{x}}_2 - \underline{\mathrm{x}}_1$.

Nota-se que $\mathbf{\Sigma}(\underline{\mathrm{x}}_2 - \underline{\mathrm{x}}_1)$ e $\mathbf{\Gamma}(\underline{\mathrm{x}}_2 - \underline{\mathrm{x}}_1)$ são as matrizes função de covariância cruzada e de variogramas e variogramas cruzados.

Acrescenta-se, ainda, que $\underline{\mathrm{Y}}(\underline{\mathrm{x}})$ é log razão estacionária de 2^a ordem (LR estacionária) se o conjunto das log razões entre todos os pares é estacionário de 2^a ordem e, respectivamente, é log razão intrínsica (LR intrínsica) se o conjunto das log razões entre todos os pares é intrínsico.

2.5.10 Estrutura de Covariância Espacial

Como citado anteriormente, o interesse deste trabalho está na metodologia de geoestatística aplicada à dados composicionais. Em se tratando de geoestatística, tem-se a existência de dependência espacial entre os locais de observação. Por outro lado, observa-se que sempre existirá correlação entre os componentes de uma composição de forma que a estrutura de covariância é essencial na modelagem.

Pawlowsky-Glahn e Olea (2004) definem a estrutura de covariância espacial de

uma composição regionalizada $\underline{Y}(\underline{x})$ como o conjunto de funções B^4

$$\sigma_{ij\cdot kl}(\underline{h}) = Cov\left(\ln\left(\frac{Y_i(\underline{x})}{Y_k(\underline{x})}\right), \ln\left(\frac{Y_j(\underline{x}+\underline{h})}{Y_l(\underline{x}+\underline{h})}\right)\right), \quad i,j,k,l \in \{1,2,...,B\}$$

para $\underline{x}, \underline{x}+\underline{h} \in \Omega$.

Neste caso, as seguintes propriedades são válidas considerando-se $i,j,k,l,m,n \in \{1,2,...,B\}$:

a) em geral, $\sigma_{ij\cdot kl}(\underline{h}) \neq \sigma_{ji\cdot lk}(\underline{h})$;

b) em geral, $\sigma_{ij\cdot kl}(\underline{h}) \neq \sigma_{ji\cdot kl}(\text{-}\underline{h})$;

c) $\sigma_{ij\cdot kl}(\underline{h}) = \sigma_{ji\cdot lk}(\text{-}\underline{h})$;

d) $\sigma_{ij\cdot kl}(\underline{h}) = -\sigma_{kj\cdot il}(\underline{h}) = \sigma_{kl\cdot ij}(\underline{h}) = -\sigma_{il\cdot kj}(\underline{h})$;

e) $\sigma_{ij\cdot il}(\underline{h}) = \sigma_{ij\cdot kj}(\underline{h}) = \sigma_{ij\cdot ij}(\underline{h}) = \sigma_{ii\cdot ii}(\underline{h}) = 0$;

f) $\sigma_{ij\cdot kl}(\underline{h}) = \sigma_{ij\cdot mn}(\underline{h}) + \sigma_{in\cdot ml}(\underline{h}) + \sigma_{mj\cdot kn}(\underline{h}) + \sigma_{mn\cdot kl}(\underline{h})$.

Do exposto, segue que para quaisquer dois componentes $Y_i(\underline{x}), Y_j(\underline{x}), \quad i,j \in \{1,2,...,B\}$ de uma composição $\underline{Y}(\underline{x})$:

a) a autocovariância das log razões, denominada LR autocovariância é

$$\tau_{i\cdot j}(\underline{h}) = \sigma_{ii\cdot jj}(\underline{h})$$

que são os elementos da matriz de LR autocovariâncias (Matriz Variação), $\mathbf{T}(\underline{h})$, de dimensão $B \times B$.

b) a ALR covariância cruzada é a função

$$\sigma_{ij}(\underline{h}) = \sigma_{ij\cdot BB}(\underline{h})$$

que são os elementos da matriz de ALR covariâncias cruzadas, $\mathbf{\Sigma}(\underline{h})$, de dimensão $(B-1) \times (B-1)$.

c) a CLR covariância cruzada é a função

$$\xi_{ij}(\underline{h}) = Cov\left(\ln\left(\frac{Y_i(\underline{x})}{g\big(\underline{Y}(\underline{x})\big)}\right), \ln\left(\frac{Y_j(\underline{x}+\underline{h})}{g\big(\underline{Y}(\underline{x}+\underline{h})\big)}\right)\right),$$

que são os elementos da matriz de CLR covariâncias cruzadas, $\boldsymbol{\Xi}(\underline{h})$, de dimensão $B \times B$.

2.5.11 Estrutura de Covariância Espacial Intrínsica

Em continuação ao item anterior, a estrutura de covariância espacial intrínsica de uma composição regionalizada $\underline{Y}(\underline{x})$ é definida como o conjunto de funções

$$V_{ij\cdot kl}(\underline{h}) = \frac{1}{2}\,Cov\left(\ln\left(\frac{Y_i(\underline{x})}{Y_k(\underline{x})}\right) - \ln\left(\frac{Y_i(\underline{x}+\underline{h})}{Y_k(\underline{x}+\underline{h})}\right), \ln\left(\frac{Y_j(\underline{x})}{Y_l(\underline{x})}\right) - \ln\left(\frac{Y_j(\underline{x}+\underline{h})}{Y_l(\underline{x}+\underline{h})}\right)\right),$$

para $i, j, k, l \in \{1, 2, ..., B\}$.

Como propriedades, apresenta-se:

a) $V_{ij\cdot kl}(0) = 0$;

b) $V_{ij\cdot kl}(\underline{h}) = V_{ji\cdot lk}(\underline{h}) = V_{ij\cdot kl}(\text{-}\underline{h})$;

c) $V_{ij\cdot il}(\underline{h}) = V_{ij\cdot kj}(\underline{h}) = V_{ij\cdot ij}(\underline{h}) = V_{ii\cdot ii}(\underline{h}) = 0$;

d) $V_{ij\cdot kl}(\underline{h}) = -V_{kj\cdot il}(\underline{h}) = V_{kl\cdot ij}(\underline{h}) = -V_{il\cdot kj}(\underline{h})$;

e) $|V_{ij\cdot kl}(\underline{h})| \leq \sqrt{V_{ii\cdot kk}(\underline{h})}\sqrt{V_{jj\cdot ll}(\underline{h})}$ (Desigualdade de Cauchy-Schwarz).

Agora, em se tratando de variograma, e considerando-se quaisquer dois componentes $Y_i(\underline{x})$, $Y_j(\underline{x}), i, j \in \{1, 2, ..., B\}$ de uma composição $\underline{Y}(\underline{x})$ tem-se:

a) o variograma da log razão, LR variograma dado por

$$V_{i\cdot j}(\underline{h}) = V_{ii\cdot jj}(\underline{h})$$

que são os elementos da matriz de LR variogramas (Matriz Variação Intrínsica), $\mathbf{\Gamma}(\underline{\mathrm{h}})$, de dimensão $B \times B$.

b) o ALR variograma cruzado como a função

$$\psi_{ij}(\underline{\mathrm{h}}) = V_{ij \cdot BB}(\underline{\mathrm{h}})$$

que são os elementos da matriz de ALR covariâncias cruzadas intrínsicas, $\mathbf{\Psi}(\underline{\mathrm{h}})$, de dimensão $(B-1) \times (B-1)$.

c) a CLR covariância cruzada é a função

$$\begin{aligned} \delta_{ij}(\underline{\mathrm{h}}) \;=\; \mathrm{Cov} \Bigg(& \ln\left(\frac{Y_i(\underline{\mathrm{x}})}{g\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)}\right) - \ln\left(\frac{Y_i(\underline{\mathrm{x}}+\underline{\mathrm{h}})}{g\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}}+\underline{\mathrm{h}})\big)}\right), \\ & \ln\left(\frac{Y_j(\underline{\mathrm{x}})}{g\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}})\big)}\right) - \ln\left(\frac{Y_j(\underline{\mathrm{x}}+\underline{\mathrm{h}})}{g\big(\underline{\mathrm{Y}}(\underline{\mathrm{x}}+\underline{\mathrm{h}})\big)}\right)\Bigg), \end{aligned}$$

que são os elementos da matriz de CLR covariâncias cruzadas, $\mathbf{\Delta}(\underline{\mathrm{h}})$, de dimensão $B \times B$.

3 Material e Métodos

3.1 Material

Estudo de Caso:

Neste trabalho serão utilizados principalmente conjuntos de dados contendo os valores observados de areia, silte e argila que são frações granulométricas do solo e que formam uma composição.

Será analisado um conjunto contendo medidas de areia, silte e argila de 39 amostras de sedimentos colhidas em diferentes profundidades de água em um lago Ártico (AITCHISON, 1986). Outra análise será feita em dados coletados por Gener e constituídos de 412 amostras contendo mensurações de areia muito grossa, areia média, areia fina e areia muito fina, além de altimetria e forma da superfície, se côncava ou convexa.

Uma terceira análise será feita em dados coletados por Bassoi que contém informações de cota, areia grossa, silte, argila, pHAgua, pHKCl, Ca, Mg, K, Al, H, C, N, CTC, S, V, M, NC, CEC, CN.

Também serão analisados dados provenientes de Gonçalves (1997) cujo trabalho foi conduzido no campo experimental de irrigação do Departamento de Engenharia Rural, situado nas coordenadas 22º42' de latitude sul, longitude oeste de 47º38' e altitude média de 546 m acima do nível do mar. Esta área em estudo consistiu de um quadrante irrigado por um sistema pivô-central, com declividade média de aproximadamente 2% na sua direção bissetriz. Esse quadrante correspondeu ao topo da encosta onde foi instalado o pivô. A superfície desse solo foi submetida a uma gradagem e, em seguida, foi demarcada uma transeção segundo um raio da área, com 230 metros em solo nu na direção da menor declividade do terreno. Os pontos foram marcados em transeção longa a cada $2,0$ m e as amostras, num total de 115 foram selecionadas em cada um destes pontos e analisadas segundo a granulometria à $0,20$ m de profundidade. Em seguida, foi definido um quadrante dessa área onde a transeção passou a ser sua bissetriz. Ali foram selecionadas amostras a cada 2 m totalizando 59 pontos. Com relação a esse conjunto de dados, a partir da constatação de estrutura de dependência espacial em distâncias inferiores a 20

m construiu-se uma malha quadrada ou grade de amostragem de 20 em 20 m onde foram analisadas 75 amostras.

3.2 Método

Como já mencionado, o objetivo do estudo é a construção de um modelo geoestatístico para dados composicionais que tenha uma função de covariância válida. Para isto, serão utilizados como referências básicas os trabalhos na linha de Diggle e Ribeiro Jr. (2007), Schmidt e Sansó (2006) e Banerjee, Carlin e Gelfand (2004) em que a geoestatística foi baseada na declaração explícita de modelos e o de Aitchison (1986) que apresentou a teoria de dados composicionais sem levar em consideração a dependência espacial. Desta forma, será seguido o paradigma de modelagem diferente do proposto por Pawlowsky-Glahn e Olea (2004) que considerou a dependência espacial mas não fez inferência estatística baseada na função de verossimilhança e tão pouco fez uso de inferência bayesiana. Obage (2005) apresentou uma análise bayesiana para dados composicionais mas não considerou a dependência espacial.

O que se pretende será a integração destas metodologias. Uma possibilidade de investigação será verificar se a imposição de restrição no espaço paramétrico, por exemplo a restrição de que a soma dos componentes deve ser igual a 1, poderá ser acoplada à especificação de funções de covariância válidas. Essas restrições no espaço paramétrico poderão ser impostas pela especificação de uma distribuição de probabilidades para a *priori* no paradigma da inferência bayesiana.

Esta integração será feita através de estudos sobre métodos geoestatísticos e teoria de dados composicionais separadamente; estudos sobre a teoria de verossimilhança e de inferência bayesiana para buscar soluções sobre como fazer tal integração; e estudar formas de implementar as restrições induzidas no espaço paramétrico. A partir do modelo obtido, espera-se poder aplicá-lo aos dados disponíveis e construir mapas de classificação do solo.

A performance de tais metodologias vai ser verificada através da análise de dados reais que motivam este trabalho.

Recursos Computacionais:

Será utilizado neste trabalho o ambiente operacional GNU/Linux. O pacote estatístico R (R Development Core Team, 2006), o pacote geoestatístico geoR (RIBEIRO JR.; DIGGLE, 2001) e o pacote *compositions* (BOOGAART, 2005), todos de uso livre e sob a licença GPL (*General Public Licence*) serão utilizados para análise geoestatística permitindo ajuste de modelos lineares, estimação linear, predição, krigagem finalizando com a construção de mapas temáticos, construção de diagramas ternários e cálculos de estatísticas descritivas para dados composicionais. Eventualmente será utilizado outros pacotes ou, até mesmo, poderá ser desenvolvido aplicações próprias.

Análise Estatística:

Será realizada uma análise descritiva dos dados referentes a areia, silte e argila com o objetivo exploratório de verificação dos pressupostos do modelo geoestatístico, como por exemplo, verificação da normalidade dos dados, identificação de pontos discrepantes e possível tendência direcional.

A avaliação da gaussianidade dos dados, será feita através da avaliação do perfil da função log-verossimilhança para o parâmetro λ de transformação.

Serão utilizados métodos baseados em verossimilhança para ajuste do modelo geoestatístico. A função de correlação adotada será da família de Matérn.

Eventualmente serão utilizados métodos bayesianos através da especificação da função de verossimilhança conforme Equação 26 através da qual será possível obter a distribuição *posteriori* dada pela Equação 28 com *prioris* a serem determinadas. Com isto será possível determinar a distribuição preditiva $P(\underline{Y}_0, \underline{Y})$ como na Equação 29.

Tentar-se-á adaptar os modelos bivariados apresentados na seção 2.4 aos dados composicionais incluindo a restrição de que a soma dos componentes deve ser igual a 1.

Análise de dados composicionais:

Esta análise inclui a construção do diagrama ternário para os dados de areia, silte e argila, a interpretação deste gráfico bem como o cálculo de estatísticas descritivas

como o centro da distribuição, a variância total, matriz de LR autocovariâncias, CRL covariâncias cruzadas servindo como uma análise exploratória dos dados.

Estudos serão feitos para a escolha da transformação a ser adotada bem como o componente que será considerado como a base para o cálculo das razões. Desta forma, a composição após transformação será formada por 2 componentes que justifica o modelo bivariado. O método de krigagem será empregado na construção do mapa de classificação do solo.

Modelos Iniciais:

O modelo bivariado composicional será dado por:

$$\begin{cases} Y_{i1} &= \mu_1(\underline{x}_i) + Z_1(\underline{x}_i) \\ Y_{i'2} &= \mu_2(\underline{x}_{i'}) + Z_2(\underline{x}_{i'}) \end{cases} \tag{41}$$

com $i, i' = 1, ..., n_j,\ j = 1, 2$ e $Z_i \sim N(0;\, \tau_i^2)$. Assim,

$$\begin{aligned} Cov(Y_{i1}, Y_{i1}) &= Cov(Y_1(\underline{x}_i); Y_1(\underline{x}_i)) \\ &= Cov(Z_1(\underline{x}_i); Z_1(\underline{x}_i)) \\ &= \rho_c(\underline{x}_i, \underline{x}_i)\tau_1\tau_1 \\ &= \tau_1^2\ ; \end{aligned}$$

de forma equivalente, $Cov(Y_{i'2}) = \tau_2^2$ e

$$\begin{aligned} Cov(Y_{i1}; Y_{i'2}) &= Cov(Z_1(\underline{x}_i); Z_1(\underline{x}_{i'})) \\ &= \rho_c(\underline{x}_i, \underline{x}_{i'})\tau_1\tau_2. \\ &= \rho_c(\underline{x}_i, \underline{x}_{i'})\tau_1\tau_2 I_1(i, i') \end{aligned}$$

onde I é a variável indicadora dada por:

$$I_1(i, i') = \begin{cases} 1 & ,\ \text{se} \quad i = i' \\ 0 & ,\ \text{se} \quad i \neq i'. \end{cases}$$

Logo, a matriz de covariância será:

$$\mathbf{\Sigma} = \begin{bmatrix} \tau_1^2 & \rho_c\tau_1\tau_2 & 0 & 0 & \cdots & 0 & 0 \\ \rho_c\tau_1\tau_2 & \tau_2^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \tau_1^2 & \rho_c\tau_1\tau_2 & \cdots & 0 & 0 \\ 0 & 0 & \rho_c\tau_1\tau_2 & \tau_2^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \tau_1^2 & \rho_c\tau_1\tau_2 \\ 0 & 0 & 0 & 0 & \cdots & \rho_c\tau_1\tau_2 & \tau_2^2 \end{bmatrix}.$$

que pode ser reescrita como $\mathbf{\Sigma} = \tau_1^2 G$, com $g = \tau_2/\tau_1$.

A função de verossimilhança para o modelo dado em (41) será:

$$L(\underline{\theta}) = \prod_{i=1}^{n} \left[(2\pi)^{-1/2} |\tau_1^2 G|^{-1/2} \exp\left\{ -\frac{1}{2\tau_1^2} (\underline{Y} - \underline{\mu}_{\underline{Y}})' G^{-1} (\underline{Y} - \underline{\mu}_{\underline{Y}}) \right\} \right]$$

e a função de log-verossimilhança:

$$l(\underline{\theta}) = \sum_{i=1}^{n} \left(log \left[(2\pi)^{-1/2} |\tau_1^2 G|^{-1/2} \exp\left\{ -\frac{1}{2\tau_1^2} (\underline{Y} - \underline{\mu}_{\underline{Y}})' G^{-1} (\underline{Y} - \underline{\mu}_{\underline{Y}}) \right\} \right] \right) \qquad (42)$$

Fazendo $Q = (\underline{Y} - \underline{\mu}_{\underline{Y}})' G^{-1} (\underline{Y} - \underline{\mu}_{\underline{Y}})$ onde $\underline{\mu}_{\underline{Y}} = D\underline{\mu} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$; derivando a Equação 42 em relação aos parâmetros $\underline{\mu}$ e $\underline{\tau}_1$ e igualando-as a zero obtém-se as seguintes estimativas:

$$\underline{\hat{\mu}} = \left(\sum_{i=1}^{n} D' G^{-1} D \right)^{-1} \left(\sum_{i=1}^{n} D' G^{-1} \underline{Y} \right)$$

e

$$\hat{\tau}_1 = \frac{1}{n} \sqrt{\sum_{i=1}^{n} Q}. \qquad (43)$$

Substituindo a Equação 43 na Equação 42 obtém-se a log-verossimilhança con-

centrada:

$$l(\underline{\theta}^*) \;=\; -\frac{n}{2}\left(log(2\pi) + n\,log\left(\sum_{i=1}^{n}\hat{Q}\right) + 2n\,log(n) + n\,log|G| + n^2\right). \qquad (44)$$

O modelo bivariado espacial composicional será dado por:

$$\begin{cases} Y_{i1} &= \mu_1(\underline{x}_i) + S_1(\underline{x}_i) + Z_1(\underline{x}_i) \\ Y_{i'2} &= \mu_2(\underline{x}_{i'}) + S_2(\underline{x}_{i'}) + Z_2(\underline{x}_{i'}) \end{cases} \qquad (45)$$

em que $S_j(\underline{x}) \sim N(0;\sigma_j^2)$ e $Z_j(\underline{x}) \sim N(0;\tau_j^2)$, $j = 1,2$. Os efeitos aleatórios S_1 e S_2 serão substituídos por U, um efeito aleatório padronizado com vetor de médias iguais a zero e matriz de covariância com variâncias unitárias na diagonal principal e covariâncias cruzadas dadas pela função de correlação adotada considerando o mesmo parâmetro de alcance e as unidades serão preservadas nas constantes padronizadoras σ_1 e σ_2. O modelo em (45) será então reescrito como:

$$\begin{cases} Y_{i1} &= \mu_1(\underline{x}_i) + \sigma_1 U(\underline{x}_i;\phi) + Z_1(\underline{x}_i) \\ Y_{i'2} &= \mu_2(\underline{x}_{i'}) + \sigma_2 U(\underline{x}_{i'};\phi) + Z_2(\underline{x}_{i'}) \end{cases}$$

Para este modelo,

$$\begin{aligned} Cov(Y_{i1};Y_{i1}) &= Cov(Y_1(\underline{x}_i);Y_1(\underline{x}_i)) \\ &= \sigma_1^2 Cov(U(\underline{x}_i;\phi);U(\underline{x}_i;\phi)) + Cov(Z_1(\underline{x}_i);Z_1(\underline{x}_i)) \\ &= \sigma_1^2 + \tau_1^2\,; \end{aligned}$$

$$\begin{aligned} Cov(Y_{i1};Y_{i'1}) &= Cov(Y_1(\underline{x}_i);Y_1(\underline{x}_{i'})) \\ &= \sigma_1^2 Cov(U(\underline{x}_i;\phi);U(\underline{x}_{i'};\phi)) + Cov(Z_1(\underline{x}_i);Z_1(\underline{x}_{i'})) \\ &= \sigma_1^2 \rho_U(\underline{x}_i;\underline{x}_{i'})\,; \end{aligned}$$

similarmente, $Cov(Y_{i2};Y_{i2}) = \sigma_2^2 + \tau_2^2$; $Cov(Y_{i2};Y_{i'2}) = \sigma_2^2\rho_U(\underline{x}_i;\underline{x}_{i'})$, e

$$
\begin{aligned}
Cov(Y_{i1}; Y_{i'2}) &= Cov(Y_1(\underline{x}_i); Y_2(\underline{x}'_i)) \\
&= \sigma_1\sigma_2 Cov(U(\underline{x}_i;\phi); U(\underline{x}_{i'};\phi)) + Cov(Z_1(\underline{x}_i); Z_1(\underline{x}_{i'})) \\
&= \sigma_1\sigma_2\rho_U(\underline{x}_i;\underline{x}_{i'})I_2(i,i') + \rho_c(\underline{x}_i;\underline{x}_{i'})\tau_1\tau_2 I_3(i,i')\,; \\
&= \sigma_1\sigma_2 I_2(i,i') + \tau_1\tau_2 I_3(i,i')\,;
\end{aligned}
$$

onde

$$
I_2(i,i') = \begin{cases} 1 & , \text{ se } \quad i = i' \\ \rho_U(\underline{x}_i;\underline{x}_{i'}) & , \text{ se } \quad i \neq i'. \end{cases}
$$

$$
I_3(i,i') = \begin{cases} \rho_U(\underline{x}_i;\underline{x}_{i'}) & , \text{ se } \quad i = i' \\ 0 & , \text{ se } \quad i \neq i'. \end{cases}
$$

Portanto, a matriz de covariância será dada por

$$
\boldsymbol{\Sigma} = \begin{bmatrix}
\sigma_1^2+\tau_1^2 & \sigma_1\sigma_2+\rho_c\tau_1\tau_2 & \sigma_1^2\rho_U & \rho_U\sigma_1\sigma_2 & \cdots & \sigma_1^2\rho_U & \rho_U\sigma_1\sigma_2 \\
\sigma_1\sigma_2+\rho_c\tau_1\tau_2 & \sigma_2^2+\tau_2^2 & \rho_U\sigma_1\sigma_2 & \sigma_2^2\rho_U & \cdots & \rho_U\sigma_1\sigma_2 & \sigma_2^2\rho_U \\
\sigma_1^2\rho_U & \rho_U\sigma_1\sigma_2 & \sigma_1^2+\tau_1^2 & \sigma_1\sigma_2+\rho_c\tau_1\tau_2 & \cdots & \sigma_1^2\rho_U & \rho_U\sigma_1\sigma_2 \\
\rho_U\sigma_1\sigma_2 & \sigma_2^2\rho_U & \sigma_1\sigma_2+\rho_c\tau_1\tau_2 & \sigma_2^2+\tau_2^2 & \cdots & \rho_U\sigma_1\sigma_2 & \sigma_2^2\rho_U \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\sigma_1^2\rho_U & \rho_U\sigma_1\sigma_2 & \sigma_1^2\rho_U & \rho_U\sigma_1\sigma_2 & \cdots & \sigma_1^2+\tau_1^2 & \sigma_1\sigma_2+\rho_c\tau_1\tau_2 \\
\rho_U\sigma_1\sigma_2 & \sigma_2^2\rho_U & \rho_U\sigma_1\sigma_2 & \sigma_2^2\rho_U & \cdots & \sigma_1\sigma_2+\rho_c\tau_1\tau_2 & \sigma_2^2+\tau_2^2
\end{bmatrix}
$$

que pode ser reescrita como $\boldsymbol{\Sigma} = \sigma^2 V$, ao se fazer as reparametrizações: $\sigma_1 = \sigma$; $\sigma_2 = \eta\sigma$; $\nu_1 = \tau_1/\sigma$; $\nu_2 = \tau_2/\sigma$. A função de log-verossimilhança é dada pela Equação 12 e as estimativas como nas Equações 14 e 16.

Na sequência deste estudo, esses modelos serão tratados sob o paradigma bayesiano.

Resultados preliminares:

Os dados obtidos na malha quadrada consistiram de 76 localizações amostrais com coordenadas mínimas iguais a (0; 0) e máximas iguais a (180; 180). A distância mínima entre duas localizações foi igual a 20 e a máxima igual a $254,5584$. A transformação

logit,$log(x/(1-x))$, adequada para dados expressos em proporções foi aplicada aos dados de areia, silte e argila e através da Tabela 2 observa-se que os maiores percentuais ocorreram para areia e os dados de silte se apresentaram mais homogêneos.

Tabela 2: Estatísticas descritivas do logit da porcentagem de areia, silte e argila.

Componente	Min	Max	Média	Mediana	Q1	Q3	Desv.Pad.
Areia	$0,10450$	$1,87300$	$1,23500$	$1,32400$	$1,04300$	$1,48700$	$0,37848$
Silte	$0,08673$	$0,97400$	$0,60480$	$0,61300$	$0,49640$	$0,72270$	$0,17556$
Argila	$0,03541$	$1,59100$	$0,69050$	$0,66360$	$0,42080$	$0,94430$	$0,35100$

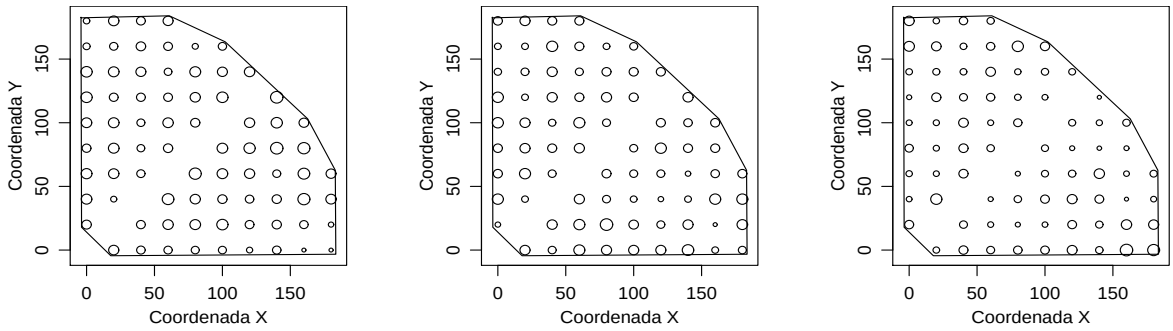

Figura 7: Gráfico de círculos do logit da porcentagem de areia, silte e argila.

A Figura 7 aponta que tanto a coordenada Y como a coordenada X não interferem no valor médio do logit da porcentagem de areia, silte e argila pois ocorrem valores distintos para todas as coordenadas.

Tabela 3: Estimativas de máxima verossimilhança.

Componente	$\hat{\beta}$	$\hat{\tau}^2$	$\hat{\sigma}^2$	$\hat{\phi}$	k	log-vero
Areia	$0,2659$	0	$0,1503$	$4,6867$	$4,5$	$-23,75804$
Silte	$-0,3427$	0	$0,0168$	$7,0253$	$0,5$	$26,28311$
Argila	$-0,3946$	0	$0,2057$	$4,2024$	$3,5$	$-24,69256$

Analisando a areia, através da Figura 8 no canto esquerdo superior, pode-se observar a possível existência de dependência espacial entre as localizações. Os diagramas de dispersão não revelam a existência de tendência nas coordenadas. O histograma e a Figura 9 indicam a necessidade de transformação dos dados. Pela Figura 10, tem-se que o intervalo de 95% de confiança para λ compreende valores aproximadamente entre $1,2$ e $2,2$ considerando-se $\lambda = 1,5$. O valor de k que maximiza o logaritmo da função de verossimilhança é $4,5$ obtendo-se como estimativas dos parâmetros os valores

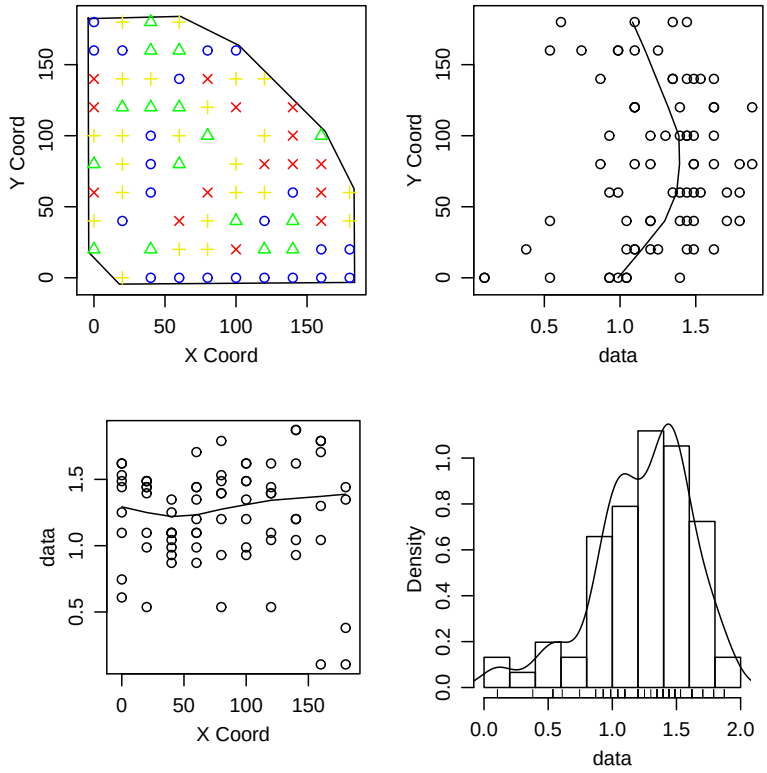

Figura 8: Localizações (topo à esquerda), logit(porcentagem) vs coordenadas (topo à direita e baixo à esquerda), e histograma (baixo à direita) da areia.

apresentados na Tabela 3 juntamente com o valor maximizado do logaritmo da função de verossimilhança. O mapa para os dados de areia transformados para a escala original é apresentado na Figura 17 (esquerda).

Em relação a silte, pela Figura 11 no canto esquerdo superior, também se observa a possível existência de dependência espacial entre as localizações. Os diagramas de dispersão não revelam a existência de tendência nas coordenadas. O histograma e a Figura 12 indicam a necessidade de transformação dos dados. Através da Figura 13, tem-se que o intervalo de 95% de confiança para λ compreende valores aproximadamente entre $0,9$ e $2,1$ considerando-se $\lambda = 1,5$. O valor $k = 0,5$ também foi considerado neste caso obtendo-se como estimativas dos parâmetros os valores apresentados na Tabela 3 juntamente com o valor maximizado do logaritmo da função de verossimilhança. O mapa para os dados de silte transformados para a escala original é apresentado na Figura 17(centro).

Em se tratando dos dados de argila, pela Figura 14, são válidos os mesmos

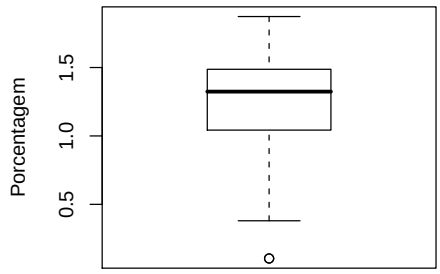

Figura 9: Boxplot do logit da porcentagem de areia.

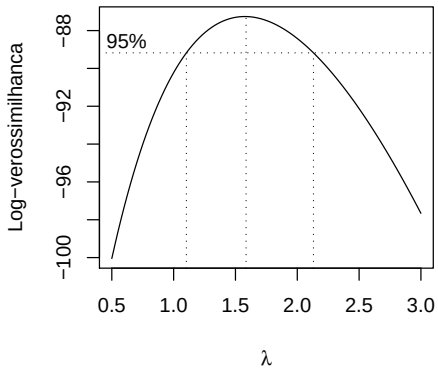

Figura 10: Perfil do log da verossimilhança para o parâmetro λ de transformação de Box-Cox para areia

comentários feitos para areia e silte com excessão do histograma, em que a distribuição da argila se apresenta levemente assimétrica à direita e que juntamente com a Figura 15 indicam a necessidade de uma transformação dos dados. Através da Figura 16, tem-se que o intervalo de 95% de confiança para λ compreende valores aproximadamente entre $0,4$ e $1,0$. Considera-se $\lambda = 0,5$ e $k = 3,5$ (Figura 16) e as estimativas dos parâmetros com o valor que maximiza a log-verossimilhança encontram-se na Tabela 3. O mapa de krigagem para os dados de argila transformados para a escala original é apresentado na Figura 17 (direita).

Por outro lado, como areia, silte e argila forma uma composição, tem-se na Figura 18 o respectivo diagrama ternário. A disposição dos pontos no diagrama indica que a proporção de argila é maior do que os outros dois componentes e a menor proporção foi obtida para o componente silte. Por outro lado, a variabilidade da razão argila/areia e maior que a variabilidade da razão silte/areia. Da mesma forma, a variabilidade da razão argila/silte é maior que a variabilidade da razão areia/silte. Isto também pode ser observado na Figura 20. Pode-se observar pela Figura 19 a localização do centro da distribuição e que a região de confiança de 2-sigma contém a maioria das observações.

O centro da distribuição é igual a $(0,2785999; 0,2324486; 0,4889515)$, correspondendo a areia, silte e argila, respectivamente, e a matriz de variâncias na estrutura

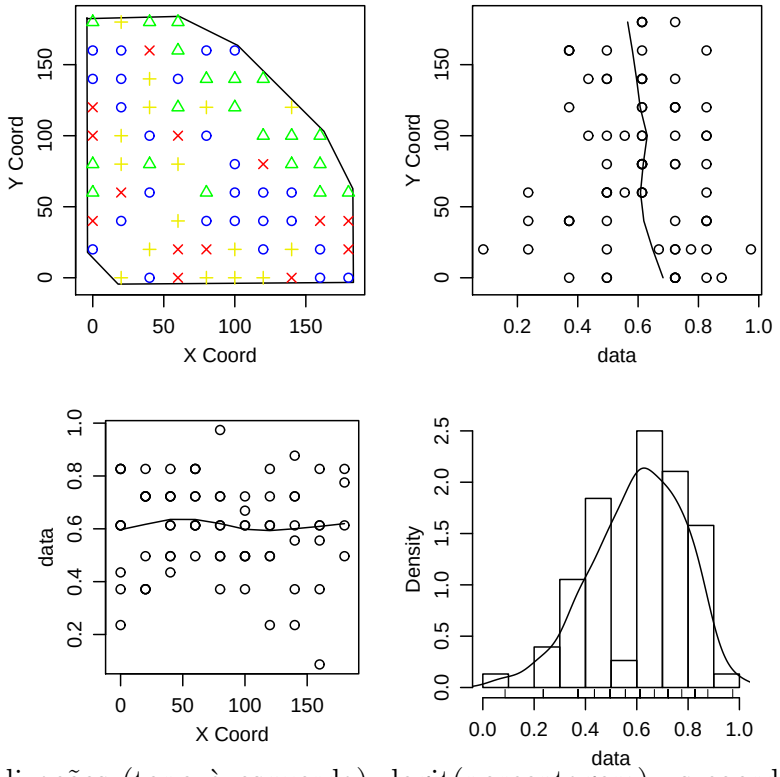

Figura 11: Localizações (topo à esquerda), logit(porcentagem) vs coordenadas (topo à direita e baixo à esquerda), e histograma (baixo à direita) do silte.

CLR do espaço euclidiano é dada por:

$$Var = \begin{bmatrix} 0,051986106 & -0,004567419 & -0,047418688 \\ -0,004567419 & 0,011556683 & -0,006989264 \\ -0,047418688 & -0,006989264 & 0,054407952 \end{bmatrix}$$

A variância métrica que nada mais é do que o traço da matriz CLR é igual a $0,1179507$. Considerando a média geométrica, os desvios padrão clássicos para areia, silte e argila são iguais $0,252532$, $0,130244$ e $0,175178$ indicando a baixa variabilidade dos componentes. A matriz variação $\big(Var\big(log(x_i/x_j)\big)\big)$ é igual a:

$$Var = \begin{bmatrix} 0 & 0,07267763 & 0,20123143 \\ 0,07267763 & 0 & 0,07994316 \\ 0,20123143 & 0,07994316 & 0 \end{bmatrix}$$

Por fim, a variância total é igual a $0,1179507$.

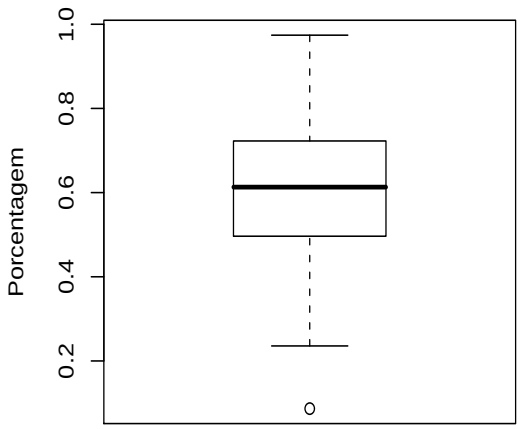

Figura 12: Boxplot do logit da porcentagem de silte.

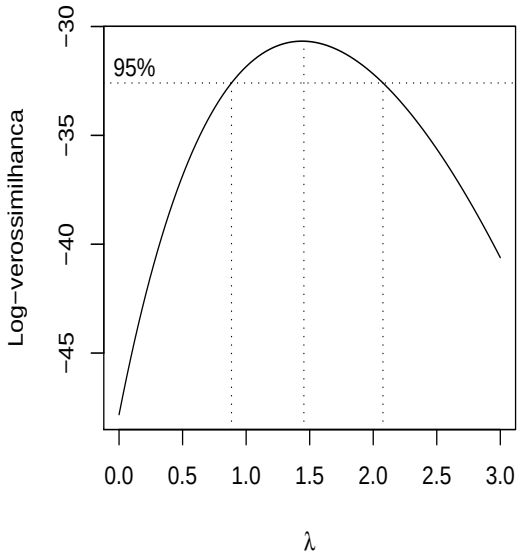

Figura 13: Perfil do log da verossimilhança para o parâmetro λ de transformação de Box-Cox para silte.

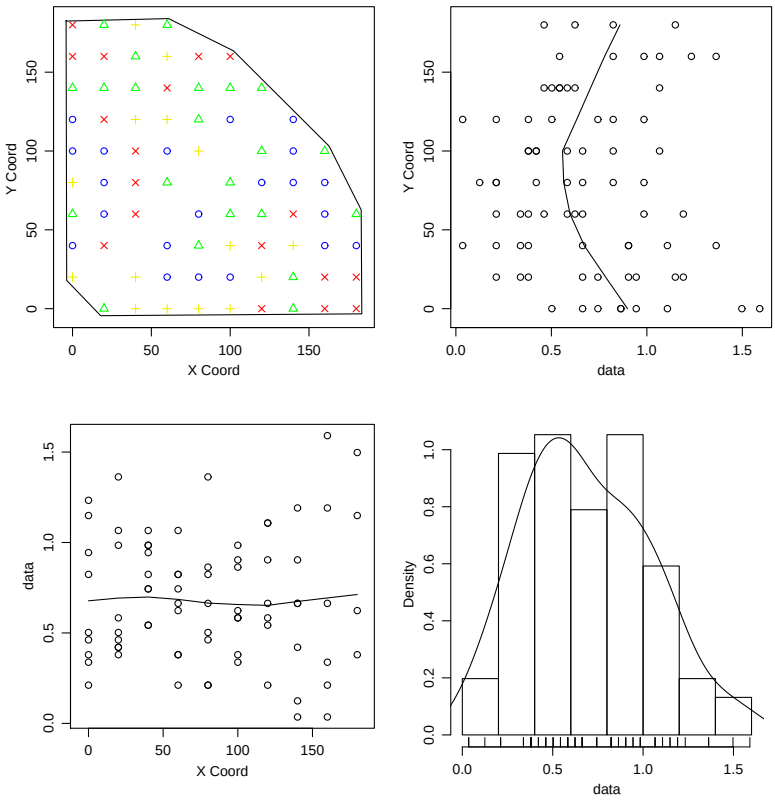

Figura 14: Localizações (topo à esquerda), logit(porcentagem) vs coordenadas (topo à direita e baixo à esquerda), e histograma (baixo à direita) da argila.

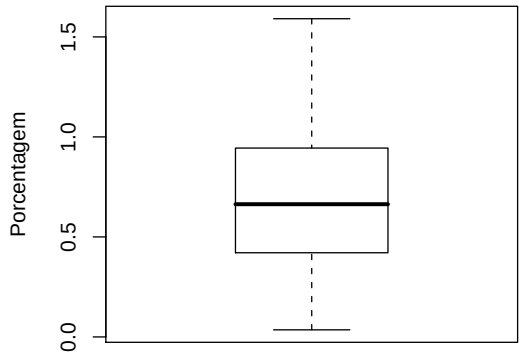

Figura 15: Boxplot do logit da porcentagem de argila

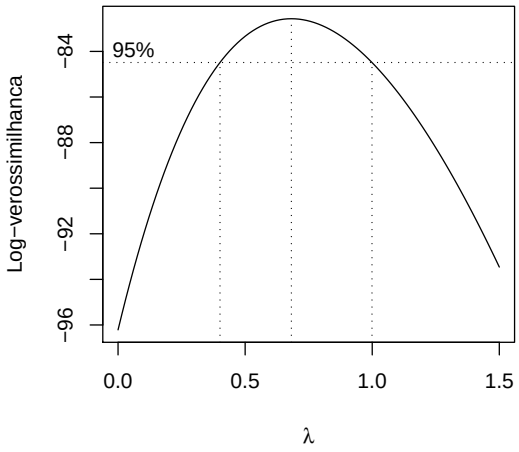

Figura 16: Perfil do log da verossimilhança para o parâmetro λ de transformação de Box-Cox para argila.

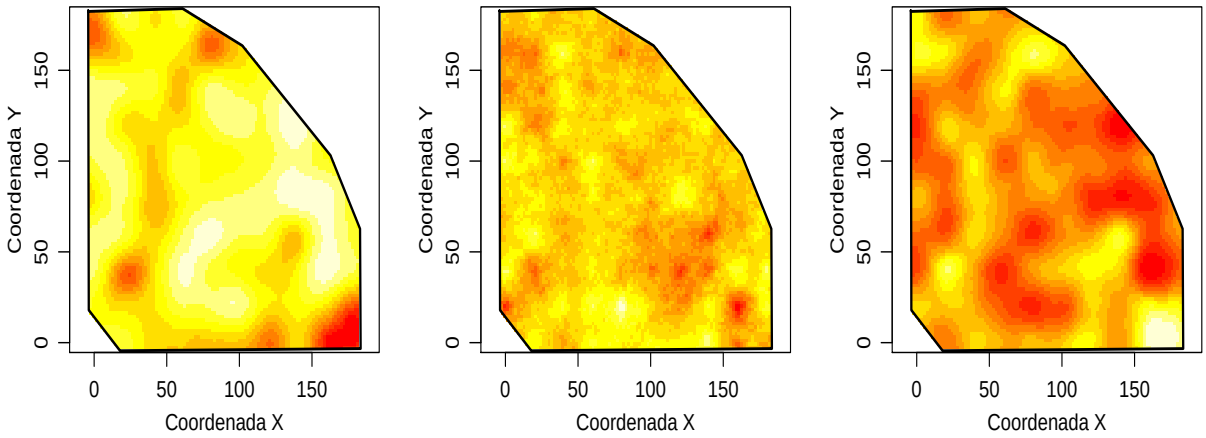

Figura 17: Mapas da porcentagem de areia (à esquerda), silte (centro) e argila (à direita).

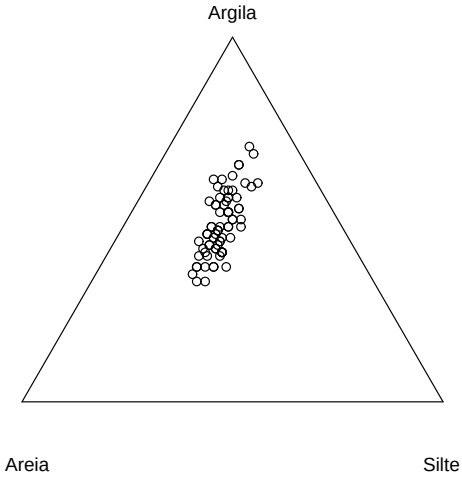

Figura 18: Diagrama ternário para areia, silte e argila.

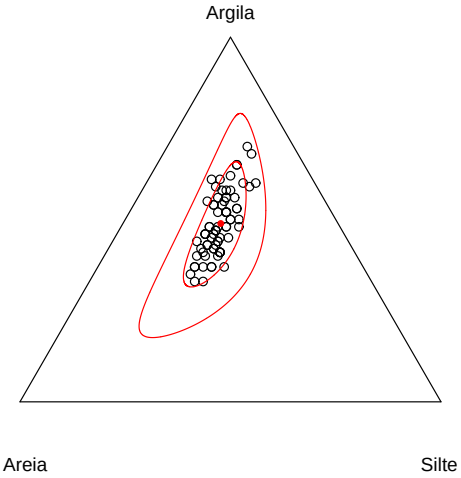

Figura 19: Diagrama Ternário para areia, silte e argila incluindo o centro da distribuição e regiões de confiança de 2 e 4 sigma.

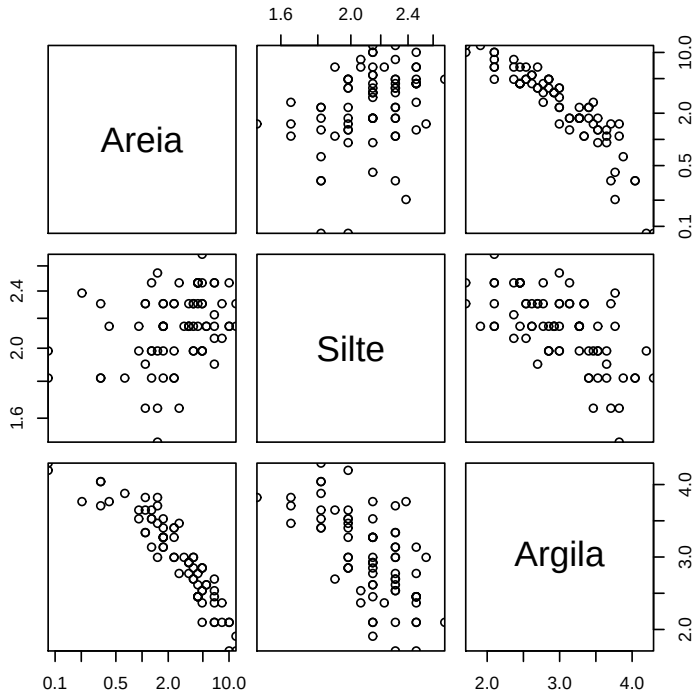

Figura 20: Diagrama de dispersão para areia vs silte, areia vs argila e silte vs argila.

4 Resultados Esperados

- Dado que a aplicação de uma transformação resultará numa composição com dois componentes espera-se que a imposição no modelo bivariado de uma restrição induzida pelo modelo de dados composicionais torne o modelo bivariado compatível com a estrutura de dados correlacionados;

- Avaliar o modelo proposto em comparação com o modelo considerando as variáveis separadamente;

- Desenvolver programas computacionais para a implementação e análise do modelo proposto;

- Aplicar a metodologia em dados reais que motivam este trabalho, por exemplo, na contrução de mapa temático de classificação espacial do solo segundo concentrações de areia, silte e argila.

5 Cronograma

Ano	2008				2009												2010	
Mês	09	10	11	12	01	02	03	04	05	06	07	08	09	10	11	12	01	02
1. Qualificar o projeto junto ao PPGMNE	X																	
2. Completar a revisão da literatura e Elaborar Metodologia de análise de dados	X	X	X	X	X													
3. Aplicar metodologia aos conjuntos de dados					X	X	X	X	X	X	X							
4. Elaborar relatório preliminar com resultados e discussão								X	X	X								
5. Elaborar versão preliminar da tese											X	X	X	X				
6. Fazer revisão ortográfica e gramatical da tese															X	X		
7. Redação e submissão de artigos			X	X	X	X	X	X	X	X	X	X	X	X	X			
8. Confecção da tese	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		
9. Defesa da tese																	X	X

Referências

AITCHISON, J. The statistical analysis of compositional data. **Royal Statistical Society, Series B**, v. 44, p. 139–177, 1982. ISSN 0882-8121.

AITCHISON, J. (Ed.). **The Statistical analysis of compositional data**. First. New Jersey: The Blackburn Press, 1986. ISBN 1-930665-78-4.

AITCHISON, J. Logratios and natural laws in compositional data analysis. **Mathematical Geology**, v. 31, p. 563–580, 1999. ISSN 0882-8121. Subject Collection: Earth and Environment Science.

AITCHISON, J. et al. Logratios analysis and compositional distance. **Mathematical Geology**, v. 32, p. 563–580, 2000. ISSN 0882-8121.

AITCHISON, J.; EGOZCUE, J. J. Compositional data analysis: Where are we and where should we be hesding? **Mathematical Geology**, v. 37, p. 829–850, October 2005. Subject Collection Earth and Environmental Science. DOI: 10.1007/s11004-005-7383-7.

AITCHISON, J.; GREENACRE, M. Biplot of compositional data. **Journal of the Royal Statistical Society, Series C**, v. 51, p. 375–392, October 2002.

BAILEY, T. C.; GATRELL, A. C. **Interactive spatial data analysis**. Harlow: Longman, 1995. ISBN 0-582-24493-5.

BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. **Hierachical modelling and analysis for spatial data**. Boca Raton: Chapman and Hall, 2004. ISBN 1-58488-410-X.

BARCELÓ-VIDAL, C.; MARTÍN-FERNÁNDEZ, J. A.; PAWLOWSKY-GLAHN, V. Mathematical foundations of compositional data analysis. 2001. Acesso em 08/06/08. Disponível em: <http://ima.udg.edu/∼jamf/carles_martin_vera_cancun.pdf>.

BOOGAART, K. G. v. d. Using the r package "compositions". p. 1–17, June 2005. Acesso em 09/02/08. Disponível em: <http://www.stat.boogaart.de/compositions>.

BOX, G.; COX, D. An analysis of transformation. **Journal of the Royal Statistical Society, Series B**, p. 211–252, 1964.

BUTLER, A.; GLASBEY, C. A latent gaussian model for compositional data with zeros. **Journal of the Royal Statistical Society, Series C, (in press)**, 2008. Disponível em: <http://www.bioss.ac.uk/staff/adam/documents/ButlerGlasbey_resubmit.pdf>.

CRESSIE, N. A. C. (Ed.). **Statistics spatial data**. New York: Wiley, 1993.

DEGROOT, M. H.; SCERVISH, M. J. **Probability and Statistics**. Third. Boston: Addison Wesley, 2002. ISBN 0-201-52488-0.

DIGGLE, P.; LEITE, R. M. D. M.; SU, T.-L. Geoestatistical analysis under preferencial sampling. v. 1, 2007. Acesso em: 25/07/08. Disponível em: <http://www.maths.lancs.ac.uk/∼sut3/CAPAPER/geopref_tingli.pdf>.

DIGGLE, P. J.; RIBEIRO JR., P. J. **Model-based geostatistics**. USA: Springer Series in Statistics, 2007.

DIGGLE, P. J.; RIBEIRO JR, P. J.; CHRISTENSEN, O. F. An introduction to model-based geostatistics. In: MØLLER, J. (Ed.). **Spatial Statistics and Computational Methods**. New York: Springer, 2003. p. 43–86.

DIGGLE, P. J.; TAWN, J. A.; MOYEED, R. A. Model-based geostatistics. **Applied Statistics**, v. 47, n. 3, p. 299–350, 1998. Disponível em: <citeseer.ist.psu.edu/diggle98modelbased.html>.

EYNATTEN, H. v.; BARCELÓ-VIDAL, C.; PAWLOWSKY-GLAHN, V. Modelling compositional change: the example of chemical weathering of granitoid rocks. **Mathematical Geology**, v. 35, p. 231–251, April 2003. ISSN 0882-8121. Subject Collection Earth and Environmental Science. DOI:10.1023/A1023835513705.

GAMERMAN, D.; LOPES, H. F. **Markov chain Monte Carlo: stochastic simulation for bayesian inference**. 2a. ed. Londres: Chapman & Hall/CRC, 2006.

GONÇALVES, A. C. A. (Ed.). **Variabilidade espacial de propriedades físicas do solo para fins de manejo da irrigaçã**. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz". Universidade de São Paulo: Piracicaba, 1997.

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

GRAF, M. Precision of compositional data in a stratified two-stage cluster sample: comparison of the swiss earnings structure survey 2002 and 2004. In: **Survey Research Methods Section, ASA**. [s.n.], 2006. Disponível em: <http://www.amstat.org/Sections/Srms/Proceedings/>.

ISAAKS, E. H.; SRISVASTAVA, R. M. **An introduction to applied geostatistics**. New York: Oxford University Press, 1989.

JOHNSON, R. A.; WICHERN, D. W. **Applied statistical analysis**. Fourth. USA: Prentice Hall, 1998.

KITANIDIS, P. **Introduction to geostatistics: applications in hydrogeology**. Third. New York: Cambridge University Press, 1997.

LABUS, M. Compositional data analysis as a toll for interpretation of rock porosity parameters. **Geological Quarterly**, v. 49, p. 347–354, 2005.

MATÉRN, B. **Spatial Variation**. Stockholm, 1960.

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

MATHERON, G. (Ed.). **Les variables of régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature**. Paris: Masson et Cie. 305 p., 1965.

MATHERON, G. **Pour une Analyse Krigeante des Données Régionaliées**. Ecole Nationale Supérieure des Mines de Paris, 1982. Technical Report.

MCBRATNEY, A. B.; DE GRUIJTER, J. J.; BRUS, D. J. Spatial prediction and mapping of continuous soil classes. **Geoderma**, p. 39–64, 1992.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. (Ed.). **Introduction to the theory of statistics**. Third. USA: McGraw-Hill, 1974.

OBAGE, S. C. (Ed.). **Uma análise bayesiana para dados composicionais**. Dissertação (Mestrado em Estatística). Universidade Federal de São Carlos: São Carlos, 2005. Disponível em: <hdl.handle.net/10229/37845>.

ODEH, I. O. A.; TOOD, A. J.; TRIANTAFILIS, J. Spatial prediction of soil particle-size fractions as compositional data. **Soil Science**, v. 168, n. 7, p. 501–515, July 2003. ISSN 0882-8121. DOI:10.1097/01.ss.0000080335.10341.23.

OLIVEIRA, J. B. Classificação de solos e seu emprego agrícola e não agrícola. 2008? Disponível em: <http://jararaca.ufsm.br/websites/dalmolin/download/textospl/classif.pdf>.

PAWLOWSKY-GLAHN, V.; OLEA, R. A. (Ed.). **Geostatistical Analysis of Compositional Data**. New York: Oxford University Press, Inc., 2004. (Studies in Mathematical Geology;7). ISBN 0-19-517166-7.

PAWLOWSKY-GLAHN, V.; OLEA, R. A.; DAVIS, J. C. Estimation of regionalized compositions: a comparison of three methods. **Mathematical Geology**, v. 27, n. 1, p. 105–127, 1995. ISSN 0882-8121.

REIS, E. **Estatística Multivariada Aplicada**. Lisboa: Edições Sílabo, LDA, 1997.

REYMENT, R. A.; SAVAZZI, E. (Ed.). **Aspects of Multivariate Statistical Analysis in Geology**. [S.l.]: Elsevier, 1999.

RIBEIRO JR, P. J.; DIGGLE, P. J. **Bayesian inference in Gaussian model-based geostatistics**. Lancaster, 1999. Relatório Técnico.

RIBEIRO JR., P. J.; DIGGLE, P. J. geoR: a package from geostatistical analysis. **R-NEWS**, v. 1, p. 15–18, June 2001. ISSN 1609-3631. Disponível em: <http://cran.R-project.org/doc/Rnews>.

SCHABENBERGER, O.; GOTWAY, C. A. **Statistical Methods for Spacial Data Analysis**. Boca Raton: Chapman and Hall, 2005.

SCHABENBERGER, O.; PIERCE, F. J. (Ed.). **Contemporary Statistical Models for the Plant and Soil Sciences**. Boca Raton: Taylor and Francis, 2001.

SCHMIDT, A. M.; GELFAND, A. E. A bayesian coregionalization approach for multivariate pollutant data. **Journal of Geophysical Research**, v. 108, p. 10–1–10–8, 2003. ISSN 8783.

SCHMIDT, A. M.; SANSÓ, B. Modelagem bayesiana da estrutura de covariância de processos espaciais e espaço temporais. In: *17 SINAPE e ABE-Associação Brasileira de Estatística*. Caxambu: Associação Brasileira de Estatística, 2006. **Minicurso**.

STEIN, M. L. **Interpolation of Spatial Data: Some Theory for Kriging**. New York: Springer Verlag, 1999.

TOLOSANA-DELGADO, R.; OTERO, N.; PAWLOWSKY-GLAHN, V. Some basic concepts of compositional geometry. **Mathematical Geology**, v. 37, n. 7, p. 563–580, October 2005. ISSN 0882-8121. DOI: 10.1007/s11004-005-7374-8.

WACKERNAGEL, H. (Ed.). **Multivariate Geostatistics: An Introduction with Applications**. Second. Germany: Springer, 1998.

WILLIAMS, C. K. I. Gaussian processes. 2002. Disponível em: <http://www.gaussianprocess.org/>.