**Universidade de São Paulo**
**Escola Superior de Agricultura "Luiz de Queiroz"**

# Abordagens para análise de dados composicionais

**Naimara Vieira do Prado**

Tese apresentada para obtenção do título de Doutora em Ciências. Área de concentração: Estatística e Experimentação Agronômica

**Piracicaba**
**2017**

**Naimara Vieira do Prado**
**Engenheira Agrícola**

# Abordagens para análise de dados composicionais

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **PAULO JUSTINIANO RIBEIRO JR**

Tese apresentada para obtenção do título de Doutora em Ciências. Área de concentração: Estatística e Experimentação Agronômica

**Piracicaba**
**2017**

**Dados Internacionais de Catalogação na Publicação**
**DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Prado, Naimara Vieira do

Abordagens para análise de dados composicionais / Naimara Vieira do Prado. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2017 .

83 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Equações generalizadas 2. Composição regionalizada 3. Geoestatística 4. Dependência espacial I. Título.

# AGRADECIMENTOS

Em primeiro lugar quero agradecer a Deus, por proporcionar saúde, força e determinação para finalizar esse trabalho.

Aos meus pais, ao Fabrício, Thaís e Silvio por todo amor e apoio nos momentos difíceis que passei para chegar até aqui. Muito, muito obrigada.

Agradeço ao meu orientador Prof. Paulo Justiniano Ribeiro Junior pela orientação, ensinamentos e paciência. Agradeço também a sua família pelo acolhimento, em especial ao Luca pelos momentos alegres e amizade.

Aos meus amigos da Esalq, Simone Grego e Everton Batista, muito obrigada pela amizade sincera .

Ao pessoal do LEG (UFPR) agradeço por toda contribuição durante a temporada que estive em Curitiba.

Aos meus amigos e colegas de profissão da Universidade Tecnológica Federal do Paraná (UTFPR), campus de Francisco Beltrão, muito obrigada pelo apoio para finalizar este trabalho.

Às secretárias Solange de Assis Paes Sabadin e Mayara Segatto, obrigada pela ajuda durante todos esses anos.

# SUMÁRIO

# RESUMO

## Abordagens para análise de dados composicionais

Dados composicionais são vetores, chamados de composições, cujos componentes são todos positivos, satisfazem a soma igual a 1 e possuem um espaço amostral próprio chamado Simplex. A restrição da soma induz a correlação entre os componentes. Isso exige que os métodos estatísticos para análise desses conjuntos de dados considerem esse fato. A teoria para dados composicionais foi desenvolvida inicialmente por Aitchison na década de 80. Desde então, várias técnicas e métodos têm sido desenvolvidos para a modelagem dos dados composicionais. Este trabalho apresenta as principais abordagens para a análise estatística de dados composicionais independentes. Sendo, regressão Dirichlet (distribuição natural aos dados composicionais) ou o uso de transformações em razões logarítmicas que saem do espaço simplex para o espaço real. Também descreve os métodos para os casos em que a suposição de independência não pode ser atendida. Por exemplo, dados composionais com dependência espacial. Para esses casos, há na literatura métodos baseados nas teorias desenvolvidas para análise geoestatística de dados univariados; ou, no uso de transformações em razões logarítmicas com a inclusão da dependência espacial. Além de revisitar os métodos já difundidos, propõe-se o uso do método de Equações de Estimação Generalizadas (EEG) como alternativa para a análise de dados composicionais independentes e com dependência espacial. A principal vantagem é que as equações de estimação necessitam apenas da especificação de funções que descrevam a média e a estrutura de covariância. Assim, não é necessário atribuir uma distribuição de probabilidade aos dados ou fazer o uso de transformações. A aplicação do método EEG para dados composicionais independentes apresentou resultados tão eficientes quanto a regressão Dirichlet ou transformação em razões logarítmicas. Para os dados composicionais com dependência espacial, o método baseado em verossimilhança foi o que apresentou valores preditos mais próximos aos valores reais. O método EEG foi mais eficaz do que a abordagem geoestatística dos componentes individuais, porém, comparado com os demais métodos, foi o que apresentou maior valor residual.

**Palavras-chave:** Equações de Estimação Generalizadas, Composição regionalizada, Geoestatística, Dependência espacial.

# ABSTRACT

## Approaches to compositional data analysis

ompositional data are vectors, called compositions, whose components are all positive, it satisfies the sum equal one and has a Simplex space. The sum constraint induces the correlation between the components and this requires that the statistical methods for the analysis of datasets consider this fact. The theory for compositional data was developed mainly by Aitchison in the 1980s, and since then, several techniques and methods have been developed for compositional data modelling. This work presents the main approaches for the statistical analysis of independent compositional data, such as Dirichlet regression (natural distribution to compositional data) or the use of transformations log-ratios that aim to leave the simplex space for to Euclidean space. Also describes the methods for cases where the assumption of independence cannot be satisfied, for example, spatial dependence compositional data. For these cases, there are in the literature methods of analysis based on the theories developed for univariate geostatistics analysis or use of log-ratios transformations with the inclusion of the spatial dependence generated by the distance between the points. In addition, to revisiting the already diffused methods, this work propose the use of the Generalized Estimation Equation (GEE) method as an alternative for the analysis of independent compositional data and with spatial dependence. The GEE only requires the specification of functions that describe the mean and correlation matrix (covariance structure, therefore, it is not necessary to assign a probability distribution to the data or transformations. The application of the GEE method for independent compositional data presented results as efficient as Dirichlet regression or log-ratios transformation. Compositional data with spatial dependence, log-ratios transformations presented predicted values close to the real values. GEE method was more effective than the traditional geostatistical approach, however, compared with the other methods, It was the one that presented the high residual values.

**Keywords:** Generalized Estimation Equations, Regionalized compositions, Geostatistics, Spatial dependence.

## LISTA DE FIGURAS

# LISTA DE TABELAS

# 1 INTRODUÇÃO

Nas mais diversas áreas, agricultura, ciências sociais, medicina, ciências biológicas, geologia, dentre outras, surgem conjuntos de dados que possuem uma estrutura na forma de "composição". Em dados de textura do solo, as frações de areia, silte e argila, formam uma "composição" com três elementos para essa propriedade do solo; na descrição centesimal de alimentos, a quantidade de cinzas, umidade, proteínas, lipídeos e carboidratos formam a "composição" do total do alimento; esses são exemplos de dados que podem ser abordados como uma composição, cujos elementos, no seu total, irão formar todo o conjunto de dados. A essa característica dá-se o nome de dados composicionais.

Os dados composicionais possuem certas peculiaridades: cada elemento/categoria do conjunto, chamado componente é estritamente positivo e a soma de todos os componentes deve ser 100% ou 1, na escala decimal. Apesar dessa estrutura ser facilmente encontrada em várias áreas de estudo, a análise de dados composicionais foi sistematizada somente na década de 80, por Aitchison (AITCHISON, 1982, 1986).

Por todas as restrições, a modelagem para dados composicionais tem evoluído ao longo do tempo para levar em conta a correlação existente entre os componentes. Os dados composicionais tem um espaço amostral próprio, chamado Simplex (AITCHISON, 1982). Ao trabalhar com dados composicionais no simplex, uma classe conhecida de distribuição de probabilidades pode ser utilizada, chamada distribuição Dirichlet (BARNDORFF e JORGENSEN, 1991).

A distribuição Dirichlet é a distribuição natural e bem estruturada para dados no espaço simplex. Inicialmente, houve uma grande utilização dessa classe para modelagem dos dados e de possíveis generalizações que são consideradas até hoje. Porém, na prática, ela se torna inadequada para descrição da variabilidade de dados composicionais, pois sua estrutura de correlação é ocasionalmente inapropriada aos dados com correlação positiva, o que geralmente ocorre em dados composicionais (AITCHISON, 1986).

Aitchison (1986) propôs uma maneira de analisar os dados composicionais, no caso de amostras independentes, com o uso de transformações adequadas que saem do espaço amostral simplex $\mathbb{S}^D$ e passam ao espaço real $\mathbb{R}^D$. Porém, existem situações em que assumir a independência pode não ser adequado, por exemplo, dados longitudinais com dependência temporal ou dados geoestatísticos com dependência espacial. Nestes casos, as metodologias tradicionais apresentadas por Aitchison precisam ser estendidas.

Para os casos com dependência espacial, a abordagem geoestatística dos componentes individuais não considera a covariância entre os componentes, assim, não garantem que os valores interpolados satisfaçam a soma igual a 1, o que pode gerar valores interpolados irreais (ODEH *et al.*, 2003). Outra abordagem aplicada nesses casos sugere modelar a dependência espacial do conjunto de dados composicionais $\mathbf{Y}$, com a introdução de efeitos aleatórios que induzem uma estrutura de dependência. Assim, define-se a distri-

buição de $\mathbf{Y}$ condicional aos efeitos aleatórios e especificam-se as condicionais de $\mathbf{Y}$. Essa abordagem é apresentada no trabalho de Martins *et al.* (2009).

A proposta deste trabalho é apresentar algumas das diversas abordagens para análise de dados composicionais e propor o uso do método de Equações de Estimação Generalizadas - EEG (ZEGER e LIANG, 1986; PETER e SONG, 2007), no qual, não é necessária a suposição de distribuições aos dados. O método consiste em especificar as equações de momentos para os dados, normalmente, uma função que descreva a média e uma estrutura para a matriz de variâncias e covariâncias. Com isso, é possível, estimar os parâmetros de interesse, sem a construção da função de verossimilhança associada ao modelo. Deste modo, o intuito é desenvolver uma aproximação em que os momentos dos próprios dados são usados para obter as estimativas dos parâmetros, sem a necessidade de transformações.

O trabalho está estruturado da seguinte forma: no segundo capítulo há uma revisão dos principais conceitos associados aos dados composicionais, definição, representação e operações, assim como, algumas das técnicas para modelagem de dados composicionais sem estrutura de dependência e alternativas para o caso com dependência espacial. O terceiro e o quarto capítulos descrevem os métodos e os resultados das análises nos dois casos (independente e dependência espacial). O quinto capítulo apresenta um estudo de caso da proporção de casos de hepatite virais no estado do Paraná, por faixa etária no período de 2007 a 2015, com o uso de equações de estimação generalizadas. No quinto e último capítulo constam as principais considerações a respeito das abordagens apresentadas.

# 2 REVISÃO BIBLIOGRÁFICA

## 2.1 Dados composicionais

Um conjunto de dados tem sido chamado composicional se apresenta porções de um total, tais como, porcentagens de trabalhadores em diferentes setores, partes dos elementos químicos em um mineral, concentração de diferentes tipos de células no sangue de um paciente, porções de espécies num ecossistema, concentração de nutrientes em uma bebida, porções de tempo de trabalho gasto em tarefas diferentes, porções de tipos de falhas, porcentagens de votos para partidos políticos, frações de partículas na textura de um solo, etc.

Nesses casos, o conjunto de dados só pode ser considerado composição se apresentar pelo menos dois componentes. Caso contrário, não há uma parte em um total. Isso implica em uma diferença substancial entre os dados de composição e outros conjuntos de dados multivariados. A maioria das análises multivariadas começa com uma análise univariada das variáveis individuais, ao passo que, em várias situações a análise dos componentes individuais pode não ter significado isolada do resto.

As primeiras recomendações relacionadas com a análise estatística dos dados composicionais, remete a um artigo de Karl Pearson de 1897 sobre correlações espúrias. O artigo aponta problemas decorrentes do uso de métodos estatísticos tradicionais em proporções, como partes de um todo. Mas suas advertências foram ignoradas até por volta de 1960, quando o geólogo Felix Chayes (1960) também alertou contra a aplicação da análise multivariada padrão para dados composicionais, a fim de evitar inconsistências pela restrição de soma unitária.

Foi somente a partir de 1980 que John Aitchison, estatístico escocês e professor na Universidade de Hong Kong, sistematizou a teoria para os dados composicionais. Apresentou um espaço amostral adequado às suas restrições, com distribuição de probabilidade própria e as possíveis tranformações para o espaço amostral real. Assim, segundo Aitchison (1986), um conjunto de $n$ vetores $\mathbf{y}_i$, $i = 1, \ldots, n$, cujos elementos apresentam as restrições $y_{i1} > 0, y_{i2} > 0, \ldots, y_{iD} > 0$ e $y_{i1} + y_{i2} + \ldots + y_{iD} = 1$, em que $D$ é o número de componentes (partes) do vetor. Esses conjuntos de vetores são definidos como dados composicionais e apresentam variabilidade de vetor para vetor.

Num conjunto de dados composicionais, cada vetor é denominado uma composição e cada elemento do vetor é chamado de componente. Por razões de notação, ao longo do trabalho, os conjuntos de dados composicionais (ou matrizes) são representados por letras maiúsculas em negrito ($\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{W}$, etc); as composições (ou vetores) são descritas por letras minúsculas em negrito ($\mathbf{y}$, $\mathbf{z}$, $\mathbf{w}$, etc).

Os dados composicionais podem ser descritos por subcomposições, na forma de razão de um dos componentes ou uma função destes (AITCHISON, 1986). Uma composição com $D$ componentes, $\mathbf{y} = (y_1, \ldots, y_d, y_D)$, em que, $D = d + 1$, está completamente

determinada por uma subcomposição com $d$ componentes, ou seja,

$$\mathbf{y} = (y_1, \ldots, y_d)$$

pois, se $y_1 + \ldots + y_d + y_{d+1} = 1$ (restrição de dados composicionais), logo:

$$y_{d+1} = 1 - y_1 - \ldots - y_d.$$

Deste modo, fazendo $d = D - 1$, tem-se $y_D = 1 - y_1 - \ldots - y_d$. Uma composição $D$-particional é um vetor com dimensão $d$ e só pode ser representado em algum conjunto $d$-dimensional conveniente, essa representação diminui a dimensão do espaço amostral.

Uma outra maneira de determinar uma composição é baseada em razões dos componentes. Seja uma composição $\mathbf{y}_i = y_{i1}, \ldots, y_{iD}$, os componentes podem ser apresentados da seguinte forma:

$$r_i = \frac{y_i}{y_D}, \qquad \text{com} \qquad D = 1 + d \qquad \text{e} \qquad i = 1, \ldots, d, \text{ logo:}$$

$$y_i = \frac{r_i}{1 + r_i + \ldots + r_d} \qquad \text{e} \qquad y_D = \frac{1}{1 + r_i + \ldots + r_d}.$$

Segundo Aitchison e Greenacre (2002), uma outra forma de representar os dados composicionais é em termos da média geométrica dos componentes:

$$i = 1, \ldots, D \text{ e } s_i = \frac{y_i}{g(\mathbf{y})}, \text{ com } g(\mathbf{y}) = \sqrt[D]{y_1 \times y_2 \times \ldots \times y_D}$$

em que $g(\mathbf{y})$ é a média geométrica dos componentes.

Para representar os dados composicionais, o espaço amostral natural que aborda as restrições impostas é o Simplex, denotado por $\mathbb{S}^D$. Seguindo a mesma notação, qualquer vetor no simplex é chamado composição, cada elemento desse vetor é chamado de componente e qualquer coleção de tais vetores, dados composicionais.

## 2.2 Representação gráfica

Um vetor aleatório $\mathbf{y}$ com valores em $\mathbb{S}^D$ é chamado de vetor aleatório simplex, ou uma composição aleatória, ou vetor de proporções contínuas (Aichison, 1986; Pawlowsky-Glahn e Buccianti, 2011). O simplex $\mathbb{S}^D$ é definido por:

$$\mathbb{S}^D = \{(y_1, y_2, \ldots, y_D) \in \mathbb{R}^D : y_i > 0 \text{ para todo } i = 1, \ldots, D, \text{ e } y_1 + \ldots + y_D = 1\}.$$

em que, $\mathbb{R}^D$ denota o espaço real $D$-dimensional e $\mathbb{S}^D$ o simplex $D$-dimensional.

O simplex é um espaço amostral natural para os dados composicionais, porém provou ser um espaço amostral complicado para ser tratado estatisticamente devido a

ausência de classes paramétricas satisfatórias (PAWLOWSKY-GLAHN e BUCCIANTI, 2011).

Buchanan *et al.*, (2012) definem o espaço simplex como uma representação geométrica do espaço de atributos, onde uma composição de $D$ partes é representada por um número mínimo de vértices correspondente ao número de dimensões.

O diagrama ternário, triângulo cujos vértices representam os três componentes da composição é uma técnica gráfica adotada para representação de dados composicionais. O diagrama ternário pode ser construído com altura unitária (Figura 2.1) ou com lado unitário (Figura 2.2). Um mesmo ponto, pode ocupar posições diferentes no diagrama ternário de acordo com a configuração escolhida. Isso pode ser observado pela composição $\mathbf{x}$ com 3 componentes, $\mathbf{x} = (0,2; 0,2; 0,6)$ representado nas figuras 2.1 e 2.2.

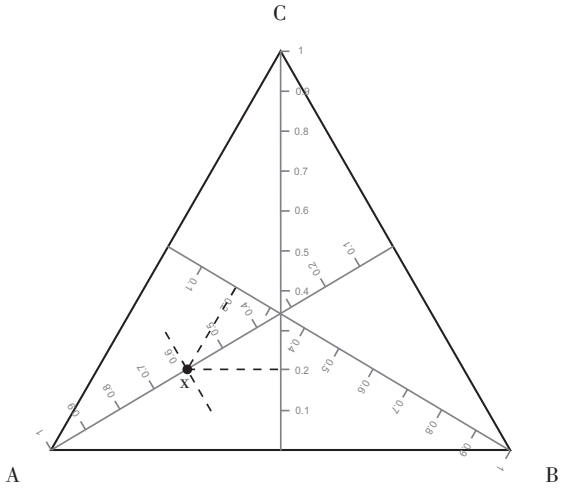



**Figura 2.1.** Composição $\mathbf{x} = (0,2; 0,2; 0,6)$ em $\mathbb{S}^3$ representado no triângulo de altura unitária.

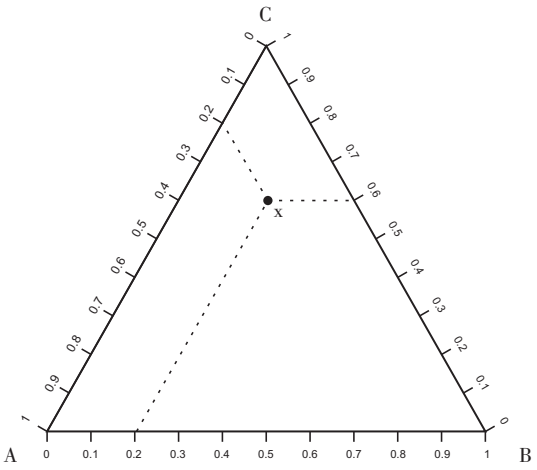



**Figura 2.2.** Composição $\mathbf{x} = (0,2; 0,2; 0,6)$ em $\mathbb{S}^3$ representado no triângulo de lado unitário.

Conforme aumenta o número de componentes a visualização fica comprometida devido ao aumento no número de dimensões no espaço simplex. Os conjuntos de dados composicionais podem ser representados de diversas formas: diagramas de dispersão com dois componentes, diagramas ternários (apresentados anteriormente) aplicado para três componentes, diagramas de dispersão em escala log para os componentes, ou sequências de barras, porém, as duas primeiras formas são as mais comuns (BOOGAART e TOLOSANA-DELGADO, 2013). A descrição de outras técnicas gráficas podem ser encontradas, por exemplo, em Aitchison e Egozcue (2005).

Ao utilizar o simplex como espaço amostral natural para os dados composicionais, para um conjunto de dados composicionais $\mathbf{Y}$ com $D$ componentes e $n$ linhas, cada linha é um vetor com $D$ partes, representado por um ponto em $\mathbb{S}^D$. Assim, o conjunto de dados $\mathbf{Y}$ é representado por $n$ pontos em $\mathbb{S}^D$.

Na representação por diagramas de dispersão para dois componentes, cada componente $D_i$ com $i = 1, \ldots, D$ é um vetor com $n$ elementos positivos e pode ser representado com um ponto em $\mathbb{R}^n_+$ então o conjunto $\mathbf{Y}$ pode ser representado por $D$ pontos em $\mathbb{R}^n_+$.

## 2.3 Operações

Para garantir que um conjunto de dados composicionais $\mathbf{Y} \in \mathbb{S}^D$ seja um corpo com operações bem estruturadas, deve-se garantir que, ao operar dois elementos desse conjunto, os resultados da operação continuam pertencendo ao espaço amostral $\mathbb{S}^D$.

O operador de restrição $\zeta$ transforma cada vetor de dados $\mathbf{x}$ com $D$ elementos, todos positivos, em um vetor de soma unitária. Em outras palavras, transforma um vetor de dados reais positivos em uma composição $\mathbf{s}$ do espaço amostral simplex. O operador define a seguinte transformação:

$$\zeta : \mathbb{R}^D \to \mathbb{S}^D$$

$$\zeta(x_1, x_2, \ldots, x_D) = (s_1, s_2, \ldots, s_D)$$

de tal forma:

$$\frac{(x_1, x_2, \ldots, x_D)}{(x_1 + x_2 + \ldots + x_D)} = (s_1, s_2, \ldots, s_D).$$

Por exemplo:

$\mathbf{x} = (20, 50, 10) \in R^3 \to \zeta(20, 50, 10) = \dfrac{20, 50, 10}{20 + 50 + 10} = (0,250, 0,625, 0,125) = \mathbf{s} \in \mathbb{S}^3.$

### 2.3.1 Perturbação

Seja $\mathbf{y}$ uma composição com $D$-componentes e $\mathbf{u}$ um vetor com dimensão $D$ e elementos reais positivos. Então, a operação

$$\mathbf{w} = \mathbf{u} \oplus \mathbf{y} = \zeta(x_1 u_1, \ldots, x_D u_D) = \frac{(x_1 u_1, \ldots, x_D u_D)}{(x_1 u_1 + \ldots + x_D u_D)}$$

é denominada perturbação com a composição original $\mathbf{y}$ sendo operada pelo vetor perturbador $\mathbf{u}$ para formar a composição perturbada $\mathbf{w}$.

Essa operação é equivalente a adição, ou seja, a operação entre um vetor de dados composicionais $\mathbf{y}$ e um vetor de números reais positivos $\mathbf{u}$ de mesma dimensão, o resultado continua sendo um elemento do espaço amostral simplex, com soma unitária. Por exemplo:

$$\mathbf{y} = (0,3; 0,5; 0,2) \in \mathbb{S}^3 \text{ e } \mathbf{u} = (1, 4, 7) \in \mathbb{R}^3$$

$$\mathbf{y} \oplus \mathbf{u} = \frac{(0,3 \times 1); (0,5 \times 4); (0,2 \times 7)}{0,3 \times 1 + 0,5 \times 4 + 0,2 \times 7} = \frac{(0,3; 2; 1,4)}{0,3 + 2 + 1,4} = (0,081; 0,541; 0,378)$$

### 2.3.2 Potência

Segundo Boogaart e Tolosana-Delgado (2013), a operação potência é equivalente a multiplicação por "escalar", e é definida da seguinte forma:

$$\mathbf{y} = \lambda \odot \mathbf{x} = \zeta({x_1}^{\lambda}, \ldots, {x_D}^{\lambda})$$

com $\lambda \in \mathbb{R}$.

O resultado dessa operação será uma composição pertencente ao Simplex. Por exemplo, seja a constante $\lambda = 3 \in \mathbb{R}$ e $\mathbf{y} = (0,40; 0,15; 0,25, 0,20) \in \mathbb{S}^4$:

$$\lambda \odot \mathbf{y} = \zeta(0,40^3; 0,15^3; 0,25^3; 0,20^3)$$

$$= \frac{0,4^3; 0,15^3; 0,25^3; 0,20^3}{0,4^3 + 0,15^3 + 0,25^3 + 0,20^3}$$

$$= (0,703; 0,037; 0,172; 0,088).$$

## 2.4 Estatística descritiva

### 2.4.1 Média composicional

Para obter o valor médio de um conjunto de dados composicionais, é usada a operação perturbação e a operação potência, que indica o centro ou média composicional de um conjunto $\mathbf{Y}$ com $n$ composições e $D$ componentes é a composição:

$$\overline{\mathbf{Y}} = \frac{1}{n} \odot \zeta \bigoplus_{i=1}^{n} \mathbf{y}_i$$

em que $\mathbf{y}_i$ é $i$-ésima composição de $\mathbf{Y}$.

Para ilustrar o cálculo da média composicional, considere as composições $\mathbf{y_1} = (0,2; 0,2; 0,6)$ e $\mathbf{y_2} = (0,4; 0,3; 0,3)$, o centro será obtido por:

$$\overline{\mathbf{Y}} = \frac{1}{n} \odot \zeta \bigoplus_{i=1}^{n} \mathbf{y}_i = \frac{1}{2} \odot \zeta \left( \frac{(0,2 \times 0,4); (0,2 \times 0,3); (0,6 \times 0,3)}{(0,2 \times 0,4) + (0,2 \times 0,3) + (0,6 \times 0,3)} \right)$$

$$= \frac{1}{2} \odot (0,25; 0,188; 0,562) = \frac{0,25^{1/2}; 0,188^{1/2}; 0,562^{1/2}}{0,25^{1/2} + 0,188^{1/2} + 0,562^{1/2}}$$

$$= (0,297; 0,257; 0,446)$$

### 2.4.2 Variância composicional

Existem várias medidas de variabilidade para dados composicionais, mas como medida global de variação, dentre elas, pode ser usada a variância métrica (PAWLOWSKY-GLAHN e EGOZCUE, 2001), também conhecida como variância total ou variância generalizada,

$$\text{mvar}(\mathbf{Y}) = \frac{1}{n-1}\sum_{i=1}^{n} d^2(\mathbf{y}_i, \overline{\mathbf{Y}})$$

isto é, a distância quadrática média das composições, $y_i$, ao centro do conjunto de dados $\overline{\mathbf{Y}}$, com graus de liberdade corrigidos como acontece com a variância convencional.

Em dados reais, é difícil interpretar quantitativamente as variâncias, e tende-se a trabalhar com desvios-padrão. Segundo Boogaart e Tolosana-Delgado (2013), não há uma definição direta de um desvio padrão da composição, optou-se, portanto, definir um desvio padrão métrico, para apoiar a interpretação

$$\text{msd}(\mathbf{Y}) = \sqrt{\frac{1}{D-1}\text{mvar}(\mathbf{Y})}, \text{ para } D \text{ componentes.}$$

Se a variação for a mesma em todas as direções, o msd pode ser interpretado como o desvio padrão radial em uma escala logarítmica. Quando a variância não é a mesma em todas as direções, ainda assim, o desvio padrão métrico pode ser entendido como algum tipo de variação média.

### 2.4.3 Matriz variância

A variância métrica não contém nenhuma informação sobre a codependência dos componentes. De acordo com Boogaart e Tolosana-Delgado (2013) e Chayes (1975), não se pode descrever esta correspondência por correlações ou covariâncias brutas devido aos efeitos espúrios do fechamento unitário. Em vez disso, Aitchison (1986) recomenda a matriz variância. Essa matriz $D \times D$ tem os elementos definidos da seguinte forma:

$$\tau_{ij} = \text{var}\left(\ln\frac{y_i}{y_j}\right)$$

e estimada por

$$\hat{\tau}_{ij} = \frac{1}{n-1}\sum_{i=1}^{n} \ln^2\frac{y_i}{y_j} - \ln^2\frac{\bar{y}_i}{\bar{y}_j}.$$

A matriz variância é uma matriz simétrica, dado que $\ln(a/b) = -\ln(b/a)$ e $\text{var}(-c) = \text{var}(c)$. Como $\tau_{ij} = \tau_{ji}$, há uma pequena variância do $\ln(y_i/y_j)$, isso indica uma "boa proporcionalidade" $y_i \propto y_j$. Quanto menor for o elemento de variação, melhor será a proporcionalidade entre os dois componentes. Para ajudar na interpretação, Aitchision (1986) sugere interpretá-la como um coeficiente de correlação.

## 2.5 Modelagem para dados composicionais independentes

De forma geral, a regressão linear visa propor e estimar um modelo linear a partir de dados (variável resposta) que dependem linearmente de uma ou mais covariáveis. No caso dos dados composicionais, as composições podem surgir como variáveis dependentes ou independentes em modelos lineares. A maioria dos métodos para modelos lineares clássicos tem um análogo próximo para os modelos lineares composicionais.

Para contemplar todas as restrições e particularidades dos dados composicionais algumas estratégias para a modelagem estatística foram implementadas. De início, as técnicas propostas para modelagem consideram apenas dados composicionais independentes. Nesses casos, a ideia é abordar o dado composicional com a estrutura de correlação entre os componentes, devido a restrição de soma. Porém, cada composição do conjunto de dados composicional não possui correlação com as demais composições e depende apenas das covariáveis associadas ao modelo.

Este capítulo revisita as principais abordagens para modelos lineares composicionais independentes, em que a variável resposta é composicional. Existem situações, em que as covariáveis do modelo são composicionais. Exemplos podem ser encontrados nos trabalhos de Pawlowsky-Glahn et al. (2015), Boogaart e Tolosana-Delgado (2013) e Ark (2005).

Supondo uma variável composicional no simplex $\mathbb{S}^D$, denotada por $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$, em que cada composição $\mathbf{y}_i, i = 1, 2, ..., n$, está associada a uma ou mais variáveis externas ou covariáveis $\mathbf{x}_i = [x_{i0}, x_{i1}, \ldots, x_{ir}]$, onde $x_{i0} = 1$ por convenção. Assim, temos um problema de regressão linear composicional cujo objetivo é estimar os coeficientes $\boldsymbol{\beta}_j$ de uma curva ou superfície em $\mathbb{S}^D$. Os coeficientes de composição do modelo, $\boldsymbol{\beta}_j \in \mathbb{S}^D$, são estimados a partir dos dados.

Em geral, para regressão linear, o método mais usado é o método dos mínimos quadrados. Como a resposta $\mathbf{y}_i$ é composicional, é natural medir os desvios também no simplex usando os conceitos para essa geometria. Além disso, os dados composicionais não devem ser diretamente associados a uma distribuição normal de probabilidade, assim, uma das tentativas para modelar dados composicionais foi usar a distribuição de Dirichlet (CONNOR e MOSIMANN, 1969) que é a distribuição natural para o espaço simplex.

### 2.5.1 Regressão Dirichlet

Uma classe conhecida de distribuições em $\mathbb{S}^D$ é a classe Dirichlet com parâmetros $\boldsymbol{\alpha} \in \mathbb{R}_+^D$ que tem função densidade dada por:

$$f(\mathbf{Y}, \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_D)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_D)} \mathbf{y}_1^{\alpha_1 - 1} \cdots \mathbf{y}_D^{\alpha_D - 1} = \frac{\Gamma(\sum_{i=1}^D \alpha_i)}{\prod_{i=1}^D \Gamma(\alpha_i)} \prod_{i=1}^D \mathbf{y}_i^{\boldsymbol{\alpha}_i - 1}$$

em que $\Gamma(\cdot)$ é a função Gama e $\mathbf{y}_i = (y_1, \ldots, y_D) \in \mathbb{S}^D$ são vetores composicionais com $D$ componentes.

A distribuição Dirichlet foi apresentada por Connor e Mosimann (1969) para modelar dados em proporções e independentes. Quando $D$ =2, a distribuição Dirichlet se reduz à distribuição beta, no intervalo $[0, 1]$, ou em $\mathbb{S}^2$.

A distribuição Dirichlet é uma classe bem estruturada no simplex. Inicialmente houve uma grande utilização dessa classe e de possíveis generalizações que são consideradas até hoje para modelagem de dados composicionais. No entanto, a distribuição de Dirichlet tem algumas propriedades muito restritivas, como a independência composicional completa. Isso torna improvável modelar qualquer estrutura de dependência entre as composições usando esta distribuição. Além disso, a estrutura de correlação de uma composição Dirichlet possui a seguinte forma:

$$\text{cov}(y_i, y_j) = \frac{-\alpha_i \alpha_j}{\alpha_+^2(\alpha_+ + 1)} \quad \text{e} \quad \text{corr}(y_i, y_j) = -(\alpha_i \alpha_j)^{1/2}[(\alpha_+ - \alpha_i)(\alpha_+ - \alpha_j)]^{-1/2}, \text{ para } i \neq j$$

$\alpha_+ = \sum_{i=1}^{D} \alpha_i$. Na prática, essa estrutura se torna inadequada para descrição da variabilidade de dados composicionais, pois sua estrutura de correlação induz a correlação negativa, sendo inapropriada aos dados com correlação positiva, o que geralmente ocorre em dados composicionais (AITCHISON, 1982).

Poucos modelos trabalham diretamente com as composições, pelo fato da distribuição Dirichlet ser a única distribuição para dados composicionais definida diretamente em $\mathbb{S}^D$. Connor e Mosimann (1969) propuseram originalmente a distribuição de Dirichlet como um modelo para dados composicionais. Outros exemplos de trabalhos com distribuição Dirichlet para dados composicionais independentes, podem ser encontrados em Hijazi e Jernigan (2007) e Camargo (2011).

Uma alternativa aos entraves dos modelos Dirichlet é aplicar aos dados certas transformações que saem do espaço amostral simplex $\mathbb{S}^D$ para o espaço euclidiano. Aitchison (1985) sugere análise com as distribuições de Dirichlet e Log-normal. Porém afirma, que o potencial para testar Dirichlet contra distribuições log-normais dentro da mesma classe seja diminuído na fronteira do espaço paramétrico. Trabalhos de Brehm et al. (1998) e Buccianti (2012) fazem comparações entre modelos Dirichlet e transformações para o espaço dos números reais.

### 2.5.2 Regressão com variáveis transformadas

As possibilidades para as análises estatísticas de dados de composição seriam consideravelmente estendidas se, além de uma estrutura de covariância adequada, houvesse uma classe paramétrica de distribuições em $\mathbb{S}^D$ rica o suficiente para capturar os padrões de variabilidade observados no espaço amostral simplex. Conforme apresentado

anteriormente, a classe familiar de Dirichlet e suas generalizações até o momento não são suficientes para analisar os dados composicionais. Aitchison e Shen (1980) propõem encontrar meios adequados de descrever padrões de variabilidade composicional no conjunto $\mathbb{R}_+$.

A ideia de induzir uma classe de distribuições bem estabelecida sobre um espaço amostral complicado é uma ideia antiga. McAlister (1879), ao tentar descrever padrões de variabilidade sobre os reais positivos, considerou $y \in \mathbb{R}$ com uma distribuição $N(\mu, \sigma^2)$, para induzir uma distribuição $\Delta(\mu, \sigma^2)$ em $\mathbb{R}_+$ e expressou suas ideias em termos de inversa. Ou seja, $y = \log w$, com $w \in \mathbb{R}_+$, a essa nova classe de distribuição atribuiu-se o nome de Log-normal.

Aitchison e Shen (1980) propuseram uma maneira de analisar os dados composicinais, no caso de amostras independentes, com o uso de transformações que saem do espaço amostral simplex $\mathbb{S}^D$ e passam ao espaço bem definido $\mathbb{R}^D$, com o uso de transformações adequadas entre $\mathbb{S}^D$ e $\mathbb{R}^D$. Dentre as transformações viáveis, as transformações em razão de logarítmos apresenta um bom desempenho, porém com limitações em valores extremos do intervalo (0 ou 1). A transformação apresentada por Aitchison e Shen (1980) provém do log da razão entre os componentes. Essa transformação foi denominada por razão log aditiva, do inglês *adittive log ratio* (*alr*), e é descrita da seguinte forma:

$$\mathbf{Y} = \{\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{iD}\} \in \mathbb{S}^D \xrightarrow{alr} \left\{ \log\left(\frac{y_{i1}}{y_{iD}}\right), \log\left(\frac{y_{i2}}{y_{iD}}\right), \ldots, \log\left(\frac{y_{i(D-1)}}{y_{iD}}\right) \right\} \in \mathbb{R}^{D-1}.$$

Aitchison (1986) sugere o uso de transformações para descrever padrões de variabilidade dos dados composicionais. Assim torna-se possível aproximar os dados a uma distribuição de probabilidades normal multivariada e usar todas as ferramentas desenvolvidas para essa distribuição. Isso pode simplificar a análise de dados e permitir sua aplicação em uma variedade de problemas com dados composicionais, além de diminuir uma dimensão na escala dos dados.

Há diversas transformações que levam do espaço simplex $\mathbb{S}^D$ para $\mathbb{R}^{D-1}$, entretanto, poucas transformações possuem a forma inversa simples.

A transformação *alr* exige que um componente seja escolhido como categoria de referência. Assim, podem surgir problemas com as composições extremas (quando um ou mais componentes são muito próximos de zero), pois, os valores em razão log tendem ao infinito. Detalhes e demonstração desta afirmação podem ser encontrados em Aitchison (1986, capítulo 5).

Conforme descrito em Aitchison (1986) a análise dos dados composicionais com o uso de transformações pode ser realizada baseando-se nos seguintes passos:

i) formular o problema em termos dos componentes da composição;

ii) transformar os dados composicionais em log das razões (com a escolha adequada da classe de referência);

iii) analisar os dados transformados por meio de uma análise estatística multivariada;

iv) aplicar a transformação inversa para os termos das composições obtidas em iii).

Análises com o uso de transformações em logarítmos da razão entre componentes podem ser encontradas numa série de outros trabalhos de Aitchison (1982, 1985, 1986), Aitchison e Bacon-Shone (1984), Aitchison e Shen (1980) e aparecem em diversos contextos para análise de dados composicionais. No entanto, há técnicas alternativas que não necessitam diretamente de transformações nos dados.

Com isso, surgem outras formas para análise de dados composicionais, tais como a abordagem bayesiana para dados composicionais independentes apresentada nos trabalhos de Iyengar e Dey (1996) e Obage (2005). Ou alternativas que não especificam uma distribuição de probabilidade aos dados, por exemplo, as equações generalizadas.

## 2.6 Modelagem para dados composicionais com dependência espacial

Uma maneira simplificada para análise estatística é abordar as composições de forma independente. Porém, há momentos em que a suposição de independência não pode ser satisfeita. Por exemplo, quando as composições são obtidas ao longo do espaço ou medidas repetidas ao longo do tempo, deve-se considerar a dependência espacial ou temporal entre as composições. Nesses casos, as abordagens estatísticas propostas anteriormente precisam ser estendidas.

Estudos geológicos comumente apresentam dados em forma de composições (textura do solo, composição química de rochas, etc). Se adicionar a suposição de que as localizações amostrais (posição geográfica das composições) são relevantes para o estudo, tem-se, um exemplo de dados composicionais com dependência espacial. Tradicionalmente, as técnicas geoestatísticas usadas quando há uma evidente estrutura de correlação espacial entre os dados, buscam descrever a estrutura de dependência espacial por meio de uma função de correlação conhecida.

Baseado nos trabalhos de Clark (1979), Matheron (1971), Cressie (1991) e Diggle et al. (1998) que descrevem conceitos referentes a análise geoestatística, e nas teorias de Aitchison (1982, 1985 e 1986) para dados composicionais, surgem abordagens para análise geoestatística de dados composicionais ou análise de dados composicionais com dependência espacial, também conhecidos como composição regionalizada.

O conceito de composição regionalizada deriva da teoria de variáveis regionalizadas desenvolvida por Matheron (1971), aliada às particularidades dos conjuntos de dados composicionais. As abordagens teóricas relativas a composições regionalizadas são encon-

tradas principalmente nos trabalhos de Pawlowsky e Burger (1992), Olea et al. (1993), Pawlosky-Glahn e Olea (2004) e Martins et al. (2016).

### 2.6.1 Modelo geoestatístico

Segundo Tolosana-Delgado (2005), geoestatística é um conjunto de técnicas e ferramentas para analisar conjuntos de dados regionalizados, que são conjuntos em que cada amostra foi recolhida num local conhecido do espaço (isto é, inclui coordenadas geográficas para cada amostra). Assim, as amostras apresentam uma dependência mútua, proveniente da proximidade espacial entre os locais de amostragem.

De forma geral, seguindo notação e definições apresentadas em Tolosana-Delgado et al. (2011), conjuntos de dados regionalizados são modelados com o conceito de que um vetor aleatório $\mathbf{Z}$ amostrado nas localizações $\mathbf{x}$, com $\mathbf{x} \in \mathbb{R}^2$ (grid de amostragem), assume estacionaridade de segunda ordem, expressa por:

$$E(\mathbf{Z}(\mathbf{x})) = \boldsymbol{\mu}(\mathbf{x}) = \boldsymbol{\mu} \text{ e } \mathrm{Cov}[\mathbf{Z}(x_1), \mathbf{Z}(x_2)] = \mathbf{C}|x_1 - x_2|$$

isto é, o valor esperado é constante e independente da localização. A matriz de covariância, denotada por $\mathbf{C}|x_1 - x_2|$ é uma função aleatória que depende apenas da distância entre os pontos $x_1$ e $x_2$. Desta hipótese, surge a estacionariedade intrísica,

$$E[\mathbf{Z}(x_1) - \mathbf{Z}(x_2)] = \mathbf{0} \text{ e } \mathrm{Var}[\mathbf{Z}(x_1) - \mathbf{Z}(x_2)] = \boldsymbol{\Gamma}|x_1 - x_2|$$

em que, a diferença da função aleatória em dois pontos é zero e a variância da diferença depende apenas do distância entre eles. As funções $\mathbf{C}(\cdot)$ e $\boldsymbol{\Gamma}(\cdot)$ são, respectivamente, chamadas de função de covariância (multivariada) e variograma. Nos casos em que ambas existem, estão relacionadas através de

$$\Gamma(\mathbf{h}) = 2\mathbf{C}(\mathbf{0}) - [\mathbf{C}(\mathbf{h}) + \mathbf{C}(-\mathbf{h})]$$

em que, $\mathbf{C}(\mathbf{h})$ é a covariância para pontos separados por distâncias $\mathbf{h}$.

Os termos diagonais são chamados autovariâncias ou variogramas direto e mostram a continuidade espacial de uma dada variável. Os termos fora da diagonal são chamados de covariância cruzada ou variogramas cruzados e explicam como variáveis tomadas em dois locais diferentes estão relacionadas. A maneira típica de trabalhar com essas funções é estimar versões empíricas dos variogramas diretos e cruzados e, ajustar-lhes um modelo linear de corregionalização que descreva a continuidade espacial conforme aumenta a distância entre os pontos amostrais (TOLOSANA DELGADO, 2005).

Uma vez que o modelo espacial ideal é escolhido, pode-se usá-lo para interpolar e estimar a variável aleatória em um local não amostrado, isso é chamado de krigagem.

A krigagem é um método de interpolação que produz os melhores estimadores e preditores lineares não tendenciosos. O estimador de krigagem ordinária $\hat{\mathbf{z}}_0$ em um local não amostrado $x_0$ é uma combinação linear dos dados $\{\mathbf{z}_i = \mathbf{z}(x_i), i = 1, \ldots, n\}$,

$$\hat{\mathbf{z}}_0 = \sum_{i=1}^{n} \mathbf{\Lambda}_i \mathbf{z}_i.$$

Os pesos $\mathbf{\Lambda}_i$ são matrizes da mesma dimensão que $\mathbf{C}(\cdot)$ ou $\mathbf{\Gamma}(\cdot)$ que mostram a influência da amostra $\mathbf{z}_i$ na previsão $\mathbf{Z}(x_0)$. Estes pesos são encontrados resolvendo o sistema de krigagem,

$$\mathbf{\Lambda} = \mathbf{S}^{-1}\mathbf{S}_0,$$

em que todas as matrizes são definidas por blocos, tomando $\mathbf{\Lambda}_{ij} = \mathbf{\Lambda}(x_i - x_j)$:

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \vdots \\ \boldsymbol{\lambda}_n \\ \nu \end{bmatrix}, \mathbf{S} = \begin{bmatrix} \mathbf{\Gamma}_{11} & \ldots & \mathbf{\Gamma}_{1n} & \mathbf{I} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{\Gamma}_{n1} & \ldots & \mathbf{\Gamma}_{nn} & \mathbf{I} \\ \mathbf{I} & \ldots & \mathbf{I} & \mathbf{0} \end{bmatrix}, \mathbf{S}_0 = \begin{bmatrix} \mathbf{\Gamma}_{10} \\ \vdots \\ \mathbf{\Gamma}_{n0} \\ \mathbf{I} \end{bmatrix}.$$

Os mesmos resultados são obtidos se considerar os blocos $\mathbf{\Gamma}_{ij} = \mathbf{C}|x_i - x_j|$, usando covariâncias em vez de variogramas.

De acordo com Tolosana Delgado (2005) essas técnicas não podem ser aplicadas diretamente a composições regionalizadas, pois, a análise estatística deve considerar que as composições transmitem apenas informações relativas. Qualquer incremento ou decremento absoluto nas composições são passíveis de gerar correlação espúria entre os componentes, a correlação espacial não é uma exceção.

Pawloswsky-Glahn e Olea (2004) relatam que problemas com correlação espacial espúria e singularidade da matriz de covariância estão relacionados à suposição básica de que o espaço amostral é irrestrito e à suposição de que a distribuição do erro de estimação é gaussiano. Isto é devido ao fato de que a soma de proporções ou percentagens é fixa e, portanto, $\mathbf{\Gamma}(h)$ ou $\mathbf{C}(h)$ são singulares em qualquer distância $h$.

Ao analisar os componentes individuais, surge a singularidade da matriz de covariância de uma composição, isso é uma conseqüência direta da restrição de soma constante. Numericamente, o problema pode ser resolvido deixando um componente fora para evitar a singularidade das matrizes. No entanto, a soma dos valores estimados difere frequentemente da constante correspondente, embora os próprios valores individuais possam ser razoáveis (PAWLOWSKY-GLAHN e OLEA, 2004).

Além disso, de acordo com resultados de Tolosana-Delgado et al. (2011) a krigagem separada de cada componente não garante que os resultados interpolados sejam positivos, ou somam até 1 (ou 100%) e pode fornecer somas maiores do que o máximo permitido. Por outro lado, a cokrigagem (método em que diversas variáveis regionalizadas

podem ser estimadas em conjunto, com base na correlação espacial entre si), preserva a soma total, mas não evita interpolações negativas.

Boezio et al. (2012) alertam para a utilização de cokrigagem aplicada direto aos dados composicionais completos (sem tranformação) ou obter as predições por diferenças. A determinação de uma variável por diferença fornece estimativas negativas que devem ser substituídas por valores válidos.

Como os problemas de geoestatística com composições são similares aos problemas da estatística clássica, as possíveis soluções para dados composicionais seguem os mesmos princípios, porém, com as modificações necessárias para "driblar" todas as particularidades dos dados composicionais.

### 2.6.2 Modelagem conjunta com ajuste em variograma

Na análise geoestatística dos componentes individuais, fica implícita a suposição de que a distribuição do erro de estimação em cada ponto da região de amostragem é gaussiano. Um modelo gaussiano requer uma amostra sem restrições no espaço amostral. Esse fato já inviabiliza o uso da distribuição gaussiana para dados composicionais. A análise dos componentes individuais não leva em consideração as restrições e utilizam alternativas que forçam o fechamento das somas (BOOGAART e TOLOSANA-DELGADO, 2013).

Autores como Pawlowsky-Glahn e colaboradores produzem desde meados da década de 80, vasta literatura para dados composicionais, incluindo também os casos com dependência espacial (PAWLOWSKY, 1984; PAWLOWSKY e BURGUER, 1992; PAWLOWSKY-GLAHN e EGOZCUE, 2001; 2015).

A ideia apresentada por Pawlowsky-Glahn e Olea (2004) e Tolosana-Delgado (2005) é a construção de variogramas baseados nas razões logarítmicas dos componentes e, pelo método de mínimos quadrados realizar o ajuste de modelos de correlação espacial. Após a escolha do modelo espacial, as predições são obtidas por krigagem ordinária. A krigagem ordinária é o método que fornece a melhor interpolação linear para uma localização não amostrada, com base no variograma conhecido.

Trabalhos dos mesmos autores apontam que a análise com ajuste baseado nos variogramas, permite investigar as dependências espaciais sem a distorção de correlações espaciais espúrias. Com isso, obtêm-se resultados melhores quando comparados aos resultados obtidos por cokrigagem dos dados originais, com estimativas no intervalo original das amostras, com somas fechadas e sem necessidade de pós-processamento.

### 2.6.3 Modelagem conjunta baseada em verossimilhança

Aitchison (1986) apresenta a teoria de dados composicionais considerando a independência das observações, enquanto Pawlowsky-Glahn e Olea (2004), Tolosana-Delgado

(2005) fazem análise geoestatıstica de dados composicionais com o ajuste de modelos baseado nos variogramas da razão logarítmica entre os componentes.

Para estudos de dados composicionais no contexto geoestatístico, com $D = 3$ componentes, os trabalhos de Martins (2009) e Martins et al. (2016) fazem uso da transformação *alr*. Com isso, especifica um modelo geoestatístico bivariado paramétrico para os dados composicionais transformados, supondo termos latentes comuns, caracterizados por uma função de correlação que induz à estrutura espacial. Neste caso, o método de inferência utilizado baseia-se na função de verossimilhança de uma distribuição normal multivariada.

Essa abordagem apresenta uma maneira de sair das restrições impostas para os dados composicionais, conforme recomendações de Aitchison (1986). Avalia-se o log das razões dos componentes, ao invés de avaliar estatisticamente a própria composição. O uso de tranformações assume um novo espaço amostral, geralmente o espaço real $\mathbb{R}^{D-1}$. Assim, a vantagem desse método, principalmente por usar a transformação *alr*, é que as variáveis transformadas assumem distribuição normal multivariada.

Essa vantagem permite o uso das ferramentas estatísticas desenvolvidas para a distribuição normal. O trabalho de Martins et al. (2009) apresenta detalhadamente a metodologia para análise geoestatística de dados composicionais, envolvendo a definição do modelo para composições com três componentes; estimação dos parâmetros via verossimilhança e a predição espacial.

A desvantagem desse método é a necessidade de pós processamento dos dados. Exige a transformação para a escala original dos dados e isso envolve operações inversas com matrizes de alta dimensão ou simulações, o que gera um alto custo computacional.

## 2.7 Equações de estimação generalizadas (EEG)

Os modelos de regressão possuem uso limitado em dados correlacionados devido a suposição de independência entre os indíviduos. Mesmo os modelos lineares generalizados (MLG) que são mais flexíveis, se forem utilizados para dados correlacionados, é provável a obtenção de distorções nas estimativas dos parâmetros e de seus erros padrões, gerando inferências inadequadas (PETER e SONG, 2007).

O método de equações de estimação generalizadas (EEG) foi proposto por Zeger e Liang (1986) com o objetivo de estimar parâmetros de regressão, especialmente quando os dados não possuem distribuição normal de probabilidade e são correlacionados. As equações de estimação generalizadas são uma extensão dos MLG's. Foram desenvolvidas para produzir estimativas mais eficientes e não viciadas para os parâmetros do modelo de regressão com dados correlacionados. Isso só é possível, porque, o método considera a estrutura de correlação entre as observações.

Em alguns casos pode haver uma estrutura natural de correlação entre os sujeitos,

devido a medidas repetidas (dados longitudinais), indivíduos em um mesmo grupo (dados agrupados), dados geoestatísticos com dependência espacial ou dados composicionais com correlação entre os componentes.

De acordo com Peter e Song (2007), o método EEG deve ser utilizado quando o objetivo da análise estatística é descrever a esperança matemática da variável resposta, em função de um conjunto de covariáveis considerando a correlação entre as observações. Assim, a esperança matemática da variável resposta pode ser especificada como uma função linear das covariáveis e a variância como uma função conhecida da média. Esse método é a base do trabalho de Bonat e Jorgensen (2016) com modelos lineares generalizados para diversas variáveis respostas multivariadas e com potenciais aplicações para dados composicionais.

### 2.7.1 Formulação geral do método EEG

A formulação apresentada tem o objetivo de descrever em linhas gerais os passos para a aplicação do método EEG. Esses conceitos constam no trabalho de Verbeke e Molenberghs (2000). A formulação específica para os dados composicionais será apresentada no capítulo 3. Para escrever as equações de estimação generalizadas, supõe-se que:

i) a relação entre a média da variável resposta $\boldsymbol{\mu}_i$ e as variáveis explicativas $\mathbf{X}$, pode ser expressa sob forma linear, por uma função de ligação conhecida $g$. Esta função é tal que:

$$g(\boldsymbol{\mu}_i) = X_i\boldsymbol{\beta}$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros.

ii) a variância da variável resposta pode ser expressa por uma função conhecida da média desta variável. As relação entre a média da variável resposta e as variáveis explicativas e, entre a variância e a média da variável resposta, são definidas como em um modelo linear generalizado.

Para utilizar as EEG para dados correlacionados, Zeger e Liang (1986) sugerem especificar uma matriz de correlação,também conhecida como matriz de trabalho, incorporada no termo de variância da equação. É importante ressaltar que nas EEGs, apesar das observações pertencentes a um mesmo grupo, ou a uma composição, serem correlacionadas, supõe-se que observações em grupos diferentes são independentes. Para dados composicionais independentes essa observação não gera problemas, porém, para dados composicionais com dependência espacial é necessário usar a matriz de trabalho para especificar a correlação em função da distância dos pontos amostrais.

As EEGs desenvolvidas por Liang e Zeger (1986) visam analisar dados com medidas repetidas utilizando MLG's e apresentam resultados robustos, mesmo diante de uma má especificação da matriz de correlação (para um tamanho amostral suficiente-

mente grande). Pode-se dizer que a sua popularidade deve-se ao fato de combinar, de forma simples e flexível, a modelagem do valor médio da variável resposta em função das covariáveis.

Prentice e Zhao (1991) utilizaram EEGs para obter estimativas consistentes dos parâmetros de regressão e de correlação para dados multivariados contínuos e discretos. Já Hanley et al. (2003) apresentam exemplos de aplicações do método na área de epidemiologia, em conjuntos de dados reais com respostas binárias ou quantitativas.

Song e Tang (2000) apresentaram o método EEG para dados univariados proporcionais contínuos e longitudinais assumindo um parâmetro de dispersão constante. Mais tarde, Song et al. (2004) estenderam o modelo, para que a heterogeneidade do parâmetro de dispersão possa ser avaliado e contabilizado para conduzir uma melhor estimativa dos parâmetros do modelo.

Para dados composicionais, Zhang (2012) descreve equações de estimação em um modelo multivariado simplex longitudinal. Também faz comparação entre os modelos simplex e os modelos multivariados log-normais e conclui que os modelos multivariados simplex, superam as estimativas dos modelos log-normais multivariados.

No contexto espacial, Bishop et al. (2000) apresentaram o método EEG para investigar o impacto da tecnologia no desempenho dos navios em uma pesca de arrasto. Busca explicar as correlações espaciais e temporais nos dados de esforço de captura. Para modelar o componente espacial, os autores usaram estrutura da matriz de correlação permutável ou autorregressiva (AR1).

Baseado nos resultados e considerações apresentadas nos trabalhos anteriores para o método EEG, surge a ideia do uso de equações de estimação para dados composicionais com dependência espacial, com função de correlação baseada na distância entre os pontos de amostragem.

# 3 MATERIAL E MÉTODOS

## 3.1 Material

Para apresentar e ilustrar os métodos de modelagem com dados composicionais independentes foi utilizado um conjunto de dados formado por três componentes na presença de uma covariável, apresentado originalmente por Coakley e Rust (1968) e revisitado porAitchison (1986).

O conjunto de dados trata da composição dos sedimentos do lago Stanwell-Fletcher situado na Ilha Somerset, no arquipélago Ártico no Canadá. Esse conjunto é conhecido na literatura como sedimentos do Lago Ártico e é composto pelas frações de areia, silte e argila para 39 profundidades, conforme apresentado na Tabela 3.1:

**Tabela 3.1.** Composição textural dos sedimentos do Lago Ártico (em %), para cada profundidade

| | Areia $y_1$ | Silte $y_2$ | Argila $y_3$ | Profundidade $x$ (m) |
|---|---|---|---|---|
| 1 | 77,5 | 19,5 | 3,0 | 10,4 |
| 2 | 71,9 | 24,9 | 3,2 | 11,7 |
| 3 | 50,7 | 36,1 | 13,2 | 12,8 |
| 4 | 52,2 | 40,9 | 6,6 | 13,0 |
| 5 | 70,0 | 26,5 | 3,5 | 15,7 |
| 6 | 66,5 | 32,2 | 1,3 | 16,3 |
| 7 | 43,1 | 55,3 | 1,6 | 18,0 |
| 8 | 53,4 | 36,8 | 9,8 | 18,7 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

A tabela completa com os dados encontra-se no Anexo A.

São considerados dados composicionais independentes quando a composição na profundidade $x_i$ é independente da composição na profundidade $x_j$.

Para as análises e exposição dos métodos para dados composicionais com dependência espacial, foi utilizado o conjunto de dados apresentado e descrito nos trabalhos de Gonçalves (1997) e Martins (2009). Esse conjunto de dados apresenta a composição textural do solo em uma área irrigada por sistema pivô central na Fazenda Areão, pertencente ao campus da Escola Superior de Agricultura - “Luiz de Queiroz” - ESALQ - USP, na cidade de Piracicaba - SP (Figura 3.1).


FONTE: Gonçalves (1997); Martins (2009).

**Figura 3.1.** Foto aérea do campo experimental de irrigação da ESALQ-USP com área de estudo correspondente ao quadrante irrigado por um sistema pivô central.

Nessa área foi demarcado um quadrante na porção mais elevada (topo da encosta), no qual foram obtidas 82 amostras de solo na profundidade entre 0 e 0,2 m em uma malha regular quadrada de amostragem, de lado igual a 20 metros. Em cada amostra, foram medidos os valores das frações granulométricas ou composição textural de areia, silte e argila. A tabela completa com os dados encontra-se no Anexo B.

## 3.2 Métodos: Dados composicionais independentes

As análises para os dados composicionais foram realizadas no software `R` versão 3.3.1 (R CORE TEAM, 2016). No `R` um dos pacotes específicos para os dados composicionais, com funções para descrição, representação e modelagem dos dados é o pacote `compositions` (Boogaart et al., 2014).

O pacote `compositions` apresenta os conceitos básicos dos dados composicionais, operações, estatísticas descritivas e modelos lineares para composições; também apresenta as diversas escalas e geometrias de composição, além disso, permite selecionar a geometria correta para o problema em questão.

Esta seção apresenta a descrição dos métodos usados para ajuste de modelos de regressão para dados composicionais independentes, os métodos são apresentados seguindo a ordem descrita no capítulo de revisão. Os códigos fonte com as funções usadas nas análises, estão disponíveis no Anexo C.

### 3.2.1 Modelo de regressão de Dirichlet

Seja $\mathbf{Y}_i = \{\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{iD}\}$ o conjunto de dados composicionais com $y_{ij} > 0$ e $\sum_{j=1}^{D} \mathbf{y}_{ij} = 1$ e seja $\mathbf{x}_i$ a covariável correspondente observada para $i = 1, \ldots, n$ e $j = 1, \ldots, D$, $D$ componentes. A distribuição condicional de $\mathbf{Y}_i|\mathbf{x}_i$ é uma distribuição Dirichlet com parâmetros desconhecidos, $\boldsymbol{\alpha}(x_i) = (\alpha_1(x_i), \ldots, \alpha_D(x_i))$. A função verossimilhança é dada por:

$$L(\mathbf{Y}|\boldsymbol{\alpha}) = \prod_{i=1}^{n} \left\{ \Gamma\left(\sum_{j=1}^{D} \alpha_j(x_i)\right) \prod_{j=1}^{D} \frac{y_{ij}^{\alpha_j(x_i)-1}}{\Gamma(\alpha_j(x_i))} \right\}.$$

Assim, o modelo de regressão de Dirichlet com vetor de parâmetros $\boldsymbol{\alpha}$ depende da covariável $\mathbf{X}$ por meio de uma função polinomial, conforme descrito em Hijazi e Jernigan (2007). Os parâmetros são descritos de forma reparametrizada: $\alpha_{ij} = \alpha_j(x) = \sum_{k=0}^{p} \beta_{jk} x_i^k$ e $\sum_{j=1}^{D} \alpha_{ij} = \sum_{k=0}^{p} \beta_k x_i^k$ em que $\beta_k = \sum_{j=1}^{D} \beta_{jk}$, com $k = 1, \ldots, p$ dimensão do modelo de regressão.

Neste caso, $\mathbf{Y} = \{\mathbf{y}_{i1}, \mathbf{y}_{i2}, \mathbf{y}_{i3}\} = \{\text{areia}_i, \text{silte}_i, \text{argila}_i\}$, com $i = 1, \ldots, 39$ vetores, com 3 componentes. A covariável associada ao modelo é a profundidade de amostragem $\mathbf{x}_i$. Os parâmetros de regressão serão estimados maximizando a função de verossimilhança Dirichlet.

Para análise de dados supondo uma regressão de Dirichlet foram utilizadas as funções do pacote `DirichletReg` (MAIER, 2015) que permite representações e o ajuste de modelos de regressão no simplex, usando a distribuição Dirichlet para amostras independentes.

### 3.2.2 Modelo de regressão com transformação

Aitchison (1986) propõe o uso de transformações convenientes que saem do espaço simplex e passam para o espaço euclidiano $\mathbb{R}^{D-1}$. A transformação apresentada por Aitchison e Shen (1980) provém do log da razão entre os componentes, razão log aditiva ($alr$), obtida por:

$$\mathbf{Y} = \{\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{iD}\} \in \mathbb{S}^D \xrightarrow{alr} (\log(\mathbf{y}_{i1}/\mathbf{y}_{iD}), \log(\mathbf{y}_{i2}/\mathbf{y}_{iD}), \ldots, \log(\mathbf{y}_{i(D-1)}/\mathbf{y}_{iD})) \in \mathbb{R}^{D-1}.$$

Essa abordagem consiste em transformar os dados do simplex $\mathbb{S}^D$ para $\mathbb{R}^{D-1}$, ajustar os modelos de regressão e retornar ao simplex por operações inversas. Para um conjunto de dados composicionais com três componentes (areia, silte e argila) em função de uma covariável $\mathbf{X}$ (profundidade), considerando o componente argila como referência, o vetor de médias é expresso por:

$$\boldsymbol{\mu}(x) = \begin{bmatrix} \log(y_{\text{areia}}/y_{\text{argila}})(x) & = \beta_{11} + \beta_{12}(x) + \ldots + \beta_{1j}(x^p) \\ \log(y_{\text{silte}}/y_{\text{argila}})(x) & = \beta_{21} + \beta_{22}(x) + \ldots + \beta_{2j}(x^p) \end{bmatrix} = \mathbf{X}^T\boldsymbol{\beta}, \mathbf{X} = (\mathbf{1}, \mathbf{x}, \ldots, \mathbf{x}^p)$$

$\boldsymbol{\mu}(x)$ representa o valor esperado de uma distribuição normal multivariada $N^d(\boldsymbol{\mu}(x), \boldsymbol{\Sigma})$ em que $\boldsymbol{\Sigma}$ não depende de $\mathbf{X}$. Os parâmetros da distribuição são obtidos pelo método de máxima verossimilhança. A transformação *alr* das composições de sedimentos produz vetores bidimensionais no espaço real e assim pode-se considerar uma regressão multivariada (bivariada) em função da profundidade.

Para análise de dados composicionais com transformação *alr* foram utilizadas as funções do pacote `compositions` (Boogaart et al., 2014) disponível para o software `R` versão 3.3.1. O pacote possui outras transformações disponíveis e para todas elas, além de estimar os vetores de parâmetros, há funções que realizam a operação inversa para o espaço simplex.

### 3.2.3 Modelo de regressão com equações generalizadas independentes

As equações apresentadas nessa seção foram obtidas de acordo com os trabalhos de Peter e Song (2007), Song et al. (2004) e Zhang (2012). Seja $\mathbf{Y_i}$ uma resposta composicional, com as seguintes restrições: $\mathbf{Y}_i = [y_{i1}, y_{i2}, \ldots, y_{iD}]$, $y_{ij} > 0$, $i = 1, \ldots, n$ observações independentes, $j = 1, \ldots, D$ ($D$ componentes) e $[y_{i1} + y_{i2} + \ldots + y_{iD}] = 1$.

Seja $\mathbf{X}(n \times p)$ a matriz de covariáveis. Assim, $\boldsymbol{\mu}_i$ segue um modelo linear generalizado dado por:

$$\boldsymbol{\eta}_i = g(\boldsymbol{\mu}_i) = \mathbf{X}^T\boldsymbol{\beta}$$

em que $\boldsymbol{\beta}$ é a matriz de parâmetros do modelo de regressão e $g$, a função de ligação de $\mathbb{S}^D \rightarrow \mathbb{R}^{D-1}$. A função de ligação $g$ deve ser escolhida para satisfazer a restrição $0 < \sum_{i=1}^{n} \mu_{ij} \leq 1$. Assumindo que a função de ligação logit multivariada tem a seguinte forma:

$$\boldsymbol{\eta}_i = g(\boldsymbol{\mu}_i) = \left( \log\left(\frac{\mu_{i1}}{\mu_{iD}}\right), \ldots, \log\left(\frac{\mu_{i(D-1)}}{\mu_{iD}}\right) \right). \tag{3.1}$$

$$\mu_{ij} = \frac{\exp\{\eta_{ij}\}}{1 + \sum_{j=1}^{D} \exp\{\eta_{ij}\}}, \tag{3.2}$$

A matriz $\boldsymbol{\beta}$ será estimada baseada nos efeitos das covariáveis no valor médio populacional. Seja $u_{ij} = \dfrac{\partial l_{ij}}{\partial \beta}$ vetor escore para a observação $i$ e componente $j$, $\mathbf{u}_i = (\mathbf{u}_{i1}^T, \ldots, \mathbf{u}_{iD}^T)^T$.

Peter e Song (2007) demonstram que $E(\boldsymbol{\mu}_i) = 0$ e $\mathbf{V}_i = \text{Var}(\mathbf{u}_i)$. A matriz $\mathbf{V}_i$ é decomposta como:

$$\mathbf{V}_i = \text{diag}\{\text{Var}(u_{ik})\}^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\text{diag}\{\text{Var}(u_{ik})\}^{1/2} \tag{3.3}$$

em que $\text{diag}\{\text{Var}(u_{ik})\}$ denota a matriz $p \times p$ diagonal com elementos especificados pelo vetor $\mathbf{v}_i = (\text{Var}(u_{ij1}), \ldots, \text{Var}(u_{ijp}))^T$ e $\mathbf{R}_i(\boldsymbol{\alpha})$ é a matriz de correlação (matriz de trabalho) de $u_{ij}$.

Segundo resultados de Song et al. (2004) a equação generalizada que maximiza a informação de Godambe:

$$\Psi(\beta, \alpha) = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{u}_i = 0 \tag{3.4}$$

com $\mathbf{D}_i^T = \mathbf{X}_i \text{diag}\left\{\dfrac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i}\right\} \text{diag}\{\text{Var}(u_{ij})\}$ e $\mathbf{X}_i = (\mathbf{X}_{i1}^T, \mathbf{X}_{i2}^T, \ldots, \mathbf{X}_{ip}^T)^T$.

Em que, $\text{diag}\left\{\dfrac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i}\right\}$ e $\text{Var}(\mathbf{u}_i)$ são matrizes blocos diagonal com as diagonais compostas por $\left\{\dfrac{\partial \mu_{ij}}{\partial \eta_{ij}}\right\}$ e $\text{Var}(u_{ij})$, $j = 1, \ldots, D$, respectivamente.

A estimativa $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ é definida como a solução da equação 3.4. Segundo resultados de Zhang (2012), o parâmetros de correlação $\boldsymbol{\alpha}$ pode ser estimado como, $r_{ij} = \dfrac{u_{ij}}{\sqrt{\text{Var}(u_{ij})}}$.

As análises de dados composicionais com equações generalizadas independentes, foram realizadas com auxílio das funções dos pacotes `geeM` (McDANIEL e HENDERSON, 2016) e `mmm` (ASAR e ILK, 2014) disponíveis para o software `R` versão 3.3.1 (R CORE TEAM, 2016).

### 3.2.4 Comparação entre os modelos ajustados

Os ajustes dos modelos para dados composicionais independentes, foram comparados pela soma de quadrados dos resíduos (SQR) para cada componente. A escolha do método de ajuste mais adequado será pelo menor valor de SQR. A SQR foi calculada da seguinte forma:

$$\text{SQR}_j = \sum_{i=1}^{n} (\hat{y}_{ij} - y_{ij})^2$$

em que: $\hat{y}_{ij}$ é o valor obtido pelo modelo de regressão (predito) para a $i$-ésima observação no $j$-ésimo componente; $y_{ij}$ é o valor observado na $i$-ésima observação e $j$-ésimo componente. Sendo, $j = 1, \ldots, D$ componentes e, $i = 1, \ldots, n$ composições.

Para o cálculo da SQR, os valores preditos pelo modelo e os valores observados, estão na escala original dos dados. Nos modelos em que foram aplicadas transformações, a SQR foi obtida após as operações inversas.

## 3.3 Métodos: Dados composicionais com dependência espacial

### 3.3.1 Análise geoestatística para os componentes individuais

Na área do campo experimental de irrigação da ESALQ, foram coletadas 82 amostras de solo em uma área com malha amostral de espaçamento regular, com distâncias entre os pontos variando de 20 a 256 metros. As amostras foram coletadas na profundidade de 0 a 0,2 m. Para cada amostra foram obtidos os percentuais dos componentes areia, silte e argila.

Com os dados de textura e as coordenadas geográficas dos pontos, foram construídos os semivariogramas experimentais para avaliar tendência direcional e indícios de dependência espacial. Para cada componente, assumindo distribuição normal de probabilidades, foram ajustados o modelo espacial exponencial e os parâmetros obtidos pelo método da máxima verossimilhança (MV).

Com as estimativas dos parâmetros do modelo espacial, foi realizada a krigagem ordinária e construção dos mapas temáticos para avaliar a distribuição espacial de cada componente do solo. As análises foram realizadas com auxílio do pacote `geoR` (RIBEIRO JUNIOR e DIGLLE, 2016).

Todos os códigos fonte com as funções usadas para as análises de dados composicionais com dependência espacial, estão disponíveis no Anexo D.

### 3.3.2 Modelagem conjunta com ajuste baseado em variograma

Nessa abordagem a dependência espacial é analisada e modelada por descrições de momento de segunda ordem, como funções de covariância ou variogramas. O variograma é a variância das diferenças dos valores $\mathbf{Y}(x_i)$ e $\mathbf{Y}(x_j)$ nos locais $x_i$, $x_j$ separados por uma distância $h$, isto é, para $h = ||x_i - x_j||$, tem-se:

$$2\gamma(h) = \text{Var}(\mathbf{Y}(x_i) - \mathbf{Y}(x_j)).$$

Assim, assume-se que, o valor esperado é constante e não depende da localização. A variância da diferença depende apenas da distância $h$ entre os pontos. Este conceito é chamado de estacionariedade intrínseca e garante que o variograma seja bem definido. Mais detalhes podem ser obtidos em Cressie (1991).

Segundo Pawlowsky-Glahn e Olea (2004) e Tolosana Delgado (2005) os variogramas de todos os pares de razões logarítmicas dos componentes ($alr$) podem ser estimados a partir dos dados, calculando a média do quadrado das diferenças de pares de observações em diferentes classes de distância:

$$\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{(i,j)\in N(h)} (\mathbf{Y}(x_i) - \mathbf{Y}(x_j))^2$$

em que $N(h)$ denota o conjunto de pares de observações com distância aproximada por:

$$N(h) = \{(i,j) : ||x_i - x_j|| \approx h\}.$$

Assim, modelos espaciais válidos para variogramas devem ser ajustados baseado apenas na configuração do variograma empírico $\hat{\gamma}(h)$.

Como o ajuste do variograma é feito pelo método dos mínimos quadrados, o ajuste deve ser tipicamente inspecionado visualmente. Uma vez obtido um modelo de variograma satisfatório, faz-se interpolações por krigagem ordinária. Os resultados são exibidos em mapas temáticos. Para obter os resultados para esse método foram usadas as funções disponíveis no pacote `compositions` (BOOGAART et al., 2014).

### 3.3.3 Modelagem conjunta baseada em verossimilhança

Seguindo metodologia proposta por Martins et al. (2009), Martins et al. (2016) e, de acordo com os princípios de Aitchison (1986), foi usada a transformação razão log aditiva (*alr*). O vetor $\mathbf{Y}$ no espaço amostral simplex de dimensão $D$, é transformado em um vetor do espaço real $\mathbb{R}^{D-1}$, nas localizações $\{x_1, \ldots, x_i, \ldots, x_n\}$ , de tal forma que:

$$\text{alr}(\mathbf{Y}) = \left(\log\left(\frac{y_{i1}}{y_{iD}}\right), \ldots, \log\left(\frac{y_{i(D-1)}}{y_{iD}}\right)\right).$$

O modelo geoestatístico bivariado composicional:

$$\begin{cases} \mathbf{Y}_1(\mathbf{x}_i) = \boldsymbol{\mu}_1(\mathbf{x}_i) + \mathbf{S}_1(\mathbf{x}_i) + \mathbf{Z}_1(\mathbf{x}_i) \\ \mathbf{Y}_2(\mathbf{x}_{i'}) = \boldsymbol{\mu}_2(\mathbf{x}_{i'}) + \mathbf{S}_2(\mathbf{x}_{i'}) + \mathbf{Z}_2(\mathbf{x}_{i'}) \end{cases}$$

em que, $\mathbf{Y}_1 = \log(\text{areia/argila})$, $\mathbf{Y}_2 = \log(\text{silte/argila})$, $i, i^{'} = 1, \ldots, 82$, $\mathbf{S}_j(\mathbf{x}_i) \sim N(0, \sigma_j^2)$ e $\mathbf{Z}_j(\mathbf{x}_i) \sim N(0, \tau_j^2), j = 1, \ldots, D$.

Os termos $\mathbf{S}_j$ modelam os efeitos espaciais e possuem uma função de correlação comum $\rho\mathbf{U}(\mathbf{x}, \phi)$. São escalonados por $\sigma_j$ definindo o termo comum padronizado $\mathbf{U}$ com vetor de médias iguais a zero e matriz de variâncias/covariâncias induzidas pela função de correlação exponencial. Por outro lado, os efeitos aleatórios $\mathbf{Z}_1$ e $\mathbf{Z}_2$ representam os efeitos composicionais do modelo. Desta forma o modelo pode ser reescrito como:

$$\begin{cases} \mathbf{Y}_1(\mathbf{x}_i) = \boldsymbol{\mu}_1(\mathbf{x}_i) + \sigma_1 U(\mathbf{x}_i, \phi) + \mathbf{Z}_1(\mathbf{x}_i) \\ \mathbf{Y}_2(\mathbf{x}_{i'}) = \boldsymbol{\mu}_2(\mathbf{x}_{i'}) + \sigma_2 U(\mathbf{x}_{i'}, \phi) + \mathbf{Z}_2(\mathbf{x}_{i'}) \end{cases}$$

os termos da matriz de covariâncias $\boldsymbol{\Sigma}$ são:

$$\text{Cov}(\mathbf{Y}_j(\mathbf{x}_i), \mathbf{Y}_j(\mathbf{x}_{i'})) = \sigma_j^2 + \tau_j^2 \rightarrow \text{Cov}(\mathbf{Y}_j(\mathbf{x}_i), \mathbf{Y}_j(\mathbf{x}_{i'})) = \sigma_j^2 \rho\mathbf{U}(x_i, x_{i'})$$

assim,

$$\text{Cov}(\mathbf{Y}_1(\mathbf{x}_i), \mathbf{Y}_2(\mathbf{x}_{i'})) = \sigma_1\sigma_2\mathbf{I}_2(i,i') + \tau_1 + \tau_2\mathbf{I}_3(i,i')$$

$$\mathbf{I}_2 = \begin{cases} 1, & \text{se } i = i' \\ \rho\mathbf{U}(x_i, x_{i'}) & \text{se } i \neq i' \end{cases} \qquad \mathbf{I}_3 = \begin{cases} \rho\mathbf{U}(x_i, x_{i'}) & \text{se } i = i' \\ 0, & \text{se } i = i'. \end{cases}$$

A inferência sobre o vetor de parâmetros $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)^T$ é feita usando a teoria da verossimilhança cuja função é

$$L(\mathbf{Y}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-0,5(\mathbf{Y} - \boldsymbol{\mu}_y)^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_y)\}.$$

Fazendo $\boldsymbol{\Sigma} = \sigma_1^2\mathbf{R} + \tau_1^2\mathbf{I}_b = \sigma_1^2\mathbf{V}$, em que $\mathbf{R}$ é uma matriz de covariâncias relacionada aos efeitos espaciais; $\mathbf{I}_b$ é a matriz bloco diagonal com elementos relacionados às covariâncias entre as composições com a função de log verossimilhança:

$$l(\boldsymbol{\theta}) = -0,5\left[n\log(2\pi) + 2n\log(\sigma_1) + \log(|\mathbf{V}|) + (1/\sigma_1^2)\mathbf{Qe}\right] \tag{3.5}$$

em que, $\widehat{\mathbf{Q}}\mathbf{e} = (\mathbf{Y} - \boldsymbol{\mu}_y)^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_y), \boldsymbol{\mu}_y = \mathbf{X}\mu$ e $\mathbf{X}$ é a matriz associada às médias das variáveis. Expressões analíticas fechadas podem ser obtidas para os estimadores de máxima verossimilhança de $\sigma_1$ e $\mu$:

$$\hat{\sigma}_1 = \sqrt{\mathbf{Qe}/n} \text{ e } \hat{\mu} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{V}^{-1}Y).$$

Substituindo estes estimadores na expressão 3.5 com a reparametrização $\eta = \sigma_2/\sigma_1$, $\nu_1 = \tau_1/\sigma_1$ e $\nu_2 = \tau_2/\sigma_1$, obtém-se a função de log verossimilhança concentrada

$$l(\boldsymbol{\theta}') = -0,5[\log(|\mathbf{V}|) + n\log(2\pi) + \log(\widehat{\mathbf{Q}}\mathbf{e}) - \log(n+1)],$$

em que $\boldsymbol{\theta}^T = (\eta, \nu_1, \nu_2, \phi, \rho)^T$ será maximizada numericamente.

Para a obtenção dos erros padrão na escala original dos parâmetros é utilizado o método delta por quadratura de Gauss Hermite. Os resultados apresentados foram implementados no software `R` e compactadas num pacote chamado `geoComp` (MARTINS et al., 2009), ainda não disponível no repositório do `R`.

Após as estimativas dos parâmetros e predição na escala original dos dados, foram construídos mapas temáticos para a área por meio de cokrigagem. Os valores preditos para o terceiro componente foram obtidos por diferença.

### 3.3.4 Modelagem por equações generalizadas com dependência espacial

Essa abordagem sugere o uso de equações de estimação generalizadas (EEG) visto que, para o seu uso não é necessário a suposição de distribuições aos dados. Basta especificar as equações que descrevam a média e a estrutura da matriz de variâncias e covariâncias. Com isso é possível, estimar os parâmetros de interesse, sem a construção da função de verossimilhança associada ao modelo. Deste modo, o intuito é desenvolver uma aproximação em que os momentos dos próprios dados são usados para modelar, sem a necessidade de transformações.

No contexto dos dados espaciais o valor médio da variável aleatória não depende da sua localização. Assim, as equações de estimação generalizada (EEG) para os parâmetros médios é representada pelas equações 3.1 e 3.2 definidas na seção 3.2.3.

Para descrever a função de variância no método EEG será usada a equação 3.3 descrita da seção 3.2.3, em que a matriz de trabalho $\mathbf{R}_i\alpha$ no caso com dependência espacial, será denominada $\mathbf{R}_i(\rho, \mathbf{D})$, com $i = 1, \ldots, n$, $\rho$ é o parâmetro de alcance e $\mathbf{D}$ a matriz de distâncias das composições.

Seja $\mathbf{R}_i(\rho, \mathbf{D})$ a matriz de correlação $n \times n$ para toda a amostra, e seja $\mathbf{R}_j(\rho, \mathbf{D}_j)$, com $j = 1, \ldots, D$ a matriz de correlação $D \times D$ para o componente $j$. Uma suposição comum do $ij$-ésimo elemento de $\mathbf{R}_i(\rho; \mathbf{D})$ é que

$$\mathbf{R}_{jj} = 1 - \gamma(d_{ij}, \rho), \text{ em que: } \gamma(d_{ij}, h) = \begin{cases} 0 & \text{se } d_{ij} = 0, \\ c + b[1 - exp(-d_{ij}/\rho)] & \text{caso contrário.} \end{cases}$$

O vetor de parâmetros espaciais $h = (c, b, \rho), c \geq 0, b \geq 0, \rho \geq 0,$ e $c + b \leq 2$ (ver Cressie (1993) para mais exemplos).

Se $b = c = 1$ sem perda de generalidade, então:

$$\mathbf{R}_{jj} = \begin{cases} 1 & \text{se } d_{ij} = 0, \\ exp(-d_{ij}/\rho) & \text{caso contrário.} \end{cases}$$

Outro exemplo seria,

$$\mathbf{R}_{jj} = \begin{cases} 1 & \text{se } d_{ij} = 0, \\ d_{ij}/\rho & \text{caso contrário.} \end{cases}$$

Embora a especificação acima não represente todas as possiblilidades, ela fornece pelo menos uma maneira de parametrizar a correlação espacial e fornece a base para testar a correlação espacial.

Para a implementação do método EEG para dados composicionais com dependência espacial, foram feitas adaptaçãos em funções disponíveis no pacote `JGEE` (INAN, 2015).

### 3.3.5 Método para comparação dos resultados

Sempre que são aplicados métodos diferentes, é necessário definir quais critérios serão utilizados para comparar os resultados finais. Na geoestatística univariada, é comum, a reutilização preditiva de amostras baseada na técnica de "sair-um-fora", conhecida como validação cruzada.

A validação cruzada, consiste na retirada do conjunto de dados, um ponto de cada vez. Remove-se o ponto e executa-se a interpolação para obter o valor predito do ponto removido a partir dos pontos restantes. Então, calcula-se a diferença (resíduo) entre o valor real do ponto removido e a estimativa para este ponto. Esse procedimento é repetido até que todos os pontos do conjunto tenham sido removidos (TOMCZAK, 1998). A medida utilizada para comparação será a soma absoluta dos resíduos da validação, denota por EA (erro absoluto), quanto menor o EA, melhor o ajuste.

Os métodos também foram comparados pelo quadrado médio dos resíduos (QMR) para cada componente, quanto menor os valor de QMR, melhor. Os valores de QMR foram obtidos da seguinte forma:

$$\mathrm{QMR}_j = \sum_{i=1}^{n} (\hat{\mathbf{y}}_{ij} - \mathbf{y}_{ij})^2 / n$$

em que: $\hat{\mathbf{y}}_{ij}$ é o valor predito pelo método para a $i$-ésima localização no $j$-ésimo componente e $\mathbf{y}_{ij}$ é o valor observado na $i$-ésima localização e $j$-ésimo componente, com $j = 1, \ldots, D$, total de componentes e, $i = 1, \ldots, n$ pontos amostrais.

# 4 RESULTADOS E DISCUSSÕES

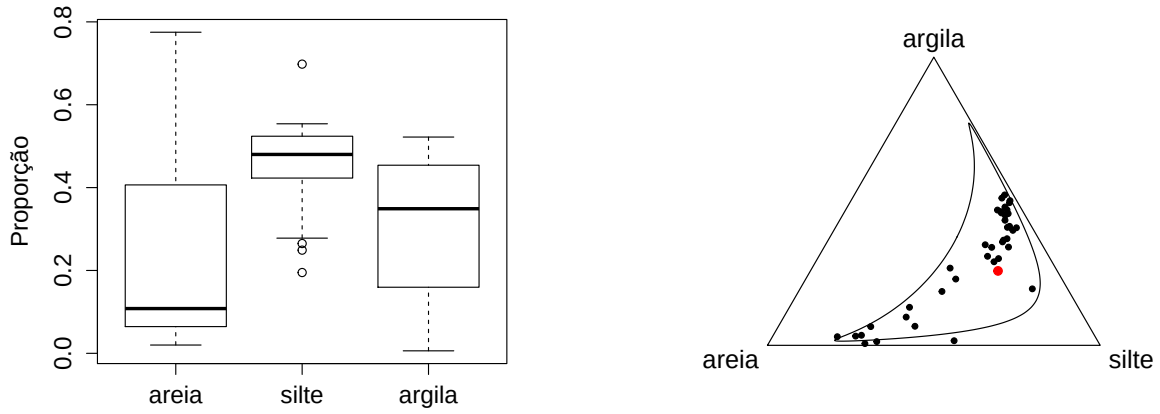
## 4.1 Dados independentes: Composição textural dos sedimentos do lago Ártico (Aitchison, 1986)

As três componentes, areia, silte e argila mensuradas nas 39 profundidades do lago, formam o conjunto de dados composicionais independentes. A seguir, são apresentadas as estatísticas descritivas (Tabela 4.1) da composição textural dos sedimentos do lago Ártico.

**Tabela 4.1.** Estatísticas descritivas das proporções de areia, silte e argila dos sedimentos do lago Ártico

| | Mínimo | Média | Máximo | Desvio padrão | Variância |
|---|---|---|---|---|---|
| Areia | 0,020 | 0,242 | 0,775 | 0,245 | 0,060 |
| Silte | 0,195 | 0,457 | 0,698 | 0,102 | 0,010 |
| Argila | 0,006 | 0,301 | 0,522 | 0,171 | 0,029 |

Nota-se que, em média, os sedimentos do lago apresentam uma maior proporção de silte. As proporções de areia, foram as que apresentaram maior desvio padrão. A variabilidade dos componentes, pode ser observada no gráfico boxplot da Figura 4.1. A assimetria dos componentes observada no boxplot, dá indícios de que os componentes não possuam distribuição normal de probabilidades.




(a) (b)

**Figura 4.1.** Gráfico boxplot para cada componente e diagrama ternário das composições.

Na Figura 4.1 (a), as proporções de silte apresentaram menor variabilidade, porém, com alguns pontos discrepantes. No diagrama ternário da Figura 4.1 (b), destaca-se a média composicional, ou centro ($\overline{\mathbf{Y}} = 0,178; 0,563; 0,258$) e observa-se que uma região de confiança de 2-desvios padrão, contempla a grande maioria das composições.

As proporções dos componentes (areia, silte e argila) dos sedimentos do Lago Ártico em cada profundidade são apresentadas na Figura 4.2. O interesse principal é investigar se há alguma dependência entre estas proporções e as profundidades para cada componente, ou seja, se é possível explicar as frações dos sedimentos em função da profundidade do lago.

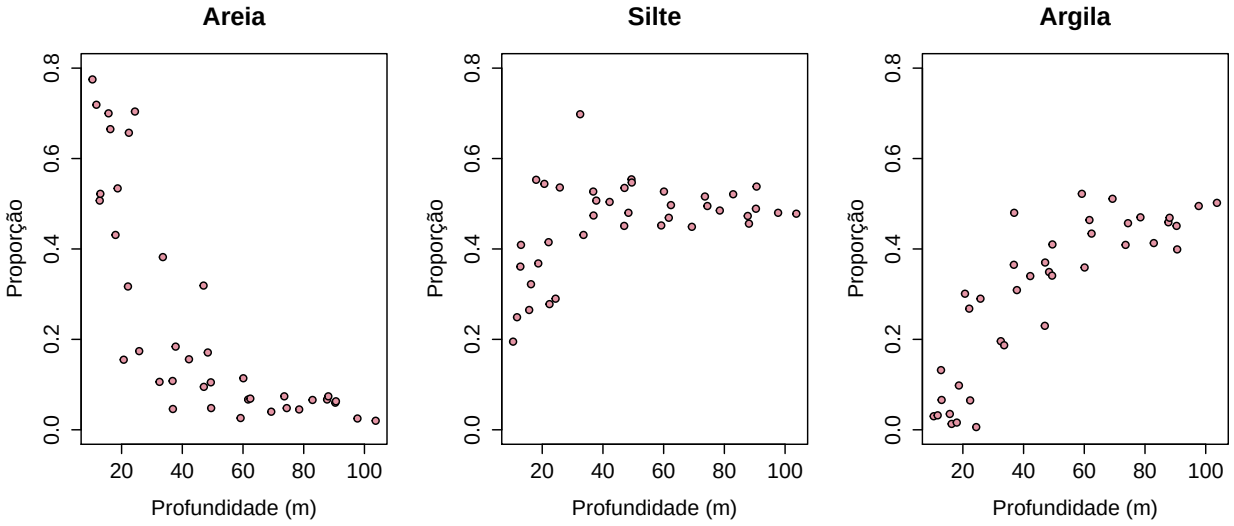



**Figura 4.2.** Proporções de cada composição em função da profundidade

Ao observar a Figura 4.2 torna-se evidente que há uma tendência decrescente nas proporções da areia conforme aumenta a profundidade. Para as proporções de silte e argila, percebe-se que estas seguem uma tendência crescente conforme aumenta a profundidades da água. Nesse caso, temos um problema de regressão, em que o vetor com as proporções dos componentes é a variável resposta e a profundidade é a variável explicativa ou covariável. O aspecto diferente do problema é a restrição de soma igual a 1 para os componentes do vetor da variável resposta. Com base nessa restrição, a primeira abordagem foi utilizar um modelo de regressão Dirichlet.

### 4.1.1 Modelo de regressão de Dirichlet

A distribuição Dirichlet por ser uma generalização da distribuição beta, com valores no intervalo (0,1) é a distribuição natural para os dados composicionais.

Para o ajuste dos modelos de regressão foi considerado que a variável composicional é formada pelos três componentes (areia, silte e argila), medidos em 39 amostras ao

longo da profundidade do lago. A principal vantagem da regressão Dirichlet é manter o espaço amostral original dos dados, o simplex e, não ser necessário aplicar mudanças de escala ou tranformações aos dados.

A Figura 4.3 apresenta os componentes com os modelos de regressão quadrática Dirichlet ajustados.

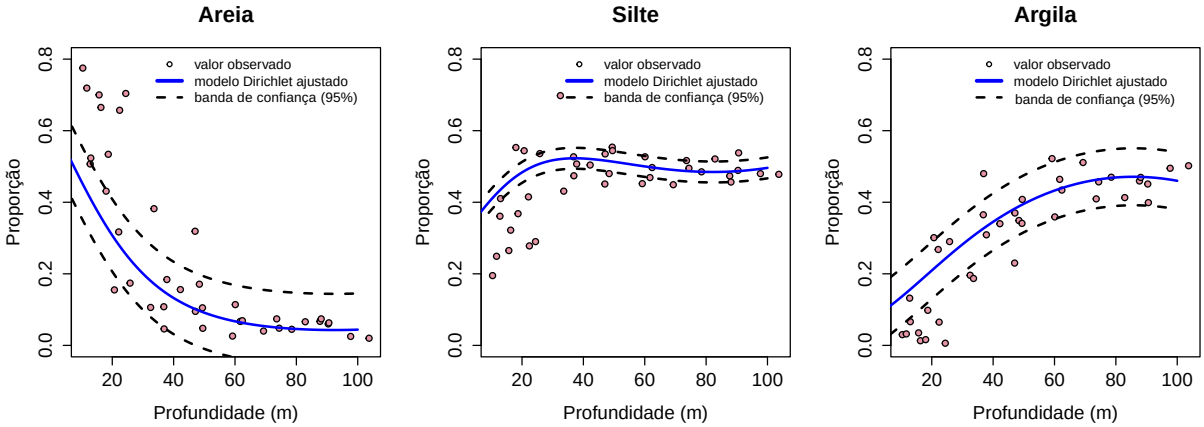



**Figura 4.3.** Modelos de regressão Dirichlet ajustados para cada componente

Foram testados os modelos linear simples e quadrático. O modelo quadrático foi o que obteve melhor ajuste, com os coeficientes apresentados na Tabela 4.2. As estimativas dos parâmetros do modelo foram obtidas pelo método de máxima verossimilhança (MV) para a função densidade de probabilidade Dirichlet.

**Tabela 4.2.** Estimativas dos parâmetros do modelo de Dirichlet quadrático estimados por MV

| | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | SQR* |
|---|---|---|---|---|
| Areia | 1,436 | $-0,007$ | $1,31 \times 10^{-4}$ | 0,538 |
| Silte | $-0,026$ | 0,072 | $-2,73 \times 10^{-4}$ | 0,206 |
| Argila | $-1,793$ | 0,111 | $-4,9 \times 10^{-4}$ | 0,232 |

* Soma de quadrado dos resíduos.

A soma dos quadrados dos resíduos foi escolhida para avaliar os resultados dos ajustes. Na Tabela 4.2 pode-se observar que o componente areia, foi o que apresentou maior SQR. Para os demais componentes, os valores obtidos foram menores e bem similares.

De forma geral, com a regressão Dirichlet foi possível fazer os ajustes para descrever as frações do solo em função da profundidade. Para efeitos comparativos, também foi realizada a análise de regressão transformando as variáveis composicionais em log da razão dos componentes, transformação ($alr$).

### 4.1.2 Modelo de regressão com transformação

A transformação em log das razões do componentes (alr) facilita a estimação dos parâmetros, pois, aproveita os recursos desenvolvidos para a distribuição normal. Além disso, há uma redução na dimensão dos dados; conjuntos composicionais com três componentes, após a transformação passam a dois "novos" componentes transformados. De tal forma que:

$$(areia, silte, argila) = \log\left(\frac{areia}{argila}\right), \log\left(\frac{silte}{argila}\right).$$

Para obter as estimativas dos parâmetros foi usada uma distribuição normal bivariada. A Figura 4.4 apresenta o ajuste dos modelo de regressão alr.

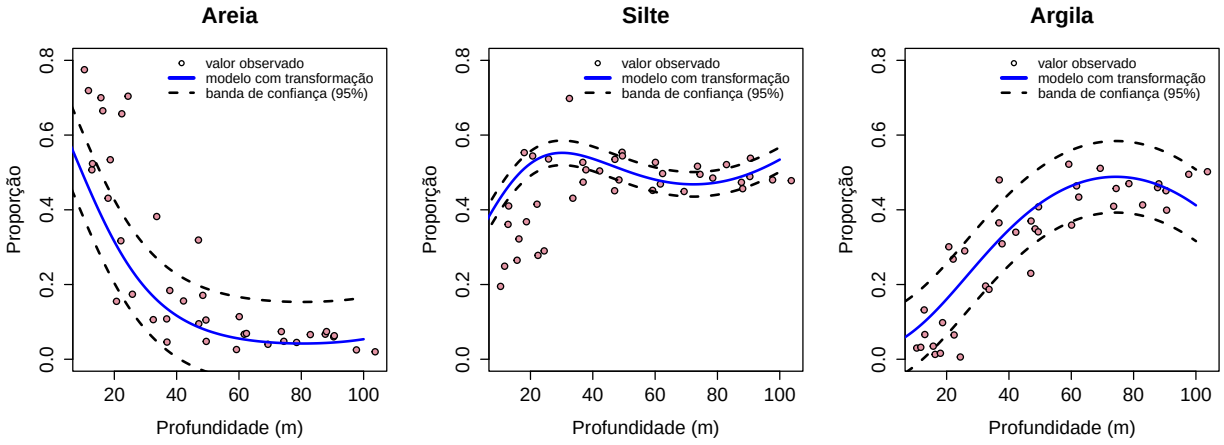



**Figura 4.4.** Modelos quadráticos ajustados com transformação *alr* para cada componentes

Os coeficientes do modelo de regressão alr (Tabela 4.3) foram obtidos pelo método de máxima verossimilhança. Após o ajuste com os dados transformados, foi aplicada a operação inversa para voltar na escala original. Os valores na escala original para o terceiro componente são obtidos por diferença a partir dos valores preditos para os demais componentes. O inconveniente do uso da tranformação é que os dados não são invariantes à transformação, ou seja, a escolha da classe de referência gera diferentes resultados.

**Tabela 4.3.** Estimativas do parâmetros do modelo de regressão alr quadrático estimados por MV

| | $\beta_0$ | $\beta_1$ | $\beta_2$ | SQR* |
|---|---|---|---|---|
| log(Areia) | 4,671 | $-0,178$ | 0,001 | 0,506 |
| log(Silte) | 2,941 | $-0,076$ | $4.86 \times 10^{-4}$ | 0,229 |
| Argila | - | - | - | 0,258 |

* Soma de quadrado dos resíduos.

Os valores de SQR para o modelo com transformação alr (Tabela 4.3) são muito próximos aos obtidos no modelo de Dirichlet, isso indica que os dois modelos tiveram desempenho similar para esse conjunto de dados. Porém, Hijazi e Jernigan (2007) ao analisarem dados com a distribuição Dirichlet e comparar com a regressão em razão log aditiva, afirmam que, em geral, o modelo em razões log superestima a curvatura dos dados, enquanto o modelo de Dirichlet o subestima. Para o conjunto de dados aqui apresentado, essa afirmação fica mais evidente apenas para o componente silte (Figura 4.4).

### 4.1.3 Modelo de regressão com equações generalizadas

Para estimar os parâmetros de regressão, as equações generalizadas foram construídas supondo função de ligação *logit*, similar a especificação de um GLM e a estrutura de covariância para dados independentes foi a matriz identidade. A Figura 4.5 apresenta os modelos ajustados pelo método de equações generalizadas (EEG) e os coeficientes do modelo de regressão seguem na Tabela 4.4.

**Tabela 4.4.** Estimativas dos parâmetros do modelo de regressão quadrática estimados por equações generalizadas

| | $\beta_0$ | $\beta_1$ | $\beta_2$ | SQR* |
|---|---|---|---|---|
| Areia | 0,783 | $-0,019$ | $1,23 \times 10^{-4}$ | 0,579 |
| Silte | $-1,281$ | 0,020 | $-1,56 \times 10^{-4}$ | 0,244 |
| Argila | $-3,253$ | 0,067 | $-4,37 \times 10^{-4}$ | 0,269 |

* Soma de quadrado dos resíduos.

Nota-se na Tabela 4.4 que a SQR para o componente areia foi superior em relação silte e argila. O trabalho de Song et al. (2004) sugere que o uso de equações generalizadas é flexível e eficiente na modelagem de dados composiconais.

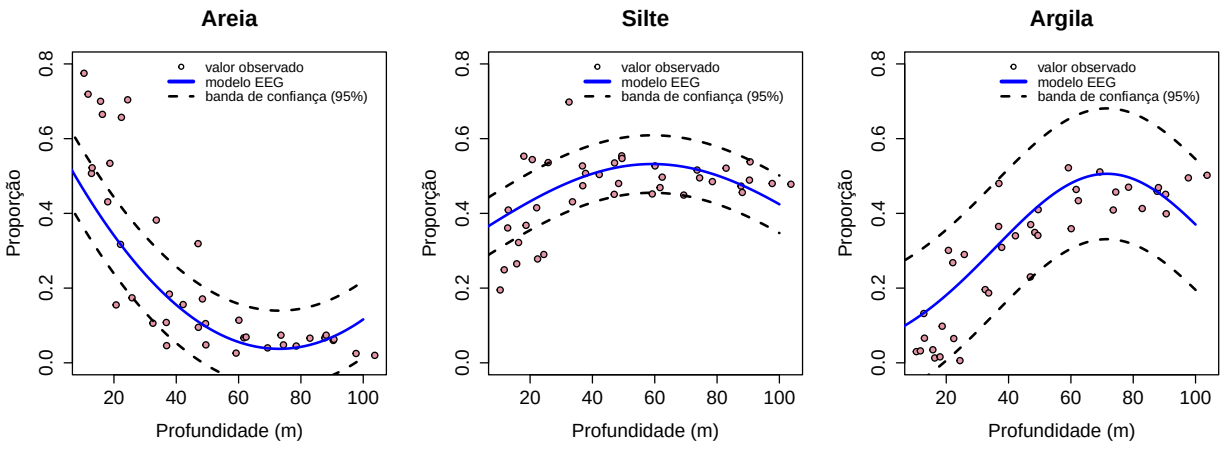



**Figura 4.5.** Modelos quadráticos ajustados pelo método de equações generalizadas

Devido a maior variabilidade dos valores preditos, o componente argila (Figura 4.5) apresentou uma banda de confiança (95%) mais ampla. Com isso, o modelo ajustado para esse componente foi o que incluiu o maior número de observações, na região das bandas de confiança.

### 4.1.4 Comparação entre os modelos de regressão independentes

Para facilitar a comparação visual dos modelos ajustados, a Figura 4.6 apresenta a sobreposição de todos os modelos ajustados para cada componente.

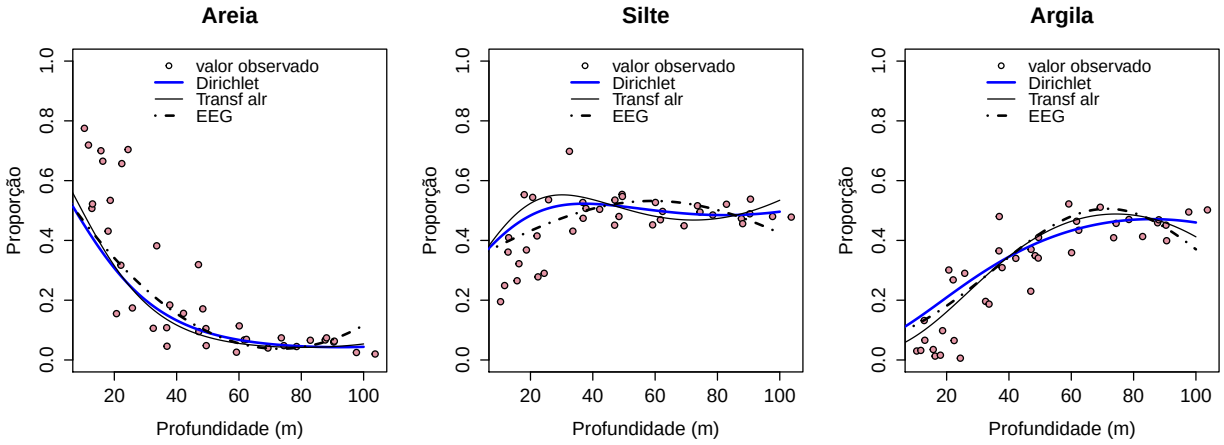



**Figura 4.6.** Sobreposição dos modelos de regressão ajustados para cada componente

Os modelos apresentados na Figura 4.6 quando comparados, evidenciaram maiores diferenças no componente silte. Para esse componente, o método EEG ajustou uma parábola mais côncava que os demais métodos, decrescendo os valores com o aumento da profundidade. Para a componente argila também há uma queda mais acentuada com o acréscimo nos valores de profundidade.

Para avaliar o ajuste pelos três métodos propostos, a Tabela 4.5 apresenta todas as somas de quadrados de resíduos (SQR) para os componentes. Pode-se notar que todos os métodos apresentaram maiores valores de SQR para o componente areia e o menor valor foi encontrado com o método de regressão Dirichlet para o componente silte.

**Tabela 4.5.** Comparação das somas de quadrado dos resíduos (SQR) para cada método

| Métodos | Areia | Silte | Argila |
|---|---|---|---|
| Regressão Dirichlet | 0,538 | 0,206 | 0,232 |
| Transformação (alr) | 0,506 | 0,229 | 0,258 |
| Regressão com EEG | 0,579 | 0,244 | 0,269 |

Percebe-se que métodos usados tiveram desempenhos bem semelhantes, porém, os valores de SQR para o método EEG foram superiores para todos os componentes. Vale ressaltar que no método EEG não é necessário especificar a distribuição dos dados. Basta apenas apresentar uma função que descreva a média e especificar a estrutura de covariância, isso torna o método simples para ser usado.

Deste modo, com os critérios utilizados percebe-se que o método EEG apresentou estimativas próximas aos métodos comumente utilizados na análise de dados composicionais independentes (regressão Dirichlet e regressão com transformação alr). As pequenas diferenças observadas nas estimativas dos coeficientes de regressão e nos valores de SQR entre os modelos, indicam que as equações generalizadas, podem ser uma opção para análise de dados composicionais independentes.

## 4.2 Dados com dependência espacial: Composição textural do solo em área irrigada com pivô central (Martins, 2009)

Os valores percentuais para cada fração do solo (areia, silte e argila) foram amostrados na localizações apresentadas na Figura 4.7, numa área com *grid* regular com distância mínima entre os pontos de 20 metros e distância máxima de 254,56 metros. A Figura 4.8 mostra o diagrama ternário para as composições, incluíndo a média composicional ($\overline{\mathbf{Y}} = (0,281; 0,233; 0,486)$) e as elipses com 2 e 4-desvios padrão. Percebe-se que os dados possuem baixa variabilidade dentro do diagrama ternário.

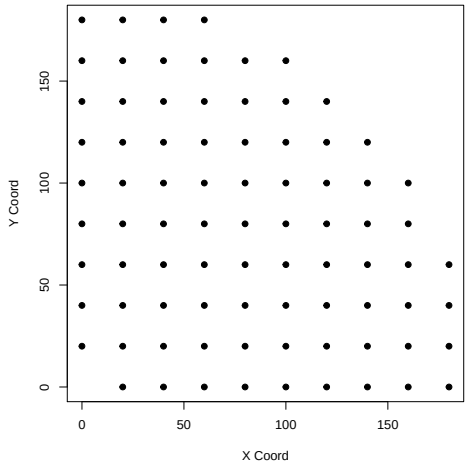



**Figura 4.7.** Localização e distribuição dos pontos amostrais na área em estudo

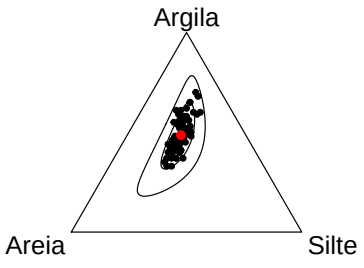
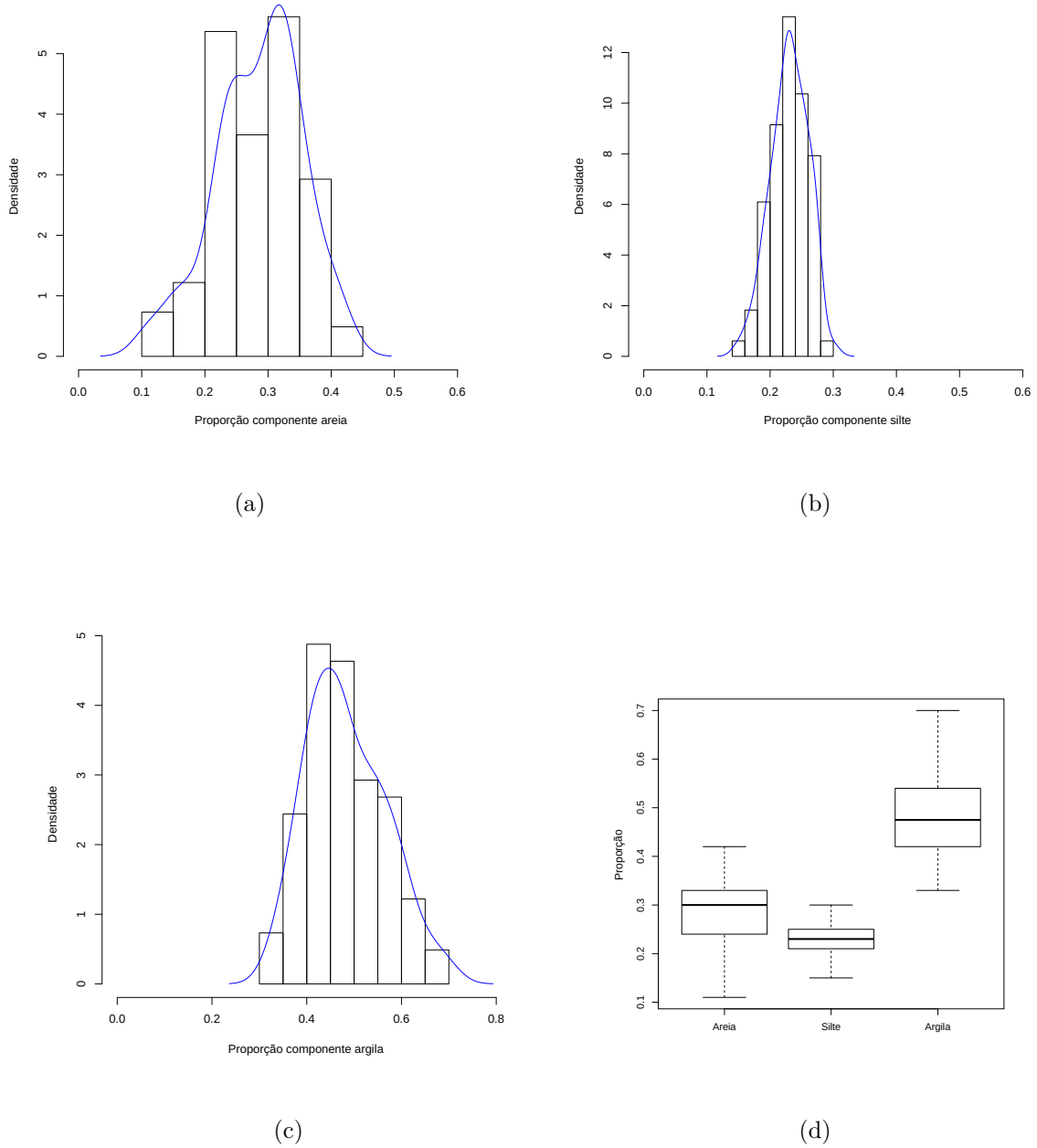



**Figura 4.8.** Diagrama ternário das composições

Com as estatísticas descritivas para cada uma das frações do solo (Tabela 4.6) é possível notar que, os maiores teores são de argila, a fração do solo que apresenta menor variabilidade é o silte.

**Tabela 4.6.** Estatísticas descritivas para as frações do solo em uma área irrigada por pivô central

| | Mínimo | Média | Máximo | Desvio padrão | Variância |
|---|---|---|---|---|---|
| Areia | 0,110 | 0,285 | 0,420 | 0,068 | $4,72 \times 10^{-3}$ |
| Silte | 0,150 | 0,231 | 0,300 | 0,029 | $8,74 \times 10^{-4}$ |
| Argila | 0,330 | 0,488 | 0,700 | 0,083 | $6,69 \times 10^{-3}$ |




(a) (b)

(c) (d)

**Figura 4.9.** Distribuição empírica das proporções dos atributos de solo e gráfico boxplot.

Na Figura 4.9 (a) (b) e (c) percebe-se que cada uma das frações do solo possui um comportamento aproximado a uma distribuição normal de probabilidades, com maior densidade em torno da média. A Figura 4.9 (d) apresenta os gráficos boxplots para cada atributo, não há pontos atípicos e os teores de argila são os que possuem maior variação.

A seguir são apresentadas alternativas para análise dos dados composicionais geoestatísticos, com intuito de descrever a dependência espacial dos componentes em função da distância de amostragem. Para efeitos comparativos, o primeiro passo foi realizar uma análise geoestatística univariada (componentes avaliados separadamente), que consiste no ajuste de três modelos espaciais, um para cada componente (areia, silte e argila).

### 4.2.1 Análise geoestatística para os componentes individuais

Cada composição foi analisada com os recursos geoestatísticos disponíveis no pacote `geoR` versão 1.7-5.2 (RIBEIRO JR e DIGGLE, 2016), com a suposição de que cada composição segue uma distribuição normal de probabilidades. Na Figura 4.10 tem-se os semivariogramas com o modelo exponencial ajustado para cada composição; assim, pode-se ter uma ideia do comportamento da distribuição espacial com o aumento da distância entre os pontos.

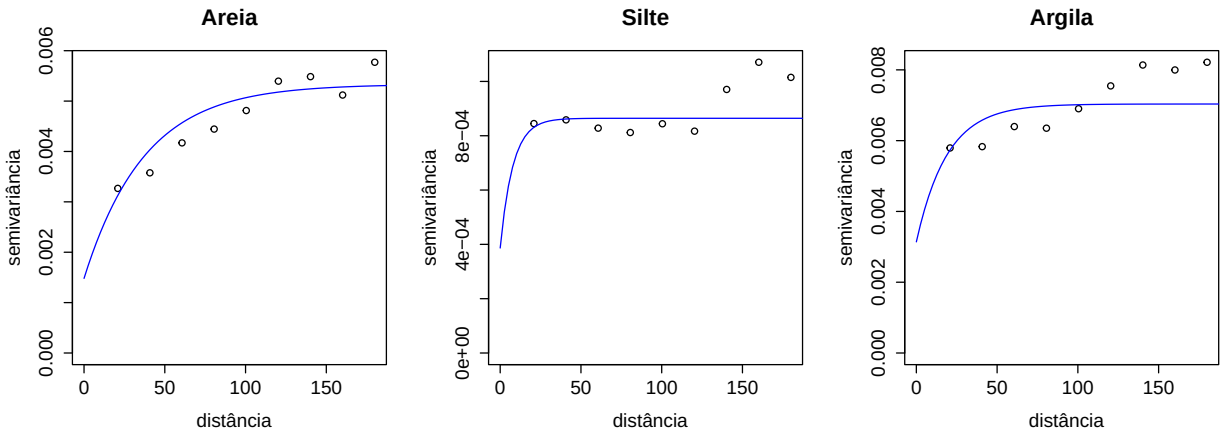



**Figura 4.10.** Semivariogramas experimentais com modelo exponencial ajustado para cada componente

As estimativas dos parâmetros dos modelos ajustados, estão apresentados na Tabela 4.7. Observa-se que os alcance prático da areia foi o maior entre os demais componentes. Isso indica que o raio de dependência espacial para essa composição é superior às outras duas.

O efeito pepita ($\tau$) descreve a descontinuidade a distâncias muito curtas. O $\sigma$ representa as diferenças espaciais entre os valores de uma variável, tomadas em dois pontos

separados por distâncias cada vez maiores. Para avaliar a qualidade do ajuste foi usado o quadrado médio do resíduo (QMR) e nota-se que os componentes areia e silte apresentam um ajuste melhor em relação a argila.

**Tabela 4.7.** Estimativas dos parâmetros para o modelo exponencial ajustado a cada composição

| Parâmetros | Areia | Silte | Argila |
|---|---|---|---|
| $\beta$ | 0,270 | 0,231 | 0,489 |
| $\tau$ | $1,23 \times 10^{-3}$ | $1,01 \times 10^{-4}$ | $3,09 \times 10^{-3}$ |
| $\sigma$ | $4,45 \times 10^{-3}$ | $3,92 \times 10^{-4}$ | $4,13 \times 10^{-3}$ |
| $\phi$ | 37,529 | 7,840 | 19,003 |
| Alcance prático (m) | 112,428 | 23,488 | 56,928 |
| QMR | $6,52 \times 10^{-4}$ | $2,29 \times 10^{-4}$ | $1,58 \times 10^{-3}$ |

QMR: Quadrado médio do resíduo.

A Figura 4.11 apresenta os mapas de predições por krigagem ordinária para os três componentes do solo. Por ser uma análise clássica, seus conceitos estão bem difundidos e esta poderia ser usada como uma análise preliminar para os dados composicionais, pois, não considera a correlação intrínseca existente entre os componentes do solo. Porém, isto pode gerar valores estimados irreais, cuja soma dos componentes estimados não necessariamente fecha na constante unitária.

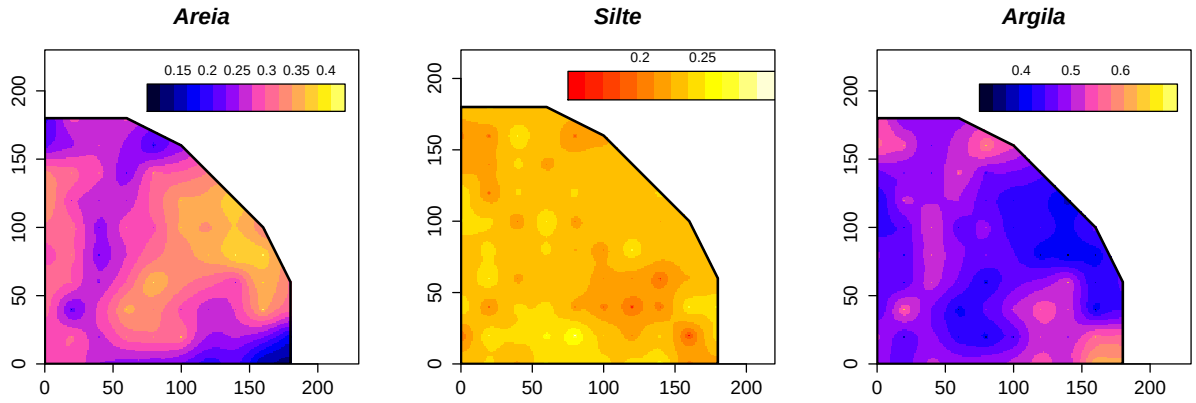



**Figura 4.11.** Mapas das predições das proporções de areia, silte e argila obtidas por krigagem ordinária para cada componente separadamente

### 4.2.2 Análise conjunta com ajuste baseado em variograma

Pawlowsky-Glahn e Olea (2004) e Tolosana-Delgado (2005) reforçam que a análise individual dos componentes não é a mais adequada para os dados composicionais. Porém, a análise direta dos dados composicionais faz com que a matriz de covariâncias cruzadas

seja singular, o que inviabiliza a cokrigagem direta dos dados originais. Assim, os autores propõem a krigagem de razões logarítmicas como alternativa para análise espacial de dados composicionais.

Essa abordagem baseia-se no ajuste de modelos espaciais a partir do variograma construído com a transformação *alr* para todos os pares de componentes. Os modelos são ajustados pelo método dos mínimos quadrados logarítmicos. Na Figura 4.12 tem-se os variogramas com os modelos ajustados.

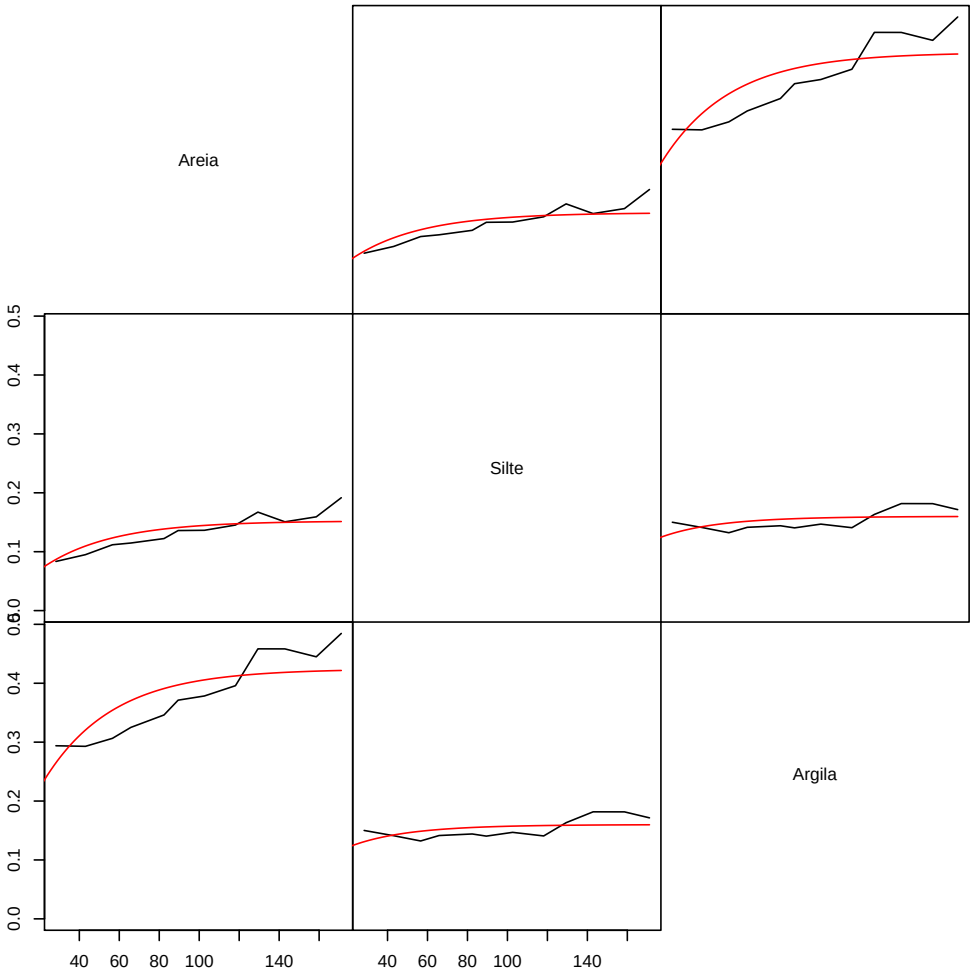



**Figura 4.12.** Variogramas com transformação alr dos componentes com o modelo exponencial ajustado.

A Figura 4.12 apresenta a matriz de variogramas *alr* para todos os pares de componentes. Segundo Tolosana-Delgado (2005) o ajuste pode ser visualizado na matriz de variogramas onde cada painel pode ser interpretado separadamente, como um variograma univariado. Assim, após o ajuste do modelo exponencial e uso de transformações inversas, as estimativas dos parâmetros estão apresentados na Tabela 4.8.

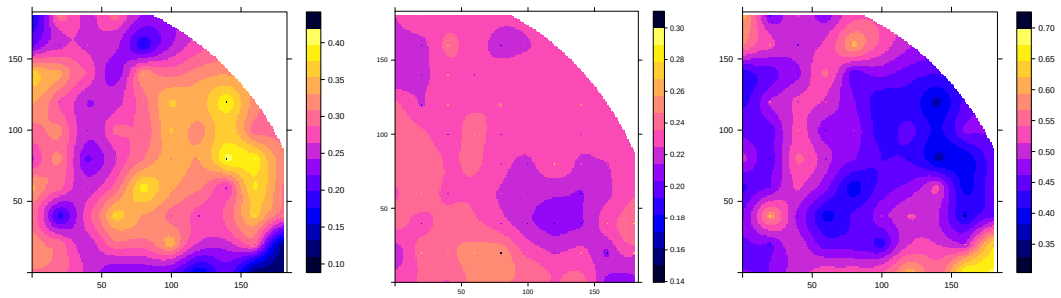
**Tabela 4.8.** Estimativas dos parâmetros para o modelo exponencial, ajustado aos variogramas

| Parâmetros | Areia | Silte | Argila |
|---|---|---|---|
| $\tau$ | 0,043 | 0,025 | 0,051 |
| $\sigma$ | $3,58 \times 10^{-3}$ | $3,01 \times 10^{-3}$ | $4,75 \times 10^{-3}$ |
| $\phi$ | 23,71 | 12,53 | 17,28 |
| Alcance prático (m) | 71,13 | 37,59 | 51,85 |
| QMR | $1,23 \times 10^{-4}$ | $1,52 \times 10^{-4}$ | $1,53 \times 10^{-3}$ |

QMR: Quadrado médio do resíduo.

Na Tabela 4.8 nota-se que a componente silte foi a que apresentou menor alcance prático. Ao comparar esses resultados com os obtidos pela análise geoestatística clássica (Tabela 4.7) constata-se que as estimativas baseada nos variogramas em razão log foram superiores para as variâncias espaciais ($\sigma$) em todos os componentes do solo. Além disso, exceto para a argila, todos os QMR foram menores.

A Figura 4.13 apresenta os mapas com as predições para os três componentes. Observa-se que o alcance prático para o componente areia foi a maior entre os demais componentes, isso indica que a areia possui maior raio de dependência espacial.



(a) Areia (b) Silte (c) Argila

**Figura 4.13.** Mapas das predições das proporções de areia, silte e argila, por krigagem ordinária.

Segundo resultados de Tolosana-Delgado et al. (2009), o uso do método baseado em variograma seguido por krigagem ordinária, aplicado aos teores de minérios de ferro em rochas, apresentou melhores resultados do que por cokrigagem dos dados originais. Além disso, relatam que o método apresentou as estimativas no intervalo original das amostras, satisfazendo as somas fechadas.

### 4.2.3 Análise conjunta baseada em verossimilhança

Para essa análise, os componentes foram transformados usando razão log, transformação alr. O componente argila foi usado como classe de referência.

A principal diferença entre esse método e a abordagem anterior é que a modelagem geoestatística é construída supondo uma distribuição normal bivariada aos dados e os parâmetros são estimados pelo método de máxima verossimilhança de tal distribuição. Fora do contexto composicional, trabalhos com dados univariados de Diggle et al. (1998) e Diggle e Ribeiro Jr (2007) apontam vantagens aos métodos de máxima verossimilhança em relação aos métodos de mínimos quadrados baseados em variogramas, nos casos em que a suposição de normalidade é razoável.

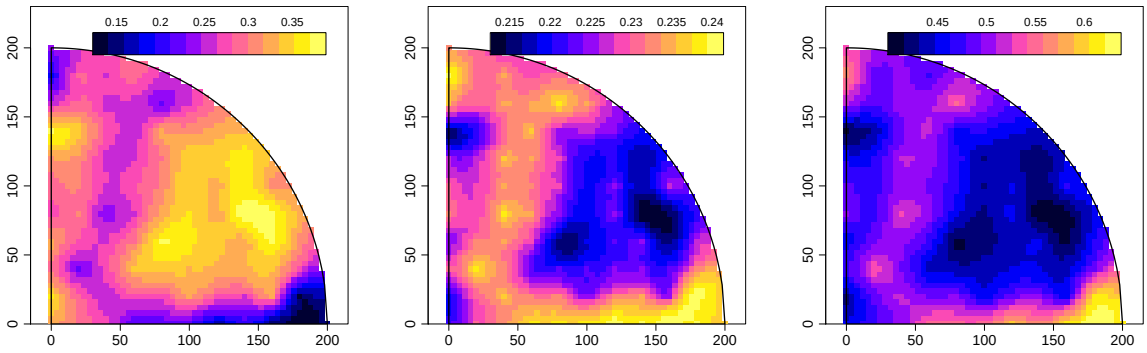
A Tabela 4.9 apresenta as estimativas dos parâmetros do modelo exponencial ajustado para os componentes $\log(areia/argila)$ e $\log(silte/argila)$, mas, os coeficientes não devem ser interpretados por estarem na escala transformada.

**Tabela 4.9.** Estimativas dos parâmetros para o modelo exponencial ajustado

| Parâmetros | log(areia/argila) | log(silte/argila) |
|---|---|---|
| $\beta$ | $-0,786$ | $-0,794$ |
| $\tau$ | 0,284 | 0,262 |
| $\sigma$ | 0,470 | 0,117 |
| $\phi$ | 78,437 | 81,437 |

Após as estimativas dos parâmetros, os valores preditos são obtidos na escala original dos dados por transformação inversa. O terceiro componente foi obtido por diferença dos valores preditos dos demais componentes.

A Figura 4.14 apresenta os mapas de predições para os três componentes. Os valores para os quadrados médios dos resíduos (QMR) foram $1,11 \times 10^{-4}$, $1,02 \times 10^{-4}$ e $1,19 \times 10^{-3}$, para os componentes areia, silte e argila, respectivamente. Isso aponta que, baseado nesse critério, a componente silte foi a que apresentou melhor ajuste.




(a) Areia (b) Silte (c) Argila

**Figura 4.14.** Mapas das predições das proporções de areia, silte e argila na escala original.

### 4.2.4 Análise por equações generalizadas

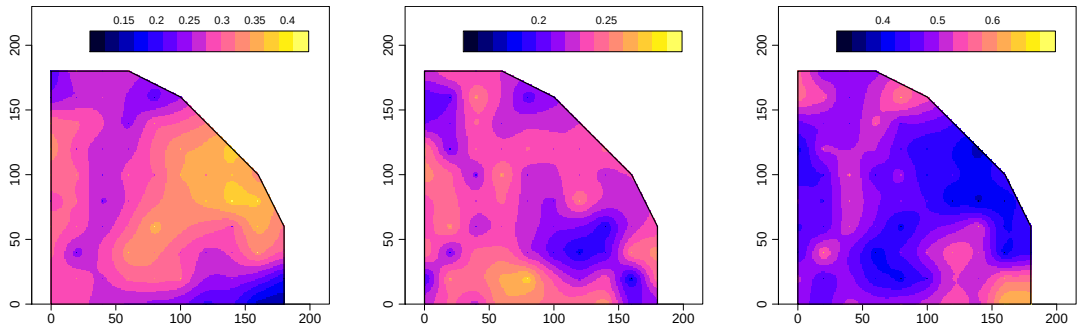
Baseado na ideia de usar as equações de estimação generalizadas para os dados composicionais independentes, foram aplicados os mesmos conceitos para as composições regionalizadas, sem a necessidade de especificar uma distribuição aos dados ou aplicar transformação diretamente aos dados. Para estimar os parâmetros de média foi usada a função de ligação logit e como função de covariância foi aplicada o modelo espacial exponencial. Na Tabela 4.10 tem-se as estimativas dos parâmetros do modelo exponencial ajustado a composição textural do solo.

**Tabela 4.10.** Estimativas dos parâmetros para o modelo exponencial ajustado com EEG

| Parâmetros | Areia | Silte | Argila |
|---|---|---|---|
| $\beta$ | 0,264 | 0,253 | 0,521 |
| $\tau$ | 0,003 | 0,001 | 0,004 |
| $\sigma$ | 0,005 | 0,001 | 0,007 |
| $\phi$ | 48,635 | 18,417 | 25,456 |
| Alcance prático | 145,904 | 55,251 | 76,368 |
| QMR | $4,25 \times 10^{-4}$ | $1,98 \times 10^{-4}$ | $2,23 \times 10^{-3}$ |

QMR: Quadrado médio do resíduo.

Nota-se que o componente areia foi o que apresentou maior alcance prático, seguido por argila e silte. Para o método EEG o QMR dos componentes areia e silte foram menores em relação ao QMR para argila. A Figura 4.15 apresenta os mapas de predições para os três componentes pelo método de krigagem ordinária.




(a) Areia (b) Silte (c) Argila

**Figura 4.15.** Mapas das predições das proporções de areia, silte e argila obtidas por krigagem ordinária.

### 4.2.5 Comparação dos resultados

A Tabela 4.11 apresenta os valores do Erro Absoluto (EA) obtidos por validação cruzada. Nota-se que, os menores valores de EA foram para o componente silte em todos os métodos. Além disso, o método baseado em verossimilhança foi o que apresentou melhor desempenho.

**Tabela 4.11.** Comparação dos valores de EA da validação cruzada para cada método

| Componentes/Métodos | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Areia | 4,11 | 3,05 | 2,97 | 3,98 |
| Silte | 1,96 | 1,29 | 1,21 | 1,91 |
| Argila | 5,78 | 3,28 | 3,42 | 5,53 |

1 - Análise geoestatística para os componentes individuais; 2 - Análise conjunta com ajuste baseado em variograma; 3 - Análise conjunta baseada em verossimilhança; 4 - Análise por equações generalizadas.

Para efeitos de comparação entre os métodos, a Tabela 4.12 apresenta os valores dos quadrados médios dos resíduos (QMR) de cada componente.

**Tabela 4.12.** Comparação dos valores de QMR para cada método

| Métodos | Areia | Silte | Argila |
|---|---|---|---|
| Análise geoestatística para os componentes individuais | $6,52 \times 10^{-4}$ | $2,29 \times 10^{-4}$ | $1,58 \times 10^{-3}$ |
| Análise conjunta com ajuste baseado em variograma | $1,23 \times 10^{-4}$ | $1,52 \times 10^{-4}$ | $1,53 \times 10^{-3}$ |
| Análise conjunta baseada em verossimilhança | $1,11 \times 10^{-4}$ | $1,02 \times 10^{-4}$ | $1,19 \times 10^{-3}$ |
| Análise por equações generalizadas | $4,25 \times 10^{-4}$ | $1,98 \times 10^{-4}$ | $2,23 \times 10^{-3}$ |

Pode-se notar que todos os métodos apresentaram maiores valores de QMR para o componente argila. Os valores de QMR para o método com transformação alr foram menores para todos os componentes.

Percebe-se que o método baseado em variograma e a análise conjunta baseada em verossimilhança tiveram desempenhos satisfatórios e semelhantes entre si.

O método EEG apresentou melhores QMR e EA apenas quando comparado com o uso de técnicas geoestatísticas clássicas. Deste modo, nota-se que, baseado nesses critérios, as análises conjuntas baseadas em variogramas ou em verossimilhança foram mais adequadas do que o método EEG para análise de dados composicionais com dependência espacial.

# 5 ESTUDO DE CASO: PROPORÇÃO DE HEPATITES VIRAIS NO ESTADO DO PARANÁ

O conjunto de dados a seguir foi escolhido apenas como ilustração didática para análise de dados composicionais com dependência espacial por equações generalizadas (EEG).

Hepatites virais são doenças infecciosas sistêmicas que afetam o fígado. Cinco diferentes vírus são reconhecidos como agentes etiológicos da hepatite viral humana: o vírus da hepatite A (HAV), o vírus da hepatite B (HBV), o vírus da hepatite C (HCV), o vírus da hepatite D ou Delta (HDV) e o vírus da hepatite E (HEV).

Por serem doenças infecciosas, é de extrema relevância o acompanhamento dos casos notificados. Indivíduos infectados pelo vírus da hepatite B têm 5% a 10% de risco de tornarem-se doentes crônicos. Na hepatite C, o risco é de 85% (HOUGHTON, 1995).

O tratamento das hepatites B e C é feito com agentes antivirais, com 70% e 35% de sucesso, respectivamente. As medidas preventivas incluem o saneamento básico, as boas práticas de higiene pessoal, o uso de preservativos, o uso de agulhas e seringas descartáveis, o não compartilhamento de objetos pérfuro-cortantes (barbeadores, instrumentos de manicure/pedicure, etc). Existem vacinas para as hepatites A e B; a vacina para hepatite B pode ser obtida nos postos de saúde da rede pública (SECRETARIA DE SAÚDE DO ESTADO DO PARANÁ, 2015).

No Brasil, os dados referentes às hepatites virais estão baseados em um sistema universal de notificação e investigação epidemiológica de todos os casos suspeitos, Sistema de Investigação de Agravos de Notificação (SINAN).

No Paraná, durante o período de 1972 a 1992, o acompanhamento dos casos de hepatites virais, ficou restrito apenas à notificação de casos, sem investigação e classificação. A partir de 1993, a notificação passou a incluir a classificação etiológica, faixa etária e local de residência do paciente acometido por este agravo (SECRETARIA DE SAÚDE DO ESTADO DO PARANÁ, 2015).

Em 2004, com incentivo de campanhas nacionais do Ministério da Saúde, foi implantado o Programa Estadual de Controle Hepatites Virais, desde então, somaram-se esforços para ampliar o diagnóstico nos municípios, implantando a rede de atendimento especializado ambulatorial, laboratorial e farmacêutico.

Este capítulo apresenta a distribuição da proporção para cada faixa etária dos casos das hepatites virais nos municípios do estado do Paraná. Baseados na série histórica de 2007 até 2015, seguindo o padrão do Boletim Epidemiológico do Ministério da Saúde disponível no DATASUS (Departamento de Informática do Sistema Único de Saúde).

Os dados disponíveis no sítio do DATASUS (`http://datasus.saude.gov.br/`), foram selecionados os casos confirmados por ano de diagnóstico em cada município. Os dados estão divididos nas faixas etárias de 0 a 14 anos, 15 a 59 anos e acima de 60 anos.

Assim, obteve-se a proporção relativa ao total do município para cada faixa etária. Os municípios incluídos no estudo são os que possuem casos notificados em todas as faixas etárias; logo, do total de 399 municípios no estado do Paraná, fazem parte da pesquisa 127 municípios.

O número de casos confirmados por faixa etária foram transformados em dados composicionais com o operador de restrição apresentado na seção 2.3. Além disso, a proporção de casos de hepatites virais foram ponderadas pela proporção de habitantes em cada faixa etária (IBGE - censo 2010). Para essa ponderação foi usada a operação perturbação, entre as composições de casos notificados por faixa etátria e as composições da população em cada faixa etária.

A Figura 5.1 apresenta a posição dos 127 municípios no mapa do estado. Os municípios são representados pelas coordenadas central da área do município, expressas em UTM.

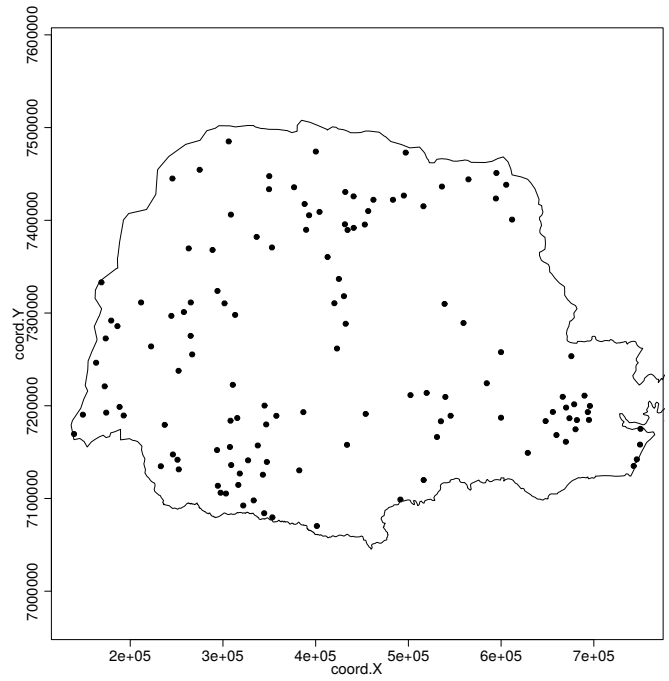
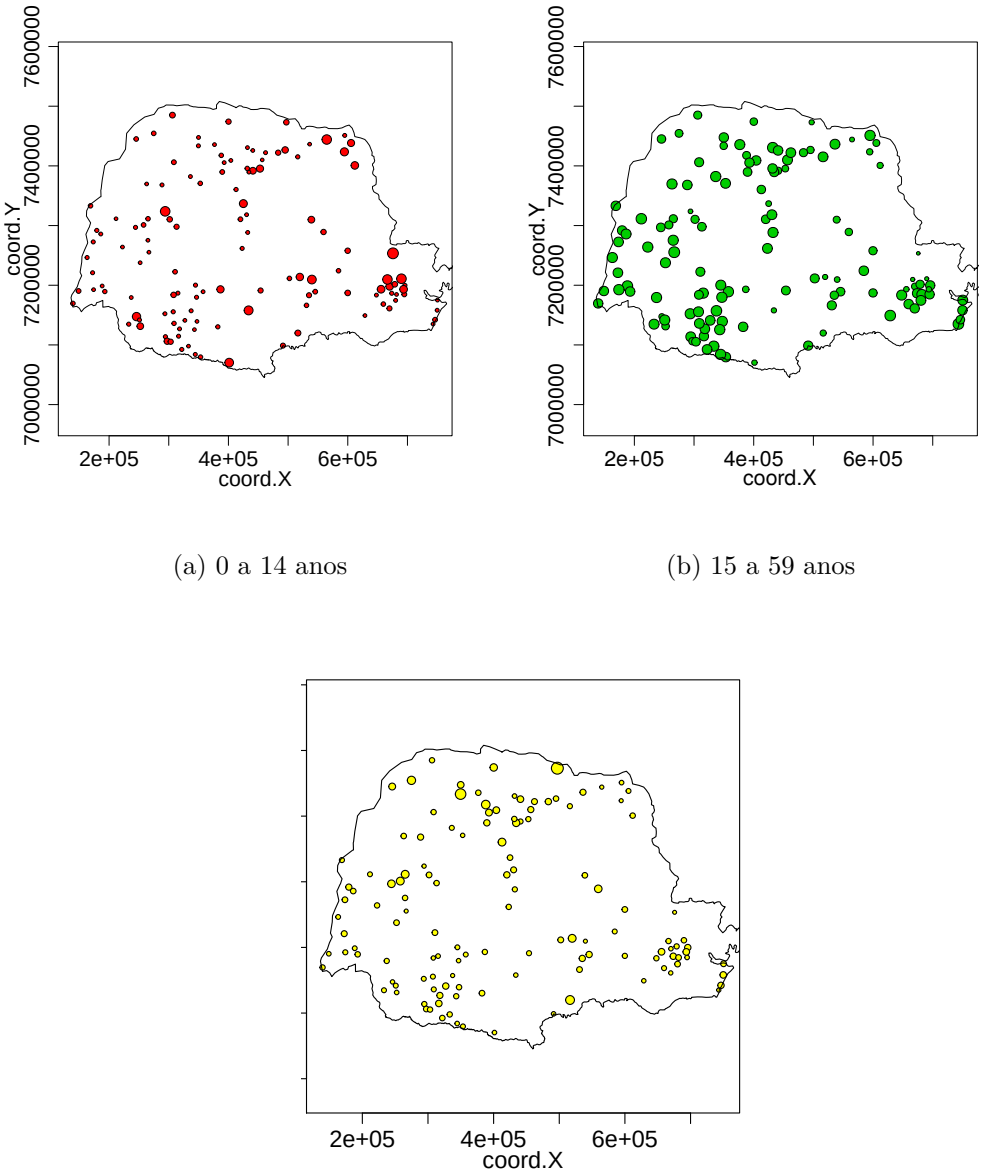



**Figura 5.1.** Localização dos municípios com casos notificados de hepatite viral no Paraná

A Figura 5.2 apresenta os municípios com casos notificados de hepatites virais em cada faixa etária. Os círculos maiores representam os municípios com maior proporção de casos de hepatites naquela faixa etária, os círculos menores representam menor proporção.

Segundo dados da Secretaria de Saúde do Paraná (2015) no Boletim Epidemiológico de Hepatites Virais do Estado do Paraná, as maiores concentrações de casos por 100.000 habitantes nas regionais de Francisco Beltrão, Foz do Iguaçu, Pato Branco, Cascavel e Toledo (regiões oeste e sudoeste do estado). Ao observar a Figura 5.2 (b), percebe-se que nessas regiões, a faixa de 15 a 59 anos apresenta a maior proporção de casos quando

comparada com as demais faixas etárias. Na faixa de 0 a 14 anos, as maiores proporções estão na cidade de Curitiba e região metropolitana (Figura 5.2 (a)).




(a) 0 a 14 anos

(b) 15 a 59 anos

(c) acima de 60 anos

**Figura 5.2.** Proporções de hepatite viral, relativas ao total de cada município paranaense para cada faixa etária.

A variação das proporções, em cada faixa etária (componentes) pode ser observada na Figura 5.3: as faixas etárias de 0 a 14 anos e acima de 60 apresentam as menores proporções.

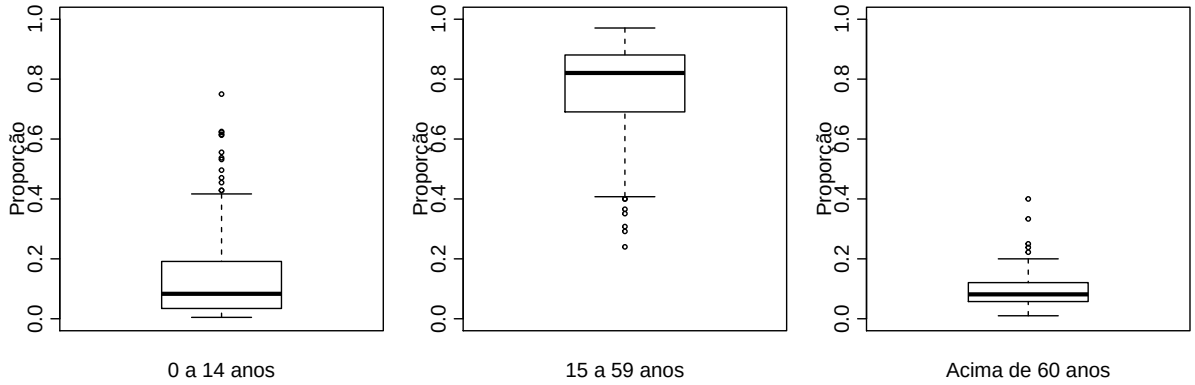



**Figura 5.3.** Boxplot das proporções de casos de hepatite viral em cada faixa etária.

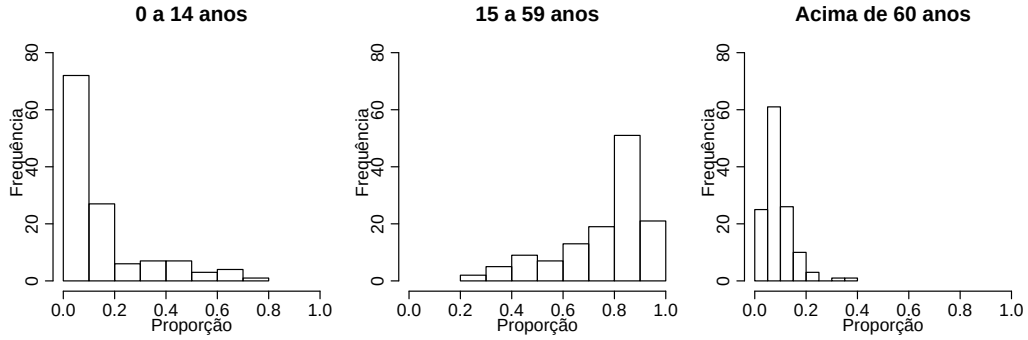



**Figura 5.4.** Distribuições empíricas das proporções de casos por faixa etária.

Na Figura 5.4 nota-se que as distribuições empíricas das proporções em cada faixa etária não possuem distribuições simétricas ou aproximadamente normal. Neste caso, o uso de equações generalizadas possui a vantagem de não necessitar a especificação uma distribuição de probabilidade aos dados.

Na Tabela 5.1, os valores de desvio padrão e variância indicam que a faixa acima dos 60 anos foi a que apresentou menor variabilidade em relação a proporção de cada município.

Os resultados apresentados na Tabela 5.1 complementam os resultados anteriores que apontam a faixa de 15 a 59 anos com as maiores proporções de casos nos municípios, enquanto as faixas de 0 a 14 anos e acima de 60 anos, as menores proporções, respectivamente. Isso se deve ao fato de que, para todas as regiões, a proporção de pessoas na faixa de idade de 15 a 59 anos é superior às demais taxas.

**Tabela 5.1.** Estatísticas descritivas das proporções de hepatite viral.

| Estatísticas | 0 a 14 anos | 15 a 59 anos | Acima de 60 anos |
|---|---|---|---|
| Mínimo | 0,0046 | 0,2400 | 0,0100 |
| Quartil 1 | 0,0343 | 0,6906 | 0,0574 |
| Mediana | 0,1330 | 0,8205 | 0,0814 |
| Quartil 3 | 0,1913 | 0,8806 | 0,1206 |
| Máximo | 0,7500 | 0,9706 | 0,4000 |
| Média | 0,1498 | 0,7559 | 0,0942 |
| Desvio padrão | 0,1688 | 0,1693 | 0,0602 |
| Variância | 0,0284 | 0,0286 | 0,0036 |

A Figura 5.5 mostra o diagrama ternário das proporções de hepatite para cada faixa etária. Fica evidente que na faixa de idades de 15 a 59 anos são maiores. O ponto que representa a média composicional está destacado no diagrama ternário. As elipses expressam a variabilidade de 2 e 4-desvios padrão. Mesmo no espaço simplex, nota-se a maior variabilidade na faixa de 15 a 59 anos.

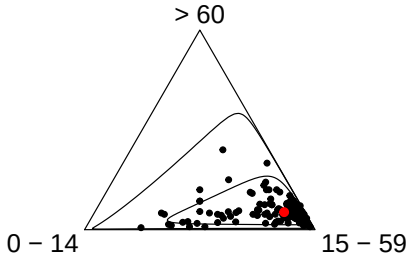



**Figura 5.5.** Diagrama ternário para as proporções de hepatite viral por faixa etária.

A seguir, a matriz de correlação mostra que as proporções da faixa de 0 a 14 anos é altamente correlacionada negativamente com as faixas de 15 a 59 anos e acima de 60 anos ($r = -0,7962$ e $r = -0,8051$). Esse fato reforça os outros resultados apresentados e indica que os municípios com altas proporções de casos na faixa de 0 a 14 anos apresentam menores proporções nas demais faixas.

Para descrever a dependência espacial entre as proporções de hepatite viral nos municípios foi usado o método EEG. Na Tabela 5.2 tem-se as estimativas dos parâmetros do modelo exponencial ajustado para as três faixas etárias. As estimativas foram obtidas usando equações generalizadas com matriz de correlação especificada $\mathbf{R}_{ij} = \exp\{-d_{ij}/\phi\}$, em que, $d_{ij}$ é a distância entre 2 municípios e $\phi$ é o parâmetro de alcance.

| | 0 a 14 anos | 15 a 59 anos | Acima de 60 anos |
|---|---|---|---|
| 0 a 14 anos | $1,0000$ | $-0,7962$ | $-0,8051$ |
| 15 a 59 anos | $-0,7962$ | $1,0000$ | $0,2821$ |
| Acima de 60 anos | $-0,8051$ | $0,2821$ | $1,0000$ |

Usando o modelo geoestatístico exponencial baseado na distância entre os municípios, os parâmetros foram estimados com funções do pacote `JGEE`(INAN, 2015) disponível para o software `R` (R CORE TEAM, 2016). A faixa de 15 a 59 anos foi a que apresentou maior alcance prático, isso indica que as taxas de ocorrência para essa faixa etária possuem um alcance maior ao longo do Estado.

**Tabela 5.2.** Estimativas dos parâmetros para o modelo exponencial ajustado na matriz de correlação

| Parâmetros | 0 a 14 anos | 15 a 59 anos | Acima de 60 anos |
|---|---|---|---|
| $\beta$ | 0,1686 | 0,7393 | 0,0957 |
| $\tau$ | 0,0181 | 0,0195 | 0,0029 |
| $\sigma$ | 0,0216 | 0,0197 | 0,0021 |
| $\phi$ | $0,6 \times 10^5$ | $1,2 \times 10^5$ | $0,8 \times 10^5$ |
| Alcance prático | 188,15 Km | 345,94 Km | 241,48 Km |

Para ter uma ideia da distribuição espacial das proporções, conforme o aumento da distância entre os pontos, a Figura 5.6 apresenta os variogramas experimentais com os modelos cujos parâmetros estão na Tabela 5.2. Os modelos ajustados foram construídos a partir dos dados, por meio da estimação por EEG.

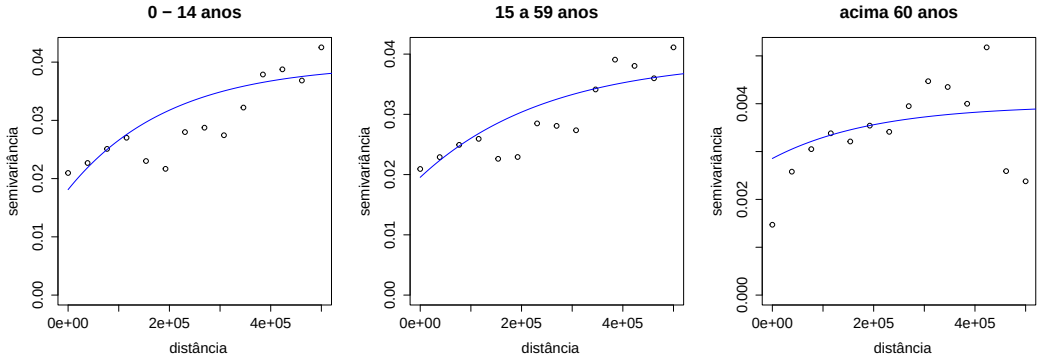


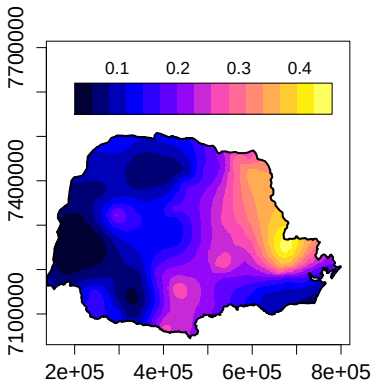

**Figura 5.6.** Variogramas experimentais com modelos exponencial ajustado.

Com os mapas de predição para as três faixas etárias (Figura 5.7), pode-se constatar que na faixa de 15 a 59 anos, as regiões oeste e sudeste no estado apresentam

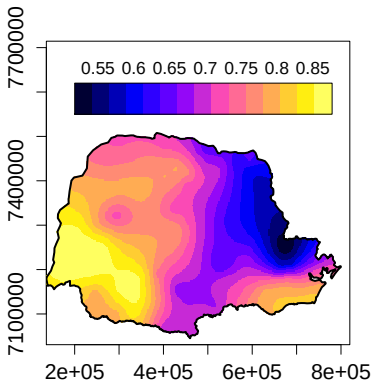
maiores proporções. Isso corrobora com os resultados apontados no boletim da Secretaria de Saúde do Estado.

Para minimizar, cada vez mais, as proporções de hepatite viral no estado, a Secretaria de Saúde do Estado do Paraná (2015) reforça que a população nascida a partir de meados da década de 1980 foi imunizada no nascimento ou na primeira infância. No ano de 1999, a vacinação para hepatite B foi ampliada até 14 anos de idade, em 2001 até 19 anos, em 2010 até 29 anos e em 2013 até 49 anos.

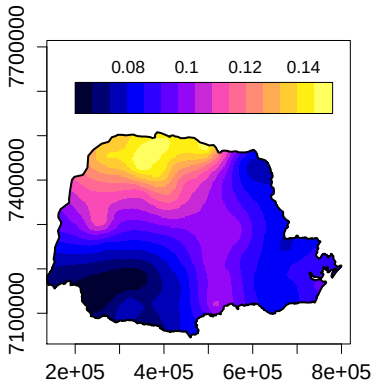
A faixa etária acima de 60 anos, em geral, apresenta baixa proporção de hepatite viral em todos os municípios, exceto para as regiões norte/noroeste do Estado. No entanto, algumas ressalvas devem ser feita a respeito das deficiências diversas na base de dados de notificação, que impõem cautela na interpretação dos valores encontrados. Pessoas acima de 60 anos estão mais suscetíveis a outras doenças infecciosas. Os que possuem algum tipo de hepatite viral podem ser inicialmente notificados por outras doenças ou morbidades, o que pode interferir nos resultados apresentados.




(a) 0 a 14 anos




(b) 15 a 59 anos




(c) acima de 60 anos

**Figura 5.7.** Mapas das predições das proporções de hepatite viral por faixa etária no estado do Paraná.

# 6 CONSIDERAÇÕES FINAIS

A análise de dados composicionais tem evoluído ao longo dos últimos 40 anos, com técnicas e recursos computacionais que permitem a análise dos dados composicionais que considera a correlação intrínseca existente entre os componentes. Diversos métodos têm sido aplicados nos mais variados contextos, seja para os dados composicionais independentes ou dados composicionais com dependência espacial ou temporal.

Para dados composicionais com a suposição de independência, as abordagens apresentadas e aplicadas neste trabalho tiveram um bom desempenho para descrever o comportamento dos dados composicionais em função de covariáveis.

Já no contexto com dependência espacial, dentre as técnicas apresentadas, ficou evidente que as ferramentas geoestatísticas univariadas podem ser usadas para análise de dados composicionais, porém, o ideal é que seus resultados sejam considerados como preliminares para o ajuste de outras técnicas mais especificas a esses conjuntos de dados.

O método de Equações de Estimação Generalizadas (EEG) tem sido um recurso promissor para analisar conjuntos de dados com algum tipo de correlação entre componentes. A estimação dos parâmetros por EEG também foi efetiva para os dados composicionais com independência.

Para as composições regionalizadas, o uso de equações generalizadas apresentou resultados mais precisos que a utilização da geoestatística univariada. Porém, ao ser comparada com as demais técnicas (análise conjunta baseada em variograma e análise conjunta baseada em verossimilhança) apresentou maiores erros residuais em relação às outras duas técnicas. Vale ressaltar que, por ser uma técnica de fácil uso, sem a suposição de distribuição de probabilidades ou necessidade transformações, a aplicação desse método pode ser aprimorado para esse tipo de dados.

Como sugestões de estudos futuros, pode-se destacar a investigação para encontrar uma famılia de distribuições no simplex que os satisfazem as restrições dos dados composicionais. Ou ainda especificar uma estrutura de covariância geral para expressar a variabilidade espacial em dados composicionais.

# REFERÊNCIAS

AITCHISON, J. The statistical analysis of compositional data. **Journal of the Royal Statistical Society**. Series B. Methodological, London, v. 44, n. 2, p. 139-177, 1982.

AITCHISON, J. A general class of distributions on the simplex. **Journal of the Royal Statistical Society**. Series B. Methodological, London, v. 47, n. 1, p. 136-146, 1985.

AITCHISON, J. **The Statistical of Compositional Data**. New York: Chapman and Hall, 1986. 415 p.

AITCHISON, J; EGOZCUE, J.J. Compositional data analysis: where are we and where should we be hesding? **Mathematical Geology**, New York, v. 37, n. 7, p. 829–850, 2005.

AITCHISON, J; GREENACRE, M. Biplots of compositional data. **Appl. Statist.**, London, v. 51, n. 4, p. 375-392, 2002.

AITCHISON, J; BACON-SHONE, J. Log contrast models for experiments with mixtures. **Biometrika**, London, v. 71, n. 2, p. 323-330, 1984.

AITCHISON, J; SHEN, S.M. Logistic-normal distributions: Some properties and uses. **Biometrika**, London, v. 67, n. 2, p. 261-272, 1980.

ARK, L.A. van der. Regression Analysis of Compositional Data When Both Dependent Variable and Independent Variable are Components. 2005. Disponível em: `http://spitswww.uvt.nl/~avdrark/research.html`. Acesso em 20 de agosto de 2015.

ASAR, O. `mmm`: an R package for analyzing multivariate longitudinal data with multivariate marginal models, 2014. R package version 1.4. Disponível em: `https://CRAN.R-project.org/package=mmm`. Acesso em 10 outubro de 2016.

BARNDORFF, N; JØRGENSEN, B. Some parametric models on the simplex. **Journal of Multivariate Analysis**, v. 39, n. 1, p. 106-116, 1991.

BISHOP, J.; DIE, D.; WANG, Y.G. A generalized estimating equations approach for analysis of the impact of new technology on a trawl fishery. **Aust. N. Z. J. Statist**, v. 42, n. 2, p. 159–177, 2000.

BOEZIO, M.N.M.; COSTA, J.F.C.L.; KOPPE, J.C. Cokrigagem de razões logarítmicas aditivas (alr) na estimativa de teores em depósitos de ferro. **Rev. Esc. Minas**, Ouro Preto, v. 65, n. 3, p. 401-411, 2012.

BONAT, W.H.; JORGENSEN, B. multivariate covariance generalized linear models. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 65, n. 5, p. 649-675, 2016.

BOOGAART, K.G. van den; TOLOSANA-DELGADO, R. **Analyzing Compositional Data with R**. New York: Springer, 2013. 258 p.

BOOGAART, K.G. van den; TOLOSANA-DELGADO, R.; BREN, M. `compositions`: Compositional Data Analysis, 2014. R package version 1.40-1. Disponível em: `https://CRAN.R-project.org/package=compositions`. Acesso em 15 de setembro de 2014.

BREHM, J.; GATES, S.; GOMEZ, B. A comparison of methods of compositional data analysis. **In Political Methodology Society annual meetings**, San Diego, 1998.

BUCHANAN, S.; TRIANTAFILIS, J.; ODEH, I.O.A. Digital soil mapping of compositional particle-size fractions using proximal and remotely sensed ancillary data. **Geophysics**, v. 77, n. 4, p. WB201-WB211, 2012.

BUCCIANTI, A. Is compositional data analysis a way to see beyond the illusion? **Computers & Geosciences**, v. 50, p. 165-173, 2013.

CAMARGO, A.P. **Modelos de regressão sobre dados composicionais**. 2011. 61 p. Dissertação (Mestrado em Matemática Aplicada) - Universidade de São Paulo - SP, São Paulo, 2011.

CAMPBELL, G.; MOSIMANN, J. **Modelling continous proportional data with the Dirichlet distribution**. Manuscrito não publicado, p. 12, 1987.

CHAYES, F. On correlations between variables of constant sum. **Journal of Geophysical Research**, v. 65, n. 12, p. 4185-4193, 1960.

CHAYES, F. A priori and experimental approximation of simple ratio correlations. In R. B. McCAMMON (Ed.), **Concepts in geostatistics**, New York: Springer, p. 168, 1975.

CLARK, I. **Pratical Geostatistics**. London: Applied Science Publishers, 1979, 129 p.

COAKLEY, J.P.; RUST, B.R. Sedimentation in an Arctic lake. **Journal of Sedimentary Petrology**, v. 38, n. 4, p. 1290-1300, 1968.

CONNOR, J.R.; MOSIMANN, J.E. Concepts of independence for proportions with a generalization of the Dirichlet distibution. **Journal of the American Statistical Association**, v. 64, p. 194-206, 1969.

CRESSIE, N. **Statistics for Spatial Data**. New York: John Wiley & Sons. 1991, 900 p.

DIGGLE, P.J.; TAWN, J.A.; MOYEED, R.A. Model-based geostatisticas (with discussion). **Journal of the Royal Statistical Society**, Series C. Applied Statistics, London, v. 47, n. 3, p. 299-350, 1998.

GONÇALVES, A.C.A. **Variabilidade espacial de propriedades físicas do solo para fins de manejo da irrigação**. 1997. 189 p. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz". Universidade de São Paulo, Piracicaba. 1997.

HANLEY, J.A.; NEGASSA, A.; EDWARDES, M.D.B.; FORRESTER, J.E. Statistical analusys of correlated data using generalized estimating equations: An orientation. **American Journal of Epidemiology**, v. 157, n. 5, p. 364-375, 2003.

HIJAZI, R.H.; JERNIGAN, R.W. Modelling compositional data using Dirichlet regression models. **Journal of Applied Probability & Statistics**, v. 4, n. 1, p. 77-91, 2007.

HOUGHTON, M. Hepatitis C virus. In: Fields et al. (eds.), **Virology**, 3ed, 1995.

INAN, G. `JGEE`: Title Joint Generalized Estimating Equation Solver, 2015. R package version 1.1. Disponível em: `https://CRAN.R-project.org/package=JGEE`. Acesso em 01 outubro de 2016.

IYENGAR, M.; DEY, D.K. **Bayesian Analysis of Compositional Data**. Manuscrito não publicado, 1996. Disponível em: `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.3146`. Acesso em 15 de maio de 2016.

MAIER, M.J. `DirichletReg`: Dirichlet Regression in R, 2015. R package version 0.6-3. Disponível em: `http://dirichletreg.r-forge.r-project.org/`. Acesso em 05 de julho de 2016.

MARTINS, A.B.T.; RIBEIRO JR, P.J.; BONAT, W.H. Um modelo geoestatístico para dados composicionais. **Revista Brasileira de Biometria**, v. 27, n. 3, p. 456-477, 2009.

MARTINS, A.B.T. **Análise geoestatística de dados composicionais**. 2009. 217 p. Tese (Doutorado em Métodos Numéricos em Engenharia) - Universidade Federal do Paraná, Curitiba, 2009.

MARTINS, A.B.T.; RIBEIRO JR, P.J.; BONAT, W.H. Likelihood analysis for a class of spatial geostatistical compositional models. **Spatial Statistics**, v. 17, p. 121-130, 2016.

MATHERON, G. The theory of regionalized variables and its applications. **Technical Report C-5**, École Nationale Supérieure de Mines de Paris, Centre de Geostatistique et de Morphologie Mathematique, Fontainebleau, p. 211, 1971.

MCALISTER, D. The law of the geometric mean. **Proceedings of the Royal Society of London**, v. 29, p. 367-376, 1879.

MCDANIEL, L.; HENDERSON, N. `geeM`: Solve Generalized Estimating Equations, 2015. R package version 0.10.0. Disponível em: `https://CRAN.R-project.org/package=geeM`. Acesso em 15 de julho de 2016.

MINISTÉRIO DA SAÚDE DO PARANÁ. **Secretaria de Vigilância em Saúde (SVS): base de dados do Sistema Nacional de Vigilância Epidemiológica** [boletins de notificação semanal e Sistema de Informação de Agravos de Notificação – Sinan (a partir de 1998)]. 2015.

ODEH, I.O.A.; TODD, A.J.; TRIANTAFILIS, J. Spatial prediction of soil particle-size fractions as compositional data. **Soil Science**, v. 168, n. 7, p. 1397-1403, 2003.

OBAGE, S.C. **Uma análise bayesiana para dados composicionais**, 2005. 69 p. Dissertação (Mestrado em Estatística). Universidade Federal de São Carlos - SP, São Carlos, 2005.

OLEA, R.A.; PAWLOWSKY, V.; DAVIS, J.C. Volumetric calculations in an oil field: The basis method. **Computers & Geosciences**, v. 19, n. 10, p. 1517–1527, 1993.

PAWLOWSKY, V. On spurious spatial covariance between variables of constant sum. **Science de la Terre**, Série Informatique, v. 21, p. 107-113.

PAWLOWSKY, V.; BURGER, H. Spatial structure analysis of regionalized compositions. **Math. Geology**, v. 24, n. 6, p. 675–691, 1992.

PAWLOWSKY-GLAHN, V.; BUCCIANTI, A. **Compositional Data Analysis: Theory and Applications**. Wiley, 2011, 400 p.

PAWLOWSKY-GLAHN, V.; EGOZCUE, J.J. Geometric approach to statistical analysis on the simplex. **Stochastic Environmental Research and Risk Assessment (SERRA)**, v. 15, n. 5, p. 384-398, 2001.

PAWLOWSKY-GLAHN, V.; EGOZCUE, J.J.; TOLOSANA-DELGADO, R. **Modeling and Analysis of Compositional Data**. Wiley, 2015, 247 p.

PAWLOWSKY-GLAHN, V.; OLEA, R.A. **Geostatistical Analysis of Compositional Data**. New York: Oxford University Press, Inc., 2004, p. 265.

PEARSON, K. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. **Proceedings of the Royal Society of London**, London, v. LX, p. 489-502, 1897.

PETER, X.; SONG, K. **Correlated Data Analysis: Modeling, Analytics and Applications**. New York: Springer, 2007, 346 p.

PRENTICE, R.L.; ZHAO, L.P. Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. **Biometrics**, v. 47, n. 3, p. 825-839, 1991.

R CORE TEAM. `R`: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. Disponível em: `https://www.R-project.org/`.

RIBEIRO JR, P.J.; DIGGLE, P. `geoR`: Analysis of Geostatistical Data, 2016. R package version 1.7-5.2. Disponível em: `https://CRAN.R-project.org/package=geoR`. Acesso em 15 de julho de 2016.

SECRETARIA DE SAÚDE PR. **Boletim Epidemiológico de Hepatites Virais do Estado do Paraná - Ano 2015**. Disponível em: `http://www.saude.pr.gov.br/arquivos/File/boletim_heptaites_virais.pdf`. Acesso em 10 de janeiro de 2017.

SONG, P.X.K.; TAN, M. Marginal models for longitudinal continuous proportional data. **Biometrics**, v. 56, n. 2, p. 496–502, 2000.

SONG, P.X.K.; ZHENGUO, Q.; TAN, M. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. **Biometrical Journal**, v. 46, n. 5, p. 540-553, 2004.

TOLOSANA-DELGADO, R. **Geostatistics for constrained variables: positive data, compositions and probabilities. Application to environmental hazard monitoring**. 2005. 215 p. Tese (Medi Ambient - Environmental Sciences) - Universidade de Girona - Catalonia, 2005.

TOMCZAK, M. Spatial interpolation and its uncertainty using automated anisotropic inverse distance weighting (IDW) - cross-validation/jackknife approach. **Journal of Geographic Information and Decision Analysis**, v.2, p.18-30, 1998.

VERBEKE, G.; MOLENBERGHS, G. **Linear Mixed Models for Longitudinal Data**. New York: Springer, 2000, 570 p.

ZHANG, B. **On Compositional Data Modeling and its Biomedial Application**. 2012. 103 p. Tese (Doctor of Philosophy) - Columbia University, 2012.

ZEGER, S.L.; LIANG, K.Y. Longitudinal Data Analysis Using Generalized Linear Models. **Biometrika**, London, v. 73, n. 1, p. 13-22, 1986.

# ANEXOS

## Anexo A - Composição textural dos sedimentos do Lago Ártico (em %), para cada profundidade.

| | Areia $y_1$ | Silte $y_2$ | Argila $y_3$ | Profundidade $x$ (m) |
|---|---|---|---|---|
| 1 | 77,5 | 19,5 | 3,0 | 10,4 |
| 2 | 71,9 | 24,9 | 3,2 | 11,7 |
| 3 | 50,7 | 36,1 | 13,2 | 12,8 |
| 4 | 52,2 | 40,9 | 6,6 | 13,0 |
| 5 | 70,0 | 26,5 | 3,5 | 15,7 |
| 6 | 66,5 | 32,2 | 1,3 | 16,3 |
| 7 | 43,1 | 55,3 | 1,6 | 18,0 |
| 8 | 53,4 | 36,8 | 9,8 | 18,7 |
| 9 | 15,5 | 54,4 | 30,1 | 20,7 |
| 10 | 31,7 | 41,5 | 26,8 | 22,1 |
| 11 | 65,7 | 27,8 | 6,5 | 22,4 |
| 12 | 70,4 | 29,0 | 0,6 | 24,4 |
| 13 | 17,4 | 53,6 | 29,0 | 25,8 |
| 14 | 10,6 | 69,8 | 19,6 | 32,5 |
| 15 | 38,2 | 43,1 | 18,7 | 33,6 |
| 16 | 10,8 | 52,7 | 36,5 | 36,8 |
| 17 | 18,4 | 50,7 | 30,9 | 37,8 |
| 18 | 4,6 | 47,4 | 48,0 | 36,9 |
| 19 | 15,6 | 50,4 | 34,0 | 42,2 |
| 20 | 31,9 | 45,1 | 23,0 | 47,0 |
| 21 | 9,5 | 53,5 | 37,0 | 47,1 |
| 22 | 17,1 | 48,0 | 34,9 | 48,4 |
| 23 | 10,5 | 55,4 | 34,1 | 49,4 |
| 24 | 4,8 | 54,7 | 41,0 | 49,5 |
| 25 | 2,6 | 45,2 | 52,2 | 59,2 |
| 26 | 11,4 | 52,7 | 35,9 | 60,1 |
| 27 | 6,7 | 46,9 | 46,4 | 61,7 |
| 28 | 6,9 | 49,7 | 43,4 | 62,4 |
| 29 | 4,0 | 44,9 | 51,1 | 69,3 |
| 30 | 7,4 | 51,6 | 40,9 | 73,6 |
| 31 | 4,8 | 49,5 | 45,7 | 74,4 |
| 32 | 4,5 | 48,5 | 47,0 | 78,5 |
| 33 | 6,6 | 52,1 | 41,3 | 82,9 |
| 34 | 6,7 | 47,3 | 45,9 | 87,7 |
| 35 | 7,4 | 45,6 | 46,9 | 88,1 |
| 36 | 6,0 | 48,9 | 45,1 | 90,4 |
| 37 | 6,3 | 53,8 | 39,9 | 90,6 |
| 38 | 2,5 | 48,0 | 49,5 | 97,7 |
| 39 | 2,0 | 47,8 | 50,2 | 103,7 |

## Anexo B - Composição textural (em %) do solo do campo experimental de irrigação da ESALQ-USP, com as coordenadas de cada ponto amostrado

| | Coord.X | Coord.Y | Areia | Silte | Argila | | Coord.X | Coord.Y | Areia | Silte | Argila |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 20 | 0 | 31 | 25 | 44 | 42 | 40 | 80 | 21 | 23 | 56 |
| 2 | 40 | 0 | 24 | 21 | 55 | 43 | 60 | 80 | 27 | 25 | 48 |
| 3 | 60 | 0 | 23 | 27 | 50 | 44 | 80 | 80 | 30 | 22 | 48 |
| 4 | 80 | 0 | 22 | 25 | 53 | 45 | 100 | 80 | 33 | 21 | 46 |
| 5 | 100 | 0 | 22 | 25 | 53 | 46 | 120 | 80 | 36 | 27 | 37 |
| 6 | 120 | 0 | 16 | 25 | 59 | 47 | 140 | 80 | 42 | 23 | 35 |
| 7 | 140 | 0 | 24 | 28 | 48 | 48 | 160 | 80 | 40 | 23 | 37 |
| 8 | 160 | 0 | 11 | 19 | 70 | 49 | 0 | 100 | 32 | 27 | 41 |
| 9 | 180 | 0 | 11 | 21 | 68 | 50 | 20 | 100 | 33 | 25 | 42 |
| 10 | 0 | 20 | 28 | 17 | 55 | 51 | 40 | 100 | 22 | 20 | 58 |
| 11 | 20 | 20 | 33 | 27 | 40 | 52 | 60 | 100 | 32 | 27 | 41 |
| 12 | 40 | 20 | 25 | 25 | 50 | 53 | 80 | 100 | 27 | 21 | 52 |
| 13 | 60 | 20 | 32 | 27 | 41 | 54 | 100 | 100 | 36 | 23 | 41 |
| 14 | 80 | 20 | 33 | 30 | 37 | 55 | 120 | 100 | 31 | 23 | 46 |
| 15 | 100 | 20 | 36 | 24 | 40 | 56 | 140 | 100 | 36 | 22 | 42 |
| 16 | 120 | 20 | 25 | 21 | 54 | 57 | 160 | 100 | 29 | 23 | 48 |
| 17 | 140 | 20 | 27 | 25 | 48 | 58 | 0 | 120 | 36 | 27 | 37 |
| 18 | 160 | 20 | 24 | 15 | 61 | 59 | 20 | 120 | 25 | 19 | 56 |
| 19 | 180 | 20 | 14 | 26 | 60 | 60 | 40 | 120 | 25 | 25 | 50 |
| 20 | 0 | 40 | 33 | 27 | 40 | 61 | 60 | 120 | 25 | 23 | 52 |
| 21 | 20 | 40 | 16 | 19 | 65 | 62 | 80 | 120 | 31 | 25 | 44 |
| 22 | 40 | 40 | 29 | 25 | 46 | 63 | 100 | 120 | 36 | 23 | 41 |
| 23 | 60 | 40 | 38 | 25 | 37 | 64 | 120 | 120 | 34 | 23 | 43 |
| 24 | 80 | 40 | 31 | 21 | 48 | 65 | 140 | 120 | 42 | 25 | 33 |
| 25 | 100 | 40 | 27 | 19 | 54 | 66 | 0 | 140 | 36 | 20 | 44 |
| 26 | 120 | 40 | 24 | 17 | 59 | 67 | 20 | 140 | 33 | 21 | 46 |
| 27 | 140 | 40 | 27 | 19 | 54 | 68 | 40 | 140 | 30 | 25 | 45 |
| 28 | 160 | 40 | 40 | 27 | 33 | 69 | 60 | 140 | 21 | 21 | 58 |
| 29 | 180 | 40 | 32 | 27 | 41 | 70 | 80 | 140 | 34 | 23 | 43 |
| 30 | 0 | 60 | 34 | 23 | 43 | 71 | 100 | 140 | 30 | 23 | 47 |
| 31 | 20 | 60 | 32 | 27 | 41 | 72 | 120 | 140 | 32 | 23 | 45 |
| 32 | 40 | 60 | 23 | 21 | 56 | 73 | 0 | 160 | 19 | 19 | 62 |
| 33 | 60 | 60 | 29 | 23 | 48 | 74 | 20 | 160 | 23 | 19 | 58 |
| 34 | 80 | 60 | 40 | 23 | 37 | 75 | 40 | 160 | 28 | 27 | 45 |
| 35 | 100 | 60 | 33 | 21 | 46 | 76 | 60 | 160 | 25 | 23 | 52 |
| 36 | 120 | 60 | 31 | 21 | 48 | 77 | 80 | 160 | 16 | 19 | 65 |
| 37 | 140 | 60 | 22 | 17 | 61 | 78 | 100 | 160 | 23 | 21 | 56 |
| 38 | 160 | 60 | 38 | 22 | 40 | 79 | 0 | 180 | 17 | 23 | 60 |
| 39 | 180 | 60 | 30 | 23 | 47 | 80 | 20 | 180 | 32 | 25 | 43 |
| 40 | 0 | 80 | 25 | 23 | 52 | 81 | 40 | 180 | 25 | 23 | 52 |
| 41 | 20 | 80 | 33 | 25 | 42 | 82 | 60 | 180 | 30 | 23 | 47 |

## Anexo C - Código fonte das análises estatísticas de dados composicionais independentes

```
#-------------------------------------------------------------------------
# Regressão Dirichlet para dados composicionais
#-------------------------------------------------------------------------

#packages
require(compositions)
require(dirichletReg)

#dados
data(ArcticLake)
head(ArcticLake)

AL <- DR_data(ArcticLake[, 1:3], base=1)
AL[1:6, ]

lake1 <- DirichReg(AL ~ depth, ArcticLake, model="alternative")
summary(lake1)
coef(lake1)

lake2 <- DirichReg(AL ~ depth + I(depth^2), ArcticLake,  model = c("common"))
summary(lake2)
coef(lake2)  ####### coeficientes modelo quadrático

#Gráficos
Xnew <- data.frame(depth = seq(min(ArcticLake$depth),
                               max(ArcticLake$depth), length.out = 100))

# Areia
par(mfrow=c(1,3))
plot(ArcticLake$depth, as.numeric(AL[,1]), pch = 21, bg = "#E495A5",
     xlab = "Profundidade (m)", ylab = "Proporção", ylim = c(0,0.8),
     main = "Areia")
pred1_d <-matrix(predict(lake2, Xnew)[,1], 100, 1)
lines(pred1_d, col="blue", lwd = 2)
legend("top", legend=c("valor observado", "modelo Dirichlet ajustado"),
           lty=c(NA,1), col=c(1, 4), lwd=c(1, 3), bty="n", pch=c(1, NA),
           cex=(0.8))
# Silte
plot(ArcticLake$depth, as.numeric(AL[,2]), pch = 21, bg = "#E495A5",
     xlab = "Profundidade (m)", ylab = "Proporção", ylim = c(0,0.8),
     main = "Silte")
pred2_d <-matrix(predict(lake2, Xnew)[,2], 100, 1)
lines(pred2_d, col="blue", lwd = 2)
legend("top", legend=c("valor observado", "modelo Dirichlet ajustado"),
           lty=c(NA,1), col=c(1, 4), lwd=c(1, 3), bty="n", pch=c(1, NA),
           cex=(0.8))

# Argila
plot(ArcticLake$depth, as.numeric(AL[,3]), pch = 21, bg = "#E495A5",
     xlab = "Profundidade (m)", ylab = "Proporção", ylim = c(0,0.8),
     main = "Argila")
pred3_d <-matrix(predict(lake2, Xnew)[,3], 100, 1)
lines(pred3_d, col="blue", lwd = 2)
```

```
legend("top", legend=c("valor observado", "modelo Dirichlet ajustado"),
           lty=c(NA,1), col=c(1, 4), lwd=c(1, 3), bty="n", pch=c(1, NA),
           cex=(0.8))

#-------------------------------------------------------------------------------------
#       Modelo regressão com transformação alr (Aitchison, 1986)
#-------------------------------------------------------------------------------------

# packages
require(compositions)

# dados
head(ArcticLake)
attach(ArcticLake)
y <- acomp(ArcticLake[,1:3])
x <- depth
# ajuste de modelos
mylm <- lm(alr(y) ~ x + I(x^2), data=ArcticLake)
summary(mylm)
alrInv(predict(mylm)) # fç para transformar para escala original,

Xnew <- data.frame(x = seq(min(depth), max(depth), length.out = 100))

# dados com modelos ajustados
# Areia
plot(depth, as.numeric(y[,1]), pch = 21, bg = "#E495A5",
     xlab = "Profundidade (m)", ylab = "Proporção", ylim = c(0,0.8),
     main = "Areia")
pred1 <- matrix(alrInv(predict(mylm, Xnew))[,1], 100, 1)
lines(pred1, col="blue", lwd = 2)
legend("top", legend=c("valor observado", "modelo log-normal ajustado"),
           lty=c(NA,1), col=c(1, 4), lwd=c(1, 3), bty="n", pch=c(1, NA),
           cex=(0.8))

# Silte
plot(ArcticLake$depth, as.numeric(AL[,2]), pch = 21, bg = "#E495A5",
     xlab = "Profundidade (m)", ylab = "Proporção", ylim = c(0,0.8),
     main = "Silte")
pred2 <- matrix(alrInv(predict(mylm, Xnew))[,2], 100, 1)
lines(pred2, col="blue", lwd = 2)
legend("top", legend=c("valor observado", "modelo log-normal ajustado"),
           lty=c(NA,1), col=c(1, 4), lwd=c(1, 3), bty="n", pch=c(1, NA),
           cex=(0.8))

# Argila
plot(ArcticLake$depth, as.numeric(AL[,3]), pch = 21, bg = "#E495A5",
     xlab = "Profundidade (m)", ylab = "Proporção", ylim = c(0,0.8),
     main = "Argila")
pred3 <- matrix(alrInv(predict(mylm, Xnew))[,3], 100, 1)
lines(pred3, col="blue", lwd = 2)
legend("top", legend=c("valor observado", "modelo log-normal ajustado"),
           lty=c(NA,1), col=c(1, 4), lwd=c(1, 3), bty="n", pch=c(1, NA),
           cex=(0.8))
```

```
#-------------------------------------------------------------------------------------------
# Ajuste com GEE composicional = multivariado e iid
#-------------------------------------------------------------------------------------------

# packages
require(compositions) #fçs dados composicionais
require(geepack)  # univariado - especif. matriz correlacao
require(mmm)      # multivariado - especif matriz correl

# cj de dados
data(ArcticLake)
head(ArcticLake)
y <- acomp(ArcticLake[,1:3])
ArcticLake <- data.frame(depth = ArcticLake[,4], sand=y[,1], silt=y[,2], clay=y[,3])

# ajuste modelo linear independente
k <- rep(seq(1:13), times=3)
depth2 <- depth^2
fit3 <- mmm(formula= cbind(sand, silt, clay) ~ depth + depth2, id=k,
              data=ArcticLake, family = binomial("log"))
summary(fit3)

xnew <- data.frame(x =seq(min(depth), max(depth), length.out = 234))
fitted(fit3)

plot(depth, sand, pch = 21, bg = "#E495A5",
       xlab = "Profundidade (m)", ylab = "Proporção", ylim = 0:1,
       main = "Areia")
lines(depth, pred[1:39], type="l", col="darkgreen")
```

## Anexo D - Código fonte das análises estatísticas de dados composicionais com dependência espacial

```
#-------------------------------------------------------------------------
#           Análise geoestatística tradicional (clássica)
#-------------------------------------------------------------------------
# packages
require(geoR)
require(splancs)

# Arquivo de dados
dados<- read.geodata("pivo.txt", head=T, coords.col=4:5,data.col=c(1,2,3))
#dados

# Calcula as Estatísticas descritivas dos dados
# média, mediana, Q1, Q3, min, max
summary(dados$data)
var(dados$data) # Apresenta (matriz) a Variancia
max(dist(dados$coords))
min(dist(dados$coords))

#Apresenta o grafico Box-plot
boxplot(dados$data,main=' ', ylab="%")

#Histograma
hist(dados$data[,1], breaks=4, xlab='Areia',ylab='Frequência',main=' ')
hist(dados$data[,2], breaks=4, xlab='Silte',ylab='Frequência',main=' ')
hist(dados$data[,3], breaks=4, xlab='Argila',ylab='Frequência',main=' ')

#Graficos
areia <- read.geodata("pivo.txt", head=T, coords.col=4:5, data.col=1)
plot(areia)

silte <- read.geodata("pivo.txt", head=T, coords.col=4:5, data.col=2)
plot(silte)

argila <- read.geodata("pivo.txt", head=T, coords.col=4:5, data.col=3)
plot(argila)

# Grafico Post-plot para o estudo de tendência direcional
#Post Plot
par(mfrow=c(1,3))
points(areia, xlab="", ylab="",main='Areia',font.main = 3,pt.div="quartile",col=gray(seq(0.8,0,l=5)))
points(silte, xlab="", ylab="",main='Silte',font.main = 3,pt.div="quartile",col=gray(seq(0.8,0,l=5)))
points(argila, xlab="", ylab="",main='Argila',font.main = 3,pt.div="quartile",col=gray(seq(0.8,0,l=5)))


# Construcao do semivariograma
par(mfrow=c(1,1))
d.var.areia <- variog(areia, uvec=seq(1,180,l=10),estimador.type="classical", pairs.min=30,
                direction="omnidirectional",tolerance=pi/8)
plot(d.var.areia, xlab="distância", ylab="semivariância",main='Areia',font.main = 3)

d.var.silte <- variog(silte, uvec=seq(1,180,l=10),estimador.type="classical", pairs.min=30,
                direction="omnidirectional",tolerance=pi/8)
plot(d.var.silte, xlab="distância", ylab="semivariância",main='Silte',font.main = 3)
```

```
d.var.argila <- variog(argila, uvec=seq(1,180,l=10),estimador.type="classical", pairs.min=30,
                direction="omnidirectional",tolerance=pi/8)
plot(d.var.argila, xlab="distância", ylab="semivariância",main='Argila',font.main = 3)
max(dist(dados$coords))
min(dist(dados$coords))

#obtendo informaçoes sobre o semivariograma experimental
distancia <- c(d.var.areia$u, d.var.silte$u, d.var.argila$u)
semivariancia <- c(d.var.areia$v, d.var.silte$v, d.var.argila$v)
pares <- c(d.var.areia$n, d.var.silte$n, d.var.argila$n)
tabela <- cbind(distancia, semivariancia, pares)
tabela

# Grafico de envelopes

##envelopes (ic dentro do semivar - os pts fora do limite indica provável dependencia espacial)
set.seed(990)
dados.env.areia <- variog.mc.env(areia, obj.v=d.var.areia, nsim = 99)
plot(d.var.areia, env=dados.env.areia, xlab="distância", ylab="semivariância",
                  main='Areia',font.main = 3)
set.seed(990)
dados.env.silte <- variog.mc.env(silte, obj.v=d.var.silte, nsim = 99)
plot(d.var.silte, env=dados.env.silte, xlab="distância", ylab="semivariância",
                  main='Silte',font.main = 3)
set.seed(990)
dados.env.argila <- variog.mc.env(argila, obj.v=d.var.argila, nsim = 99)
plot(d.var.argila, env=dados.env.argila, xlab="distância", ylab="semivariância",
                   main='Argila',font.main = 3)

## Modelo ajustado
exp.ml <- likfit(areia, ini=c(20, 150), lik.method= "ML", cov.model="exp")
exp.ml
summary(exp.ml)
plot(d.var.areia, xlab='alcance',ylab='semivariância', main='')
lines(exp.ml,col="blue")

exp.ml.2 <- likfit(silte, ini=c(6, 150), lik.method= "ML", cov.model="sph")
exp.ml.2
summary(exp.ml.2)
plot(d.var.silte, xlab='alcance',ylab='semivariância',main='')
lines(exp.ml.2,col="blue")

exp.ml.3 <- likfit(argila, ini=c(50, 80), lik.method= "ML", cov.model="exp")
exp.ml.3
summary(exp.ml.3)
plot(d.var.argila, xlab='alcance',ylab='semivariância',main='Argila')
lines(exp.ml.3,col="blue")

# Construcao de mapa tematico

# Adicionando bordas
# Abre o aquivo bordas.txt
bor <- read.table("borda.txt", head=TRUE)
bor  # apresenta os dados da borda
plot(bor)
```

```
polygon(bor)

apply(bor,2,range) #Mostra o min e max das coords
gr<-expand.grid(x=seq(0,180,by=1), y=seq(0,180,by=1))
plot(gr)
points(areia, pt.div="equal", pch.seq = 20) #monta o grid de interpolacao

gi<- polygrid(gr,bor=bor)
points(gi, pch="+", col=2) #o novo grid considerando apenas a regiao limitada
plot(gi)
par(mfrow=c(1,3))

# Areia
KC = krige.control(obj=exp.ml,lam=1) #krigagem
d.k = krige.conv(areia, loc=gi, krige=KC) #Faz o mapa da vari?vel d por krigagem ordin?ria
image(d.k, loc=gr, border=bor, col=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 210), x.leg=c(100,195),
      y.leg=c(165,185),zlim=range(d.k$predict))
title(main = "Areia", font.main=4)

# Silte
KC2 = krige.control(obj=exp.ml.2,lam=1) #krigagem
d.k2 = krige.conv(silte, loc=gi, krige=KC2) #Faz o mapa da vari?vel d por krigagem ordin?ria
image(d.k2, loc=gr, border=bor, ccol=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 210), x.leg=c(100,195),
      y.leg=c(165,185),zlim=range(d.k2$predict))
title(main = "Silte", font.main=4)

# Argila
KC3 = krige.control(obj=exp.ml.3,lam=1) #krigagem
d.k3 = krige.conv(argila, loc=gi, krige=KC3) #Faz o mapa da vari?vel d por krigagem ordin?ria
image(d.k3, loc=gr, border=bor, col=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 210), x.leg=c(100,195),
      y.leg=c(165,185),zlim=range(d.k3$predict))
title(main = "Argila", font.main=4)

#-----------------------------------------------------------------------------
#     Análise geoestatística baseada em semivariograma - Pawlowsky-Glahn
#     p.200 book: Analysing Compositional with R (2013)
#-----------------------------------------------------------------------------
# packages
require(geoR)
require(compositions)
require(gstat)
require(splancs)
require(lattice)
require(sp)

# Arquivo de dados
dados<- read.geodata("pivo.txt", head=T, coords.col=4:5,data.col=c(1,2,3))
dados

X <- dados$coords  # coordenadas
max(dados$coords)

Z <- acomp(dados$data)   # dados composicionais

lrv <- logratioVariogram(Z, X, dists=seq(20,max(dist(dados$coords))*0.7,
                                         length.out=12+1))  # variogram
```

```
# Plot semivariogram
opar <- par(cex=1.25)
plot(lrv, type="l")
par(opar)

lrvModel <- CompLinModCoReg(~nugget()+R1*exp(80),Z)  # Modelo exponencial

vgmModel <- vgmFit2lrv(lrv, lrvModel, print.level=0)
vgmModel

plot(lrv,lrvg=vgram2lrvgram(vgmModel$vg)) # Plota o modelo ajustado no variograma

# Abre o aquivo bordas.txt
#bor <- read.table("borda.txt", head=TRUE)
#bor  # apresenta os dados da borda
## Border
nn = 100
bor <- cbind(c(0,seq(0, 200, l=nn), 0), c(0,sqrt(200^2-seq(0, 200, l=nn)^2), 0))

plot(bor)
polygon(bor)
apply(bor,2,range) #Mostra o m?nimo e m?ximo das coordenadas

gr<-expand.grid(x=seq(0,180,by=1), y=seq(0,180,by=1))
plot(gr)
gi<- polygrid(gr,bor=bor)
points(gi, pch="+", col=2) #o novo grid considerando apenas a regi?o limitada

KP <- compOKriging(Z,X,gi,vg=vgmModel$vg) #Krigagem
names(KP)

### Mapa temático
par(mfrow=c(1,1))

# Areia
levelplot(Z[,"Areia"]~X[,1]+X[,2], KP, col.regions=bpy.colors(15), xlim=c(0, 183), ylim=c(0, 183),
          ylab=("Y Coord"), xlab=("X Coord"), zlim=range(KP$Z[,1]))
# Silte
levelplot(Z[,"Silte"]~X[,1]+X[,2],KP,col.regions=bpy.colors(15), xlim=c(0, 183), ylim=c(0, 183),
          ylab=("Y Coord"), xlab=("X Coord"), zlim=range(KP$Z[,1]))
# Argila
levelplot(Z[,"Argila"]~X[,1]+X[,2],KP,col.regions=bpy.colors(15), xlim=c(0, 183), ylim=c(0, 183),
          ylab=("Y Coord"), xlab=("X Coord"), zlim=range(KP$Z[,1]))


#-----------------------------------------------------------------------------
#     Análise geoestatística com transformação - Modelo normal bivariado
#     Martins (2009)
#-----------------------------------------------------------------------------

# packages
require(statmod)
require(compositions)
require(geoR)
require(mvtnorm)
```

```
require(bbmle)
require(numDeriv)
# Extra functions
source("functions_compositional.r")
source("derivada_compositional.r")

## Loading the data set
load("dados.rdata")
load("pivo.rdata")

## Data manipulation
dados <- pivo[,c(6,7,8,1,2)]
dados.geo <- as.geoComp(dados)
gride <- dados.geo[[3]]
Y = c(dados.geo[[1]][,1],dados.geo[[1]][,2])

## Distance matrix
U <- dist(gride,diag=TRUE, upper=TRUE)

## Exploratory data analysis
plot.descritiva(dados)

## Fitting the compositional geostatistical model
# Initial values
ini <- inicial(dados.geo[[1]], U)

modelo <- mle2(log.Vero,start=list(s1= ini[1], s2= ini[2], t1= ini[3], t2= ini[4], phi=ini[5], rho=ini[6]),
               method="L-BFGS-B",control=list(ndeps=c(1e-03,1e-03,1e-03,1e-03,1e-10,1e-10)),
               gr = escore,
               lower=list(s1=1e-10,s2=1e-10,t1=1e-5,t2=1e-5, phi = 1e-10, rho=-0.9999),
               upper=list(s1=Inf,s2=Inf,t1=Inf,t2=Inf,phi=Inf, rho=0.9999),
               data=list(Y= Y, U = U))

## Building the Observed Information matrix
pontual <- coef(modelo)
round(pontual,4)
log.Vero2 <- function(par, Y, U){log.Vero(s1=par[1],s2=par[2],t1=par[3],t2=par[4],phi = par[5],
             rho = par[6],Y=Y, U = U)}
tt <- hessian(func = log.Vero2, x = coef(modelo), method="Richardson",
              method.args = list(eps=1e-4, d=0.0001, zero.tol=sqrt(.Machine$double.eps/7e-7),
              r=4, v=2,show.details=TRUE), Y = Y, U = U)
VCOV <- solve(tt)
std.error <- sqrt(diag(VCOV))
round(std.error,4)

## Spatial prediction
Sigma <- monta.sigma(s1=coef(modelo)[1],s2=coef(modelo)[2],t1=coef(modelo)[3],t2=coef(modelo)[4],
                     phi=coef(modelo)[5], rho=coef(modelo)[6],U = U)
n <- dim(as.matrix(U))[1]
X <- as.factor(c(rep(1,n),rep(2,n)))
D <- model.matrix(~ X -1)
DtS1 <- t(solve(Sigma,D))
DtS1D <- DtS1%*%D
DtS1Y <- DtS1%*%Y
beta.chapeu <- solve(DtS1D,DtS1Y)
esti.par <- c(beta.chapeu,coef(modelo))
```

```
# Krigagem
## Border
nn = 100
bor <- cbind(c(0,seq(0, 200, l=nn), 0), c(0,sqrt(200^2-seq(0, 200, l=nn)^2), 0))

# Grid resolution
# resol = 4 We have used a high resolution
resol = 20
gr <- pred_grid(bor, by=resol)
md.cov.ck <- cokrigagem(esti.par, loc=gr, dados.comp=dados.geo)
#preditos.gh <- volta.quad(md.cov.ck,n.pontos=7,Variancia=FALSE) ## Spent 4 hrs for high resolution
#preditos.simu <- volta.cokri(md.cov.ck,num.simu=500,retorna.tudo=FALSE,int.conf=0.95) # Spent hours
load("preditosgh.rdata")
load("preditossimul.rdata")

#xleg <- c(85, 205); yleg <- c(185, 200)
#par(mfrow=c(3,3), mar=c(1.5,7,2,.5), mgp=c(1.7,0.5, 0))

## Prediction maps - Gauss Hermite method
gr <- pred_grid(bor, by=4)
image(structure(list(predict=preditos.gh[[1]][,1]), class="kriging"),
      loc=gr, bor, col=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 230), x.leg=c(30,199),
      y.leg=c(195,211), xlab='X Coord', ylab='Y Coord')

## Silte
image(structure(list(predict=preditos.gh[[1]][,2]), class="kriging"),
      loc=gr, bor, col=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 230), x.leg=c(30,199),
      y.leg=c(195,211), xlab='X Coord', ylab='Y Coord')

## Argila
image(structure(list(predict=preditos.gh[[1]][,3]), class="kriging"),
      loc=gr, bor, col=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 230), x.leg=c(30,199),
      y.leg=c(195,211), xlab='X Coord', ylab='Y Coord')

#-----------------------------------------------------------------------------
#     Estimação por equações generalizadas - GEE
#-----------------------------------------------------------------------------

# packages
require(compositions)
require(mmm)
require(JGEE)

# Arquivo de dados
dados<- read.geodata("pivo.txt", head=T, coords.col=4:5,data.col=c(1,2,3))
dados

resposta <- acomp(dados$data)

X <- as.matrix(rep(c(1),n),1)   #Vetor de um's para o caso sem covariaveis
I <- diag(1,n,n)                #Matriz identidade
H <- as.matrix(dist(dados$coords, method="euclidean",diag=TRUE, upper=TRUE)) #Matriz de distancias

# estimativas
initial <- mmm(formula=resposta ~ X,
```

```
             data= dados, family = poisson("log"))
summary(initial)

FunList <- list(phi, dist, tau, sigma)

# especifica matriz de correlacao
## Modelo Exponencial
cor = function(phi, H) # The function to maximize.
{ R1 <- exp(-(H/phi))
        return(R1)}

#1ª derivada
cor_der = function(tau, sigma) # The function to maximize.
{    d_phi <- tau*I + sigma*(R1%*%(H/(phi)^2))
     return(d_phi)}

fit <- JGee1(resposta ~ X, data=dados, nr = lenght(dados$data),
             family = family = poisson("log"), corstr2 = "fixed",
             beta_int = initial, R2 = FunList, silent = TRUE)
summary(fit)

# variograma
# areia
plot(variog(data.geo, uvec=seq(1,180,l=10),estimador.type="classical", pairs.min=30,
                direction="omnidirectional",tolerance=pi/8))
# silte
plot(variog(data.geo, coords = dados$coords, uvec=seq(1,180,l=10),estimador.type="classical", pairs.min=30,
                direction="omnidirectional",tolerance=pi/8))
# Argila
plot(variog(data.geo, coords = dados$coords, uvec=seq(1,180,l=10),estimador.type="classical", pairs.min=30,
                direction="omnidirectional",tolerance=pi/8))

## Mapas
# Abre o aquivo bordas.txt
bor <- read.table("borda.txt", head=TRUE)
bor  # apresenta os dados da borda
plot(bor)
polygon(bor)
apply(bor,2,range) #Mostra o m?nimo e m?ximo das coordenadas

gr<-expand.grid(x=seq(0,180,by=1), y=seq(0,180,by=1))
plot(gr)
gi<- polygrid(gr,bor=bor)
points(gi, pch="+", col=2) #o novo grid considerando apenas a regi?o limitada

## KRIGAGEM
intercept <- fit[[1]]
sigma <- fit[[2]]
phi <- fit[[3]]
nugget <- fit[[4]]

# areia
KC= krige.control(beta=intercept[,1], cov.model="exponential", cov.pars=c(sigma[,1], phi[,1]),
            nugget=nugget[,1])
d.k= krige.conv(dados, coords=dados$coords, data=resposta[,1],
           locations=gi, borders=bor, krige=KC)
```

```
image(d.k, loc=gr, border=bor, col=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 230), x.leg=c(30,199),
      y.leg=c(195,211),zlim=range(d.k$predict))
title(main = "Areia", font.main=4)

# silte
KC2= krige.control(beta=intercept[,2], cov.model="exponential", cov.pars=c(sigma[,2], phi[,2]),
           nugget=nugget[,2])
d.k2= krige.conv(dados, coords=dados$coords, data=resposta[,2],
          locations=gi, borders=bor, krige=KC2)
image(d.k2, loc=gr, border=bor, col=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 230), x.leg=c(30,199),
      y.leg=c(195,211),zlim=range(d.k2$predict))
title(main = "Silte", font.main=4)

# argila
KC3= krige.control(beta=intercept[,3], cov.model="exponential", cov.pars=c(sigma[,3], phi[,3]),
           nugget=nugget[,3])
d.k3= krige.conv(dados, coords=dados$coords, data=resposta[,3],
          locations=gi, borders=bor, krige=KC3)
image(d.k3, loc=gr, border=bor, col=bpy.colors(15), xlim=c(0, 200), ylim=c(0, 230), x.leg=c(30,199),
      y.leg=c(195,211),zlim=range(d.k3$predict))
title(main = "Argila", font.main=4)
```