# Análise de Dados Composicionais Via Árvores de Regressão

Ana Beatriz Tozzo Martins - PPGMNE/UFPR-DES/UEM
Cesar Augusto Taconeli - DEST/UFPR
Paulo Justiniano Ribeiro Junior - LEG/UFPR
Antônio Carlos Andrade Gonçalves - DAG/UEM.

4 de fevereiro de 2009

# Roteiro de Apresentação

1. Introdução
2. Dados Composicionais
3. CART
4. Metodologia
5. Resultados
6. Conclusão

# Introdução

**Dados composicionais:**

- Ciências da Terra: dados expressos como frações ou porcentagens.

Aitchison (1986).

- Exemplos:
  - textura de solos;
  - composição química de uma rocha;
  - estruturas de dados resultantes de algoritmos de classificação.

Walvoort, D. J. J. e Gruijter, J.J. (2001).

# Introdução

- **Dados Composicionais:** Aitchison (1986);
- **Análise Geoestatística de Dados Composicionais:** Pawlowsky-Glahn e Olea (2004);
- **Inferência Bayesiana de Dados Composicionais Sem Efeito Espacial:** Obage (2007);
- **Inferência Bayesiana Espacial:** Tjelmeland e Lund (2003);

# Introdução

**CART -Classification and Regression Trees:**

- Modelagem não paramétrica de uma variável resposta categorizada (classificação) ou numérica (regressão) com base em um conjunto de covariáveis e interações entre as mesmas;

Breiman et al. (1984).

# Introdução

- **Árvores de Classificação e Regressão - CART:** Breiman et al. (1984);
- **CART para Análise de Dados Multivariados:** Segal (1992), Zhang (1998), De'Ath (2002) e Lee (2005), Taconeli (2008).

# Introdução

- Modelar dados composicionais via CART
  - extensão da proposta apresentada em Taconeli (2008);
  - a distância de Aitchison, no lugar dos coeficientes de dissimilaridades.

# Dados Composicionais

- Butler e Glasbey (2008): Registram informação sobre frequências relativas associadas a diferentes componentes de um sistema.
- Aitchison (1986):
  - a. Vetores cujos elementos são proporções de algum todo.
  - b. **Composição**: Vetor $\underline{Y} = (Y_1, Y_2, ..., Y_B)'$ satisfazendo:
    - $Y_1 \geqslant 0, \ldots, Y_B \geqslant 0$;
    - $Y_1 + Y_2 + \ldots + Y_B = 1$.
  - c. **Espaço Amostral**:
    $\mathbb{S}^B = \{\underline{Y} \in \mathbb{R}^B;\ Y_i > 0,\ i = 1, ..., B;\ \underline{j}'\underline{Y} = 1\}$

# Dados Composicionais

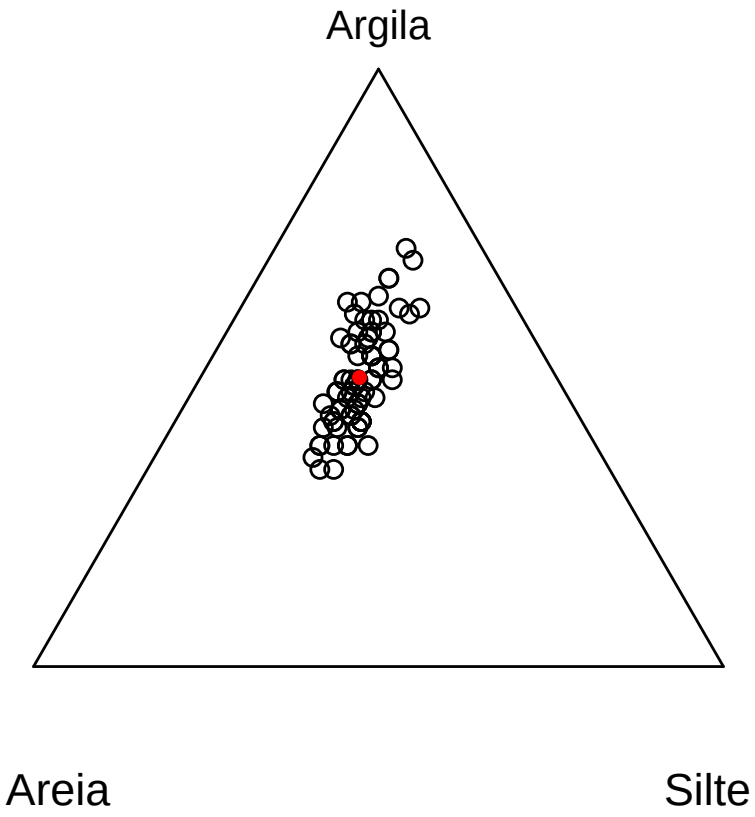



Figura: Diagrama ternário das porcentagens de areia, silte e argila.

# Dados Composicionais

**Base**: Vetor $\underline{W}(\underline{x})$, $\underline{x} \in \Omega \subset \mathbb{R}^n$ com componentes medidos na mesma escala e positivos.

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

**Operador fechamento:** Base $\Rightarrow$ Composição

$$\begin{array}{rccl} \mathcal{C}: & \mathbb{R}^B_+ & \longrightarrow & \mathbb{S}^B \\ & \underline{W}(\underline{x}) & \longrightarrow & \mathcal{C}[\underline{W}(\underline{x})] = \dfrac{\underline{W}(\underline{x})}{\underline{j}'\underline{W}(\underline{x})}, \quad \underline{j}' \text{ vetor de } 1's. \end{array}$$

# Dados Composicionais

Aitchison (1986), Aitchison e Egozcue (2005), Tolosana-Delgado et al. (2005):

Operações que definem uma estrutura de espaço vetorial de dimensão $B - 1$ no simplex.

- 1. **Perturbação** ($\oplus$) : operação comutativa;
- 2. **Potência** ($\odot$) : produto externo.

# Dados Composicionais

**Perturbação**: Multiplicação de composições componente a componente e divisão de cada componente pela soma de todos.

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

# Dados Composicionais

- **Potência**: análogo à multiplicação por um escalar no espaço real.

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

- Vetor de **diferenças** composicionais:

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

# Dados Composicionais

- **Centro**:

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

  - $g(Y_i)$ - média geométrica do $i$-ésimo componente
  - $g_s = g(Y_1) + g(Y_2) + \ldots + g(Y_B)$.

# Dados Composicionais

Efeito de **correlação espúria** (Pawlowsky e Olea, 2004):

- Covariâncias sujeitas à controles não estocásticos

  $\Downarrow$

  interpretação errônea da estrutura de covariância espacial;

- Singularidade da matriz de covariância de uma composição.

# Dados Composicionais

- **Graf (2006):** Soma constante $\Rightarrow$ correlação negativa entre os componentes.

- **Aitchison (1986):**

  - Propõe transformação que generaliza a transformação logística $\ln \frac{Y}{1-Y}$ para um vetor composicional de 2 partes;

  - Magnitudes relativas ou razões $\Rightarrow$ tratabilidade e interpretação estatística.

- **Transformações logísticas:** Aitchison (1982), Aitchison et al. (2000), Odeh et al. (2003).

# Dados Composicionais

**Transformação razão log-aditiva** (ALR):

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

Pawlowsky et al. (1995), Pawlowsky e Olea (2004).

# Dados Composicionais

- **Produto interno**:

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

- **Distância de Aitchison**:

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

# CART

**CART -Classification and Regression Trees:**

- Modelagem não paramétrica;
- Execução de sucessivas partições binárias de uma amostra, buscando a constituição de sub-amostras menos heterogêneas.
- Variável dependente:
  - Numérica – Árvore de Regressão
  - Categórica – Árvore de Classificação

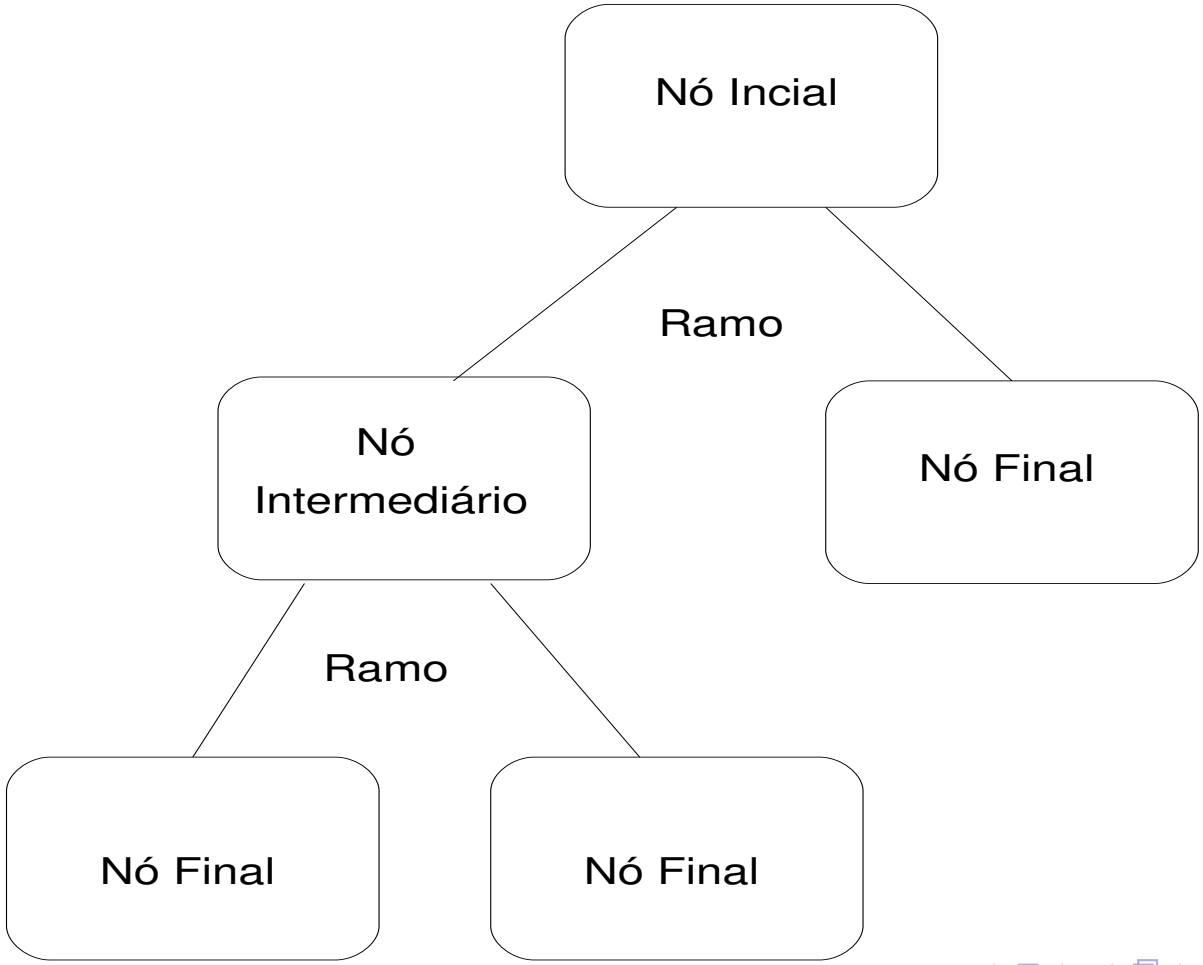
# CART

**Atrativos**

- Procedimento de simples aplicação;
- Possibilidade de modelar dados com estruturas complexas:
  - Dados desbalanceados;
  - Dados faltantes;
  - Grande número de variáveis independentes.
- Detecção de interações de ordens elevadas;
- Ausência de pressuposições paramétricas;
- Produção de resultados facilmente interpretáveis.

# CART

## Representação



# CART

## Construção do Modelo

- Partição dos nós;
  - Minimizar a heterogeneidade dos nós produzidos;
  - Baseada em uma medida de impureza.
- Poda;
  - Obtenção de uma seqüência aninhada de árvores.
- Seleção do modelo;
  - Baseada em alguma medida de qualidade preditiva.
- Caracterização dos nós finais.
  - Segundo a distribuição dos resultados em cada nó.

# Metodologia

- Dados: Gonçalves (1997)
- CART - Extensão multivariada: Taconeli (2008).

# Metodologia

- Integração das metodologias:
  - Modelagem dos dados composicionais por meio de árvores de regressão considerando a distância de Aitchison como medida de impureza e de qualidade preditiva na construção dos modelos.

- Seja $d(\underline{Y}_k, \underline{Y}_{k'})$ a distância de Aitchison calculada para duas composições $k$ e $k'$.

- **Medida de impureza** de um nó $t(\phi_{Dis}(t))$:

$$\phi_{Dis}(t) = \left(\frac{n_t(n_t-1)}{2}\right)^{-1} \sum_{k=1}^{n_t} \sum_{k<k'} d(\underline{Y}_k, \underline{Y}_{k'})$$

sendo $n_t$ o número de composições em $t$.

# Metodologia

- **Medida de qualidade de predição**:

$$\phi_{Dis}(\underline{Y}^*) = \sum_{k \subset t} \frac{d(\underline{Y}^*, \underline{Y}_k)}{n_t}.$$

- **Análise Fatorial**: estimação das cargas fatorais e escores por componentes principais - mínimos quadrados ordinários com rotação varimax.

- Estimativas dos escores fatoriais considerados **covariáveis** no modelo de regressão por árvores.

# Resultados

Cargas fatoriais

| Variável | F1 | F2 | F3 | Comunalidade |
|---|---|---|---|---|
| Ph-CaCl2 | 0,876 | | | 0,85 |
| Matéria orgânica | | −0,848 | | 0,77 |
| Fósforo | | −0,711 | | 0,61 |
| Potássio | | −0,531 | | 0,36 |
| Cálcio | 0,806 | | | 0,82 |
| Magnésio | 0,783 | | | 0,83 |
| Hidrogênio+Alumínio | −0,873 | | | 0,79 |
| Densidade global | | | 0,765 | 0,75 |
| Densidade da partícula | | | −0,807 | 0,68 |
| Porosidade total | | | −0,965 | 0,98 |
| Altura do terreno | | −0,681 | | 0,70 |
| Var. Acum | 0,29 | 0,52 | 0,74 | |

# Resultados

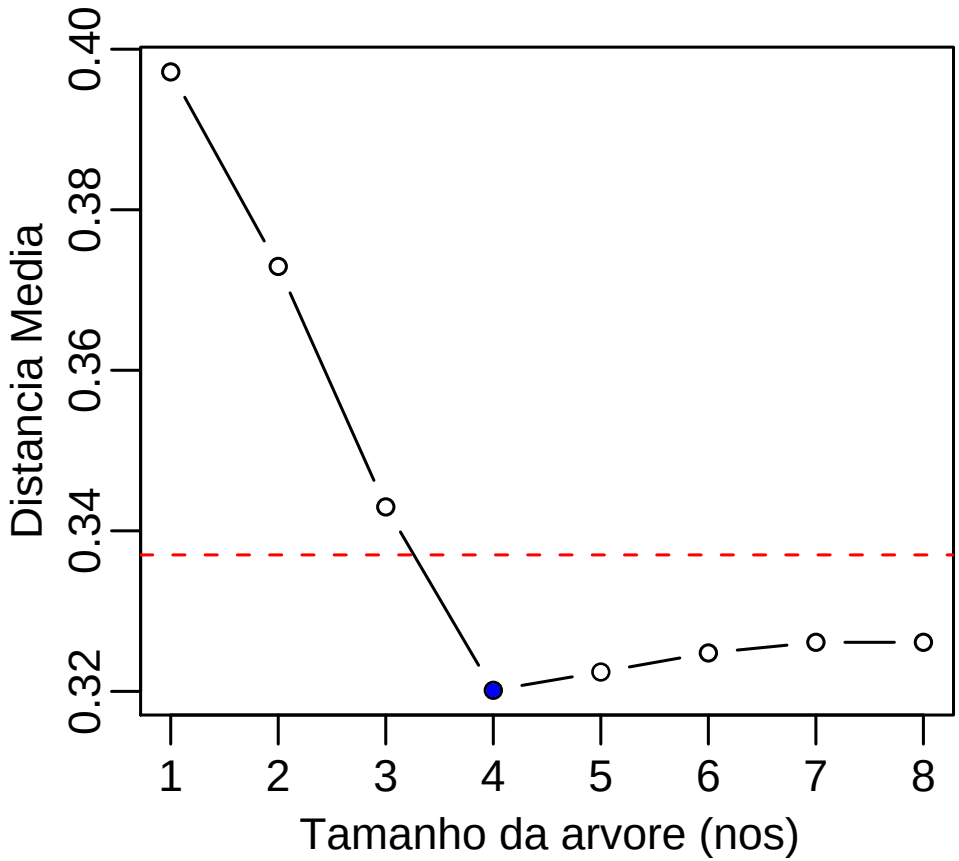



Figura: Curva de custo-complexidade.

# Resultados

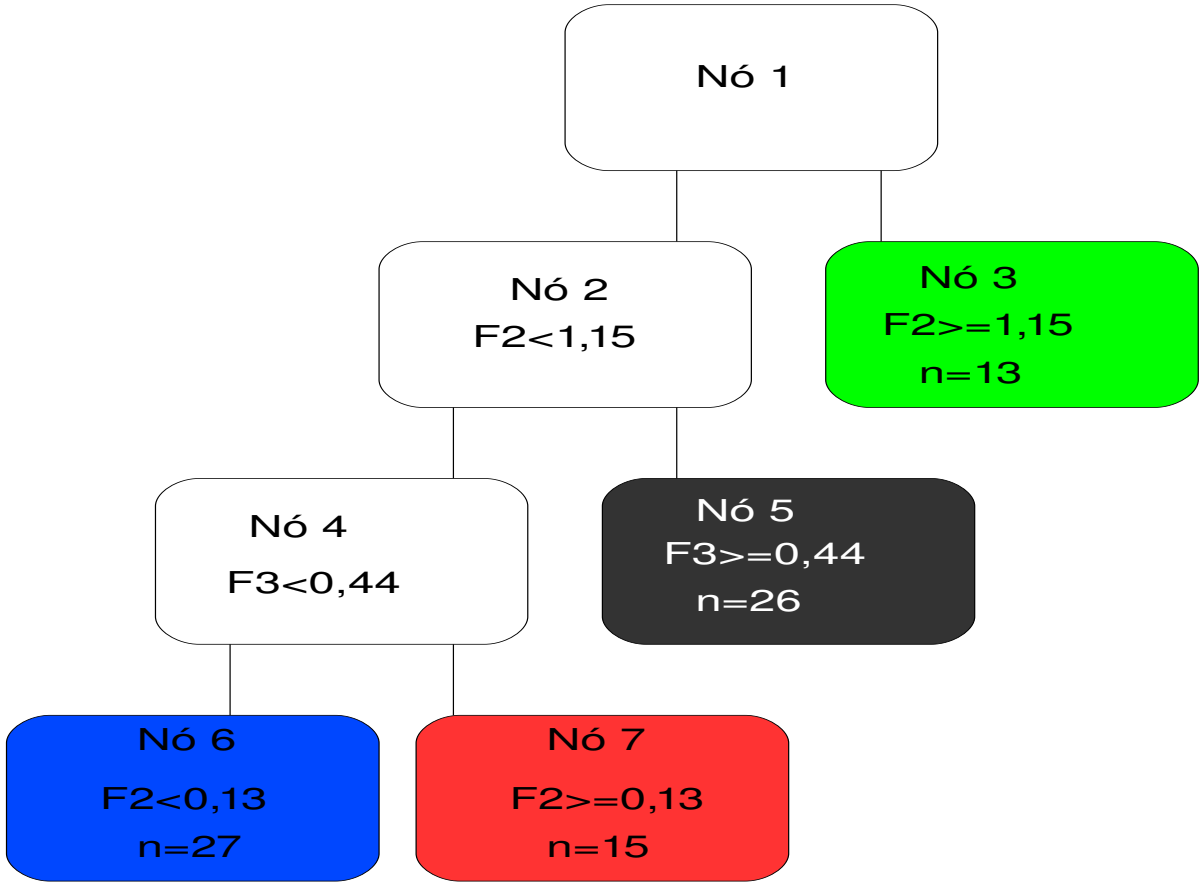



Figura: Árvore de regressão.

# Resultados

- Fatores considerados na construção da árvore: 2 e 3;
- Associação das variáveis matéria orgânica, fósforo, potássio, altura do terreno (Fator 2), densidade global, densidade da partícula e porosidade total (Fator 3) com a composição do solo;
- Variáveis do Fator 1: Ph-Cacl2, cálcio, magnésio, hidrogênio+alumínio não estão associadas à composição do solo.

# Resultados

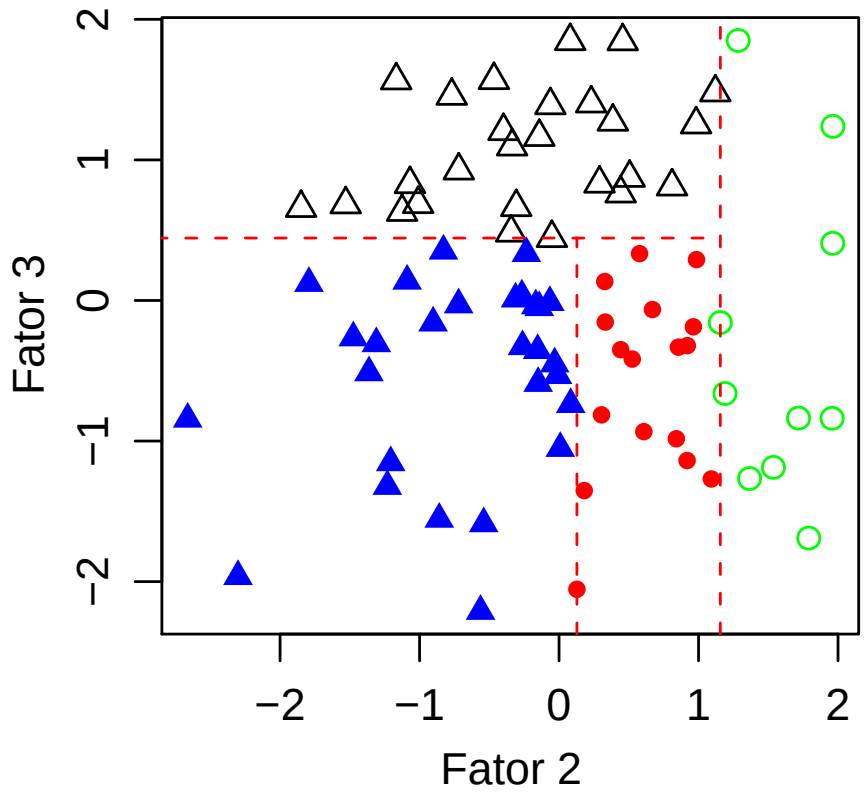



Figura: Gráfico de dispersão dos escores fatoriais para o segundo e terceiro fatores.

# Resultados

**Caracterização dos nós quanto às covariáveis:**

- **Pontos verdes - nó 3:** Matéria orgânica, fósforo, potássio e altura do terreno em quantidade pequena nas amostras de solo;
- **Pontos azuis - nó 6:** Elevados teores de matéria orgânica, fósforo, potássio, elevada altura do terreno em detrimento a baixa densidade global e altas densidade de partícula e porosidade total;

# Resultados

**Caracterização dos nós quanto às covariáveis:**

- **Pontos vermelhos - nó 7:** Baixas quantidades de matéria orgânica fósforo, potássio e baixa altura do terreno em relação à altas dendidade de partículas e porosidade total mas elevada densidade global;
- **Pontos pretos - nó 5:** Alta densidade global em detrimento à baixas densidade de partícula e porosidade total.

# Resultados

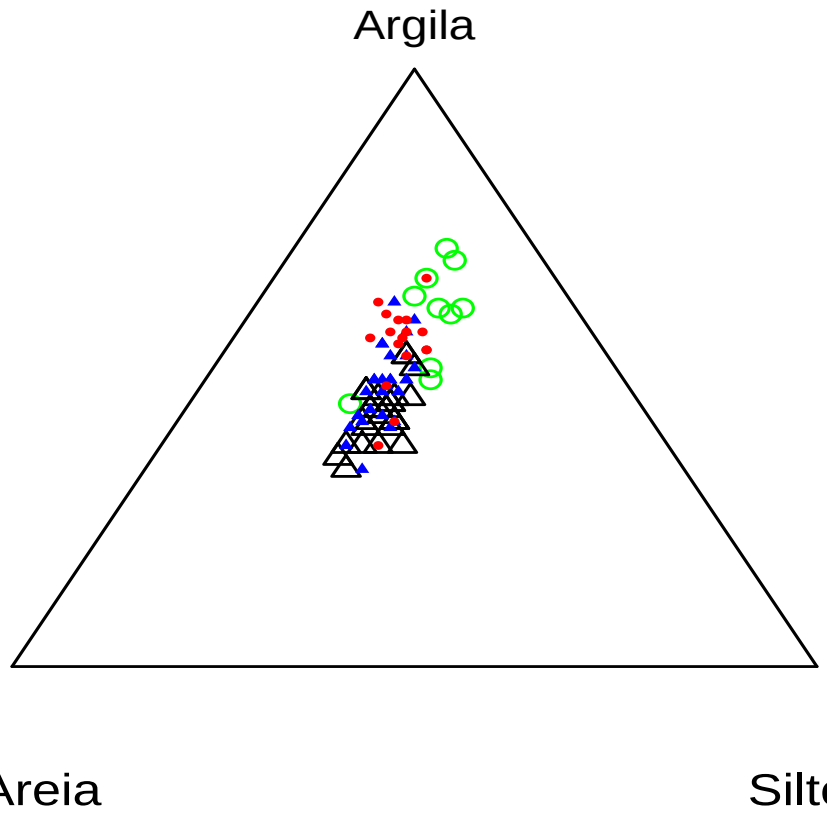



Figura: Diagrama ternário das porcentagens de areia, silte e argila.

# Resultados

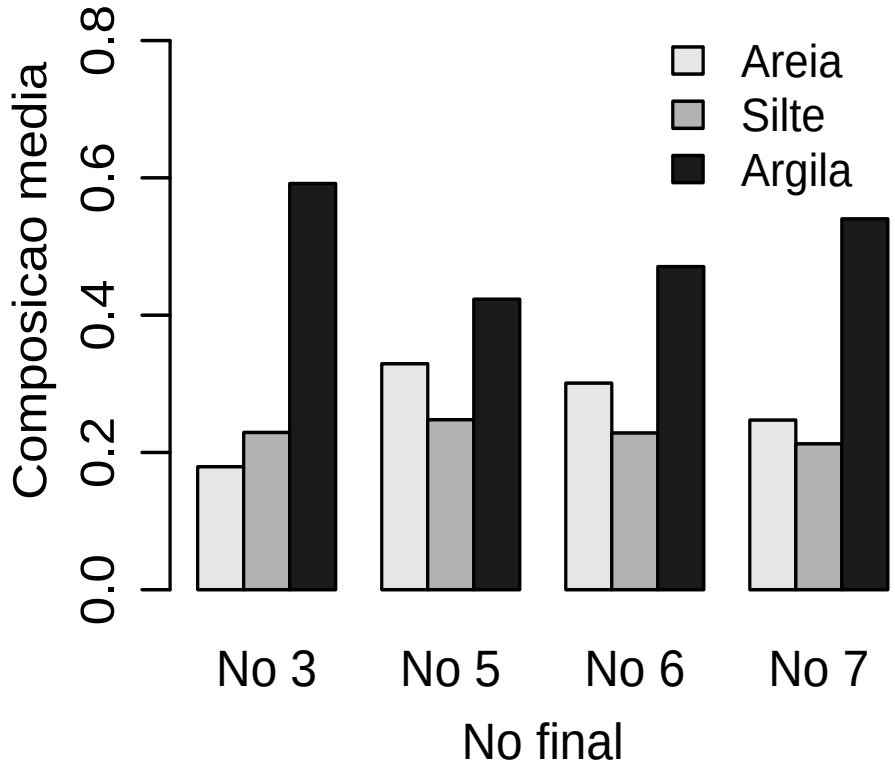



Figura: Distribuição da composição média segundo os nós.

# Resultados

**Caracterização dos nós finais quanto à composição:**

- Nó 3, pontos verdes, representam composições com maiores teores de argila e mais silte do que areia dentre todos os nós;
- Nó 5, pontos pretos, representam composições com maior equilíbrio entre os componentes. Nó com composições menos argilosas em relação aos outros;
- Nó 6, pontos azuis, não se destaca, exceto por grande quantidade de argila. Seria uma composição intermediária;
- Nó 7, pontos vermelhos, com exceção do nó 3 é composto por composições mais argilosas.

# Resultados

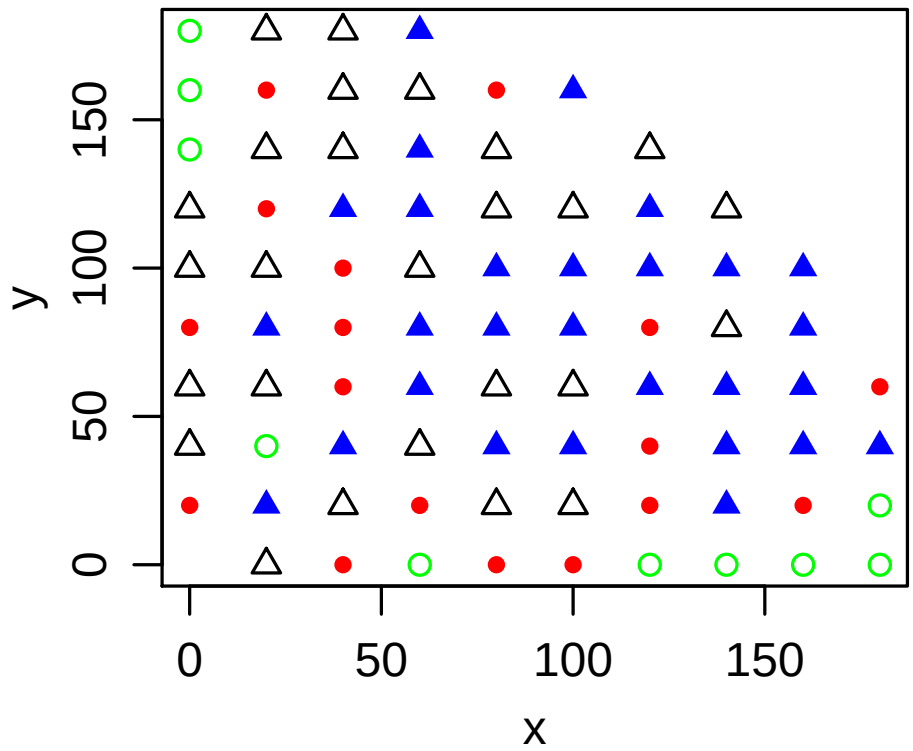



Figura: Localização espacial dos pontos amostrais em que os símbolos representam os grupos de frações granulométricas identificados pela análise

# Conclusão

Resultados produzidos permitiram identificar propriedades do solo associadas às composições, estabelecendo hierarquia entre as variáveis físico-químicas na explicação das frações granulométricas.

# Bibliografia


- AITCHISON, J. The statistical analysis of compositional data. **Journal of the Royal Statistical Society**, **Series B**, v. 44, n.2, p. 139-177, 1982.
- AITCHISON, J. **The statistical analysis of compositional data**. New Jersey: The Blackburn Press, 1986.
- BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. California: Wadsworth International Group, 1984. 358p.
- BUTLER, A.; GLASBEY, C. A latent Gaussian model for compositional data with zeros. **Journal of the Royal Statistical Society**, **Series C**, v.57, n.5, p.505-520, 2008.
- DE'ATH, G. Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. **Ecology**, Brooklin, v.83, n.4, p.1105–1117, 2002.

# Bibliografia


- GONÇALVES, A. C. A. **Variabilidade espacial de propriedades físicas do solo para fins de manejo da irrigação**. 1997. 119p. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz". Universidade de São Paulo, Piracicaba.

- GRAF, M. Precision of compositional data in a stratified two-Stage cluster sample: comparison of the swiss earnings structure survey 2002 and 2004. **Survey Research Methods Section, ASA** , Session 415: Sample Survey Quality V, p.3066–3072, 2006. Disponível em: <http://www.amstat.org/sections/SRMS/proceedings/y2006/Files/JSM2006-000771.pdf>. Acesso em: 18/05/08.

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

- LEE, S. K. On generalized multivariate decision tree by using GEE. **Computational Statistics & Data Analysis**, Amsterdan, v.49, n.4, p.1105–1119, 2005.

# Bibliografia


- OBAGE, S. C. **Uma análise bayesiana para dados composicionais**. 2007. 69p. Dissertação (Mestrado em Estatística) - Universidade Federal de São Carlos, São Carlos.
- PAWLOWSKY-GLAHN, V.; OLEA, R. A. **Geostatistical analysis of compositional data**. New York: Oxford University Press, Inc., 2004.
- R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria, 2008. Disponível em: http://www.R-project.org. Acesso em: 28 nov. 2008.
- SEGAL, M. R. Tree-structured methods for longitudinal data. **Journal of the American Statistical Association**, Boston, v.87, p.407–418, 1992.

# Bibliografia


- TACONELI, C. A. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia**. 2008. 99p. Tese (Doutorado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.
- TJELMELAND, H.; LUND, K. V. Bayesian modelling of spatial compositional data. **Journal of Applied Statistics**, v.30, n.1, p.87–100, 2003.
- ZHANG, H. P. Classification trees for multiple binary responses, **Journal of the American Statistical Association**, Boston, v.93, p.180–193, 1998.
- WALVOORT, D. J. J.; GRUIJTER, J. J. Compositional kriging: A spatial interpolation method for compositional data. **Mathematical Geology**, v.33, n.8, p. 951-966, nov 2001.

**OBRIGADA PELA ATENÇÃO!**

# Agradecimentos


- UEM/DES - Universidade Estadual de Maringá/Departamento de Estatística
- PPGMNE - Programa de Pós-Graduação em Métodos Numéricos em Engenharia
- LEG - Laboratório de Estatística e Geoinformação
- CNPQ - Conselho Nacional de Desenvolvimento Científico e Tecnológico