Análise de Dados Composicionais Via Árvore de Regressão

Ana Beatriz Tozzo Martins[1] **Cesar Augusto Taconeli**[2]
Paulo Justiniano Ribeiro Jr[3] **Antônio Carlos Andrade Gonçalves**[4]
[1]PPGMNE/UFPR- Departamento de Estatística, UEM; [2]Departamento de Estatística, UFPR;
[3]Laboratório de Estatística e Geoinformação, LEG- DES/UFPR; [4]Departamento de Agronomia, UEM
abtmartins@uem.br

1. Introdução

1.1 Dados composicionais

- Ciências da Terra: É comum os dados serem expressos como frações ou porcentagens.

  Aitchison (1986).

- Exemplo: textura de solos ou granulometria que se refere a proporção de areia, silte e argila do solo.

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

- Literatura:

  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).

- Composição: Vetor $\underline{Y} = (Y_1, Y_2, ..., Y_B)'$ satisfazendo:

  $Y_1 \geqslant 0, \ldots, Y_B \geqslant 0$;
  $Y_1 + Y_2 + \ldots + Y_B = 1$.

  Espaço Amostral:
  $\mathbb{S}^B = \{\underline{Y} \in \mathbb{R}^B;\ Y_i > 0,\ i = 1, ..., B;\ \underline{j}'\underline{Y} = 1\}$

- 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

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

- Operações que definem uma estrutura de espaço vetorial de dimensão $B-1$ no simplex: Perturbação e Potência.
- Transformação razão log-aditiva (ALR):

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

Pawlowsky e Olea (2004).

- Distância de Aitchison:

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

1.2 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).

- Literatura:

  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).

- Execução de sucessivas partições binárias de uma amostra, buscando a constituição de sub-amostras menos heterogêneas.

1.2.1 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.

1.3 Objetivo

Modelar dados composicionais via CART por meio de uma extensão da proposta apresentada em Taconeli (2008), considerando a distância de Aitchison ao invés de dissimilaridades.

2. Metodologia

- Dados: Gonçalves (1997), ESALQ-USP.
- CART - Extensão multivariada: Taconeli (2008), ESALQ-USP.
- Integração das metodologias:

  Utilização da 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.

- **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 incorporadas como **covariáveis** no modelo de regressão por árvores.

3. Resultados

Tabela 1: *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	

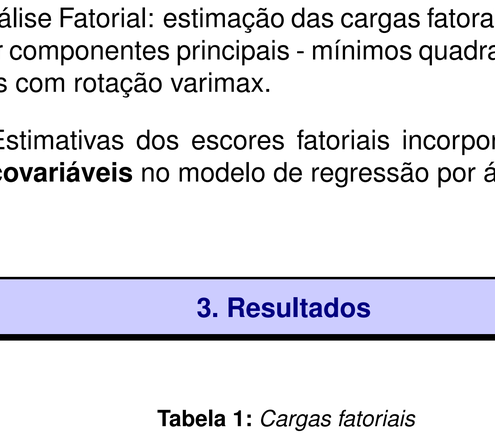

Figura 1: *Árvore de regressão.*

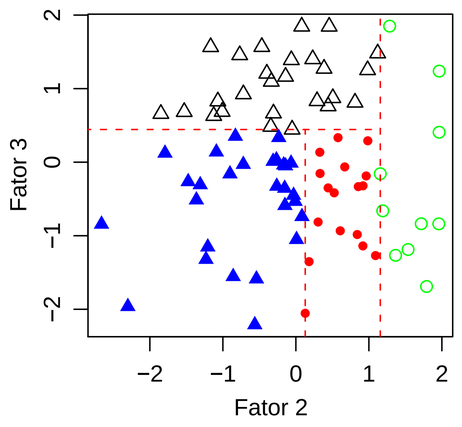

Figura 2: *Gráfico de dispersão dos escores fatoriais para o segundo e terceiro fatores.*

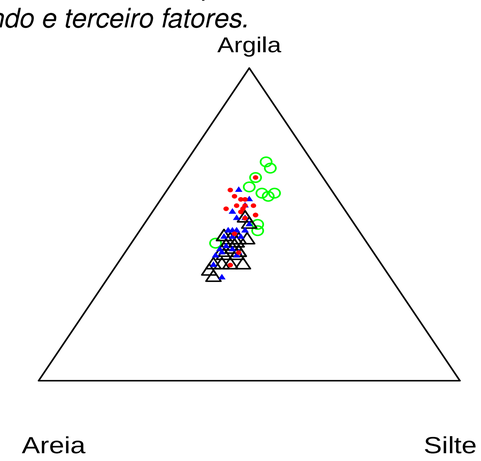

Figura 3: *Diagrama ternário das porcentagens de areia, silte e argila.*

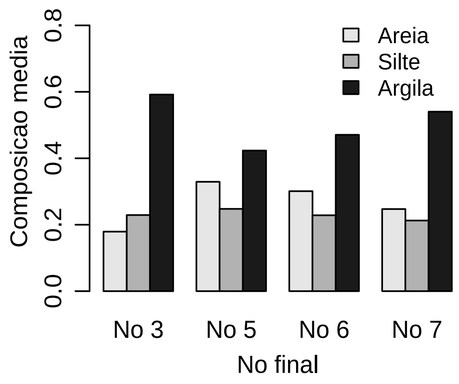

Figura 4: *Composição média segundo os nós.*

Tabela 2: *Resultados*

Nó	Técnica	Descrição do nó
3	AF	Menos matéria orgânica, menos fósforo e menos potássio e áreas com menores alturas.
	CART	Solos argilosos, mais silte do que areia.
5	AF	Elevada densidade global, reduzidas densidade de partícula e porosidade total.
	CART	Solos pouco argilosos, areia, silte e argila equilibrados
6	AF	Mais matéria orgânica, fósforo, potássio e áreas com maiores alturas. Reduzida densidade global, elevadas densidade de partícula e porosidade total.
	CART	Composição intermediária.
7	AF	Menos matéria orgânica, menos fósforo, menos potássio e áreas com menores alturas.
	CART	Mais argila e menos areia que nó 6.

4. Conclusão

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

Referências Bibliográficas

AITCHISON, J. **The statistical analysis of compositional data**. New Jersey: The Blackburn Press, 1986, 416 p.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. California: Wadsworth International Group, 1984. 358p.

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.

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.