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 - UEM

Resumo

Dados composicionais consistem de vetores (composições), cujos componentes são frações e satisfazem a restrição de soma 1. Técnicas estatísticas convencionais podem levar a resultados inconsistentes na análise deste tipo de dados. Neste sentido, a proposição de extensões dos métodos estatísticos usuais de maneira a permitir a análise de composições torna-se imprenscidível. Tais extensões são possíveis, por exemplo, mediante a transformação razão log-aditiva. Árvores de regressão permitem a modelagem de uma variável resposta numérica por meio de um conjunto de covariáveis e de suas interações, sem impor restrição paramétrica aos dados. Propõe-se a 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. Esta proposta é fundamentada em adaptações do algoritmo de árvores de classificação multivariadas construídas a partir de coeficientes de dissimilaridades. Como exemplo de aplicação utilizou-se dados de areia, silte e argila e covariáveis relacionadas às propriedades físico-químicas do solo provenientes de um trabalho conduzido no campo experimental da ESALQ-USP. O trabalho foi realizado utilizando recursos de *software* livre em ambiente operacional GNU/Linux; no ambiente estatístico R, utilizando o pacote *compositions* e rotinas específicas. Os resulta-

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

Introdução

Este estudo é motivado pelo interesse em caracterizar frações granulométricas do solo, definindo grupos e áreas de homogeneidade e investigando a relação dessas com outros atributos do solo. Do ponto de vista metodológico, de forma geral, o interesse está em combinar e conciliar na modelagem os aspectos de que as variáveis resposta são proporções das frações de solo, a distribuição espacial e as relações com potenciais covariáveis, incluindo ainda o uso de algoritmos de classificação e regressão no auxílio da caracterização do solo. Pretende-se assim, combinar a teoria de dados composicionais com análise geoestatística e algoritmos de classificação e regressão.

A análise de dados composicionais foi introduzida nos anos 80 por Aitchison (1982) e é aplicada nas ciências agrárias, geologia entre outras onde este tipo de dados ocorre com frequência. Dados composicionais consistem de vetores, denominados composições, cujos componentes $Y_1, ..., Y_B$ representam frações de algum "todo" e satisfazem a restrição de que a soma dos componentes é igual a 1 (AITCHISON, 1986), ou seja,

$$Y_1 \geq 0, Y_2 \geq 0, ..., Y_B \geq 0,$$

e

$$Y_1 + Y_2 + \cdots + Y_B = 1.$$

O espaço amostral é o simplex unitário de dimensão igual ao número de componentes dado por

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

sendo $\underline{j}'$ um vetor com elementos iguais a 1.

Um vetor $\underline{W}$ cujos componentes são positivos e medidos na mesma escala denomina-se base e pode se tornar uma composição através do operador fechamento $\mathcal{C}$ que garante que

a restrição de soma igual a 1 seja satisfeita:

$$\begin{aligned} \mathcal{C} \ : \ \mathbb{R}_+^B &\longrightarrow \mathbb{S}^B \\ \underset{\sim}{\mathrm{W}} &\longrightarrow \mathcal{C}(\underset{\sim}{\mathrm{W}}) = \frac{\underset{\sim}{\mathrm{W}}}{\underset{\sim}{\mathrm{j}}'\underset{\sim}{\mathrm{W}}}. \end{aligned}$$

Neste espaço amostral, o simplex, as operações matemáticas de soma e multiplicação definidas no espaço real equivalem às operações pertubação

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

e potência

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

respectivamente, e a média passa a ser a média geométrica $g(\underset{\sim}{\mathrm{Y}}_1) = \sqrt[B]{\prod_{j=1}^{B} Y_{1j}}$.

Uma característica desse tipo de dados é que a restrição de que a soma dos componentes deve ser igual a 1 implica em correlação negativa entre os componentes fazendo com que as correlações não sejam diretamente interpretáveis (GRAF, 2006). Neste sentido Aitchison (1986) propôs, dentre outras, a transformação razão log-aditiva (ALR) que generaliza a transformação logística para um vetor composicional de duas partes e é dada por:

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

Então, acrescentando às operações definidas anteriormente, o produto interno

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

tem-se uma estrutura de espaço Euclidiano real para o simplex. Este produto interno induz uma distância (entendida, por exemplo, como grau de alteração) no simplex, denominada distância de Aitchison, usada para calcular a distância ou diferença entre duas composições

e útil para entender a variabilidade dentro de um conjunto de dados:

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

A representação gráfica de uma amostra de composições pode ser feita através do diagrama ternário, por exemplo no caso em que $B = 3$, um triângulo equilátero cujos vértices representam os três componentes da composição (BUTLER, 2008).

A teoria de dados composicionais vêm sendo estudada e apresentada na literatura sob diferentes abordagens. Aitchison (1986) apresenta esta teoria considerando a independência entre as observações (composições), Pawlowsky-Glahn e Olea (2004) acrescentam a esta teoria o efeito espacial, Obage (2005) faz inferência bayesiana de dados composicionais sem considerar o efeito espacial e Tjelmeland e Lund (2003) tratam do aspecto da inferência bayesiana espacial. Nossos desenvolvimentos visam combinar estas abordagens sob a perspectiva da teoria de árvores de classificação e regressão e a contribuição desta para a análise de dados composicionais.

Árvores de classificação e regressão (Classification And Regression Trees – CART - BREIMAN et al., 1984) permitem a explicação de uma variável categorizada (classificação) ou numérica (regressão) com base em um conjunto de covariáveis e das eventuais interações entre as mesmas. Tais técnicas destacam-se por serem flexíveis, não impondo qualquer restrição paramétrica às variáveis sob estudo, e versáteis, dadas suas aplicações como complemento ou alternativa a diversos procedimentos estatísticos. A extensão do CART para a análise de dados multivariados (SEGAL, 1992; ZHANG, 1998; De'ATH, 2002; LEE, 2005) permite modelar conjuntamente duas ou mais variáveis respostas, mediante a definição de medidas de heterogeneidade e de qualidade preditiva adequadas. Taconeli (2008) propõe a construção de árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridades.

Propõe-se, no presente trabalho, modelar dados composicionais via CART segundo proposta apresentada em Taconeli (2008), considerando a distância de Aitchison, aplicável na análise de dados desta natureza, no lugar dos coeficientes de dissimilaridades originalmente formulados.

Metodologia

Os dados analisados são provenientes de Gonçalves (1997) cujo trabalho foi conduzido no campo experimental de irrigação do Departamento de Engenharia Rural da Escola Superior de Agricultura Luiz de Queiroz (ESALQ-USP) situado nas coordenadas 22º42' de latitude sul, longitude oeste de 47º38' e altitude média de 546 m acima do nível do mar. Esta área em estudo consistiu de um quadrante irrigado por um sistema pivô-central, com declividade média de aproximadamente 2% na sua direção bissetriz. Esse quadrante correspondeu ao topo da encosta onde foi instalado o pivô. Construiu-se uma malha quadrada ou grade de amostragem de 20 em 20 m onde foram analisadas 81 amostras de solo e medidos os percentuais de areia, silte e argila, além dos valores de ph-CaCl2, matéria orgânica, fósforo, potássio, cálcio, magnésio, hidrogênio+alumínio, densidade global, densidade da partícula, porosidade total e cota (altura do terreno).

O algoritmo proposto para a análise de dados composicionais via CART é semelhante ao apresentado em Breiman et al. (1984), baseado na extensão multivariada proposta em Taconeli (2008), diferindo apenas quanto às medidas de impureza e de qualidade preditiva empregadas. Inicialmente, seja $d(\underset{\sim}{\mathrm{Y}}_k, \underset{\sim}{\mathrm{Y}}_{k'})$ o resultado da distância de Aitchison, calculado a partir dos vetores correspondentes às composições de duas amostras de solo k e k'. Utiliza-se como medida de impureza para n_t elementos que constituem um nó t a distância média entre tais elementos, ou seja,

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

servindo como base para a partição dos nós e para a poda.

Considere, ainda, T uma árvore de regressão multivariada. Suponha que uma nova composição $\underset{\sim}{\mathrm{Y}}^*$, independente daquelas utilizadas na construção de T, seja alocada a um nó $t \in T$. Seja d_k^* a distância de $\underset{\sim}{\mathrm{Y}}^*$ em relação a uma observação $k \subset t$. Considera-se como medida de qualidade da predição a distância média entre esta nova observação e as n_t contidas em t, ou seja,

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

Dada a existência de elevadas correlações e o elevado número de covariáveis, decidiu-se aplicar antes uma análise fatorial (JOHNSON, 1998), com o objetivo de compor um número reduzido de fatores interpretáveis capazes de conservar boa parte da variabilidade associada às variáveis originais. A estimação das cargas fatoriais e dos escores, foi realizada pelo método das componentes principais usando o procedimento de mínimos quadrados ordinários com rotação varimax. A incorporação dos resultados produzidos pela análise fatorial ao modelo de regressão por árvore ocorreu ao considerar como covariáveis as estimativas dos escores fatoriais, em detrimento às variáveis originais.

Todo o trabalho foi realizado utilizando recursos de *software* livre em ambiente operacional GNU/Linux; no ambiente estatístico R (R development Core Team, 2008), utilizando o pacote *compositions* e rotinas desenvolvidas em Taconeli (2008).

Resultados

O primeiro passo da análise consistiu na execução da análise fatorial, aplicada ao conjunto de 11 covariáveis. Optou-se pela constituição de três fatores, uma vez que conjuntamente eles mostraram-se capazes de conservar 74% da variabilidade original. Na seqüência, os fatores obtidos foram caracterizados segundo suas cargas fatoriais. Dessa forma, os escores fatoriais tornam-se interpretáveis, viabilizando a utilização dos mesmos como covariáveis no modelo. A Tabela 1 apresenta as maiores cargas fatoriais e as comunalidades referentes aos três fatores sob estudo. Dentre todas as variáveis originais, apenas a quantidade de Potássio tem comunalidade inferior a $0,6$, indicando que as variáveis originais são bem representadas pelos três fatores.

As variáveis com maior carga fatorial no primeiro fator são o Ph-CaCl2, o cálcio e o magnésio (com cargas positivas) e hidrogênio+alumínio (negativa). A configuração das cargas indica correlação positiva entre as variáveis com cargas positivas, e correlação negativa dessas variáveis em relação a hidrogênio+alumínio. Amostras de solo com elevados escores para esse fator têm elevados Ph-CaCl2 e teor de cálcio e magnésio, em detrimento a uma reduzida quantidade de hidrogênio+alumínio. Já amostras com escores reduzidos têm características opostas às mencionadas.

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	$0,29$	$0,52$	$0,74$	

No segundo fator, aparecem com maiores cargas a matéria orgânica, o fósforo, o potássio e a altura do terreno, todas negativas, indicando correlação positiva entre essas quatro variáveis. Já no fator 3 a densidade global se contrapõe à densidade de partícula e a porosidade total, por ter carga positiva, ao contrário das outras duas. A interpretação dos escores fatoriais das amostras de solo segundo os sinais das cargas fatoriais obtidas é realizada de maneira semelhante à descrita para o primeiro fator.

O gráfico de custo complexidade apresentado na Figura 1 indica, segundo a regra do '1 desvio padrão' (BREIMAN et al., 1984), a seleção da árvore com quatro nós finais. A referida árvore é apresentada na Figura 2. Verifica-se que apenas os fatores 2 e 3 são responsáveis por partições, evidenciando a relação entre as variáveis associadas a estes fatores e as propriedades físico-químicas do solo.

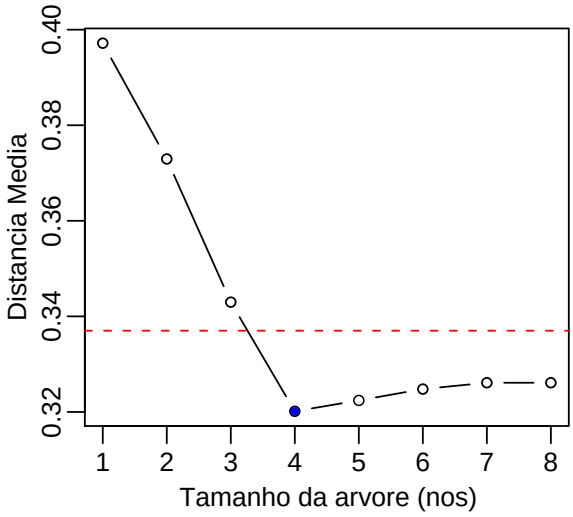

Figura 1: Curva de custo-complexidade.

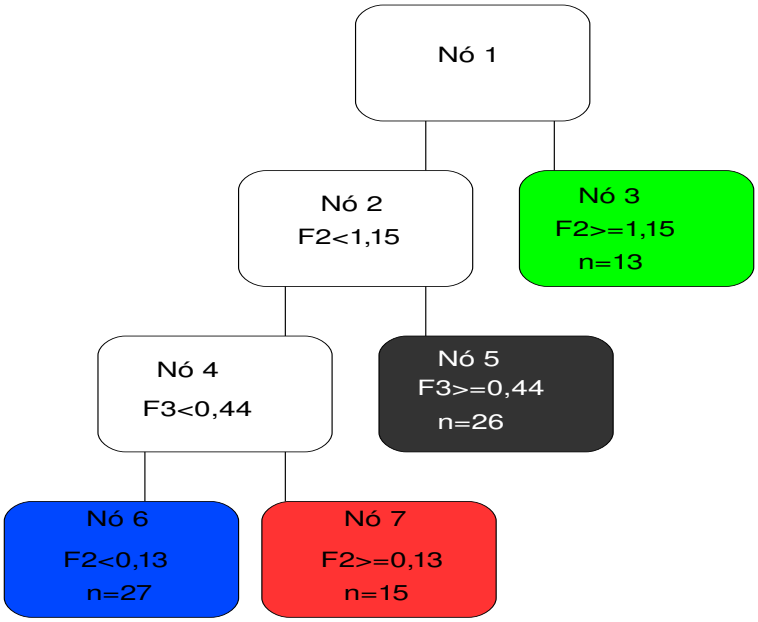

Figura 2: Árvore de regressão.

A Figura 3 apresenta o gráfico de dispersão para os escores fatoriais para o segundo e terceiro fatores (responsáveis por partições no modelo). Sua avaliação, conjugada à árvore de regressão obtida, permite interpretar as relações entre a composição do solo e as demais covariáveis consideradas. A Figura 4, por sua vez, apresenta a disposição das 81 composições de solo em um diagrama ternário. Nesse tipo de representação, quanto mais

próxima uma composição estiver de um dos vértices do triângulo, maior a concentração do componente correspondente a esse vértice na referida composição. As cores dos pontos em cada um destes gráficos indicam o nó final da árvore em que cada composição foi alocada, de acordo com a configuração de cores adotada na árvore apresentada na Figura 2.

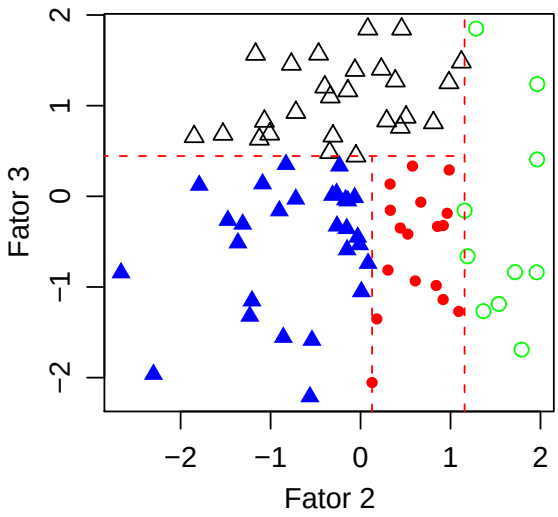

Figura 3: Gráfico de dispersão dos escores fatoriais para o segundo e terceiro fatores.

Considerando-se as composições alocadas em cada um dos nós finais, calculou-se a média geométrica de cada componentes de modo que a composição média segundo os nós está representada na Figura 5.

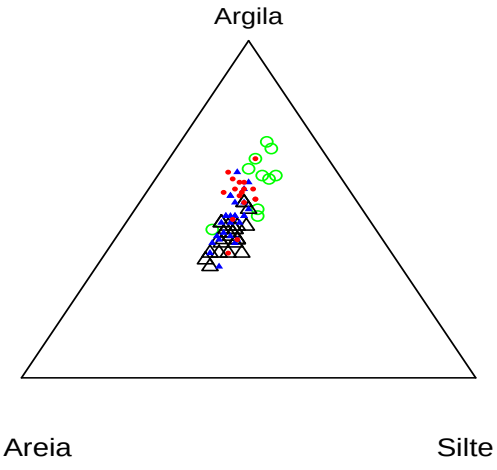

Figura 4: Diagrama ternário das porcentagens de areia, silte e argila.

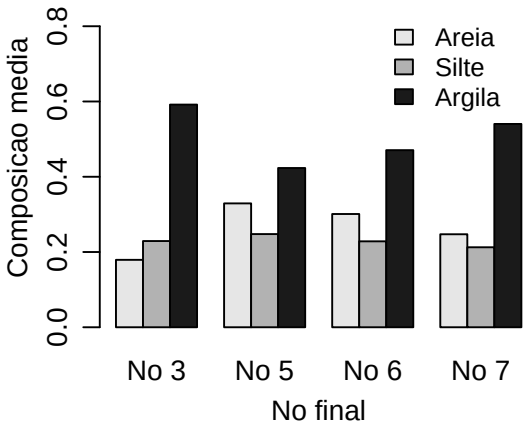

Figura 5: Distribuição da composição média segundo os nós.

A Figura 6 mostra a distribuição das coordenadas das amostras de acordo com o nó a que pertencem e permite levantar evidências quanto à possíveis padrões espaciais.

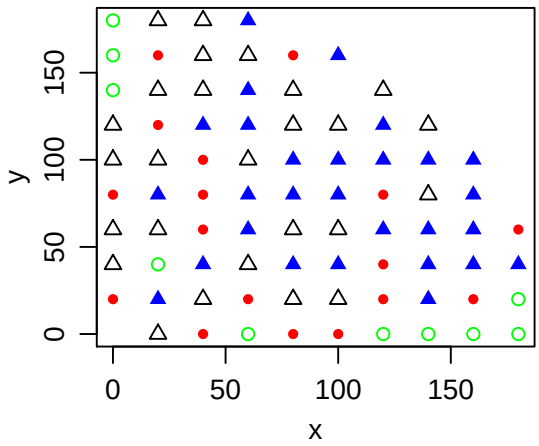

Figura 6: Localização espacial dos pontos amostrais em que os símbolos representam os grupos de frações granulométricas identificados pela análise.

A avaliação conjunta dos gráficos apresentados nas Figuras 3 e 4 e 5 permite estudar as relações entre a composição do solo e as demais covariáveis consideradas. A Tabela 2 apresenta os principais resultados extraídos da análise fatorial e da árvore de regressão multivariada executadas.

Tabela 2: Resultados

Nó	Técnica	Descrição do nó
3	Análise fatorial	Menores quantidades de matéria orgânica, fósforo e potássio e áreas com menores alturas.
	Árvore de regressão	Solos acentuadamente argilosos, com mais silte do que areia.
5	Análise fatorial	Elevada densidade global, em detrimento a reduzidas densidade de partícula e porosidade total.
	Árvore de regressão	Solos pouco argilosos, nó em que as porcentagens de areia, silte e argila são mais equilibradas.
6	Análise fatorial	Maiores quantidades de matéria orgânica, fósforo e potássio e áreas com maiores alturas. Reduzida densidade global, em detrimento a elevadas densidade de partícula e porosidade total.
	Árvore de regressão	Composição intermediária.
7	Análise fatorial	Características semelhantes às do nó 6, mas com menores quantidades de matéria orgânica, fósforo e potássio e áreas com menores alturas.
	Árvore de regressão	Mais argila e menos areia que as amostras que compõem o nó 6.

Conclusão

Os resultados produzidos até o momento permitiram identificar propriedades do solo associadas às composições, estabelecendo uma hierarquia entre as variáveis físico-químicas na explicação das frações granulométricas. Próximos passos incluem propostas para modelagem conjunta espacial, com implementações de predição bayesianas permitindo obter incertezas associadas às classificações. A metodologia deve ainda ser testada em dados adicionais da área de estudo considerada, bem como dados provenientes de estudos em outras áreas, a fim de se verificar a aplicabilidade e generalidade da proposta.

Referências Bibliográficas

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.

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.

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.

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.