

Predição Espacial Bayesiana Para Dados Composicionais

Aluna de Doutorado: Ana Beatriz Tozzo Martins

Orientador: Paulo Justiniano Ribeiro Junior

Universidade Federal do Paraná

Programa de Pós-Graduação em Métodos Numéricos em Engenharia

19 de outubro de 2009



TESE

1. Introdução
2. Revisão da Literatura:
 - Modelo Geoestatístico Gaussiano;
 - Predição Linear Espacial- Clássica e Bayesiana;
 - Matriz de Covariância, Correlação e Variograma Cruzado;
 - Modelo Multivariado;
 - Dados Composicionais;
3. Material e Métodos:
 - Modelo Geoestatístico Composicional - Clássica e Bayesiana;
 - Análise de Dados Composicionais Simulados;
4. Resultados: Análise de frações granulométricas de um solo;
5. Conclusão.

MOTIVAÇÃO

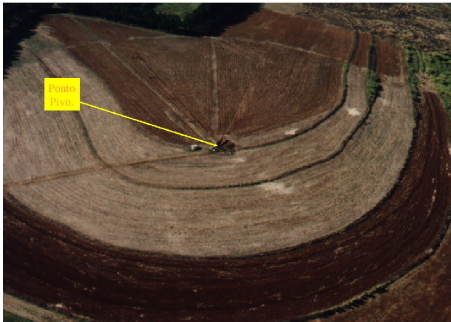


Figura: Quadrante irrigado por sistema pivô-central no campo experimental da ESALQ-USP.

SOLO

frações granulométricas

(composição)



propriedades físico-hídricas



prática agrícola



agricultura de precisão

Dados Composicionais - Aitchison (1986)



Figura: Diagrama Ternário das composições.

• **Composição:** $\underline{X} = (X_1, X_2, \dots, X_B)'$

$$X_i \geq 0; \quad X_1 + X_2 + \dots + X_B = 1$$

$$\mathbb{S}^B = \{ \underline{X} \in \mathbb{R}^B; X_i > 0, \mathbf{1}'\underline{X} = 1 \}$$

• **Base:** $\underline{W} = (W_1, W_2, \dots, W_B)'$

$$\mathbb{R}_+^B = \{ \underline{W} \in \mathbb{R}^B; W_i > 0, i = 1, \dots, B \}$$

• **Operador fechamento:**

$$\mathcal{C} : \mathbb{R}_+^B \longrightarrow \mathbb{S}^B$$

$$\underline{W} \longrightarrow \mathcal{C}[\underline{W}] = \frac{\underline{W}}{\mathbf{1}'\underline{W}}$$

LITERATURA

- Dados composicionais:
 - Aitchison (1986).
- Modelos espaciais uni e multivariados:
 - Diggle e Ribeiro Jr (2007);
 - Schmidt e Gelfand (2003);
 - Banerjee, Carlin e Gelfand (2004);
 - Schmidt e Sansó (2006).
- Combinação de dados composicionais e espaciais:
 - Pawlowsky e Olea (2004);
 - Tjelmeland e Lund (2003).



OBJETIVO

- Estender o modelo geoestatístico bivariado para dados composicionais, derivando e implementando estimação via inferência bayesiana baseada na verossimilhança;
- Obter preditores espaciais no simplex que permitam a construção de mapas de predição das frações do solo na área de estudo.



Fundamentos Para Análise de Dados Composicionais

Transformação razão log-aditiva (alr):

$$\text{alr} : \mathbb{S}^B \longrightarrow \mathbb{R}^{B-1}$$

$$\underline{X}(\underline{x}) \longrightarrow \text{alr}(\underline{X}(\underline{x})) = \left(\ln \frac{X_1(\underline{x})}{X_B(\underline{x})}, \dots, \ln \frac{X_{B-1}(\underline{x})}{X_B(\underline{x})} \right)'$$

Composição: $\underline{X}(\underline{x}) = (X_1(\underline{x}), X_2(\underline{x}), X_3(\underline{x}))' = (\text{Areia}, \text{Silte}, \text{Argila})'$

$$\underline{Y}(\underline{x}) = (Y_1(\underline{x}), Y_2(\underline{x}))' = \left(\ln \frac{X_1(\underline{x})}{X_3(\underline{x})}, \ln \frac{X_2(\underline{x})}{X_3(\underline{x})} \right)'$$

Transformação logística generalizada aditiva (agl): $\text{agl} = \text{alr}^{-1}$

MODELO ESPACIAL

Modelo geoestatístico bivariado para dados composicionais:

$$\begin{cases} Y_1(\underline{x}_i) &= \mu_1(\underline{x}_i) + \sigma_1 U(\underline{x}_i; \phi) + \tau_1 Z(\underline{x}_i; \rho) \\ Y_2(\underline{x}_{i'}) &= \mu_2(\underline{x}_{i'}) + \sigma_2 U(\underline{x}_{i'}; \phi) + \tau_2 Z(\underline{x}_{i'}; \rho) \end{cases}$$

- $\underline{x}_i, \underline{x}_{i'} \in \mathbb{R}^2$; $i, i' = 1, \dots, n_1, n_1$ tamanho amostral;
- $\underline{Y}_{n \times 1}(\underline{x}) = (Y_1(\underline{x}_1), Y_2(\underline{x}_1), \dots, Y_1(\underline{x}_{n_1}), Y_2(\underline{x}_{n_1}))'$;
- $U \sim N(\underline{0}; \Sigma_U)$, Σ_U : var. unitárias, função de $\rho_U(\phi)$ exponencial;
- $Z \sim N(\underline{0}; \Sigma_Z)$, Σ_Z : var/cov em função de ρ induzida pela estrutura composicional.

MODELO ESPACIAL

- $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)' \rightarrow \underline{\theta}^* = (\eta, \nu_1, \nu_2, \phi, \rho)'$
- **Log-Verossimilhança \rightarrow Normal multivariada com EMV:**

$$\hat{\underline{\mu}} = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}(\mathbf{D}'\mathbf{V}^{-1}\underline{Y}) \quad \text{e} \quad \hat{\sigma}_1 = \sqrt{\hat{Q}e/n}$$

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

- **Log-verossimilhança concentrada \rightarrow solução numérica:**

$$l(\underline{\theta}^*, \underline{Y}) = -0,5 \left[\ln(|\mathbf{V}|) + n \left(\ln(2\pi) + \ln(\hat{Q}e) - \ln(n) + 1 \right) \right]$$

INFERÊNCIA BAYESIANA

- $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)'$ → variáveis aleatórias
- Dist conjunta = veros. \times *priori*: $P(\underline{Y}, \underline{\theta}) = P(\underline{Y}|\underline{\theta}) P(\underline{\theta})$.
- **Teorema de Bayes** \Rightarrow distribuição à *posteriori* de $\underline{\theta}$:

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

- **Resumos estatísticos** como esperanças à *posteriori* de funções de $\underline{\theta}$:

$$E(f(\underline{\theta})|\underline{Y}) = \frac{\int f(\underline{\theta})P(\underline{\theta})P(\underline{Y}|\underline{\theta})d\underline{\theta}}{\int P(\underline{\theta})P(\underline{Y}|\underline{\theta})d\underline{\theta}} \propto \int f(\underline{\theta})P(\underline{\theta})P(\underline{Y}|\underline{\theta})d\underline{\theta}.$$

INFERÊNCIA BAYESIANA (cont.)

- **Resolução:**
 - **aproximação de Laplace:** Nott, Fielding e Leonte (2009), Gelman, Carlin, Stern e Rubin (2003), Gilks, Richardson e Spiegelhalter (1996);
 - **quadratura de Gauss-Hermite:** Paulino, Turkman e Murteira (2003), Gamerman (2006), Abramowitz e Stegun (1972);
 - **integração Monte Carlo - Cadeias de Markov de Monte Carlo (MCMC):** Gamerman (2006), Lee (2004), Gelman, Carlin, Stern e Rubin (2003), Gill (2002), Gilks, Richardson e Spiegelhalter (1996).

INFERÊNCIA BAYESIANA (cont.)

- $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)'$
- $\underline{\theta}^* = (\eta, \nu_1, \nu_2, \phi, \rho)$ e $(\underline{\mu}, \sigma_1^2)$ independentes
- **Distribuição à posteriori de $\underline{\theta}$:**

$$P(\underline{\mu}, \sigma_1^2, \underline{\theta}^* | \underline{Y}) \propto P(\underline{Y} | \underline{\mu}, \sigma_1^2, \underline{\theta}^*) P(\underline{\mu}, \sigma_1^2) P(\underline{\theta}^*)$$

PRIORIS

- **Prioris:**

- $P(\underline{\mu}, \sigma_1^2) \propto \frac{1}{\sigma_1^2};$
- $P(\underline{\mu}|\sigma_1^2) = 1;$
- $P(\eta), P(\nu_1)$ e $P(\nu_2)$: lognormais com parâmetros correspondentes às razões das estimativas de máxima verossim.;
- $P(\phi) = \text{Gama}(66, 1);$
- $P(\rho) = 1$

Distribuições à Posteriori

- $$\int_{\underline{\theta}^*} P(\underline{\mu}, \sigma_1^2, \underline{\theta}^* | \underline{Y}) d\underline{\theta}^* \propto \int_{\underline{\theta}^*} P(\underline{Y} | \underline{\mu}, \sigma_1^2, \underline{\theta}^*) P(\underline{\mu}, \sigma_1^2) P(\underline{\theta}^*) d\underline{\theta}^*$$

- $$P(\underline{\mu}, \sigma_1^2 | \underline{Y}) \propto P(\underline{Y} | \underline{\mu}, \sigma_1^2, \underline{\theta}^*) \frac{1}{\sigma_1^2} \int_{\underline{\theta}^*} P(\underline{\theta}^*) d\underline{\theta}^*.$$

$$\propto \frac{1}{\sigma_1^2} P(\underline{Y} | \underline{\mu}, \sigma_1^2, \underline{\theta}^*)$$

- $$[\underline{\mu} | \underline{Y}, \sigma_1^2] \sim N \left(\hat{\underline{\mu}}; \sigma_1^2 \left(\mathbf{D}' \mathbf{V}_Y^{-1} \mathbf{D} \right)^{-1} \right) \quad \hat{\underline{\mu}} = (\mathbf{D}' \mathbf{V}^{-1} \mathbf{D})^{-1} (\mathbf{D}' \mathbf{V}^{-1} \underline{Y})$$

$$[\sigma_1^2 | \underline{Y}] \sim \chi_{\text{Sinv}}^2 \left(n - n_{\hat{\underline{\mu}}}; S^2 \right) \quad S^2 = \frac{(\underline{Y} - \mathbf{D} \hat{\underline{\mu}})' \mathbf{V}_Y^{-1} (\underline{Y} - \mathbf{D} \hat{\underline{\mu}})}{n - n_{\hat{\underline{\mu}}}}.$$

Algoritmo Metropolis-Hastings

- $\theta^{(t)} \Rightarrow \theta^{(t+1)}$: θ' de $q(\cdot|\theta^{(t)})$ com probabilidade $\alpha(\theta^{(t)}, \theta')$
- Passos:
 1. Escolher número de simulações - n
 2. Iniciar o contador de iterações - $t=1$
 3. Valor inicial $\theta^{(t)}$
 4. Gerar θ' de $q(\cdot|\theta^{(t)})$
 5. Gerar $u \sim U(0; 1)$
 6. $\alpha(\theta; \theta') = \min \left\{ 1; \frac{P(\underline{Y}|\theta') P(\theta') q(\theta|\theta')}{P(\underline{Y}|\theta) P(\theta) q(\theta'|\theta)} \right\}$
 7. Se $u < \alpha(\theta; \theta') \Rightarrow \theta^{(t)} = \theta'$, cc $\theta^{(t)} = \theta^{(t-1)}$.
 8. Se $t = n$, FIM.
 9. Incrementar o contador t fazendo $t = t + 1$.
 10. Voltar para o passo 4. FIM.

Passos Para Inferência Bayesiana

1. Algoritmo Metropolis-Hastings \Rightarrow matriz de covariância do modelo;
2. Calcular $\hat{\underline{\mu}}$ e S^2 e amostrar σ_1^2 de $[\sigma_1^2 | \underline{Y}] \sim \chi_{Sinv}^2(n - n_{\underline{\mu}}; S^2)$;
3. Calcular $\text{Var}(\underline{\mu})$: $[\underline{\mu} | \underline{Y}, \sigma_1^2] \sim N\left(\hat{\underline{\mu}}; \sigma_1^2 \left(\mathbf{D}'\mathbf{V}_{\underline{Y}}^{-1}\mathbf{D}\right)^{-1}\right)$;
4. Amostrar um valor de $\underline{\mu}$ de $[\underline{\mu} | \underline{Y}, \sigma_1^2]$;
5. Repetir tantas vezes quanto o número de simulações;

Passos Para Inferência Bayesiana

- $$\int_{\underline{\theta}^*} P(\underline{\mu}, \sigma_1^2, \underline{\theta}^* | \underline{Y}) d\underline{\theta}^* \propto \int_{\underline{\theta}^*} P(\underline{Y} | \underline{\mu}, \sigma_1^2, \underline{\theta}^*) P(\underline{\mu}, \sigma_1^2) P(\underline{\theta}^*) d\underline{\theta}^*$$

- $$P(\underline{\mu}, \sigma_1^2 | \underline{Y}) \propto P(\underline{Y} | \underline{\mu}, \sigma_1^2, \underline{\theta}^*) \frac{1}{\sigma_1^2} \int_{\underline{\theta}^*} P(\underline{\theta}^*) d\underline{\theta}^*.$$

- $$\propto \frac{1}{\sigma_1^2} P(\underline{Y} | \underline{\mu}, \sigma_1^2, \underline{\theta}^*)$$

- $$[\underline{\mu} | \underline{Y}, \sigma_1^2] \sim N \left(\hat{\underline{\mu}}; \sigma_1^2 \left(\mathbf{D}' \mathbf{V}_Y^{-1} \mathbf{D} \right)^{-1} \right) \quad \hat{\underline{\mu}} = (\mathbf{D}' \mathbf{V}^{-1} \mathbf{D})^{-1} (\mathbf{D}' \mathbf{V}^{-1} \underline{Y})$$

- $$[\sigma_1^2 | \underline{Y}] \sim \chi_{\text{Sinv}}^2 \left(n - n_{\hat{\underline{\mu}}}; S^2 \right) \quad S^2 = \frac{(\underline{Y} - \mathbf{D} \hat{\underline{\mu}})' \mathbf{V}_Y^{-1} (\underline{Y} - \mathbf{D} \hat{\underline{\mu}})}{n - n_{\hat{\underline{\mu}}}}.$$

Passos Para Inferência Bayesiana (cont.)

6. Predição espacial:

6.1 **Cokrigagem** de \underline{Y}_0 em $\underline{x}_0 = (\underline{x}_{10}, \underline{x}_{20}, \dots, \underline{x}_{n_20})$ com cada conjunto de parâmetros simulados:

- $\underline{\mu}_{\underline{Y}_0|\underline{Y}} = \underline{\mu}_{\underline{Y}_0} + \Sigma_{\underline{Y}_0\underline{Y}}\Sigma_{\underline{Y}\underline{Y}}^{-1}(\underline{Y} - \underline{\mu}_{\underline{Y}})$
- $\Sigma_{\underline{Y}_0|\underline{Y}} = \Sigma_{\underline{Y}_0\underline{Y}_0} - \Sigma_{\underline{Y}_0\underline{Y}}\Sigma_{\underline{Y}\underline{Y}}^{-1}\Sigma_{\underline{Y}\underline{Y}_0}$

6.2 Gerar uma amostra de $NM \sim (\underline{\mu}_{\underline{Y}_0|\underline{Y}}; \Sigma_{\underline{Y}_0|\underline{Y}})$ para cada conjunto de parâmetros simulados.

6.3 Aplicar transformação agl.

6.4 Calcular a média das simulações para cada componente.

Análise de Frações Granulométricas de Um Solo

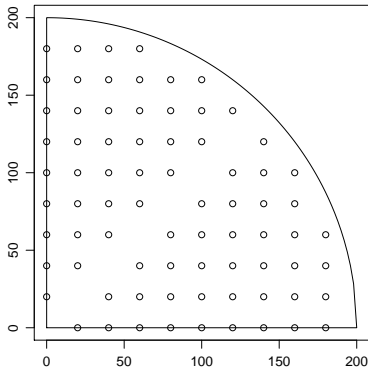


Figura: Distribuição das localizações na área de estudo.

- Gonçalves (1997):
Tese - ESALQ-USP;
- Área: quadrante irrigado por um sistema pivô-central;
- Grade de amostragem 20×20m
⇒ 76 amostras: $\left\{ \begin{array}{l} \text{Areia} \\ \text{Silte} \\ \text{Argila.} \end{array} \right.$

Análise de Frações Granulométricas de Um Solo (cont.)

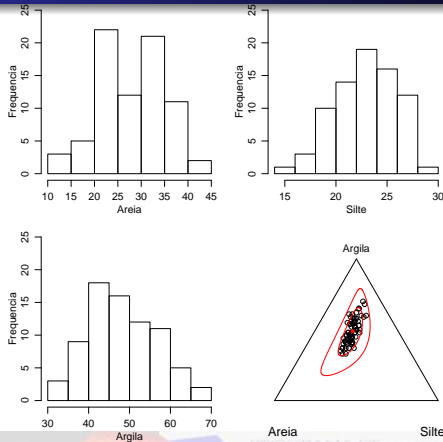


Figura: Distribuição de areia, silte e argila e diagrama ternário das composições.

Análise de Frações Granulométricas de Um Solo (cont.)

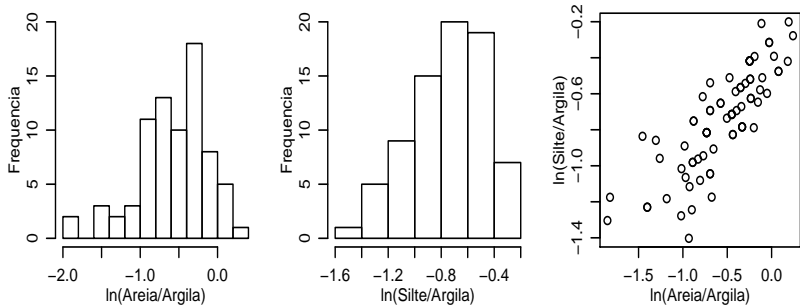


Figura: Distribuição das log-razão e correspondente diagrama de dispersão.

Análise de Frações Granulométricas de Um Solo (cont.)

Tabela: Estimativas e intervalos de 95% de confiança obtidos por inferência Bayesiana considerando-se 12000 simulações, *burn-in*= 1000 e salto= 10.

Parâmetros	Estimativas	LI	LS
μ_1	-0.7467	-1.1078	-0.4038
μ_2	-0.7900	-0.8958	-0.6887
σ_1	0.1992	0.1273	0.2889
σ_2	0.0512	0.0247	0.0890
τ_1	0.1268	0.0978	0.1625
τ_2	0.1179	0.0884	0.1510
ϕ	65.7278	53.7055	80.5381
ρ	0.9314	0.85290	0.98854

Análise de Frações Granulométricas de Um Solo (cont.)

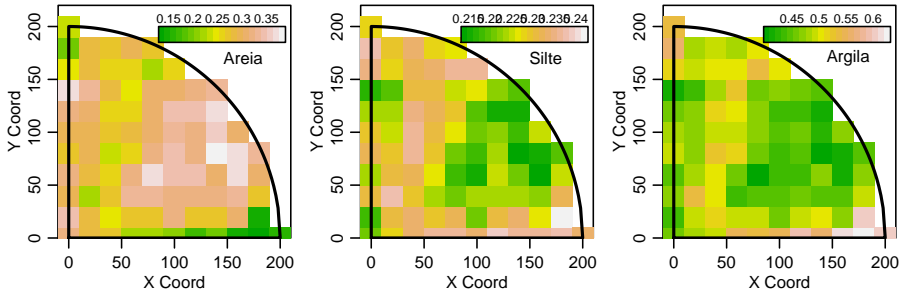


Figura: Mapas das porcentagens de areia, silte e argila obtidos por inferência Bayesiana.



Análise de Frações Granulométricas de Um Solo (cont.)

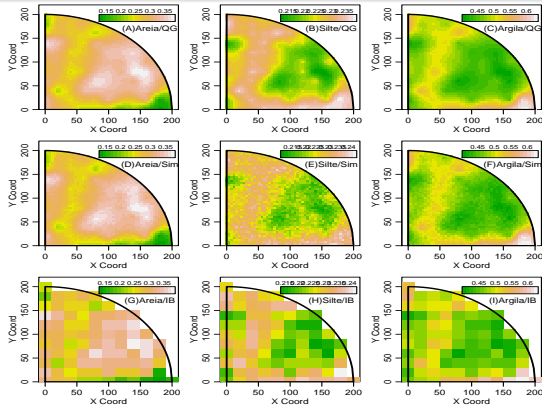


Figura: Mapas das porcentagens de areia, silte e argila por quadratura de Gauss-Hermite (A-C), simulação (D-F) e inferência Bayesiana (G-I).

Conclusão

- Construção de mapas de areia, silte e argila garantindo que as frações somem 1, nos pontos observados e preditos;
- Modelo captura variações espaciais, induzidas pelas composições e não estruturadas;
- Declaração explícita do modelo permite fazer inferências sobre parâmetros tanto pelo método clássico quanto por inferência Bayesiana considerando nas predições a incerteza associada à estimação dos parâmetros;
- Estimação de outros funcionais;
- Necessidade de investigar alternativas para computação mais eficiente e formas de especificação do modelo para um número maior de componentes.

- AITCHISON, J., *The statistical analysis of compositional data*, New Jersey: The Blackburn Press, 1986.
- DIGGLE, P.J.; RIBEIRO JR, P.J., *Model-based geostatistics*, USA: Springer Series in Statistics, 2007.
- GONÇALVES, A.C.A., *Variabilidade espacial de propriedades físicas do solo para fins de manejo da irrigação*. 1997. 119p. Tese (Doutorado em Agronomia) - ESALQ-USP, São Paulo, Piracicaba.
- PAWLOWSKY-GLAHN, V.; OLEA, R.A., *Geostatistical analysis of compositional data*, New York: Oxford University Press, Inc., 2004.
- CARLIN, B. P.; LOUIS, T. A., **Bayesian Methods for Data Analysis**, Boca Raton: Chapman and Hall, Third ed., 2009.
- GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B., **Bayesian Data Analysis**, Boca Raton: Chapman and Hall, Second ed., 2003.
- GILL, J., **Bayesian Methods For The Social And Behavioral Sciences**, London: Chapman and Hall, 2002.
- PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B., **Estatística Bayesiana**, Lisboa: Fundação Calouste Gulbenkian, 2003.

OBRIGADA PELA ATENÇÃO!

Agradecimentos

- DES - Departamento de Estatística, UEM.
- PPGMNE - Programa de Pós-Graduação em Métodos Numéricos em Engenharia, UFPR.
- LEG - Laboratório de Estatística e Geoinformação, UFPR.
- CNPQ - Conselho Nacional de Desenvolvimento Científico e Tecnológico.
- FINEP, projeto CT-INFRA/UFPR.