Uma Abordagem Bayesiana na Análise Geoestatística de Dados Composicionais

Ana Beatriz Tozzo Martins - PPGMNE/LEG, UFPR e DES, UEM [1] [2]
Paulo Justiniano Ribeiro Junior - LEG, UFPR
Wagner Hugo Bonat - LEG, UFPR
Antônio Carlos Andrade Gonçalves - DAG, UEM

Resumo: *Este trabalho é motivado pelo interesse em modelar o padrão espacial de dados composicionais utilizando métodos bayesianos. A teoria de dados composicionais foi desenvolvida nos anos 80 para observações independentes. A partir dos anos 90 surgiram resultados em geoestatística, na concepção de uma declaração explícita de modelo. Nos anos 2000 surgiram trabalhos sobre dados composicionais, sob a abordagem geoestatística tradicional e sem considerar predição espacial bayesiana. O objetivo é propor um modelo geoestatístico bivariado para dados composicionais, sob o paradigma bayesiano, visando a predição espacial com a produção de mapas temáticos. O modelo considera o efeito da locação e a natureza multivariada dos dados, na construção de uma estrutura de covariância espacial adequada, permitindo resultados satisfatórios para a elaboração de mapas de distribuição espacial. Para alguns parâmetros do modelos foi possível derivar as distribuições a posteriori analiticamente. Para os restantes utilizou-se o algoritmo Metropolis-Hastings de Monte Carlo via cadeia de Markov. Distribuições a priori dos tipos: não informativa (flat), imprópria, Log-normal e Gama foram utilizadas. As demandas computacionais foram supridas com os pacotes estatísticos R (R Development Core Team, 2009); geoR, compositions, MCMCpack, coda, e rotinas desenvolvidas para o trabalho organizadas em forma de pacote geoComp. A metodologia foi aplicada a um conjunto de dados de frações granulométricas de um solo. O resultado obtido foi a representação da distribuição espacial dos teores de areia, silte e argila. A metodologia proposta mostrou-se satisfatória para representar a distribuição espacial de frações granulométricas do solo.*

Palavras-chave: *geoestatística multivariada, dados composicionais, verossimilhança, frações granulométricas, inferência bayesiana.*

1 Introdução

Este trabalho é motivado pelo interesse em modelar o padrão espacial de dados composicionais utilizando métodos bayesianos. Neste sentido, combina-se a teoria de dados composicionais originalmente desenvolvida para observações independentes (AITCHISON, 1986) com

[1] Agradecimento à CAPES pelo apoio financeiro. Esse trabalho foi parcialmente financiado pela FINEP projeto CT-INFRA/UFPR.
[2] Contato: *abtmartins@uem.br*

métodos geoestatísticos (PAWLOWSKY-GLAHN; OLEA, 2004), adotando a declaração explícita de um modelo para descrever a espacialização das variáveis (DIGGLE; RIBEIRO JR., 2007).

Dados composicionais consistem de vetores, denominados composições, cujos componentes $X_1, ..., X_B$ representam frações de algum "todo", e satisfazem a restrição de que a soma dos componentes é igual a 1 (AITCHISON, 1986), ou seja,

$$X_1 \geq 0, X_2 \geq 0, ..., X_B \geq 0, \qquad \text{e} \qquad X_1 + X_2 + \cdots + X_B = 1.$$

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

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

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

Um vetor $\underline{\mathrm{W}}$ cujos componentes são positivos e medidos na mesma escala denomina-se base. Uma base pode se tornar uma composição através do operador fechamento, $\mathcal{C}$, que garante que a restrição de soma igual a 1 seja satisfeita :

$$\begin{array}{rccl} \mathcal{C} & : & \mathbb{R}_+^B & \longrightarrow & \mathbb{S}^B \\ & & \underline{\mathrm{W}} & \longrightarrow & \mathcal{C}(\underline{\mathrm{W}}) = \dfrac{\underline{\mathrm{W}}}{\underline{1}'\underline{\mathrm{W}}}, \end{array}$$

sendo a notação espacial omitida por simplificação de exposição. Neste espaço amostral, o simplex, as operações matemáticas de soma e multiplicação definidas no espaço real equivalem às operações pertubação

$$\underline{\mathrm{X}}_1 \oplus \underline{\mathrm{X}}_2 = (X_{11}, X_{12}, ..., X_{1B}) \oplus (X_{21}, X_{22}, ..., X_{2B}) = \mathcal{C}(X_{11}X_{21}, X_{12}X_{22}, ..., X_{1B}X_{2B}),$$

e potência

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

respectivamente, e a média passa a ser a média geométrica $g(\underline{\mathrm{X}}_1) = \sqrt[B]{\prod_{j=1}^{B} X_{1j}}$.

Uma característica desse tipo de dados é que estes apresentam um efeito de correlação espúria. 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), ou seja, as covariâncias estão sujeitas a controles não estocásticos o que implica, segundo Pawlowsky-Glahn e Olea (2004), em singularidade da matriz de covariância de uma composição. Com isto, a aplicação de técnicas estatísticas padrão podem levar a resultados inconsistentes. Para contornar este problema, Aitchison (1986) propôs, dentre outras, a transformação log-razão aditiva (ALR) que generaliza a transformação logística para um vetor composicional de duas **ou mais** partes e é dada por:

$$\begin{array}{rccl} \text{alr} & \mathbb{S}^B & \longrightarrow & \mathbb{R}^{B-1} \\ & \underline{X}(\underline{x}) & \longrightarrow & \text{alr}\big(\underline{X}(\underline{x})\big) = \left(\ln\left(\dfrac{X_1(\underline{x})}{X_B(\underline{x})}\right), \ldots, \ln\left(\dfrac{X_{B-1}(\underline{x})}{X_B(\underline{x})}\right)\right)'. \end{array}$$

Por outro lado, a transformação inversa denominada transformação logística generalizada aditiva

(AGL) é dada por

$$\begin{aligned} \text{agl}: \ & \mathbb{R}^{B-1} \longrightarrow \mathbb{S}^B \\ & \underline{Y}(\underline{x}) \longrightarrow \underline{X}(\underline{x}) = \mathcal{C}\left(\left(\exp\left\{\ln\left(\frac{X_1(\underline{x})}{X_B(\underline{x})}\right)\right\}, \ldots, \exp\{0\}\right)'\right) \end{aligned}$$

A representação gráfica de uma amostra de composições pode ser feita por meio 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 literatura contempla trabalhos sobre a teoria de dados composicionais e geoestatística. Aitchison (1986) apresenta a teoria de dados composicionais considerando a independência das observações; Pawlowsky-Glahn e Olea (2004) fazem análise geoestatística de dados composicionais seguindo a teoria clássica da geoestatística e Lark e Bishop (2007) apresentam um estudo sobre cokrigagem de frações de partículas do solo concluindo que a predição de dados composicionais pode ser feita por cokrigagem ALR com transformação de volta das predições, para a escala das composições, calculadas por quadratura de Gauss-Hermite na aproximação da esperança condicional. Por outro lado, Diggle e Ribeiro Jr (2007), Schmidt e Sansó (2006), Banerjee, Carlin e Gelfand (2004), Finley, Banerjee e Carlin (2007), Wackernagel (1998) e outros explicitam modelos geoestatísticos uni e multivariados através da especificação de uma função de correlação em que a covariância é função da distância entre pares de localizações. Sob o paradigma bayesiano, tem-se o trabalho de Obage (2007) que faz inferência bayesiana de dados composicionais sem considerar o efeito espacial; já Tjelmeland e Lund (2003) o consideram, mas não especificam uma forma fechada para a matriz de covariância. Para esta, os autores adotam uma função de correlação exponencial generalizada e uma *priori* Wishart para a estimação dos parâmetros de variância. Considera-se aqui o modelo gaussiano bivariado de componentes comum proposto por Diggle e Ribeiro Jr. (2007) e adotado por Bognola et. al. (2008) no contexto de inventário florestal utilizando uma variável física como informação secundária.

O objetivo deste trabalho é estender o uso do modelo geoestatístico bivariado para estruturas de dados composicionais, derivando e implementando estimação bayesiana e obtendo preditores espaciais. A metodologia desenvolvida é aplicada em um conjunto de dados referente a frações granulométricas do solo.

2 Metodologia

Para $\underline{X}=(X_1, ..., X_B)'$ sendo uma composição com B componentes e $\underline{Y}=\left(\ln\left(\frac{X_1}{X_B}\right), \ldots, \ln\left(\frac{X_{B-1}}{X_B}\right)\right)'$ um vetor com $B-1$ elementos, o modelo geoestatístico com componente comum pode ser obtido seguindo a formulação dada em Diggle e Ribeiro Jr. (2007). Neste trabalho, considerou-se composições de 3 componentes, (X_1, X_2, X_3) que correspondem a (Areia, Silte, Argila) em uma aplicação à fração granulométrica do solo.

A partir do modelo geoestatístico apresentado em Diggle e Ribeiro Jr. (2007), propõe-se uma adaptação (MARTINS et. al., 2009; MARTINS, 2010) e o modelo passa a ser escrito como:

$$\left\{ \begin{array}{rcl} Y_1(\underline{x}_i) & = & \mu_1(\underline{x}_i) + \sigma_1 U(\underline{x}_i; \phi) + Z_1(\underline{x}_i) \\ Y_2(\underline{x}_{i'}) & = & \mu_2(\underline{x}_{i'}) + \sigma_2 U(\underline{x}_{i'}; \phi) + Z_2(\underline{x}_{i'}). \end{array} \right.$$

em que $\underline{x}_i, \underline{x}_{i'} \in \mathbb{R}^2$, são as localizações amostrais, $i, i^{'} = 1, ..., n_1$, sendo n_1 o tamanho da amostra; $Y_1 = \ln(X_1/X_3)$, $Y_2 = \ln(X_2/X_3)$ são as variáveis resposta do modelo de modo que $\underline{Y}_{n\times 1} = \left(Y_1(\underline{x}_1), Y_2(\underline{x}_1), \ldots, Y_1(\underline{x}_{n_1}), Y_2(\underline{x}_{n_1})\right)^{'}$. Neste modelo, assume-se que U é um efeito aleatório com distribuição gaussiana multivariada com vetor de médias iguais a zero e matriz de variância/covariância, com variâncias unitárias e covariâncias dadas pela função de correlação exponencial (ρ_U). Esta funçao é caracterizada pelo parâmetro de alcance, ϕ, que controla o decaimento da correlação como função da separação espacial entre duas localizações. No modelo bivariado geral as unidades de medida são preservadas nas constantes padronizadoras σ_1 e σ_2, enquanto que no contexto considerado aqui são admensionais. Os efeitos aleatórios $Z_j \sim N(0; \tau_j^2)$, $j = 1, 2$ capturam a variabilidade não espacial incluindo a correlação, ρ, induzida pela estrutura composicional.

Sendo assim, $\underline{Y} \sim N_2(\underline{\mu}; \mathbf{\Sigma})$, com matriz de covariâncias $\mathbf{\Sigma}$ composta pelos elementos

$$\begin{array}{rclrcl} Cov(Y_1(\underline{x}_i); Y_1(\underline{x}_i)) & = & \sigma_1^2 + \tau_1^2 & Cov(Y_1(\underline{x}_i); Y_1(\underline{x}_{i'})) & = & \sigma_1^2 \rho_U(\underline{x}_i; \underline{x}_{i'}) \\ Cov(Y_2(\underline{x}_i); Y_2(\underline{x}_i)) & = & \sigma_2^2 + \tau_2^2 & Cov(Y_2(\underline{x}_i); Y_2(\underline{x}_{i'})) & = & \sigma_2^2 \rho_U(\underline{x}_i; \underline{x}_{i'}) \end{array}$$

e

$$Cov(Y_1(\underline{x}_i); Y_2(\underline{x}_{i'})) = \sigma_1 \sigma_2 I_2(i, i') + \tau_1 \tau_2 I_3(i, i')$$

em que

$$I_2(i, i') = \begin{cases} 1 & , \text{ se } \quad i = i' \\ \rho_U(\underline{x}_i; \underline{x}_{i'}) & , \text{ se } \quad i \neq i' \end{cases} \qquad I_3(i, i') = \begin{cases} \rho & , \text{ se } \quad i = i' \\ 0 & , \text{ se } \quad i \neq i'. \end{cases}$$

Desta forma, a inferência sobre o vetor de parâmetros $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)'$ que sob o paradigma bayesiano é considerado aleatório, é feita de modo que a função de verossimilhança é escrita como

$$L(\underline{\theta}, \underline{Y}(\underline{x})) \propto |\mathbf{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} \left(\underline{Y}(\underline{x}) - \underline{\mu}_{\underline{Y}(\underline{x})} \right)^{'} \mathbf{\Sigma}^{-1} \left(\underline{Y}(\underline{x}) - \underline{\mu}_{\underline{Y}(\underline{x})} \right) \right\}.$$

Esta função é reparametrizada de acordo com Martins et. al. (2009), Martins (2010) de modo

$$\mathbf{\Sigma} \;=\; \sigma_1^2 \mathbf{R} \;+\; \tau_1^2 \mathbf{I}_b \;=\; \sigma_1^2 \mathbf{V}$$

em que $\mathbf{R}$ é uma matriz de correlação espacial de ordem $n \times n$, e $\mathbf{I}_b$ é a matriz produto de Kronecker,

$$\mathbf{I}_b = \begin{bmatrix} \nu_1^2 & \nu_1 \nu_2 \rho \\ \nu_1 \nu_2 \rho & \nu_2^2 \end{bmatrix} \otimes \mathbf{I}_n.$$

Para se fazer inferência bayesiana é preciso construir a distribuição conjunta dos vetores aleatórios $\underline{Y}(\underline{x})$ e $\underline{\theta}$ que, do Teorema de Bayes, resulta na distribuição *a posteriori* para $\underline{\theta}$

$$P(\underline{\theta} | \underline{Y}(\underline{x})) = \frac{L(\underline{\theta}, \underline{Y}(\underline{x})) P(\underline{\theta})}{P(\underline{Y}(\underline{x}))} \propto L(\underline{\theta}, \underline{Y}(\underline{x})) P(\underline{\theta})$$

ou

$$P(\underline{\mu}, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho | \underline{Y}(\underline{x})) \propto P(\underline{Y}(\underline{x}) | \underline{\mu}, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho) P(\underline{\mu}, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho).$$

Observa-se que para o modelo geoestatístico bivariado composicional não é possível derivar analiticamente as distribuições marginais *a posteriori* para todos os parâmetros do modelo. Neste caso, considerando a reparametrização (MARTINS et.al., 2009; MARTINS, 2010) com $\underline{\theta}^* = (\eta, \nu_1, \nu_2, \phi, \rho)'$ e supondo $(\underline{\mu}, \sigma_1^2)$ e $\underline{\theta}^*$ independentes pode-se escrever

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

e então é possível obter expressões fechadas para as distribuições marginais para $\underline{\mu}$ e σ_1^2. Para isto, observa-se que integrando ambos os lados da Equação (1) em relação à $\underline{\theta}^*$ tem-se

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

e supondo $1/\sigma_1^2$ como *priori* para $P(\underline{\mu}, \sigma_1^2)$ obtém-se

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

Não sendo possível derivar analiticamente a distribuição *a posteriori* para $\underline{\theta}^*$ e tampouco calcular analiticamente a integral em (2), esta foi resolvida numericamente usando *Monte Carlo via cadeias de Markov*, por meio do algoritmo de Metropolis-Hastings. Foram executadas 12000 simulações, o período de aquecimento da cadeia foi de 1000 simulações usando um salto igual a 10 e uma sintonia (*tune*) igual a (0,5; 0,5; 0,5; 0,5; 0,5; 0,5; 0,5; 0,5). Com isto as cadeias foram formadas por 1200 simulações.

As distribuições *a priori* utilizadas para η, ν_1 e ν_2 foram Log-normais com parâmetros correpondentes ao logaritmo das razões das estimativas de máxima verossimilhança obtidas no processo de maximização. O algoritmo de otimização utilizado nesse processo foi o "L-BFGS-B", método de Byrd et al. (1995). Este método permite informar os limites inferior e superior de busca no espaço paramétrico, e os valores iniciais para os parâmetros foram calculados considerando-se as estimativas obtidas através da amostra. Foi adotado ρ como sendo o coeficiente de correlação de Pearson e $\phi = \min + 0.2(\max - \min)$, onde "min" e "max" são, respectivamente, a menor e maior distância entre duas localizações. Para ϕ utilizou-se uma distribuição Gama com parâmetros (66; 1) em que o valor 66 é a estimativa de máxima verossimilhança de ϕ e para ρ uma distribuição *a priori flat* definida no intervalo $[-1, 1]$. Os valores estimados obtidos foram então substituídos em (2) de modo que:

$$\begin{aligned} P(\underline{\mu}, \sigma_1^2 | \underline{\theta}^*, \underline{Y}(\underline{x})) &\propto (\sigma_1^2)^{-1} P(\underline{Y}(\underline{x}) | \underline{\mu}, \sigma_1^2, \underline{\theta}^*) \\ &\propto (\sigma_1^2)^{-(\frac{n}{2}+1)} \exp\left\{ -\frac{1}{2} \left(\underline{Y}(\underline{x}) - \underline{\mu}_{\underline{Y}(\underline{x})} \right)' \mathbf{V}^{-1} \left(\underline{Y}(\underline{x}) - \underline{\mu}_{\underline{Y}(\underline{x})} \right) \right\}. \end{aligned}$$

Da definição de probabilidade condicional obtém-se

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

de forma que, supondo uma *priori flat* para $P(\underline{\mu}|\sigma_1^2, \underline{\theta}^*)$, ou seja, $[\underline{\mu}|\sigma_1^2, \underline{\theta}^*] \propto 1$ vem

$$P(\underline{\mu}|\sigma_1^2, \underline{\theta}^*, \underline{Y}(\underline{x})) \quad \propto \quad (\sigma_1^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_1^2}\left(\underline{\mu} - \underline{\hat{\mu}}\right)'\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)\left(\underline{\mu} - \underline{\hat{\mu}}\right)'\right\},$$

e a distribuição *a posteriori* marginal para $\underline{\mu}|\sigma_1^2, \underline{\theta}^*, \underline{Y}(\underline{x})$ é

$$\left[\underline{\mu}|\sigma_1^2, \underline{\theta}^*, \underline{Y}(\underline{x})\right] \sim N\left(\underline{\hat{\mu}}; \sigma_1^2\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\right), \tag{3}$$

sendo $\underline{\hat{\mu}} = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}(\mathbf{D}'\mathbf{V}^{-1}\underline{Y})$. Como

$$P(\underline{\mu}, \sigma_1^2|\underline{\theta}^*, \underline{Y}(\underline{x})) = P(\underline{\mu}|\sigma_1^2, \underline{\theta}^*, \underline{Y}(\underline{x}))P(\sigma_1^2|\underline{\theta}^*, \underline{Y}(\underline{x})) \quad \Rightarrow \quad P(\sigma_1^2|\underline{\theta}^*, \underline{Y}(\underline{x})) = \frac{P(\underline{\mu}, \sigma_1^2|\underline{\theta}^*, \underline{Y}(\underline{x}))}{P(\underline{\mu}|\sigma_1^2, \underline{\theta}^*, \underline{Y}(\underline{x}))}, \tag{4}$$

substituindo (3) em (4) tem-se:

$$P(\sigma_1^2|\underline{\theta}^*, \underline{Y}(\underline{x})) \propto (\sigma_1^2)^{-(\frac{n-n_{\underline{\mu}}}{2}+1)} \exp\left\{-\frac{1}{2\sigma_1^2}(n - n_{\underline{\mu}})S^2\right\},$$

e a distribuição marginal *a posteriori* para $\sigma_1^2|\underline{\theta}^*, \underline{Y}(\underline{x})$ é

$$\left[\sigma_1^2|\underline{\theta}^*, \underline{Y}(\underline{x})\right] \sim \chi^2_{\text{Sinv}}\left(n - n_{\underline{\mu}}; S^2\right), \tag{5}$$

sendo χ^2_{Sinv} a distribuição qui-quadrado escalonada inversa e

$$S^2 = \frac{n}{n - n_{\underline{\mu}}}\hat{\sigma}_1^2 = \frac{\left(\underline{Y}(\underline{x}) - \mathbf{D}\underline{\hat{\mu}}\right)'\mathbf{V}^{-1}\left(\underline{Y}(\underline{x}) - \mathbf{D}\underline{\hat{\mu}}\right)}{n - n_{\underline{\mu}}}. \tag{6}$$

Uma vez que métodos de simulação de cadeias de *Markov* são utilizados, diagnósticos de convergência devem ser realizados a fim de estimar o quanto a inferência baseada em simulações de *cadeia de Markov* diferem da distribuição *a posteriori* (GILKS, RICHARDSON e SPIEGELHALTER, 1996).

Inicialmente, calculou-se as autocorrelações dos parâmetros *a posteriori* (GILL, 2002) sendo indicativos de convergência da cadeia, os decréscimos nas autocorrelações à medida que as defasagens aumentam. Gráficos da trajetória da cadeia também foram utilizados e, neste caso, espera-se que a série ocorra de forma estável em torno da linha média. A não convergência é observada quando a trajetória se afasta seguindo uma tendência que não seja a linha média. Além disso, gráficos das densidades estimadas dos parâmetros *a posteriori* auxiliaram no sentido de verificação de multimodalidade, indicativo de não convergência da cadeia.

O teste de Geweke apresentado em Gill (2002) e Gamerman (2006), por exemplo, baseia-se na comparação de alguma proporção do início da cadeia após o período de aquecimento com alguma proporção do fim da cadeia. Neste trabalho foi adotada uma fração de 90% e 50%, respectivamente. É um teste de diferença de médias usando uma aproximação assintótica para o erro padrão da diferença. Valores da estatística teste superiores a 2 em termos absolutos indicam preocupação com a falta de convergência, mas valores pequenos não indicam convergência.

Outra forma de avaliação da convergência apresentada pelos mesmos autores é o Diagnóstico de Sequência Múltipla de Gelman e Rubin que se baseia na comparação de um conjunto de

cadeias com diferentes pontos que são super dispersos com relação à distribuição *a posteriori*. Baseia-se também na teoria da aproximação gaussiana para as *posterioris* marginais usando testes como ANOVA e diagnóstico com distribuição t-*Student*. Um indicador de convergência é o fator *redução de escala estimado* que, de acordo com a literatura, valores próximos a 1, mais precisamente, inferiores a $1,2$ são aceitáveis para convergência. Maiores detalhes dos diagnósticos de convergência são discutidos em Gilks, Richardson e Spiegelhalter (1996), Gamerman (2006), Gill (2002), Carlin e Louis (2009), Martins (2010), entre outros.

O próximo passo é a realização da predição espacial (cokrigagem) de $\underline{Y}_0$ em localizações não amostradas $\underline{x}_0 = (\underline{x}_{10}, \underline{x}_{20}, ..., \underline{x}_{n_20})$. De acordo com Ribeiro Jr. e Diggle (1999), a base da predição bayesiana é a distribuição preditiva $P(\underline{S}(\underline{x})|\underline{Y}(\underline{x}))$. Esta distribuição leva em consideração a incerteza sobre os parâmetros calculando, por exemplo, a média da distribuição condicional $P(\underline{S}(\underline{x})|\underline{Y}(\underline{x}), \underline{\theta})$, sobre o espaço dos parâmetros, com pesos dados pela distribuição *a posteriori* dos parâmetros do modelo $P(\underline{\theta}|\underline{Y}(\underline{x}))$:

$$\begin{aligned} P(\underline{S}(\underline{x})|\underline{Y}(\underline{x})) &= \int P(\underline{S}(\underline{x}), \underline{\theta}|\underline{Y}(\underline{x}))\, d\underline{\theta} \\ &= \int P(\underline{S}(\underline{x})|\underline{Y}, \underline{\theta}(\underline{x}))P(\underline{\theta}|\underline{Y}(\underline{x}))\, d\underline{\theta}. \end{aligned}$$

Cabe ressaltar que podem ser calculadas outras estatísticas de interesse ou mais apropriadas, como a mediana ou moda, a partir da distribuição preditiva.

Uma vez que a transformação ALR foi aplicada aos dados originais e o procedimento de estimação e cokrigagem foi realizada com os dados transformados em $\mathbb{R}^2$, deve-se fazer a transformação de volta do vetor de médias e da matriz de covariância para o espaço amostral original, o simplex $\mathbb{S}^3$. Para isto, geram-se dados bivariados de uma distribuição gaussiana multivariada com vetor de médias e matriz de covariância iguais ao vetor de médias e matriz de covariância obtidos por cokrigagem. Aplica-se a transformação AGL nos resultados obtendo-se a distribuição de cada componente, em cada localização a partir do qual se calculam os valores esperados preditos e, por último, mapas de predição espacial bayesiana são contruídos.

Com o desenvolvimento descrito pode-se resumir os passos para uma análise bayesiana de dados composicionais espacializados, segundo o modelo proposto, da seguinte forma:

a. derivar analiticamente as distribuições a *posteriori* para $\underline{\mu}|\sigma_1^2, \underline{\theta}^*, \underline{Y}(\underline{x})$ e $\sigma_1^2|\underline{\theta}^*, \underline{Y}$;

b. executar o algoritmo Metropolis-Hastings no vetor $\underline{\theta}^*$;

c. com o vetor $\underline{\theta}^*$ estimado, calcular a matriz de covariância $\mathbf{V}$ do modelo;

d. calcular $\hat{\underline{\mu}}$ dado por $\hat{\underline{\mu}} = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}(\mathbf{D}'\mathbf{V}^{-1}\underline{Y}(\underline{x}))$;

e. calcular S^2 dado em (6);

f. executar um passo Gibbs que significa amostrar um valor de $\sigma_1^2|\underline{\theta}^*, \underline{Y}(\underline{x})$ da distribuição (5);

g. com $\hat{\underline{\mu}}$ e $\text{Var}(\underline{\mu}|\sigma_1^2, \underline{\theta}^*, \underline{Y}(\underline{x}))$, executar um passo Gibbs amostrando um valor de $\underline{\mu}|\sigma_1^2, \underline{\theta}^*, \underline{Y}(\underline{x})$ de (3);

h. repetir o procedimento até o número de simulações desejado;

i. fazer a cokrigagem com cada conjunto de parâmetros simulados;

j. gerar uma amostra de uma distribuição normal multivariada com cada resultado obtido no item (i), vetor de médias e matriz de covariância;

k. aplicar a transformação AGL nos resultados obtidos no item (j);

l. construir mapa de predição para cada componente da composição.

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 *geoR* (RIBEIRO JR. e DIGGLE, 2001), *compositions* (BOOGAART, TOLOSANA e BREN, 2009), *statmod* (SMYTH, HU e DUNN, 2009), *MCMCpack* (MARTIN, QUINN e PARK, 2009), *coda* (Rnews: PLUMER, BEST, COWLES e VINES, 2006) e rotinas desenvolvidas especificamente para o desenvolvimento deste trabalho. Estas rotinas foram organizadas e dispostas em forma de pacote *geoComp* disponível em ≪http://www.leg.ufpr.br≫.

3 Análise de Frações Granulométricas de Um Solo

Como exemplo de aplicação da metodologia proposta analisou-se um conjunto de dados obtidos a partir do trabalho de de Gonçalves (1997) conduzido em uma área irrigada por sistema tipo pivô-central na Fazenda Areão, Figura 1, pertencente ao campus da Escola Superior de Agricultura - "Luiz de Queiroz" (ESALQ-USP). Nela foi demarcado um quadrante na porção mais elevada (topo da encosta) no qual foram obtidas 76 amostras de solo na profundidade entre 0 e 0, 20m em uma malha regular quadrada de amostragem, de lado igual a 20 metros. Em cada amostra foram medidos os valores das frações granulométricas, de areia, silte e argila.

Figura 1: Foto aérea do campo experimental de irrigação da ESALQ-USP com área de estudo correspondente ao quadrante irrigado por um sistema pivô-central.
FONTE: Gonçalves (1997).

A Figura 2 apresenta a configuração das 76 localizações na área de estudo correspondente ao primeiro quadrante do sistema de coordenadas, irrigado pelo sistema tipo pivô-central.

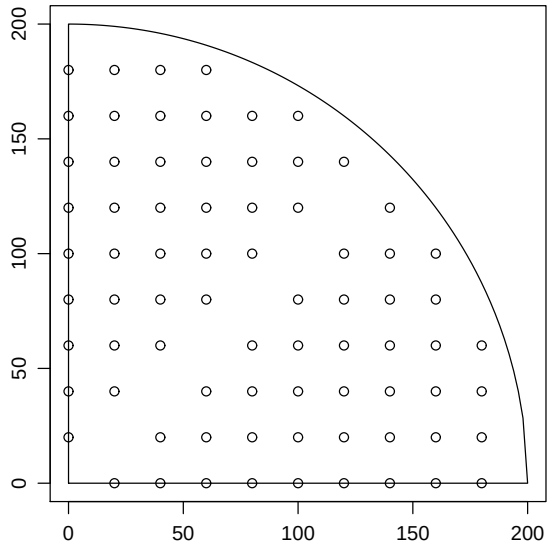

Figura 2: Distribuição das localizações na área de estudo.

Em cada uma destas localizações foram medidos os percentuais de areia, silte e argila e pelos histogramas da Figura 3, pode-se observar a distribuição destes componentes. Também, pelo diagrama ternário observa-se que nas composições amostradas, o silte é o componente que se apresenta em menor proporção e uma região de confiança de 4-desvios-padrão contemplou todas as amostras.

As distribuições dos dados após a transformação ALR, podem ser vistas na Figura 4 e o diagrama de dispersão evidencia uma correlação linear positiva entre as variáveis transformadas.

As esperanças *a posteriori* dos parâmetros bem como os respectivos intervalos de credibilidade são apresentados na Tabela 1.

Tabela 1: Esperanças de 1200 simulações da distribuição *a posteriori* de $\underline{\theta}$ e intervalos de 95% de credibilidade considerando-se 12000 simulações, *burn-in*= 1000 e salto= 10.

Parâmetros	Esperanças	LI	LS
μ_1	-0,764	-1,195	-0,334
μ_2	-0,795	-0,935	-0,674
σ_1	0,443	0,341	0,557
σ_2	0,113	0,059	0,180
τ_1	0,287	0,222	0,357
τ_2	0,266	0,226	0,309
ϕ	65,728	51,534	84,458
ρ	0,931	0,825	0,992

Os mapas dos valores esperados preditos de areia, silte e argila obtidos por inferência bayesiana são apresentados na Figura 5, cujo padrão geral assemelha-se aos obtidos por inferência clássica (MARTINS, et. al., 2009, MARTINS, 2010) (Figura 5) e observa-se que os

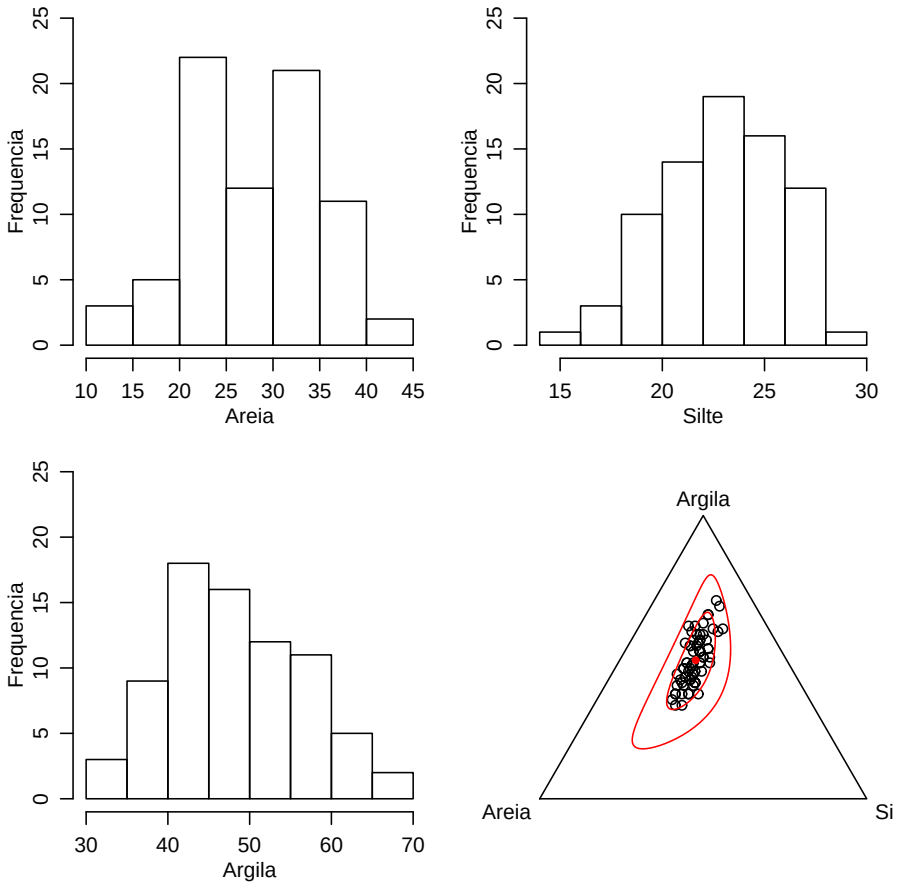

Figura 3: Distribuição dos percentuais de areia, silte e argila e diagrama ternário das composições.

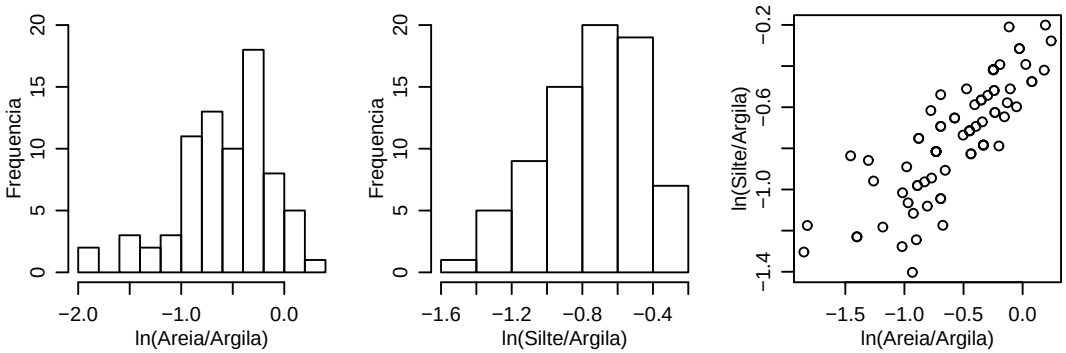

Figura 4: Distribuição das log-razão e correspondente diagrama de dispersão.

componentes areia e argila se complementam na área de estudo.

4 Conclusão

Os procedimentos adotados permitem a construção de mapas das percentagens de areia, silte e argila por uma metodologia que implicitamente garante a restrição de que as frações somem 1, não só nos pontos observados como nos pontos preditos. O modelo proposto captura variações espaciais, induzidas pelas composições e não estruturadas. A declaração explícita do modelo

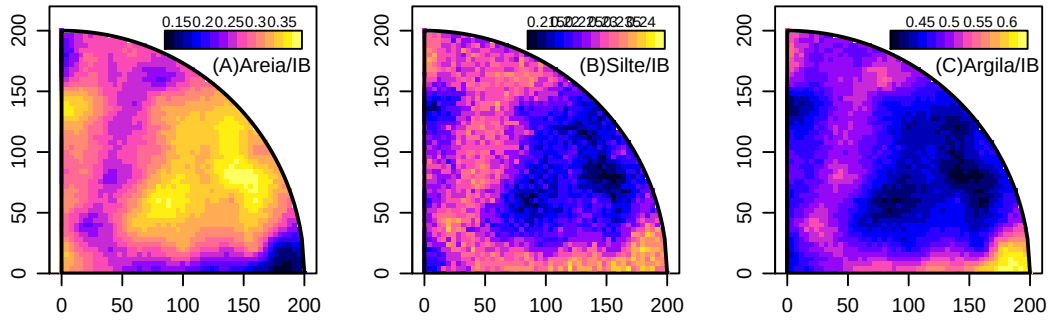

Figura 5: Valores esperados preditos dos percentuais de areia (à esquerda), silte (centro) e argila (à direita) por inferência bayesiana.

permite o tratamento bayesiano dos dados a fim de se considerar nas predições a incerteza associada à estimação dos parâmetros do modelo. As análises podem ser expandidas para estimação de outros funcionais que não necessariamente resultem em mapas de teores médios. Há a necessidade de se investigar alternativas para computação mais eficiente e considerar outras formas de especificação do modelo multivariado para o caso de maiores números de componentes.

Referências

[1] AITCHISON, J., *The statistical analysis of compositional data*, New Jersey: The Blackburn Press, 1986.

[2] BANERJEE, S.; CARLIN, B.P.; GELFAND, G.E., *Hierarchical modelling and analysis for spatial data*, Boca Raton: Chapman and Hall, 2004.

[3] BOGNOLA, I.A.; RIBEIRO Jr, P.J.; SILVA, E.A.A; LINGNAU, C.;, HIGA, A.R. Modelagem uni e bivariada da variabilidade espacial de rendimento de pinus taeda l., *Floresta*, v. 38, n. 2, p. 373-385, 2008.

[4] BOOGAART, G.v.d.; TOLOSANA, R.;BREN, M., *compositions: compositional data analysis*. Disponível em:<http://www.stat.boogaart.de/compositions>. Acesso em 25 maio 2009.

[5] 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.

[6] BYRD, R.H.; LU, P.; NOCEDAL, J.; ZHU, C., A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, v. 16, p. 1190-1208, 1995.

[7] CARLIN, B.P.; LOUIS, T.A., *Bayesian Methods for Data Analysis*, Boca Raton: Chapman and Hall, 2009.

[8] DIGGLE, P.J.; RIBEIRO JR, P.J., *Model-based geostatistics*, USA: Springer Series in Statistics, 2007.

[9] FINLEY, A.O.;BANERJEE, S. and CARLIN, B. P., spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, v. 19, n. 4, p. 1-24, 2007.

[10] GAMERMAN, D., *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, London: Chapman and Hall/CRC, 2006.

[11] GILL, J., *Bayesian Methods For The Social And Behavioral Sciences*, London: Chapman and Hall, 2002.

[12] GILL, J.; RICHARDSON,S. and SPIEGELHALTER, D.J., *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall, 1996

[13] 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, 1997.

[14] GRAF, M., Precision of compositional data in a stratified two-stage cluster sample: comparison of the swiss earnings structure survey 2002 and 2000. *Survey Research Methods Section, ASA*, Session 415: Sample Survey Quality V, p. 3066−3072, 2006.

[15] LARK, R.M. and BISHOP, T.F.A., Cokriging particle size fractions of the soil. *European Journal of Soil Science*, v. 58, p. 763−774, 2007.

[16] MARTIN, A.D., QUINN, K.M. and PARK, J.H., Markov chain Monte Carlo (MCMC) Package. *European Journal of Soil Science*, n. 1.0-4, p. 1−103, 2009. Disponível em:<http://mcmcpack.wustl.edu>. Acesso em 01 outubro 2009.

[17] MARTINS, A.B.T., *Análise geoestatística bivariado de dados composicionais.*. 2010. 184p., Tese (Doutorado em Ciências), Universidade Federal do Paraná, Curitiba, 2010.

[18] MARTINS, A.B.T.; RIBEIRO JR, P.J.; BONNAT, W.H.; GONÇALVES, A.C.A., Um modelo geoestatístico bivariado para dados composicionais. *Revista Brasileira de Biometria*, v. 27, n. 3, p. 456−477, 2009.

[19] OBAGE, S.C., *Uma análise bayesiana para dados composicionais.*. 2007. 69p., Dissertação (Mestrado em Estatística), Universidade Federal São Carlos, São Paulo, 2007.

[20] PAWLOWSKY-GLAHN, V.; OLEA, R.A., *Geostatistical analysis of compositional data*, New York: Oxford University Press, Inc., 2004.

[21] PLUMMER, M.; BEST, N.; COWLES, K. and VINES, K., CODA: Convergence Diagnosis and Output Analysis for MCMC, *R-NEWS*, v. 6, n. 1, p. 7-11, 2006. Disponível em:<http://CRAN.R-project.org/doc/Rnews>. Acesso em 15 setembro 2009.

[22] R Development Core Team.,*R: a language and environment for statistical computing*, Vienna, Austria, 2009. Disponível em:<http://www.R-project.org>. Acesso em: 25 maio 2009.

[23] RIBEIRO JR, P.J.; DIGGLE, P.J., geoR: a package for geostatistical analysis, *R-NEWS*, v. 1, n. 2, p. 14-18, 2001. Disponível em:<http://CRAN.R-project.org/doc/Rnews/>. Acesso em 25 maio 2009.

[24] SCHMIDT, A.M.; SANSÓ, B., Modelagem bayesiana da estrutura de covariância de processos espaciais e espaço temporais, In: *17 SINAPE e ABE-Associação Brasileira de Estatística*, Caxambu: Associação Brasileira de Estatística, 2006. Minicurso.

[25] TJELMELAND, H.; LUND, K.V., Bayesian modelling of spatial compositional data, *Journal of Applied Statistics*, v. 30, p. 87-100, 2003.

[26] WACKERNAGEL, H., *Multivariate Geostatistics: An Introduction with Applications*, Germany: Springer, 1998.