UM MODELO GEOESTATÍSTICO BIVARIADO PARA DADOS COMPOSICIONAIS

ANA BEATRIZ TOZZO MARTINS[1]
PAULO JUSTINIANO RIBEIRO JUNIOR[2]
WAGNER HUGO BONAT[3]

▪ RESUMO: Este trabalho é motivado pelo interesse em modelar padrões espaciais em dados composicionais. A categoria de problemas de interesse envolve, por exemplo, as frações granulométricas de um solo ou composição química de uma rocha, isto é, estruturas de dados em que as observações são partes de algum todo e referenciadas espacialmente. O interesse central está em predizer, conjuntamente, o padrão espacial dos componentes na área de estudo respeitando a restrição da soma dos componentes corresponderem ao total. Neste sentido, combina-se a teoria de dados composicionais originalmente desenvolvida para observações independentes com métodos geoestatísticos multivariados. A abordagem proposta declara explicitamente um modelo paramétrico que descreve a espacialização das variáveis. Em particular neste trabalho, o objetivo é propor e implementar um modelo geoestatístico bivariado para dados composicionais obtendo inferências via verossimilhança e expressões para predições espaciais, apresentando, discutindo e disponibilizando resultados e implementação computacional. A metodologia proposta é utilizada na análise de um conjunto de dados simulados, cujas composições são formadas por três componentes, possibilitando a obtenção de mapas de predição para cada um dos componentes.

▪ PALAVRAS-CHAVE: geoestatística multivariada; dados composicionais; verossimilhança.

[1]PPGMNE/DES, CESEC/CCE, UFPR/UEM, Rua Mariano Torres, 714, Ap64, CEP: 80060 − 120, Curitiba, Pr, Brazil, E-mail: *abtmartins@uem.br*

[2]DEst, LEG, UFPR, Caixa Postal: 19.081 , CEP: 81531-990, Curitiba, Pr, Brazil, E-mail: *paulojus@leg.ufpr.br*

[3]PPGMNE, LEG, UFPR, Caixa Postal: 19.081, CEP: 81531-990, Curitiba, Pr, Brazil, E-mail: *wbonat@leg.ufpr.br*

1 Introdução

Este estudo é motivado pelo interesse em modelar e descrever o padrão espacial de dados composicionais. Neste sentido, combina-se a teoria de dados composicionais originalmente desenvolvida para observações independentes (AITCHISON, 1986) com métodos geoestatísticos (PAWLOWSKY-GLAHN; OLEA, 2004), adotando-se a declaração explícita do modelo que descreve 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 um (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

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

Um vetor $\underset{\sim}{\mathrm{W}}$ 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 um seja satisfeita:

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

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

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

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 $g(\underset{\sim}{\mathrm{X}}_1) = \sqrt[B]{\prod_{j=1}^{B} X_{1j}}$.

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 usuais podem levar a resultados inconsistentes. Para contornar este problema, 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:

$$\begin{aligned} \text{alr} : \mathbb{S}^B &\longrightarrow \mathbb{R}^{B-1} \\ \underset{\sim}{\mathrm{X}} &\longrightarrow \text{alr}(\underset{\sim}{\mathrm{X}}) = \left(\ln\left(\frac{X_1}{X_B}\right), \ldots, \ln\left(\frac{X_{B-1}}{X_B}\right)\right)'. \end{aligned}$$

Por outro lado, a transformação inversa denominada transformação logística generalizada aditiva (agl) é dada por

$$\begin{aligned} \text{agl} : \mathbb{R}^{B-1} &\longrightarrow \mathbb{S}^B \\ \text{alr}(\underset{\sim}{\mathrm{X}}) &\longrightarrow \text{agl}(\text{alr}(\underset{\sim}{\mathrm{X}})) = \underset{\sim}{\mathrm{X}} = \left(\exp\left(\ln\left(\frac{X_1}{X_B}\right)\right), \ldots, \exp(0)\right)'. \end{aligned}$$

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). De acordo com Graf (2006), considerando que assintoticamente $\underset{\sim}{\mathrm{Y}} = (\ln(X_1/X_B), \ln(X_2/X_B), ..., \ln(X_{B-1}/X_B))'$ tem distribuição Gaussiana $N_{B-1}(\mu_L; \Sigma_L)$ em que μ_L é o vetor de médias das log-razões e Σ_L a matriz de covariâncias de ordem $(B-1) \times (B-1)$, o domínio de confiança para $\underset{\sim}{\mathrm{Y}}$ é limitado pelo elipsóide

$$B'_{1-\alpha}(\underset{\sim}{\mathrm{Y}}) = \left\{\underset{\sim}{\mathrm{Y}} \in \mathbb{R}^{B-1} | \left(\ln(\underset{\sim}{\mathrm{Y}}) - \mu_L\right)' \Sigma_L^{-1} \left(\ln(\underset{\sim}{\mathrm{Y}}) - \mu_L\right) \leq \chi^2_{B-1;1-\alpha}\right\}$$

em que $\chi^2_{B-1;1-\alpha}$ é o quantil $(1-\alpha)$ da distribuição qui-quadrado com $B-1$ graus de liberdade e o domínio correspondente para $\underset{\sim}{\mathrm{X}} = (X_1, X_2, ..., X_B)$ é um subconjunto do simplex $\mathbb{S}^B$

$$B'_{1-\alpha}(\underset{\sim}{\mathrm{X}}) = \left\{\underset{\sim}{\mathrm{X}} \in \mathbb{S}^B | \left(\ln\frac{\underset{\sim}{\mathrm{X}}_{-B}}{X_B} - \mu_L\right)' \Sigma^{-1} \left(\ln\frac{\underset{\sim}{\mathrm{X}}_{-B}}{X_B} - \mu_L\right) \leq \chi^2_{B-1;1-\alpha}\right\}.$$

A teoria de dados composicionais desenvolvida por Aitchison (1986) para amostras independentes, foi extendida por Pawlowsky-Glahn e Olea (2004) considerando a dependência espacial segundo métodos geoestatísticos em uma abordagem que evita declarar completa e explicitamente o modelo multivariado espacial associado às composições. Por outro lado, modelagens multivariadas espaciais são descritas em Diggle e Ribeiro Jr (2007), Schmidt e Gelfand (2003), Banerjee, Carlin e Gelfand (2004), Schmidt e Sansó (2006) garantindo, por construção, matrizes de covariância definidas positiva. Considera-se aqui o modelo Gaussiano bivariado de componentes comum proposto por Diggle e Ribeiro Jr (2007) e adotado por Bognola et. al. (2008) que utilizam 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 baseada na verossimilhança e obtendo preditores espaciais, na escala original dos dados no simplex, que permitam a construção de mapas de predição dos componentes, ou funcionais destes, na área de estudo.

2 Metodologia

Para $\underline{X} = (X_1, ..., X_B)'$ 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). Por simplificação, neste trabalho considera-se composições de apenas 3 componentes, X_1 = Componente 1, X_2 = Componente 2 e X_3 = Componente 3, que resultam em vetores bivariados.

A partir do modelo geoestatístico apresentado em Diggle e Ribeiro Jr (2007), propõe-se uma adaptação e o modelo pode ser escrito como:

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

em que $\underline{x}_i, \underline{x}_{i'} \in \mathbb{R}^2$, são as localizações amostrais $i, i' = 1, ..., n_1$, n_1 é 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 de média zero e variância unitária com distribuição Gaussiana multivariada com correlações espaciais dadas pela função de correlação exponencial (ρ_U), caracterizada por um parâmetro 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

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

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

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

Inferências sobre o vetor de parâmetros $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)'$ são baseadas na verossimilhança correspondente à densidade da distribuição normal multivariada com os dois primeiros elementos do vetor de parâmetros determinando o vetor de médias e os demais parametrizando a matriz de covariâncias em

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

Fazendo-se a reparametrização: $\eta = \sigma_2/\sigma_1$; $\nu_1 = \tau_1/\sigma_1$; $\nu_2 = \tau_2/\sigma_1$, pode-se escrever

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

em que $\mathbf{R}$ corresponde a parte espacial do modelo, $\mathbf{I}_b$ é uma matriz bloco diagonal correspondente a parte composicional e

$$\mathbf{V} = \begin{bmatrix} 1+\nu_1^2 & \eta+\nu_1\nu_2\rho & \rho_U(\underline{x}_1,\underline{x}_2) & \eta\rho_U(\underline{x}_1,\underline{x}_2) & \cdots & \rho_U(\underline{x}_1,\underline{x}_n) & \eta\rho_U(\underline{x}_1,\underline{x}_n) \\ \eta+\nu_1\nu_2\rho & \eta^2+\nu_2^2 & \eta\rho_U(\underline{x}_1,\underline{x}_2) & \eta^2\rho_U(\underline{x}_1,\underline{x}_2) & \cdots & \eta\rho_U(\underline{x}_1,\underline{x}_n) & \eta^2\rho_U(\underline{x}_1,\underline{x}_n) \\ \rho_U(\underline{x}_2,\underline{x}_1) & \eta\rho_U(\underline{x}_2,\underline{x}_1) & 1+\nu_1^2 & \eta+\nu_1\nu_2\rho & \cdots & \rho_U(\underline{x}_2,\underline{x}_n) & \eta\rho_U(\underline{x}_2,\underline{x}_n) \\ \eta\rho_U(\underline{x}_2,\underline{x}_1) & \eta^2\rho_U(\underline{x}_2,\underline{x}_1) & \eta+\nu_1\nu_2\rho & \eta^2+\nu_2^2 & \cdots & \rho_U(\underline{x}_2,\underline{x}_n) & \eta^2\rho_U(\underline{x}_2,\underline{x}_n) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_U(\underline{x}_n,\underline{x}_1) & \eta\rho_U(\underline{x}_n,\underline{x}_1) & \rho_U(\underline{x}_n,\underline{x}_2) & \eta\rho_U(\underline{x}_n,\underline{x}_2) & \cdots & 1+\nu_1^2 & \eta+\nu_1\nu_2\rho \\ \eta\rho_U(\underline{x}_n,\underline{x}_1) & \eta^2\rho_U(\underline{x}_n,\underline{x}_1) & \eta\rho_U(\underline{x}_n,\underline{x}_2) & \eta^2\rho_U(\underline{x}_n,\underline{x}_2) & \cdots & \eta+\nu_1\nu_2\rho & \eta^2\nu_2^2 \end{bmatrix}.$$

A função de log-verossimilhança reparametrizada é dada por

$$l(\underline{\theta},\underline{Y}) = -\frac{1}{2}\left(n\ln(2\pi) + 2n\ln(\sigma_1) + \ln(|\mathbf{V}|) + \frac{1}{\sigma_1^2}Qe\right). \qquad (1)$$

Considerando $\underline{\mu}_{\underline{Y}} = \mathbf{D}\underline{\mu}$ em que $\mathbf{D}$ é a matriz do delineamento de ordem $n \times 2$, tem-se

$$Qe = (\underline{Y}-\underline{\mu}_{\underline{Y}})'\mathbf{V}^{-1}(\underline{Y}-\underline{\mu}_{\underline{Y}}) = \underline{Y}'\mathbf{V}^{-1}\underline{Y} - 2(\underline{Y}'\mathbf{V}^{-1}\mathbf{D})\underline{\mu} + \underline{\mu}'(\underline{Y}'\mathbf{V}^{-1}\mathbf{D})\underline{\mu}.$$

Expressões analíticas fechadas podem ser obtidas para os estimadores de máxima verossimilhança de $\underline{\mu} = (\mu_1,\mu_2)'$ e σ_1 diferenciando a função (1) em relação aos respectivos parâmetros e estes são dados por

$$\underline{\hat{\mu}} = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}(\mathbf{D}'\mathbf{V}^{-1}\underline{Y}) \quad \text{e} \quad \hat{\sigma}_1 = \sqrt{\hat{Q}e/n}\,, \qquad (2)$$

e $\hat{Q}e$ pode ser escrito como

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

Ao substituir as expressões (2) em (1) obtém-se a função de log-verossimilhança concentrada

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

que é uma função do vetor de parâmetros desconhecidos $\underline{\theta}^* = (\eta,\nu_1,\nu_2,\phi,\rho)'$, e é ser maximizada numericamente.

Os algoritmos de otimização testados no processo de maximização foram "L-BFGS-B", método de Byrd et al. (1995) que permite informar os limites inferior e superior de busca no espaço paramétrico; "Nelder-Mead", uma implementação do método de Nelder e Mead (1965); "Gradiente Conjugado", baseado no método de Fletcher e Reeves (1964) e "BFGS", um método quasi-Newton. Todos encontram-se implementados no ambiente estatístico R (R Development Core Team, 2008).

Do processo de maximização obtém-se $\hat{\underline{\theta}}^* = (\hat{\eta}, \hat{\nu_1}, \hat{\nu_2}, \hat{\phi}, \hat{\rho})'$ e as respectivas variâncias através da matriz Hessiana numérica. A matriz Informação de Fisher observada é definida como o negativo da matriz Hessiana e é dada por

$$\mathbf{I}_o(\hat{\underline{\theta}}^*) = -\frac{\partial^2 l(\hat{\underline{\theta}}^*)}{\partial \hat{\underline{\theta}}^* \partial (\hat{\underline{\theta}}^*)'}$$

Para se obter $\hat{\mu}_1$, $\hat{\mu}_2$, e $\hat{\sigma}_1$, basta substituir $\hat{\underline{\theta}^*}$ nas equações em (2).

Como o interesse está na obtenção de $\hat{\underline{\theta}} = (\hat{\mu_1}, \hat{\mu_2}, \hat{\sigma_1}, \hat{\sigma_2}, \hat{\tau_1}, \hat{\tau_2}, \hat{\phi}, \hat{\rho})'$ e suas respectivas variâncias, o método delta (DEGROOT, 2002, p.284; COX e HINKLEY, 1974, p.302; AZZALINI, 1996, p.73; PAWITAN, 2001, p.226) é aplicado para obter uma aproximação da distribuição de $\hat{\underline{\theta}}$. Assintoticamente, a distribuição de $\hat{\underline{\theta}}$ será aproximadamente multivariada Gaussiana com vetor de médias $\hat{\underline{\theta}} = g(\hat{\underline{\theta}}^*)$ e variância

$$\mathrm{Var}(\hat{\underline{\theta}}) = \mathbf{I}_E(\hat{\underline{\theta}}) \geq \nabla g(\hat{\underline{\theta}}^*)' \, \mathbf{I}_E(\hat{\underline{\theta}}^*)^{-1} \, \nabla g(\hat{\underline{\theta}}^*),$$

em que

$$\nabla g(\hat{\underline{\theta}}^*) = \left(\frac{\partial g(\hat{\underline{\theta}}^*)}{\partial \eta}, \; \frac{\partial g(\hat{\underline{\theta}}^*)}{\partial \nu_1}, \; \frac{\partial g(\hat{\underline{\theta}}^*)}{\partial \nu_2}, \; \frac{\partial g(\hat{\underline{\theta}}^*)}{\partial \phi}, \; \frac{\partial g(\hat{\underline{\theta}}^*)}{\partial \rho} \right)'$$

é a função escore $\mathsf{U}(\hat{\underline{\theta}}^*)$. Assim, considerando-se $g(\hat{\underline{\theta}}^*)$ igual a $\sigma_2 = \eta\sigma_1$, $\tau_1 = \nu_1\sigma_1$, $\tau_2 = \nu_2\sigma_1$, ϕ e ρ, respectivamente, tem-se

$$\nabla g(\hat{\underline{\theta}}^*) = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \sigma_1 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

e a matriz Informação de Fisher esperada para $\hat{\underline{\theta}}$ baseada nos dados $\underline{Y}$ é substituída pela matriz $\mathbf{I}_o(\hat{\underline{\theta}}^*)$, assintoticamente equivalente, de modo que

$$\mathrm{Var}(\hat{\underline{\theta}}) \geq \nabla g(\hat{\underline{\theta}}^*)' \, \mathbf{I}_o(\hat{\underline{\theta}}^*)^{-1} \, \nabla g(\hat{\underline{\theta}}^*).$$

Para encontrar as variâncias para $\hat{\underline{\mu}}$ e $\hat{\sigma}_1$, através da função (1), obtém-se

$$\mathbf{I}_o(\underline{\mu}) = -\frac{\partial^2 l(\underline{\theta})}{\partial \underline{\mu}^2} = \frac{1}{\sigma_1^2}(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})' \quad \text{e} \quad \mathbf{I}_o(\sigma_1) = -\frac{\partial^2 l(\underline{\theta})}{\partial \sigma_1^2} = -\frac{n}{\sigma_1} + \frac{3Qe}{\sigma_1^3},$$

e portanto,

$$\begin{aligned} \mathrm{Var}(\hat{\underline{\mu}}) &= \mathbf{I}_o(\hat{\underline{\mu}})^{-1} &&= \hat{\sigma}_1^2(\mathbf{D}'\hat{\mathbf{V}}^{-1}\mathbf{D})^{-1} \\ \mathrm{Var}(\hat{\sigma}_1) &= \mathbf{I}_o(\hat{\sigma}_1)^{-1} &&= \frac{\hat{\sigma}_1^3}{3\hat{Q}e - n\hat{\sigma}_1}. \end{aligned}$$

O próximo passo é a realização da predição espacial de $\underline{Y}_0$ em localizações não amostradas $\underline{x}_0 = (\underline{x}_{10}, \underline{x}_{20}, ..., \underline{x}_{n_20})$. Neste caso, por envolver mais de uma variável,

tem-se um preditor de *cokrigagem*. O vetor de valores esperados correspondentes às variáveis Y_1 e Y_2 para todas as localizações de predição e a matriz de covariância são dadas pelo seguinte resultado da distribuição Gaussiana multivariada:

Theorem 2.1 *Seja* $\underline{Y} = (\underline{Y}_0, \underline{Y})'$ *um vetor bivariado com distribuição Gaussiana multivariada conjunta com vetor de médias* $\underline{\mu} = (\underline{\mu}_{Y_0}, \underline{\mu}_Y)'$ *e matriz de covariância*

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{\underline{Y}_0\underline{Y}_0} & \mathbf{\Sigma}_{\underline{Y}_0\underline{Y}} \\ \mathbf{\Sigma}_{\underline{Y}\underline{Y}_0} & \mathbf{\Sigma}_{\underline{Y}\underline{Y}} \end{bmatrix}$$

isto é, $\underline{Y} \sim N(\underline{\mu};\ \mathbf{\Sigma})$. *Então, a distribuição condicional de* $\underline{Y}_0$ *dado* $\underline{Y}$ *é também Gaussiana multivariada,*

$$\underline{Y}_0|\underline{Y} \sim N(\underline{\mu}_{\underline{Y}_0|\underline{Y}};\ \mathbf{\Sigma}_{\underline{Y}_0|\underline{Y}})$$

em que

$$\underline{\mu}_{\underline{Y}_0|\underline{Y}} = \underline{\mu}_{\underline{Y}_0} + \mathbf{\Sigma}_{\underline{Y}_0\underline{Y}}\mathbf{\Sigma}_{\underline{Y}\underline{Y}}^{-1}(\underline{Y} - \underline{\mu}_{\underline{Y}}) \quad e \quad \mathbf{\Sigma}_{\underline{Y}_0|\underline{Y}} = \mathbf{\Sigma}_{\underline{Y}_0\underline{Y}_0} - \mathbf{\Sigma}_{\underline{Y}_0\underline{Y}}\mathbf{\Sigma}_{\underline{Y}\underline{Y}}^{-1}\mathbf{\Sigma}_{\underline{Y}\underline{Y}_0}.$$

Sendo desconhecidos os valores de $\underline{\mu}_{\underline{Y}_0}$, estes são substituídos pelo vetor de médias estimadas obtidas no processo de otimização. A matriz $\mathbf{\Sigma}$, de onde se extrai as matrizes $\mathbf{\Sigma}_{\underline{Y}_0\underline{Y}_0}$, $\mathbf{\Sigma}_{\underline{Y}_0\underline{Y}}$, $\mathbf{\Sigma}_{\underline{Y}\underline{Y}_0}$ e $\mathbf{\Sigma}_{\underline{Y}\underline{Y}}$, é obtida substituindo-se os valores estimados para os outros parâmetros na matriz $\mathbf{V}$ multiplicada por $\hat{\sigma}_1^2$.

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 $\mathbb{S}^3$, como descrito em Pawlowsky e Olea (2004).

O objetivo é calcular para cada localização uma estimativa de

$$\underline{\mu}_{\underline{X}} = E(\underline{X}) = \int_{\mathbb{S}^B} \underline{X} f(\underline{X}) d\underline{X} \tag{3}$$

e

$$\mathbf{\Sigma}_{\underline{X}} = Cov(\underline{X}, \underline{X}) = \int_{\mathbb{S}^B} (\underline{X} - \underline{\mu}_{\underline{X}})(\underline{X} - \underline{\mu}_{\underline{X}})' f(\underline{X}) d\underline{X}. \tag{4}$$

Sabe-se que

$$f(\underline{Y}) = (2\pi)^{-\frac{B-1}{2}} |\mathbf{\Sigma}_{\underline{Y}}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left(\underline{\mathrm{Y}} - \underline{\mu}_{\underline{Y}} \right)' \mathbf{\Sigma}_{\underline{Y}}^{-1} \left(\underline{\mathrm{Y}} - \underline{\mu}_{\underline{Y}} \right) \right\},$$

segue uma distribuição Gaussiana multivariada e $\underline{Y} = \mathrm{alr}(\underline{X})$. Assim, para voltar a escala original, $\underline{X} = \mathrm{agl}(\underline{Y})$, é necessário uma transformação de variável (JAMES, 2004) cujo jacobiano é dado por

$$J_{\mathrm{alr}(\underline{X})} = \left| \frac{\partial \underline{Y}}{\partial \underline{X}} \right| = \left(\prod_{i=1}^{B} X_i \right)^{-1} \tag{5}$$

de modo que $f(\underline{X})$ é dada por

$$f(\underline{X}) = (2\pi)^{-\frac{B-1}{2}} |\mathbf{\Sigma}_{\underline{Y}}|^{-\frac{1}{2}} \left(\prod_{i=1}^{B} X_i \right)^{-1} \exp\left\{ -\frac{1}{2} \left(\text{alr}(\underline{X}) - \underline{\mu}_{\underline{Y}} \right)' \mathbf{\Sigma}_{\underline{Y}}^{-1} \left(\text{alr}(\underline{X}) - \underline{\mu}_{\underline{Y}} \right) \right\}.$$

Aitchison (1986), Pawlowsky e Olea (2004) resolvem as integrais por mátodos de quadratura (3) e (4) expressando-as por

$$\underline{\mu}_{\underline{X}} = \int_{\mathbb{R}^{B-1}} g_1(\underline{\mathsf{Z}}) f(-\underline{\mathsf{Z}}'\underline{\mathsf{Z}}) d\underline{\mathsf{Z}} \tag{6}$$

e

$$\mathbf{\Sigma}_{\underline{X}} = \int_{\mathbb{R}^{B-1}} g_2(\underline{\mathsf{Z}}) f(-\underline{\mathsf{Z}}'\underline{\mathsf{Z}}) d\underline{\mathsf{Z}} \tag{7}$$

em que $\underline{\mathsf{Z}}$ é a transformação

$$\underline{\mathsf{Z}} = \frac{\text{alr}(\underline{X}) - \underline{\mu}_{\underline{Y}}}{\sqrt{2}\mathsf{R}'} = \frac{1}{\sqrt{2}} (\mathsf{R}')^{-1} (\text{alr}(\underline{\mathrm{X}}) - \underline{\mu}_{\underline{Y}}), \tag{8}$$

com R sendo a decomposição Cholesky de $\mathbf{\Sigma}_{\underline{Y}}$, matriz triangular superior. As integrais (6) e (7) podem, então, ser aproximadas pela integração de Gauss-Hermite multivariada de ordem k:

$$\int_{\mathbb{R}^{B-1}} g(\underline{\mathsf{Z}}) f(-\underline{\mathsf{Z}}'\underline{\mathsf{Z}}) d\underline{\mathsf{Z}} \approx \sum_{i_1=1}^{\mathsf{k}} \sum_{i_2=1}^{\mathsf{k}} \cdots \sum_{i_{B-1}=1}^{\mathsf{k}} \omega_{i_1}\omega_{i_2}\cdots\omega_{i_{B-1}} g(\mathsf{Z}_{i_1}, \mathsf{Z}_{i_2}, ..., \mathsf{Z}_{i_{B-1}}). \tag{9}$$

Nesta aproximação, os pesos $\omega_{i_1}\omega_{i_2}\cdots\omega_{i_{B-1}}$ e as abscissas $\mathsf{Z}_{i_1}, \mathsf{Z}_{i_2}, ..., \mathsf{Z}_{i_{B-1}}$ são conhecidas e seus valores podem ser encontrados, por exemplo, em Abramowitz e Stegun (1972, p. 924). Segundo Gammerman (1997) ordens de quadratura de 6 a 8 são suficientes.

Explicitando o procedimento, observa-se que a integral (3) pode ser reescrita como

$$\underline{\mu}_{\underline{X}} = \int_{\mathbb{S}^B} g(\underline{X}) \left((2\pi)^{-\frac{B-1}{2}} |\mathbf{\Sigma}_{\underline{Y}}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left(\text{alr}(\underline{\mathrm{X}}) - \underline{\mu}_{\underline{Y}} \right)' \mathbf{\Sigma}_{\underline{Y}}^{-1} \left(\text{alr}(\underline{\mathrm{X}}) - \underline{\mu}_{\underline{Y}} \right) \right\} \right) d\underline{X}$$

em que

$$g(\underline{X}) = \underline{X} \left(\prod_{i=1}^{B} X_i \right)^{-1}. \tag{10}$$

Considere que $\mathbf{\Sigma}_{\underline{Y}} = \mathsf{R}'\mathsf{R}$ e $\mathbf{\Sigma}_{\underline{Y}}^{-1} = (\mathsf{R}'\mathsf{R})^{-1} = \mathsf{R}^{-1}(\mathsf{R}')^{-1} = \mathsf{R}^{-1}(\mathsf{R}^{-1})'$. Da transformação (8), tem-se

$$\underline{X} = \text{agl}(\underline{\mu}_{\underline{Y}} + \sqrt{2}\mathsf{R}'\underline{\mathsf{Z}}) = \text{agl}(\underline{Y})$$

e de (5),

$$\partial \underline{X} = \left(\prod_{i=1}^{B} X_i \right) \partial \underline{Y}. \tag{11}$$

Além disso,

$$J = \left| \frac{\partial \underline{Y}}{\partial \underline{Z}} \right| = \left| \frac{\partial (\underline{\mu}_{\underline{Y}} + \sqrt{2}\mathsf{R}'\underline{Z})}{\partial \underline{Z}} \right| = \left| \sqrt{2}\mathsf{R}' \right| = (\sqrt{2})^{B-1} |\mathsf{R}'| \, ;$$

sendo R a decomposição cholesky de $\mathbf{\Sigma}_{\underline{Y}}$,

$$|\mathsf{R}'| = |\mathsf{R}| = |\mathbf{\Sigma}_{\underline{Y}}|^{\frac{1}{2}} \ ,$$

e

$$J = \left| \frac{\partial \underline{Y}}{\partial \underline{Z}} \right| = 2^{\frac{B-1}{2}} \, |\mathbf{\Sigma}_{\underline{Y}}|^{\frac{1}{2}} \quad \Rightarrow \quad \partial \underline{Y} = 2^{\frac{B-1}{2}} \, |\mathbf{\Sigma}_{\underline{Y}}|^{\frac{1}{2}} \, \partial \underline{Z}. \tag{12}$$

Substituindo a equação (12) em (11) tem-se

$$\partial \underline{X} = \left(\prod_{i=1}^{B} X_i \right) 2^{\frac{B-1}{2}} \, |\mathbf{\Sigma}_{\underline{Y}}|^{\frac{1}{2}} \, \partial \underline{Z} \quad \Rightarrow \quad \partial \underline{Z} = 2^{-\frac{B-1}{2}} \left(\prod_{i=1}^{B} X_i \right)^{-1} |\mathbf{\Sigma}_{\underline{Y}}|^{-\frac{1}{2}} \, \partial \underline{X}.$$

Por outro lado, de (10) tem-se $g\left(\text{agl}(\underline{\mu}_{\underline{Y}} + \sqrt{2}\mathsf{R}'\underline{Z})\right) = \text{agl}\left(\underline{\mu}_{\underline{Y}} + \sqrt{2}\mathsf{R}'\underline{Z}\right)\left(\prod_{i=1}^{B} X_i\right)^{-1}$ que substituída em (6) produz

$$\underline{\mu}_{\underline{X}} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g_1(\underline{Z}) \exp\{\underline{Z}'\underline{Z}\} d\underline{Z} \tag{13}$$

com $g_1(\underline{Z}) = \pi^{-\frac{B-1}{2}} \text{agl}\left(\underline{\mu}_{\underline{Y}} + \sqrt{2}\mathsf{R}'\underline{Z}\right)$. Logo, a aproximação de Gauss-Hermite (9) de ordem 3, por exemplo, para $\underline{\mu}_{\underline{X}}$ é

$$\underline{\mu}_{\underline{X}} \approx \sum_{i_1=1}^{3} \sum_{i_2=1}^{3} \omega_{i_1} \omega_{i_2} g_1(\mathsf{Z}_{i_1}, \mathsf{Z}_{i_2})$$

Lembrando que $\underline{\mu}_{\underline{X}}$ representa a média da composição em cada localização, esta é um vetor trivariado. Então na função g_1, $\underline{\mu}_{\underline{Y}}$ é um vetor bivariado extraído de $\underline{\mu}_{\underline{Y}_0|\underline{Y}}$, cujo primeiro elemento corresponde à média da cokrigagem para a variável Y_1, o segundo para a variável Y_2 e R' de ordem 2×2 será a correspondente matriz de variância/covariâncias de Y_1, Y_2, extraída do bloco diagonal da matriz $\mathbf{\Sigma}_{\underline{Y}_0|\underline{Y}}$. Assim, em cada localização a resolução de (13) implica na resolução de 3 integrais.

O mesmo procedimento é aplicado para a obtenção de $\mathbf{\Sigma}_{\underline{X}}$ dada pela integral (7), agora reescrita como

$$\Sigma_{\underline{X}} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g_2(\underline{Z}) \exp\{\underline{Z}'\underline{Z}\} d\underline{Z} \tag{14}$$

com $g_2(\underline{Z}) = \pi^{-\frac{B-1}{2}} \left(\text{agl}(\underline{\mu}_{\underline{Y}} + \sqrt{2}\mathsf{R}'\underline{Z}) - \underline{\mu}_{\underline{X}}\right) \left(\text{agl}(\underline{\mu}_{\underline{Y}} + \sqrt{2}\mathsf{R}'\underline{Z}) - \underline{\mu}_{\underline{X}}\right)'$ e, neste caso, sendo $\mathbf{\Sigma}_{\underline{X}}$ uma matriz de ordem 3×3, a resolução da integral (14) para cada localização implica na resolução de 9 integrais. Portanto, a estimação do vetor de médias e variâncias em n_2 localizações implica na resolução de $n_2 \times 12$ integrais.

3 Análise de Dados Composicionais Simulados

Como exemplo de aplicação da metodologia proposta analisou-se três conjuntos de dados simulados, cada um com 100 valores de percentagens de três componentes, X_1, X_2 e X_3 e as localizações em um quadrado unitário encontram-se representadas na Figura 1.

O primeiro conjunto, denominado configuração 1, foi gerado a partir do vetor de parâmetros $\underline{\theta}$=$(\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho) = (-0,2; -0,5; 1; 1,5; 0,3; 0,3; 0,6; 0,9)$. Buscou-se, colocando valores das médias próximos e de mesmo sinal, fazer com que a nuvem de pontos se situasse na parte central do diagrama ternário. O fato de estarem concentrados deu-se pelo alto valor de ρ. Quanto menor o valor deste parâmetro, mais espalhados são os dados na representação no diagrama ternário. O fato das variâncias dos efeitos espaciais serem próximos com $\sigma_1 < \sigma_2$ fez com que os pontos se aproximassem mais do vértice X_2 do que do X_1 e, por último, valores iguais para as variâncias τ_1 e τ_2 justificou os pontos no interior do diagrama.

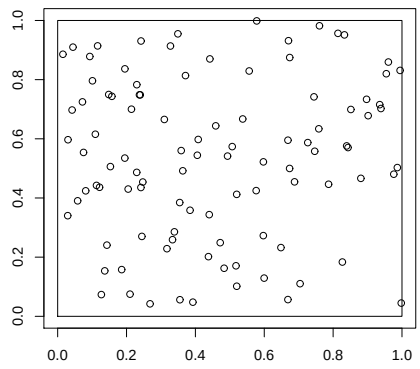

Figura - 1: Distribuição das localizações da primeira configuração no quadrado unitário.

Em cada uma destas localizações tem-se os percentuais de X_1, X_2 e X_3 e as 100 composições representadas por círculos em um diagrama ternário podem ser vistas na Figura 2. Composições mais próximas a um vértice têm altas proporções do componente correspondente àquele vértice, de modo que esta figura indica que

as maiores porcentagens na amostra ocorreram para componente X_3. Também, pode-se observar, por exemplo, que a variabilidade da razão X_3/X_1 foi maior que a variabilidade da razão X_2/X_1. Por outro lado, a variabilidade da razão X_1/X_3, X_2/X_3 foram similares. Além disso, uma região de confiança de 4-desvios-padrão contemplou todas as amostras e o centro da distribuição dado pelo fechamento da média geométrica dos três componentes está representado pelo ponto vermelho no diagrama. Pelos histogramas nota-se que as porcentagens dos três componentes não se aproximam de uma distribuição Gaussiana.

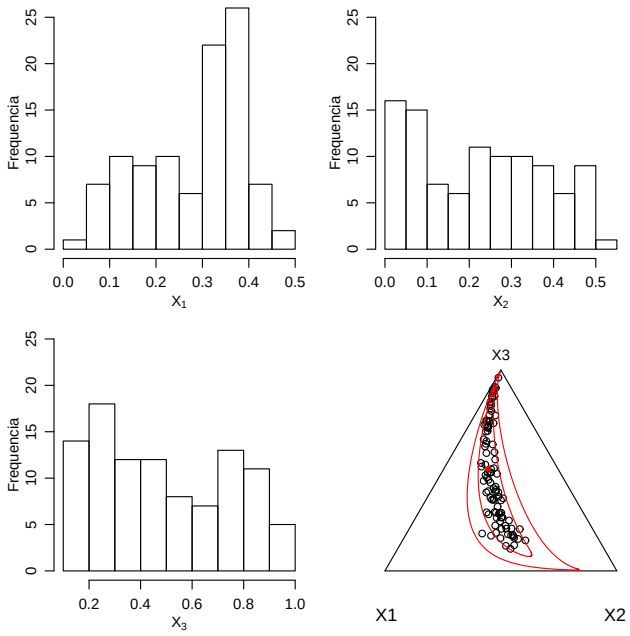

Figura - 2: Distribuição de X_1, X_2 e X_3 e diagrama ternário das composições para a primeira configuração.

Porém, ao fazer a transformação alr nos dados simulados obteve-se as variáveis $Y_1 = \ln(X_1/X_3)$ e $Y_2 = \ln(X_2/X_3)$ e os dados transformados passaram a seguir distribuição Gaussiana como pode ser visto pela Figura 3. O diagrama de dispersão, por sua vez, mostra pontos alinhados em forma linear crescente indicando uma possível correlação linear positiva.

No ajuste do modelo adotou-se o método de otimização "L-BFGS-B" por não apresentar problemas de convergência e as estimativas obtidas para os parâmetros bem como os respectivos intervalos de confiança calculados pelo método delta são apresentados na Tabela 1. Observa-se, em geral, estimativas não muito próximas aos valores dos parâmetros mas com mesmos sinais e que apenas três de oito intervalos contêm os verdadeiros valores.

No processo de otimização os valores iniciais para o vetor de médias foram

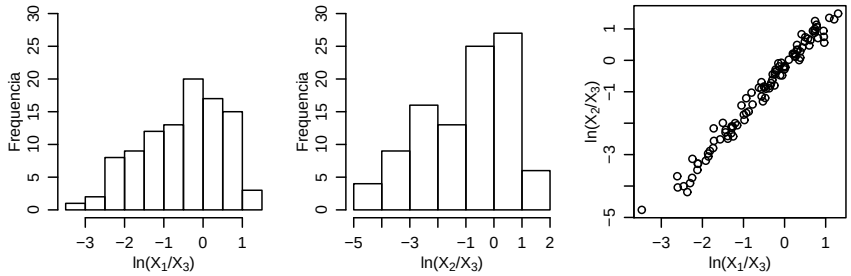

Figura - 3: Distribuição das log-razão e correspondente diagrama de dispersão para a primeira configuração.

Tabela - 1: Estimativas, erros padrão e intervalos de confiança para a primeira configuração, pelo método delta via método de otimização "L-BFGS-B".

Parâmetro	Valor	Estimativa	Erro Padrão	LI. Delta	LS. Delta
μ_1	-0,2	-0,9925955	1,60663318	-2,5992287	0,6140377
μ_2	-0,5	-1,6890643	1,98310148	-3,6721658	0,2940372
σ_1	1	1,1530662	0,10190384	1,0511623	1,2549700
σ_2	1,5	1,7581358	0,05412534	1,7040104	1,8122611
τ_1	0,3	0,4004140	0,02972426	0,3706897	0,4301382
τ_2	0,3	0,4274941	0,04045833	0,3870357	0,4679524
ϕ	0,6	0,9530621	0,51284603	0,4402160	1,4659081
ρ	0,9	0,9573298	0,02626424	0,9310656	0,9835941

considerados como as médias dos valores observados para Y_1 e Y_2. Metade da variância calculada para Y_1 foi atribuída para o efeito espacial e a outra metade para o efeito composicional. Da mesma forma, procedeu-se com Y_2. O valor incial para ρ foi calculado como o coeficiente de correlação de Pearson e $\phi = \min + 0,2(\max - \min)$, onde "min" e "max" foram, respectivamente, a menor e maior distância entre duas localizações.

Como resultado da cokrigagem obteve-se, para cada localização, o vetor de médias e a matriz de covariância no espaço $\mathbb{R}^2$. A volta dos valores preditos para o simplex $\mathbb{S}^3$ foi feita em uma grade constituída de 1156 pontos usando aproximação de Gauss-Hermite com ordem de quadratura $k = 7$. Na Figura 4 tem-se os mapas dos três componentes conforme a configuração 1 e, assim como revelou o diagrama ternário, as maiores proporções ocorreram para o componente X_3. A Figura 5, por sua vez, mostra boa concordância entre os valores observados e preditos para os três componentes.

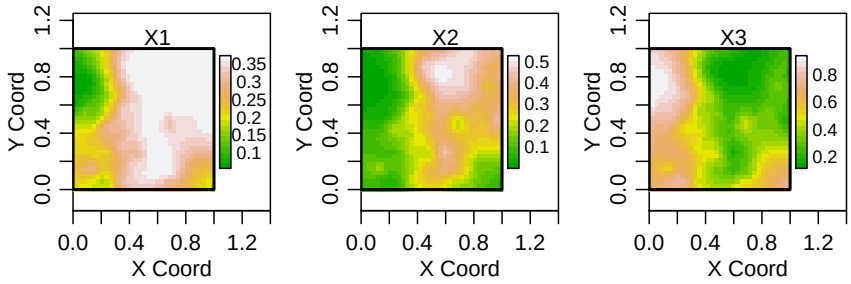

Figura - 4: Mapas das porcentagens de $X1$, $X2$ e $X3$ para dados da configuração 1.

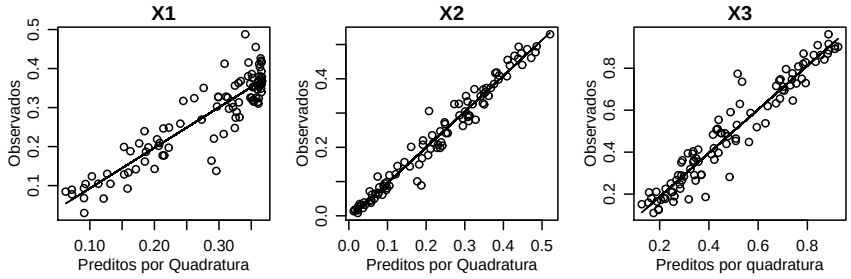

Figura - 5: Valores observados versus preditos de areia, silte e argila para a configuração 1.

O segundo conjunto de dados, configuração 2, foi gerado considerando-se o vetor de parâmetros $(\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho) = (1; 1; 1, 2; 1, 5; 0, 9; 1; 0, 6; 0, 5)$. Assim como na configuração 1, em relação às médias, buscou-se fazer com que a nuvem de pontos continuasse na parte central do diagrama; a proximidade dos

pontos ao vértice X_3 justificada pelo baixo valor de ϕ e o espalhamento pelo baixo valor de ρ. Também aqui, $\sigma_1 < \sigma_2$ fez com que os pontos se aproximassem mais de X_2 do que de X_1 e a diferença entre τ_1 e τ_2 forçou os pontos a se aproximarem mais do lado que liga os vértices X_1 e X_2, com maior aproximação para X_2 por τ_1 ser o menor valor.

Para este vetor de parâmetros a configuração das composições no diagrama ternário da Figura 6 apresenta-se mais espalhada em relação à anterior com maiores porcentagens relativas aos componentes X_2 e X_3. Ao mesmo tempo, observa-se muitas composições com porcentagens muito baixas de X_3 o que se confirma pelo respectivo histograma.

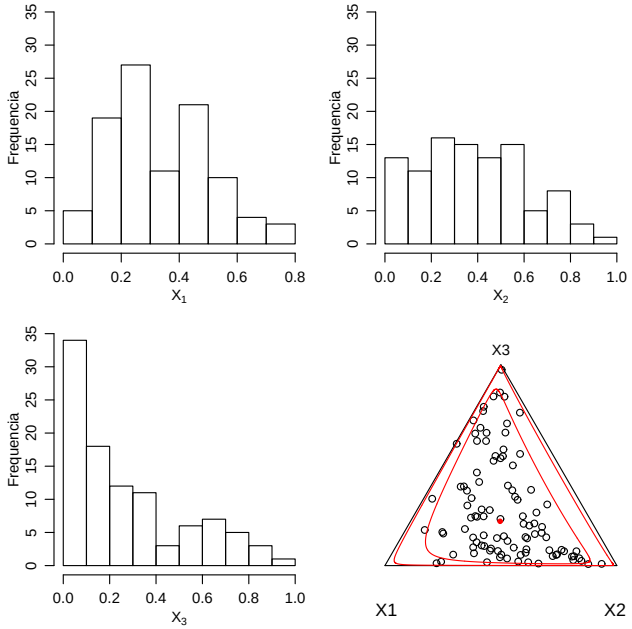

Figura - 6: Distribuição de X_1, X_2 e X_3 e diagrama ternário das composições para a segunda configuração.

A Figura 7 mostra que os dados transformados também seguem uma distribuição Gaussiana e a nuvem de pontos no diagrama de dispersão continua de forma linear crescente mas com um espalhamento maior em relação ao da configuração 1.

Na Tabela 2 tem-se os resultados do ajuste do modelo para a configuração 2 e o que se observa é que as estimativas permanecem com mesmos sinais; para as médias não são boas mas os respectivos intervalos de confiança contêm os verdadeiros valores. Somente o intervalo para σ_2 não conteve o valor do parâmetro apesar de estar próximo ao limite inferior.

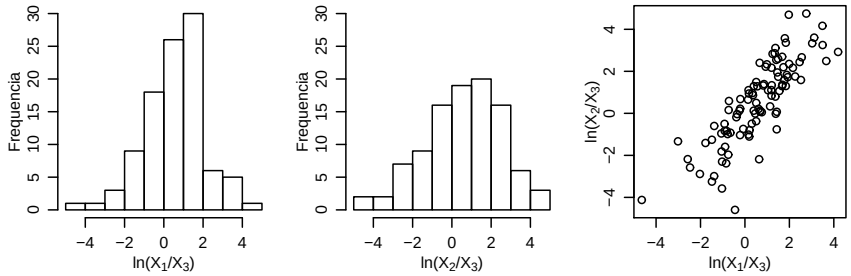

Figura - 7: Distribuição das log-razão e correspondente diagrama de dispersão para a segunda configuração.

Tabela - 2: Estimativas, erros padrão e intervalos de confiança para a segunda configuração, pelo método delta via método de otimização "L-BFGS-B".

Parâmetro	Valor	Estimativa	Erro Padrão	LI. Delta	LS. Delta
μ_1	1	0,3883751	1,4633944	-1,07501928	1,8517695
μ_2	1	0,2844116	1,7558488	-1,47143723	2,0402604
σ_1	1,2	1,2584047	0,1046900	1,15371472	1,3630947
σ_2	1,5	1,8313026	0,2648620	1,56644065	2,0961646
τ_1	0,9	0,9824042	0,1806521	0,80175207	1,1630563
τ_2	1	1,0052126	0,2132289	0,79198365	1,2184415
ϕ	0,6	0,5236357	0,4430230	0,08061266	0,9666587
ρ	0,5	0,5401549	0,1562457	0,38390917	0,6964006

Na Figura 8 tem-se os mapas onde se observam poucas mudanças em relação aos mapas obtidos para a configuração 1 e a Figura 9 mostra que concordância já não é tão boa para X_1 quanto para X_2 e X_3.

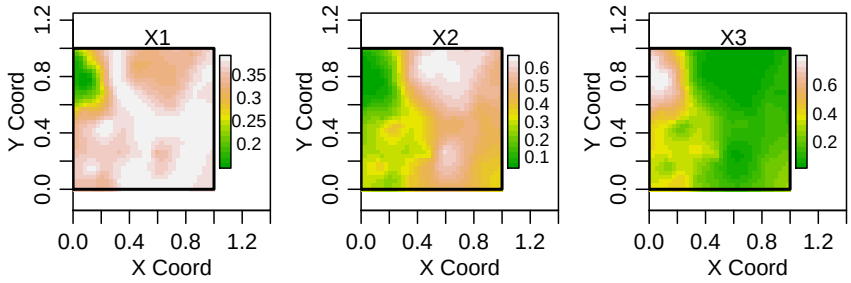

Figura - 8: Mapas das porcentagens de $X1$, $X2$ e $X3$ para dados da configuração 2.

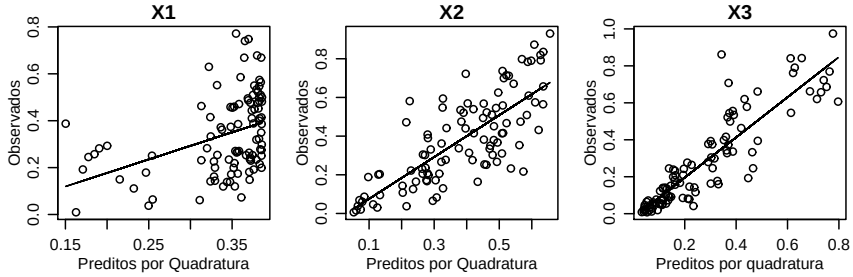

Figura - 9: Valores observados versus preditos de areia, silte e argila para a configuração 2.

A última configuração deste estudo caracterizou-se pelo conjunto de parâmetros $(\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho) = (-0,2; -1; 0.45; 0,13; 0,3; 0,3; 0,6; 0,95)$. O baixo valor absoluto de μ_1 fez com que os dados se aproximassem do lado esquerdo do diagrama com proximidade a X_1; caso contrário, a aproximação seria para o lado direito. Também neste caso, o baixo valor de ϕ aproximou os pontos ao vértice X_3 enquanto o alto valor de ρ tornou-os concentrados novamente. Como nesta configuração $\sigma_1 > \sigma_2$, os pontos se afastaram do vértice X_2 e continuaram no interior do diagrama pelos valores iguais e baixos para τ_1 e τ_2.

Como pode ser visto na Figura 10 tem-se baixos percentuais do componente X_2, e as distribuições de X_1, X_2 e X_3 se aproximam de uma distribuição Gaussiana.

De acordo com a Figura 11, os dados transformados continuam apresentando distribuição Gaussiana e os pontos no diagrama de dispersão também apresentam-se em forma linear crescente mas com espalhamento maior ainda em relação aos anteriores.

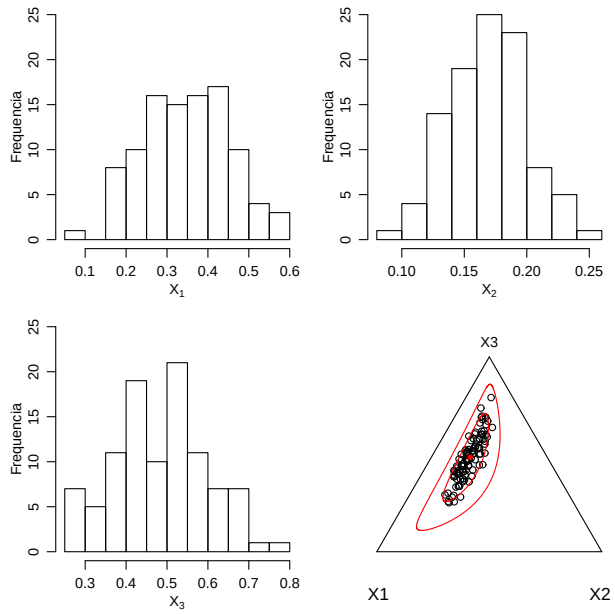

Figura - 10: Distribuição de X_1, X_2 e X_3 e diagrama ternário das composições para a terceira configuração.

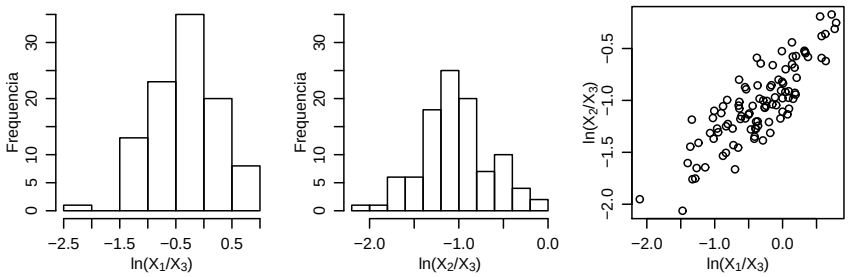

Figura - 11: Distribuição das log-razão e correspondente diagrama de dispersão para a terceira configuração.

Tabela - 3: Estimativas, erros padrão e intervalos de confiança para a terceira configuração, pelo método delta via método de otimização "L-BFGS-B".

Parâmetro	Valor	Estimativa	Erro Padrão	LI. Delta	LS. Delta
μ_1	-0,2	-0,5896071	0,69119506	-1,2808021	0,1015880
μ_2	-1	-1,1146150	0,40214445	-1,5167595	-0,7124706
σ_1	0,45	0,5305950	0,09559040	0,4350046	0,6261854
σ_2	0,13	0,1765225	0,05364894	0,1228735	0,2301714
τ_1	0,3	0,3488848	0,03599187	0,3128929	0,3848766
τ_2	0,3	0,3446375	0,03074784	0,3138897	0,3753853
ϕ	0,6	0,7403627	0,52040659	0,2199561	1,2607693
ρ	0,95	0,9383066	0,03744746	0,9008592	0,9757541

Pela Tabela 3, também neste caso as estimativas conservaram os mesmos sinais dos valores verdadeiros dos parâmetros e observa-se que apenas os intervalos de confiança para as variâncias dos efeitos aleatórios Z_1 e Z_2 não contêm o verdadeiro valor.

Na Figura 12 tem-se os mapas referentes à configuração 3 podendo-se verificar uma grande mudança das predições de areia e silte em relação às outras configurações. Neste caso, a pior concordância entre os valores observados e os preditos ocorreu para o segundo componente (Figura 13).

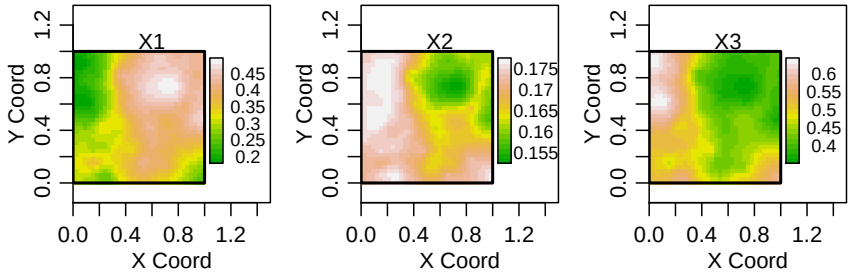

Figura - 12: Mapas das porcentagens de $X1$, $X2$ e $X3$ para dados da configuração 3.

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; DIGGLE, 2001), *compositions* (BOOGAART; TOLOSANA; BREN, 2008), *statmod* (SMYTH; HU; DUNN, 2009) e rotinas desenvolvidas especificamente para o desenvolvimento deste trabalho.

Discussão

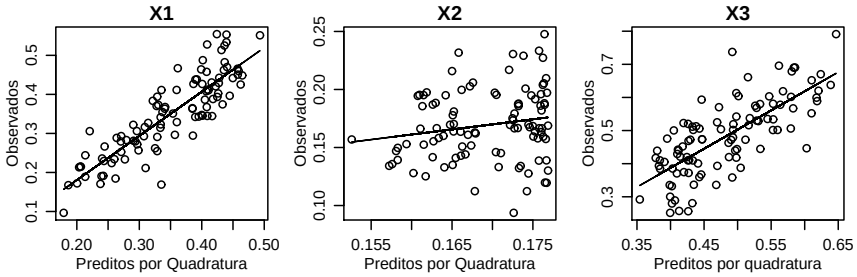

Figura - 13: Valores observados versus preditos de areia, silte e argila para a configuração 3.

Este trabalho possibilitou o estudo do comportamento do modelo proposto em três configurações distintas de dados composicionais. De forma geral, o componente X_1 apresentou-se o mais similar dentre as configurações com percentuais de médios para baixos. Com relação a X_2, a configuração 2 evidencia a maior variação nas porcentagens e as maiores porcentagens para X_3 apareceram na configuração 1.

Pode-se observar pelos diagramas de dispersão que a configuração 1 que mais se aproximou de uma forma linear foi a que apresentou as piores estimativas, enquanto que a configuração 3 que esperaria-se pior não resultou a pior em termos de qualidade de ajuste. As melhores estimativas foram obtidas com os dados da configuração 2. Os intervalos construídos via aproximação quadrática pelo método delta não foram bons principalmente para os parâmetros de variância e métodos alternativos devem ser avalidados como intervalos baseados em verossimilhança perfilhada e intervalos de credibilidade sob o enfoque da inferência bayesiana.

Verificou-se em estudos e análises não reportados aqui que os resultados obtidos considerando-se a transformação de volta do vetor de médias e da matriz de covariância para o simplex $\mathbb{S}^3$ por quadratura de Gauss-Hermite de ordem igual a 7 não diferiram dos obtidos com ordem 20. Uma outra maneira de se fazer esta transformação seria através de simulação. Neste caso, em primeiro lugar, em cada localização gerariam-se dados bivariados de uma distribuição Gaussiana padrão obtendo-se uma distribuição de percentuais para cada componente; em seguida faria-se uma decomposição da matriz de covariância obtida por cokrigagem por algum método numérico, por exemplo, fatoração Cholesky e aplicaria-se uma transformação linear, produto dos dados bivariados pela matriz Cholesky para a obtenção do vetor bivariado resposta. Aplicando-se transformação agl neste vetor resgata-se a composição. Este procedimento seria repetido para os 1156 pontos de predição. É possível que se verifique uma melhora nos resultados à medida que o número de simulações aumente. Tal método é mais geral pois permite obter a distribuição dos valores simulados em cada localização, possibilitando o cálculo de outros funcionais de interesse além da média e variância obtidas pelo método de quadratura.

Em relação aos mapas de valores preditos construídos para as três

configurações, ressalta-se que estes revelam os resultados já observados no diagrama ternário mas agora sob outra forma. De fato, as duas primeiras se mostram mais semelhantes por componente dado as nuvens de pontos localizadas na parte central do diagrama ternário. As mudanças observadas nos mapas correspondentes à configuração 3 se justificam pelo deslocamento da nuvem de pontos para o lado esquerdo implicando em baixos valores de X_2. Por último, as maiores concordâncias entre valores observados e preditos resultaram da configuração 1.

Conclusões

Os procedimentos adotados permitiram a construção de mapas de composições com três componentes por uma metodologia que implicitamente garante a restrição de que as frações somem 1, não só nos pontos simulados como nos pontos preditos.

O modelo proposto captura variações espaciais, induzidas pelas composições e não estruturadas.

A declaração explícita do modelo permite que sejam feitas inferências sobre parâmetros de forma usual. Abre-se ainda a possibilidade de tratamento Bayesiano 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 componentes 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.

Agradecimentos

Agradecimento à CAPES pelo apoio financeiro. Esse trabalho foi parcialmente financiado pela FINEP projeto CT-INFRA/UFPR.

MARTINS, A. B. T.; RIBEIRO JR, P. J.; BONAT, W. H. A bivariate geostatistical model for compositional data. *Biometrical Brazilian Journal*, São Paulo, v.xx, n.x, p.xx-xx, 2009. *Rev. Mat. Estat.* (São Paulo), v. 20, n.1, p. 1-10, 2000.

▪ *ABSTRACT: This work is motivated by the interest in modeling spatial patterns of compositional data. Target problems includes soil fractions or rock chemical composition, or, more generally, data structures with observations being parts of a whole and recorded spatial locations. The main interest is joint prediction of the components within the study area accounting for the adding to one restriction. Methods for compositional data analysis, initially developed for independent observations, are combined with a multivariate geostatistical model. An explicit parametric model is assumed for the variables. In particular, a bivariate model is presented with associated likelihood based methods of inference and results for spatial prediction as well as a computational implementation. Three simulated data with different characteristics are used to illustrate the methods of analysis and results are presented by prediction maps.*

▪ *KEYWORDS: multivariate geostatistics; compositional data; likelihood.*

Referências

1 ABRAMOWITZ, M.; STEGUN, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Washington: Milton Abramowitz and Irene A. Stegun, 1972. 1046p.

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

3 AZZALINI, A. *Statistical inference based in the likelihood*. London: Chapman and Hall, 1996. 341p.

4 BANERJEE, S.; CARLIN, B. P.; GELFAND, G. E. *Hierarchical modelling and analysis for spatial data*. Boca Raton: Chapman and Hall, 2004. 452p.

5 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.

6 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.

7 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.

8 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.

9 COX, D. R.;HINKLEY, D. V. *Theoretical Statistics*. London: Chapman and Hall, 1974. 511p.

10 DEGROOT, M. H.; SCHERVISH, M. J. *Probability and Statistics*. 3.ed. USA: Addison-Wesley, 2002. 816p.

11 DIGGLE, P. J.; RIBEIRO JR, P. J. *Model-based geostatistics*. USA: Springer Series in Statistics, 2007. 228p.

12 FLETCHER, R.; REEVES, C. M. *Function minimization by conjugate gradients*. Computer Journal, v.7, p.148−154, 1964.

13 GAMMERMAN, D. *Markov chain monte carlo*. London: Chapman and Hall, 1997. 245p.

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 JAMES, B. R. *Probabilidade: um curso em nível intermediário*. 3.ed. Rio de Janeiro: IMPA, 2004. 304p.

16 NELDER, J. A., MEAD, R. *A simplex algorithm for function minimization*. Computer Journal, v.7, p.308−313, 1965.

17 PAWITAN, Y. *In all likelihood: statistical modelling and inference using likelihood*. New York: Oxford University Press, 2001. 528p.

18 PAWLOWSKY-GLAHN, V.; OLEA, R. A. *Geostatistical analysis of compositional data*. New York: Oxford University Press, Inc., 2004. 181p.

19 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.

20 RIBEIRO JR, P. J.; DIGGLE, P. J. *geoR: a package for geostatistical analysis*. R-NEWS, v.1, n.2, p.15−18, 2001. Disponível em:<http://CRAN.R-project.org/doc/Rnews/>. Acesso em 13 junho 2009.

21 SCHMIDT, A. M.; GELFAND, A. E. *A bayesian coregionalization approach for multivariate pollutant data*. Journal of Geophysical Research, v.108, p.10−1−18−8, 2003.

22 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, 2006, Caxambu. *Minicurso*, Caxambu: Associação Brasileira de Estatística, 2006.

23 SMYTH, G.; HU, Y.; DUNN, P. *statmod: statistical modeling*. Disponível em:<http://www.statsci.org/r>. Acesso em 25 maio 2009.

Recebido em 01.01.2005.

Aprovado após revisão em 01.01.2005.