Modelo geoestatístico composicional

Ana Beatriz Tozzo Martins[a], Paulo Justiniano Ribeiro Jr[b], Wagner Hugo Bonat[b]

[a] *UEM/DEST - Universidade Estadual de Maringá, Departamento de Estatística.*
[b] *UFPR/LEG - Universidade Federal do Paraná, Laboratório de Estatística e Geoinformação.*

Abstract

Este artigo apresenta um modelo geoestatístico para dados composicionais. A inferência para os parâmetros desconhecidos do modelo é feita baseada na função de verossimilhança. Um algoritmo misto de resultados analíticos e numéricos é utilizado. Para a construção de intervalos de confiança considerou-se duas abordagens: Wald e perfil de verossimilhança. Em modelos com efeito espacial um interesse comum é a predição do processo em localizações não amostradas (krigagem), para isto, são apresentados dois algoritmos, um baseado em quadratura de Gauss-Hermite e um segundo baseado em simulação. Uma aplicação da metodologia é feita em um conjunto de dados reais referente a frações granulométricas de solo, onde o objetivo é obter um mapa da distribuição espacial das frações. No exemplo considerado, a inferência apresentou várias dificuldades, principalmente com relação a avaliação da incerteza associada as estimativas. Intervalos de Wald produziram estimativas intervalares irreais, o que foi resolvido obtendo-se intervalos via perfil de verossimilhança. A predição espacial das frações granulométricas do solo foi realizada pelos dois algoritmos, que mostraram resultados compatíveis e satisfatórios. Um estudo de simulação para avaliar o viés dos estimadores, bem como o nível de cobertura dos intervalos de Wald foi conduzido. Os resultados mostram uma tendência de subestimação para os parâmetros de variância e correlação, porém com fraca intensidade. Os intervalos de Wald apresentaram nível de cobertura pouco abaixo do nominal, por volta de 85% a 90%. De forma geral o modelo, o processo de estimação e predição espacial propostos mostraram resultados satisfatórios para a análise de dados composicionais espaciais.

Keywords: Verossimilhança, geoestatística, dados composicionais.

1. Introdução

Dados composicionais são definidos como vetores de elementos positivos e soma constante, geralmente 1 ou 100%. Esta restrição define o simplex unitário como o espaço amostral, induz correlação intrínsica entre as variáveis e impõe limitações à aplicação de técnicas estatísticas usuais para a análise e modelagem de dados.

Além da correlação natural existente entre os elementos de um vetor, denominado composição, pode-se levar em consideração a dependência que existe em decorrência dos locais onde as composições são amostradas. Descrever a distribuição espacial de composições consiste em uma informação valiosa para a descrição completa das suas características relevantes, quando se busca a otimização do processo em estudo, seja esse, de natureza qualquer.

A Estatística Espacial tem se apresentado como uma área de grande importância nas mais diversas aplicações. Trabalhos de grande relevância têm sido desenvolvidos nesta área, como os de [1], [2], [3], [4], [5], entre outros. Na área de geoestatística citam-se os trabalhos de [6], [7], [8], [9], [10], que diferem da linha tradicional no sentido, de que a análise é baseada em modelos que induzem uma estrutura de covariância, ou seja, os modelos uni e multivariados são explícitos através da especificação de uma função de correlação, em que a covariância é função da distância entre pares de localizações. Nestes modelos, é possível aplicar métodos clássicos de inferência baseados em verossimilhança, que produzem estimativas mais eficientes dos parâmetros e permitem avaliar a incerteza associada.

Trabalhos realizados por [11, 12, 13] em análise de dados composicionais, apresentam uma metodologia adequada para analisar dados, caracterizados por se apresentarem em forma de proporções complementares. O autor propôs, por exemplo, a transformação razão log-aditiva (ALR) que generaliza a transformação logística para um vetor composicional de duas partes:

$$\begin{aligned} \mathrm{alr}\ :\ \mathbb{S}^B &\longrightarrow \mathbb{R}^{B-1} \\ \underset{\sim}{X} &\longrightarrow \mathrm{alr}\big(\underset{\sim}{X}\big) = (\ln(X_1/X_B), \ldots, \ln(X_{B-1}/X_B))^\top, \end{aligned}$$

possibilitando a análise dos dados no espaço amostral dos números reais e a transformação inversa, denominada transformação logística generalizada aditiva (AGL):

$$\begin{aligned} \text{agl}: \ \mathbb{R}^{B-1} &\longrightarrow \mathbb{S}^B \\ \underset{\sim}{\mathrm{Y}} &\longrightarrow \underset{\sim}{\mathrm{X}} = \mathcal{C}\left((\exp\{\ln(X_1/X_B)\}, \ldots, \exp\{0\})^\top\right), \end{aligned}$$

devolvendo os dados à escala original, e onde $\mathcal{C}$ é o operador fechamento que garante a soma 1 dos componentes da composição.

A partir dos anos 2000, estes tipos de dados são analisados considerando as localizações amostrais, mas ainda sob a abordagem geoestatística tradicional. Por exemplo, [14] mostram que, sem considerar que a matriz de covariância do modelo composicional seja definida positiva e que os valores interpolados satisfaçam a restrição "soma um", a interpolação espacial de dados de frações de partículas de solo produzem valores interpolados irreais. [15] e [16] seguindo a teoria clássica da geoestatística, satisfazem essas exigências mas não adotam declaração explícita de modelo, no sentido de não adotarem um modelo paramétrico para os dados. [17] fazem um estudo sobre cokrigagem de frações de partículas do solo, concluindo que a predição de dados composicionais pode ser feita através da cokrigagem ALR. Segundo esses autores, a cokrigagem ALR, que considera o logaritmo das razões dos componentes, apresenta vantagens em relação à cokrigagem sem transformação, que considera apenas as razões dos componentes. Concluem ainda que existem vantagens se a transformação de volta das predições para a escala original, das composições, são calculadas por quadratura de Gauss-Hermite, para aproximar a esperança condicional.

Sob o enfoque bayesiano, [18] modelam dados composicionais espaciais sem adotar forma explícita para a função de covariância e sem fazer predição espacial. Esses autores adotam uma função de correlação exponencial generalizada e uma distribuição *a priori* do tipo Wishart para a estimação dos parâmetros de variância.

Apreciando as referências citadas, avalia-se que ainda existe espaço para novas propostas metodológicas no que diz respeito a construção de modelos, para a análise e predição espacial de dados composicionais. Considera-se que a declaração explícita de um modelo paramétrico para dados geoestatístico composicionais, com inferência formal baseada em verossimilhança, é uma importante contribuição para a análise deste tipo de dados. As interpretações provenientes das estimativas dos parâmetros do modelo podem trazer um maior entendimento dos fenômenos em estudo nas mais diversas áreas de aplicação.

A inferência formal baseada em verossimilhança traz uma abordagem integrada para a construção de intervalos de confiança, testes de hipóteses e medidas de criticidade/comparação, o que não é imediato por abordagens alternativas, como modelagem via variograma/semi-variograma. Além disso, possibilita a predição espacial e avaliação da incerteza associada.

O objetivo deste artigo é propor e implementar um modelo geoestatístico para dados composicionais utilizando estruturas multivariadas, com componentes do modelo especificados por função de correlação espacial e "composicional". A inferência para os parâmetros do modelo é feita baseada na função de verossimilhança. Para a predição espacial de dados composicionais são apresentados dois algoritmos, o primeiro fazendo uso de integração numérica pelo método de Gauss-Hermite, o segundo faz uso de técnicas de simulação. Como exemplo, aplicou-se a metodologia proposta em um conjunto de dados de solo, elaborando mapas temáticos baseados no modelo geoestatístico composicional. Além disso, é apresentado um estudo de simulação para verificar o comportamento assintótico dos estimadores de máxima verossimilhança.

O artigo está dividido em seis seções. Seção 2 apresenta o modelo geoestatístico bivariado composicional. Seção 3 contempla a estimação dos parâmetros baseada na função de verossimilhança. Seção 4, apresenta dois algoritmos para a predição espacial de dados composicionais. Seção 5 apresenta os resultados da aplicação da metodologia proposta em um conjunto de dados reais referente a frações granulométricas do solo. Nesta aplicação é apresentado todo o processo de estimação dos parâmetros, a construção de intervalos de Wald e baseados em perfil de verossimilhança. A predição espacial é feita pelos dois algoritmos e os resultados são comparados. Na sequência é apresentado um estudo de simulação, onde foi verificado o viés dos estimadores de máxima verossimilhança, bem como a cobertura dos intervalos de confiança de Wald. Seção 6 apresenta uma discussão dos métodos apresentados, relata as principais dificuldades computacionais e são indicados alguns pontos para pesquisas futuras. No Apêndice são apresentadas as principais funções desenvolvidas em [19] para as análises apresentadas no artigo.

2. Modelo geoestatístico composicional

Para $\underset{\sim}{X} = (X_1, ..., X_B)^\top$ sendo uma composição com B componentes e $\underset{\sim}{Y} = (\ln(X_1/X_B), \ldots, \ln(X_{B-1}/X_B))^\top$ um vetor com $B-1$ elementos, o modelo geoestatístico com componente comum pode ser obtido seguindo a formulação dada em [9]. Neste trabalho, considerou-se composições de três componentes, (X_1, X_2, X_3).

O modelo geoestatístico bivariado composicional [20] é definido como:

$$\begin{cases} Y_1(\underset{\sim}{x}_i) & = & \mu_1(\underset{\sim}{x}_i) + S_1(\underset{\sim}{x}_i) + Z_1(\underset{\sim}{x}_i), \\ Y_2(\underset{\sim}{x}_{i'}) & = & \mu_2(\underset{\sim}{x}_{i'}) + S_2(\underset{\sim}{x}_{i'}) + Z_2(\underset{\sim}{x}_{i'}), \end{cases}$$

em que $\underset{\sim}{x}_i, \underset{\sim}{x}_{i'} \in \mathbb{R}^2$, 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 $\underset{\sim}{Y}_{n\times 1} = \big(Y_1(\underset{\sim}{x}_1), \ldots, Y_1(\underset{\sim}{x}_{n_1}), Y_2(\underset{\sim}{x}_1), \ldots, Y_1(\underset{\sim}{x}_{n_2})\big)$, ou seja, as observações são "empilhadas" por variável; $S_j(\underset{\sim}{x}) \sim N(0; \sigma_j^2)$ e $Z_j(\underset{\sim}{x}) \sim N(0; \tau_j^2)$, $j = 1, 2$.

No modelo geoestatístico composicional os efeitos aleatórios com estrutura espacial S_1 e S_2 são substituídos por um efeito aleatório padronizado U. Supondo que este efeito tem distribuição gaussiana multivariada com vetor de médias iguais a zero e matriz de covariâncias, com variâncias unitárias e covariâncias dadas pela função de correlação exponencial, ρ_U. Esta função 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_1 e Z_2 capturam a variabilidade não espacial incluindo a correlação, ρ, induzida pela estrutura composicional. O modelo pode então ser reescrito como:

$$\begin{cases} Y_1(\underset{\sim}{x}_i) & = & \mu_1(\underset{\sim}{x}_i) + \sigma_1 U(\underset{\sim}{x}_i; \phi) + Z_1(\underset{\sim}{x}_i), \\ Y_2(\underset{\sim}{x}_{i'}) & = & \mu_2(\underset{\sim}{x}_{i'}) + \sigma_2 U(\underset{\sim}{x}_{i'}; \phi) + Z_2(\underset{\sim}{x}_{i'}). \end{cases} \tag{1}$$

Sendo assim, $\underset{\sim}{Y} \sim N(\underset{\sim}{\mu}; \Sigma)$, com $\underset{\sim}{\mu} = (\mu_1, \mu_2)^\top$ e a matriz de covariâncias Σ composta pelos elementos

$$Cov(Y_j(\underset{\sim}{x}_i); Y_j(\underset{\sim}{x}_i)) \;=\; \sigma_j^2 + \tau_j^2, \qquad Cov(Y_j(\underset{\sim}{x}_i); Y_j(\underset{\sim}{x}_{i'})) \;=\; \sigma_j^2 \rho_U(\underset{\sim}{x}_i; \underset{\sim}{x}_{i'}),$$

e

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

com as funções indicadoras I_2 e I_3 definidas como:

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

3. Inferência no modelo geoestatístico composicional

O vetor de parâmetros do modelo é $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)^\top$ e sua função de verossimilhança é obtida a partir da função de densidade da distribuição normal multivariada:

$$L(\underline{\theta}, \underline{Y}) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp^{\left(\underline{Y} - \underline{\mu}_Y\right)^\top \Sigma^{-1} \left(\underline{Y} - \underline{\mu}_Y\right)}$$

em que o termo $\underline{\mu}_Y$ é função dos parâmetros μ_1 e μ_2 e os demais parâmetros $(\sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)$ definem os termos em Σ. Considerando $\underline{\mu}_Y = D\underline{\mu}$, em que D é a matriz de delineamento de ordem $n \times 2$, tem-se que a função de log-verossimilhança é dada por:

$$l(\underline{\theta}; \underline{Y}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (\underline{Y} - D\underline{\mu})^\top \Sigma^{-1} (\underline{Y} - D\underline{\mu}) \qquad (2)$$

Os estimadores de máxima verossimilhança para $\underline{\mu}$ podem ser obtidos diferenciando 2 em relação aos respectivos parâmetros e são dados por

$$\underline{\hat{\mu}} = (D^\top \Sigma^{-1} D)^{-1} (D^\top \Sigma^{-1} \underline{Y}) \qquad (3)$$

Substituindo $\underline{\hat{\mu}}$ em 2 obtêm-se uma log-verossimilhança concentrada em $\underline{\theta}^* = (\sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)$. Claramente não é possível obter estimadores para $\underline{\theta}^*$ de forma fechada. Sendo assim, será utilizado o algoritmo "L-BFGS-B" [21], que permite informar os limites inferior e superior de busca no espaço paramétrico. Este algoritmo está implementado na função *optim*() do software [19].

A função de log-verossimilhança concentrada tem como argumento $\underline{\theta}^*$ que são os parâmetros que indexam a matriz de variância/covariância do modelo. Este passo de otimização é computacionalmente caro, uma vez que envolve a inversão de uma matriz densa (Σ). Para tornar mais rápido e estável o processo de otimização através do algoritmo "L-BFGS-B", foi obtido o gradiente analítico da função de log-verossimilhança concentrada, que é dada por:

$$l^*(\underline{\theta}^*; \underline{Y}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \underline{\hat{e}}^\top \Sigma^{-1} \underline{\hat{e}}$$

onde o termo $\underline{\hat{e}} = (\underline{Y} - D\underline{\hat{\mu}})$.

As funções escores são dadas por

$$\frac{\partial l^*(\underline{\theta}^*; \underline{Y})}{\partial \theta_i^*} = -\frac{1}{2} Tr \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i^*} \right] - \frac{1}{2} \underline{\hat{e}}^\top \left[-\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i^*} \Sigma^{-1} \right] \underline{\hat{e}}, \quad i = 1, \dots, 6.$$

onde as matrizes $\dfrac{\partial \Sigma}{\partial \theta_i^*}$ representa a matriz obtida por derivar cada elemento da matriz Σ em relação ao respectivo parâmetro.

O uso do gradiente analítico tornou o passo de otimização mais rápido (aproximadamente metade do tempo computacional do algoritmo completamente numérico) e estável computacionalmente. Para obter $\underline{\hat{\mu}}$ basta substituir as estimativas pontuais de $\underline{\theta}^*$ na equação 3.

Para a construção de intervalos de confiança assintóticos (Wald) para $\underline{\theta}^*$ é utilizada a matriz Hessiana obtida numéricamente no passo de otimização. A obtenção das variâncias para $\underline{\hat{\mu}}$ é obtida fazendo a segunda derivada da log-verossimilhança em relação aos parâmetros $\underline{\mu}$ e é dada por:

$$I_F(\underline{\mu}) = D^\top \Sigma^{-1} D$$

de onde vem,

$$V[\underline{\hat{\mu}}] = I_F(\underline{\mu})^{-1} = (D^\top \Sigma^{-1} D)^{-1}.$$

A construção de intervalos de confiança baseados em resultados assintóticos para estimativas de variância/correlação precisa ser feita com cuidado. É conhecido que tais estimadores tendem a apresentar um comportamento bastante assimétrico, principalmente perto da borda do espaço paramétrico, o que pode ocasionar até mesmo uma estimativa intervalar fora do espaço paramétrico, tornando o intervalo irreal. Além disso, o nível nominal de cobertura também pode ser substancialmente afetado, pela assimetria típica deste tipo de parâmetros. Isto será mais discutido no estudo de simulação e no exemplo de aplicação.

Uma forma alternativa é a construção de intervalos de confiança baseados em perfil de verossimilhança, esta abordagem será utilizada no exemplo com dados reais. A construção de intervalos por esta metodologia é computacionalmente cara e exige cuidados em sua implementação, mais detalhes podem ser encontrados em [22].

Um algoritmo alternativo fazendo uso de uma reparametrização deste modelo é apresentado no Apêndice A. Com a reparametrização proposta a otimização numérica é feita em cinco dimensões e não em seis como apresentada nesta seção. A princípio seria um algoritmo melhor, porém estudos de simulação não reportados aqui mostraram uma alta instabilidade numérica, por isso escolhemos o algoritmo acima que apesar de mais caro computacionalmente, mostrou-se muito estável no estudo de simulação.

4. Predição espacial no modelo geoestatístico composicional

A predição espacial no contexto multivariado, cokrigagem, é uma extensão da teoria de krigagem para o caso univariado. Considera-se a predição espacial de $\underline{\mathrm{Y}}_0(\underline{\mathrm{x}})$ em localizações não amostradas $\underline{\mathrm{x}}_0 = (\underline{\mathrm{x}}_{10}, \underline{\mathrm{x}}_{20}, ..., \underline{\mathrm{x}}_{n_2 0})$. O vetor de médias correspondentes às variáveis Y_1 e Y_2 para todas as localizações de predição, e a matriz de covariância são baseadas nos resultados da distribuição gaussiana multivariada, conforme [9]:

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

Sendo desconhecidos os valores de $\underline{\mu}_{\underline{\mathrm{Y}}_0}$, estes são substituídos pelo vetor de estimativas de verossimilhança. A matriz Σ, de onde se extrai as matrizes $\Sigma_{\underline{\mathrm{Y}}_0\underline{\mathrm{Y}}_0}$, $\Sigma_{\underline{\mathrm{Y}}_0\underline{\mathrm{Y}}}$, $\Sigma_{\underline{\mathrm{Y}}\underline{\mathrm{Y}}_0}$ e $\Sigma_{\underline{\mathrm{Y}}\underline{\mathrm{Y}}}$, é calculada substituindo-se os valores estimados, resultantes do processo de otimização.

Como o que se pretende é calcular para cada localização uma estimativa de $\underline{\mu}_{\underline{X}}$ e $\Sigma_{\underline{X}}$, integrais definidas no espaço amostral simplex com $\underline{X} = \mathrm{agl}(\underline{Y})$, utiliza-se o método do Jacobiano [23] de modo que

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

[12] e [15] propõem uma transformação de variável de modo a expressar as integrais no espaço real e seus valores são aproximados através da integração de Gauss-Hermite multivariada de ordem k

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

em que 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 conhecidos e seus valores podem ser encontrados, por exemplo, em [24]. Segundo [25] ordens de quadratura de 6 a 8 são suficientes para aproximar a integral. Maiores detalhes encontram-se em [20].

Uma outra forma de transformar os valores preditos no espaço real para o espaço simplex é por simulação, onde para cada localização geram-se dados de uma distribuição normal gaussiana multivariada com vetor de médias e matriz de covariância obtidos por cokrigagem, aplica-se a transformação AGL nesses dados, ou seja, ao vetor de valores esperados (das variáveis resposta do modelo) e calculam-se a média ou qualquer outra estatística de interesse para cada componente.

5. Resultados

Nesta seção serão apresentados a análise completa de um conjunto de dados reais, referente a frações granulométricas do solo. Na sequência será apresentado um estudo de simulação, para avaliar o viés dos estimadores e o nível de cobertura dos intervalos de Wald.

5.1. Aplication - Predição espacial de frações granulométricas de solo

Como exemplo de aplicação será analisado um conjunto de dados obtido do trabalho de [26]. O experimento foi conduzido em uma área irrigada por sistema pivô-central na Fazenda Areão, pertencente ao campus da Escola Superior de Agricultura - Luiz de Queiroz (ESALQ-USP). Nesta área foi demarcado um quadrante na porção mais elevada (topo da encosta), no qual foram obtidas 82 amostras de solo na profundidade entre 0 e 0.20 metros, 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.

AQUI TEM QUE COLOCAR ALGUMA COISA SOBRE A IMPORTANCIA DA APLICACAO.

Dentro da área a coordenada mínima foi igual a $(0, 0)$ e máxima igual a $(180, 180)$ metros. A Figura 1 apresenta uma análise exploratória para as frações granulométricas de solo.

Pode-se observar pela Figura 1 que argila apresentou maior variabilidade e também os maiores valores, por outro lado, o silte apresenta os menores valores e a menor variabilidade. Como argila determina a retenção de água no solo, pela sua alta superfície de penetração, ela foi escolhida como denominador das log razões.

Fazendo a transformação ALR nos dados originais obteve-se as variáveis $Y_1 = \ln(\text{Areia}/\text{Argila})$ e $Y_2 = \ln(\text{Silte}/\text{Argila})$, variáveis resposta do modelo. Uma vez definida as variáveis resposta, a estrutura espacial do modelo é

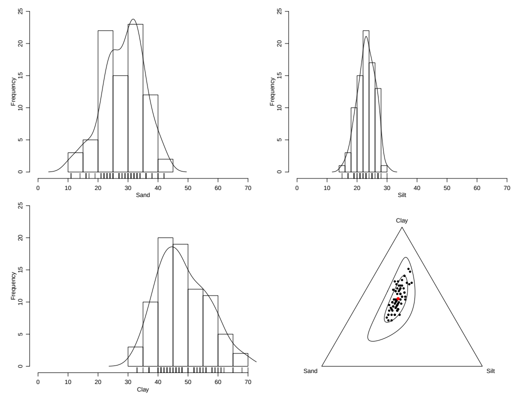

Figura 1: Distribuição dos percentuais de areia, silte e argila e diagrama ternário das composições.

determinada pelas coordenadas X e Y, através da distância euclidiana entre as observações.

Para a estimação dos parâmetros envolvidos no modelo geoestatístico composicional, utilizou-se o procedimento descrito na Seção 3, todas as funções para a estimação foram escritas em R, foi utilizado o pacote [22] para facilitar a construção de intervalos de confiança Wald e perfilhado. Este pacote também facilita a construção de testes de razão de verossimilhança e a obtenção de medidas de ajuste como AIC e BIC.

Como a verossimilhança concentrada será otimizada numéricamente é necessário um guia inicial para o algoritmo de maximização (L-BFGS-B). Diversas tentativas foram avaliadas, recomenda-se usar para σ_1 e τ_1 a metade da variância amostral de $\underline{Y}_1$. Para σ_2 e τ_2 a metade da variância amostral de $\underline{Y}_2$, para o parâmetro ϕ recomenda-se usar $min_d + 0.2 * (max_d - min_d)$, onde min_d e max_d representam a menor e maior distância respectivamente entre dois pontos amostrais. Por fim, para ρ foi utilizado o coeficiente de correlação amostral de Pearson entre as variáveis $\underline{Y}_1$ e $\underline{Y}_2$. A Tabela 1 apresenta o resultado do processo de estimação.

Pelos resultados apresentados na Tabela 1, verifica-se que a log-verossimilhança teve um grande aumento como esperado a partir do guia inicial, passou de -2746 para 5.01, além disso, o critério de parada do algoritmo numérico foi aceito, concluindo assim que o processo de maximização teve sucesso.

Analisando apenas as estimativas pontuais verifica-se que a variável $\underline{Y}_1$ apresenta uma variabilidade maior que a variável $\underline{Y}_2$, principalmente em σ_1. Com relação ao parâmetro ϕ sua estimativa pontual indica uma forte

	Inicial	Pontual	Std. Error	2.5%	97.5%.
μ_1		-0.7864	0.2561	-1.2883	-0.2845
μ_2		-0.7943	0.0694	-0.9304	-0.6583
σ_1	0.0958	0.4705	0.1827	0.1125	0.8285
σ_2	0.0381	0.1168	0.0690	-0.0185	0.2520
τ_1	0.0958	0.2838	0.0491	0.1875	0.3800
τ_2	0.0381	0.2619	0.0220	0.2187	0.3050
ϕ	66.9117	81.4365	80.4219	-76.1875	239.0606
ρ	0.8231	0.9589	0.0559	0.8492	1.0685
ll	-2746.2903	5.0194			

Tabela 1: Guia inicial, estimativas pontuais, erros padrões e intervalos de confiança de Wald.

dependência espacial. A estimativa do parâmetro ρ mostra que as variáveis transformadas são altamente correlacionadas, como era esperado devido a estrutura das composições.

Olhando para os erros padrões estimados, novamente verificamos que todas as estimativas relacionadas a variável $\underline{Y}_1$ apresentam maior variabilidade. Destaca-se a magnitude do erro padrão da estimativa de ϕ mostrando uma incerteza muito grande na estimativa deste parâmetro.

Apesar do procedimento de estimação ter tido sucesso em encontrar o máximo da função, quando constrõem-se intervalos de confiança baseados na aproximação quadrática da verossimilhança (Wald), obtem-se intervalos irreais. Nesta aplicação dos seis parâmetros estimados envolvidos na matriz de variância/covariância, três apresentam intervalos irreais, tomando valores fora do espaço paramétrico. No caso do parâmrtro ρ este fato pode ser parcialmente explicado por se tratar de um parâmetro limitado, cuja estimativa está muito próxima da borda do espaço paramétrico, onde os resultados assintóticos o EMV precisam de uma tamanho de amostra muito grande para serem adequados. O mesmo argumento pode ser usado para a estimativa de σ_2 que ficou próxima do zero, levando a um intervalo de confiança com limite inferior negativo, o que é claramente inadequado.

Esse mau desempenho dos intervalos de confiança de Wald, são esperados para parâmetros de variância e correlação, uma vez que a superfície de verossimilhança é bastante assimétrica na direção destes parâmetros. Nesta aplicação o agravante do tamanho da amostra reduzido amplifica este mau desempenho.

Apesar do intervalo de confiança para a estimativa de ρ levar a um resultado inconsistente ele não muda de forma drástica a interpretação do modelo, ou seja, a correlação entre as variáveis é um resultado bastante aparente. O mesmo já não se aplica para o parâmetro σ_2, uma vez que este representa a variabilidade devida ao efeito espacial para a variável $\underline{Y}_2$, um valor 0 para este parâmetro indicaria ausência de efeito espacial, sendo a variabilidade devida exclusivamente ao ruído τ_2 e a correlação espúria ρ entre as composições. Tem-se neste caso uma situação inconclusiva.

Além dos problemas relacionados acima, a situação mais grave na construção dos intervalos de confiança é no caso da estimativa do parâmetro ϕ, que tem uma interpretação chave para o modelo espacial, uma vez que é ele que mede o tamanho desta dependência. Apesar da sua estimativa pontual ser alta, o seu IC cobre o valor zero, o seu erro padrão é muito grande, o coeficiente de variação é de 98.75%. Da mesma forma que para σ_2 a análise se torna inconclusiva. Pelos resultados apresentados na Tabela 1 não tem-se evidências significativas da existência de dependência espacial. Porém, este resultado está condicionado a este tipo de construção de intervalo de confiança, e como estamos tratando de parâmetros de variância e correlação, devemos desconfiar destes resultados e buscar formas mais eficientes para a construção dos intervalos de confiança.

O problema com os intervalos de Wald em geral é a suposição de simetria, o que na maioria das situações é inadequada para parâmetros de variância e correlação. Uma forma de relaxar esta suposição é a construção de intervalos baseados em perfil de verossimilhança [22]. Esta abordagem é cara computacionalmente, uma vez que envolve muitos passos de otimização. Apesar disso, exploramos a construção deste tipo de intervalo nesta aplicação, já que, os intervalos assintóticos não foram satisfatórios. A Figura 2, apresenta o perfil de verossimilhança, para cada um dos seis parâmetros de indexam a matriz Σ do modelo geoestatístico composicional.

Pelos gráficos apresentados na Figura 2 é clara a forte assimetria a direita do perfil de verossimilhança dos parâmetros σ_1, σ_2 e ϕ. Para os parâmetros τ_1 e τ_2 o perfil mostra um comportamento muito próximo de uma forma quadrática, o que indica que para estes parâmetros a aproximação quadrática é uma boa alternativa. Para o parâmetro ρ, o intervalo toca a borda do espaço paramétrico no lado direito, mas não causa nenhum grave problema na estimação, só é necessário ter cuidado com o nível de confiança que deve ser menor que o nominal, pois o intervalo foi truncado. Estes resultados corrobo-

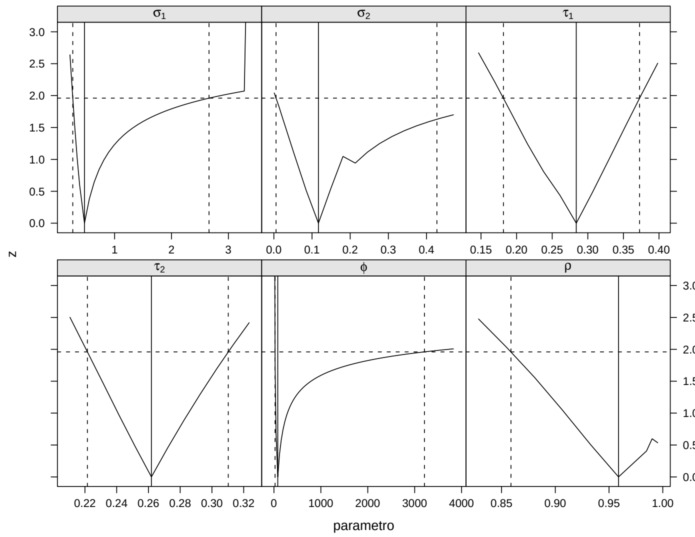

Figura 2: Perfil de verossimilhança para os parâmetros da matriz de variância/covariância do modelo geoestatístico composicional.

ram o que foi encontrado na construção dos intervalos de confiança de Wald, que principalmente para o bloco de parâmetros que indexa o efeito espacial a aproximação quadrática é ruim. Além disso, os perfis de verossimilhança são uma forma mais elegante e coerente de representar a incerteza associada a estimação dos parâmetros envolvidos no modelo proposto.

Para finalizar o processo de estimação a Tabela 2 apresenta os intervalos de confiança obtidos via aproximação quadrática (Wald) e via perfil de verossimilhança, para os parâmetros que indexam a matriz Σ, o nível de confiança adotado é de 95%.

Os resultados apresentados na Tabela 2 quantifica os encontrados da Figura 2, o intervalo de perfil para σ_1, σ_2 e ϕ é extremamente assimétrico e mais amplo que os obtidos por Wald, principalmente à direita. Para τ_1 e τ_2 as duas abordagens geram resultados próximos. No caso do ρ as duas abordagens levam a borda direita do espaço paramétrico, para o limite inferior trazem resultados parecidos. Baseados neste intervalo pode-se concluir que existe uma forte dependência espacial e esta é significativamente diferente de zero, preenchendo o vazio deixado pela abordagem anterior, o mesmo se aplica ao parâmetro σ_2.

Terminado o procedimento de estimação dos parâmetros e este tendo

	2.5 %	97.5 %	2.5 %	97.5 %
σ_1	0.1125	0.8285	0.2647	2.6599
σ_2	-0.0185	0.2520	0.0057	0.4272
τ_1	0.1875	0.3800	0.1815	0.3727
τ_2	0.2187	0.3050	0.2215	0.3103
ϕ	-76.1875	239.0606	24.3202	3207.7026
ρ	0.8492	1.0685	0.8588	

Tabela 2: Comparação entre os intervalos de confiança obtidos via aproximação quadrática e perfil de verossimilhança.

êxito, o próximo passo é a predição espacial das frações granulométricas do solo. Para isto, seguimos os procedimentos da Seção 4. A Figura 3 apresenta os mapas de predição espacial das frações granulométricas do solo. Para a construção da superfície foi criado uma malha com 2601 pontos dentro da área em estudo, no algoritmo baseado em simulação foi utilizada 500 simulações para cada ponto da malha. Para a quadratura de Gauss-Hermite foram utilizados 7 pontos de integração. Os mapas são apresentados em termos de médias, porém cabe ressaltar que pelo procedimento baseado em simulação, qualquer outro funcional poderia ser obtido de forma imediata.

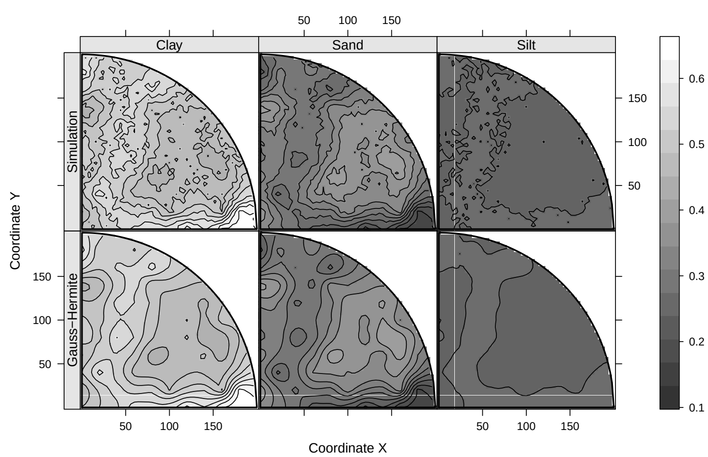

Figura 3: Mapas das predições das percentagens de areia, silte e argila obtidas por quadratura de Gauss-Hermite e por simulação.

Diante do resultados da Figura 3, pode-se observar que as duas abordagens trazem resultados muito similares, sendo que, o algoritmo baseado em simulação é uma versão mais ruidosa do procedimento via quadratura. O custo computacional da abordagem baseada em simulação é muito grande, já que, necessita simular amostras a cada passo de uma distribuição Normal multivariada em um gride fino. A abordagem via quadratura é mais rápida computacionalmente, porém não permite a avaliação de outros funcionais, como medianas, quantis ou qualquer outra função.

Os mapas mostram que os componentes areia e argila se complementam na área de estudo enquanto areia e silte são concorrentes. De acordo com os mapas apresentados e a concordância entre os dois métodos, é possível concluir que o procedimento de predição espacial de dados composicionais, teve sucesso. Foram gerados mapas coerentes com a realidade respeitando as restrições do tipo de dados analisados, trazendo resultados satisfatórios.

5.2. Estudo de simulação

Para avaliar o comportamento das estimativas de máxima verossimilhança com relação ao viés e nível de cobertura dos intervalos de Wald, foi realizado um estudo de simulação. Foram considerados dois tamanhos de amostra ($n = 100 \quad \text{e} \quad n = 225$), em um quadrado unitário com pontos regularmente espaçados. Três configurações de parâmetros foram consideradas:

- Configuração 1 - $\underline{\theta_1} = (-0.2, -0.5, 1, 1.5, 0.3, 0.3, 0.25, 0.9)$;
- Configuração 2 - $\underline{\theta_2} = (1, 1, 1.2, 1.5, 0.9, 1, 0.25, 0.5)$;
- Configuração 3 - $\underline{\theta_3} = (-0.5, -1, 0.45, 0.13, 0.3, 0.5, 0.1, 0)$.

Estas configurações foram selecionadas para serem reportadas no artigo pois geram comportamentos bastante distintos no diagrama ternário, permitindo a avaliação do procedimento de inferência em situações diferentes. A Figura 4 apresenta o diagrama ternário para uma realização do modelo geoestatístico composicional, de acordo com cada um dos conjuntos de parâmetros considerados no estudo.

Para cada configuração e tamanho de amostra foram gerados 1000 diferentes realizações do processo e conduzida a estimação pela maximização da função de log-verossimilhança. Os intervalos de confiança foram obtidos pela inversão do Hessiano numérico e são de 95% de confiança. Optou-se por apresentar os resultados apenas para os parâmetros de variância e correlação, uma vez que para os parâmetros de média os resultados (não reportados)

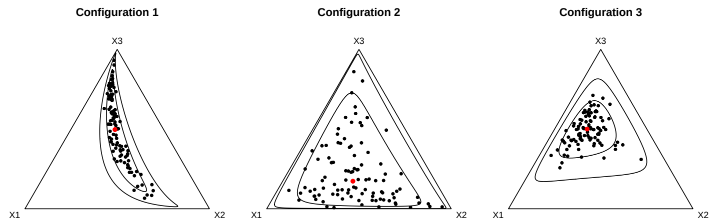

Figura 4: Diagrama ternário das composições de acordo com o conjunto de parâmetros geradores do processo.

	Configuration 1				Configuration 2				Configuration 3			
	n = 100		n = 225		n = 100		n = 225		n = 100		n = 225	
Par	BS	LC	BS	LC	BS	LC	BS	LC	BS	LC	BS	LC
σ_1	-.067	.909	-.050	.853	-.035	.966	-.059	0.912	-.048	.814	.009	.894
σ_2	-.101	.901	-.077	.834	-.046	.974	-.072	0.916	-.005	.962	.005	.957
τ_1	.004	.881	-.026	.878	-.120	.964	-.051	0.966	-.039	.788	-.062	.945
τ_2	-.004	.891	-.042	.897	-.179	.960	-.073	0.980	-.021	.956	-.007	.958
ϕ	-.001	.868	-.019	.776	-.029	.732	-.037	0.718	.039	.902	-.004	.807
ρ	-.021	.868	.004	.931	-.400	.926	-.108	0.978	-.109	.924	-.134	.973

Tabela 3: Viés (BS) e nível de cobertura (LC) de acordo com tamanho de amostra e combinações de parâmetros.

mostram um comportamento dentro do esperado. A Tabela 3 apresenta um resumo dos resultados obtidos pela simulação. São reportados o viés dos estimadores e o nível de cobertura dos intevalos de confiança, de acordo com o tamanho da amostra e configuração de parâmetros geradores do processo.

Conforme os resultados apresentados na Tabela 3 observa-se que independente da configuração de parâmetros e do tamanho da amostra os EMV tendem a subestimar os parâmetros. Com relação ao nível de cobertura dos intervalos de confiança de Wald, verifica-se que de forma geral estes tendem a apresentar nível de cobertura abaixo do nominal que é 95%.

Dado a estrutura do modelo pode-se dividir os parâmetros em dois blocos, o primeiro referente ao efeito espacial $(\sigma_1, \sigma_2, \phi)$ e o segundo referente ao erro de medida e correlação espúria induzida pelas composições (τ_1, τ_2, ρ).

Para o primeiro bloco verifica-se que não há uma diferença evidente em termos de viés com relação as configurações de parâmetros e tamanhos de amostra. As estimativas para ϕ são as que apresentam o maior viés, destaca-se o alto viés para o parâmetro ϕ na configuração 3 com $n = 100$, o viés relativo foi de 39.8%, quando o tamanho da amostra aumentou para $n = 225$ o viés relativo teve uma queda significativa ficando em 4.1%. Este resultado mostra que este parâmetro é de difícil estimação e consequentemente precisa de mais amostras para ser estimado eficientemente.

Com relação ao nível de cobertura dos intervalos de confiança, os resultados mostram que quando o $n = 225$ o nível de cobertura tende a diminuir, isto fica mais aparente para o parâmetro ϕ nas configurações 1 e 3, onde o nível de cobertura teve uma queda de aproximadamente 10% quando a amostra aumentou de $n = 100$ para $n = 225$, para os parâmetros σ_1 e σ_2 isto também acontece, porém, com menor intensidade. Este resultado não é esperado, uma vez que o nível de cobertura não deve depender do tamanho da amostra, este fato pode indicar que quando o tamanho da amostra aumenta existe uma falsa impressão de maior confiança nos dados refletida no tamanho da variância dos estimadores, fazendo o intervalo ficar mais estreito do que deveria, provavelmente devido a uma forte assimetria da verossimilhança na direção destes parâmetros.

Para o segundo bloco verifica-se que na configuração 2 o viés para os três parâmetros tende a ser bastante alto. O viés relativo para o parâmetro ρ com $n = 100$ é de -80.16% passando para -21.6% com o aumento da amostra para $n = 225$, mostrando que este parâmetro tende a ser fortemente subestimado nesta configuração, com maior intensidade se o tamanho da amostra for pequeno. Cabe ressaltar que esta configuração é a que apresenta maior erro de medida, os parâmetros τ_1 e τ_2 apresentam valores próximos as variâncias do efeito espacial, além de serem valores elevados, isto com certeza afeta o viés dos EMV. Porém este forte ruído faz com que as variâncias estimadas sejam grandes, o que reflete em intervalos de confiança com nível de cobertura maiores que o nominal. Nas demais configurações verifica-se que o padrão de subestimação continua porém com menor intensidade. O nível de cobertura ficou em torno de 90% para a configuração 1 e próximo do nominal na configuração 3 com excessão do parâmetro τ_1 para $n = 100$ onde a cobertura foi de apenas 78.80%.

Dado os resultados da simulação considera-se o procedimento de estimação por máxima verossimilhança, de acordo com o algoritmo proposto bastante satisfatório. Dada a alta complexidade do modelo, esperava-se um desempenho pior principalmente no que diz respeito a cobertura dos intervalos assintóticos. Os resultados indicam que o intervalos de confiança para o parâmetro ϕ devem ser avaliados com bastante cuidado, uma vez que devem ter menor amplitude que o nível nominal do intervalo preconiza. Além disso, estimativas para o parâmetro ρ tendem a ser fortemente subestimadas, principalmente quando o tamanho da amostra é reduzido.

6. Discussão

Este artigo apresentou um modelo geoestatístico para dados composicionais. A análise deste tipo bastante particular de dados, tem recebido pouca atenção por parte da comunidade estatística em geral, principalmente quando as composições são observados em diversas localizações espaciais, e o fenômeno em estudo é espacialmente contínuo. A modelagem de dados composicionais com dependência seja, espacial e/ou temporal é um desafio para os recursos computacionais atuais. A abordagem mais comum que é transformar os dados para o espaço dos $\Re$ para análise, induz uma distribuição Normal multivariada de grande dimensão, cuja matriz de variância/covariância é densa, que precisa ser fatorada uma grande quantidade de vezes dentro de algoritmos de otimização numérica, isto aumenta drásticamente o custo computacional envolvido na análise.

O modelo apresentado trata do caso onde as composições são de três elementos, porém generalizações deste modelo podem ser facilmente propostas. Isso no entanto não significa que a estimação e predição espacial seja trivial em modelos com mais composições. Pela estrutura proposta, a matriz de variância/covariância tem dimensão igual ao número de observações multiplicada pelo tamanho do vetor das composições menos um. Isto faz com que esta matriz tenha facilmente grande dimensão o que a torna computacionalmente restritiva. Este é um problema comum em geoestatística comumente denominado de problema do grande n. Porém, na situação composicional o problema atinge o limite computacional muito mais rapidamente devido a estrutura multivariada.

Diversas abordagens para tratar deste problema têm sido apresentadas, sem alongar muito neste assunto a abordagem proposta por [27] que apresenta uma explícita ligação entre campos aleatórios Gaussianos e campos aleatórios Gaussianos Markovianos que são de fácil tratamento computacional, parece ser a abordagem mais promissora. A extensão desta metodologia para casos multivariados como do modelo apresentado aqui, é ainda de dificil avaliação e pesquisas nesta direção devem ser consideradas.

Apesar destas restrições para problemas de tamanho moderado o modelo, junto com o processo de estimação e predição espacial apresentados indica um avanço para a análise de dados composicionais com estrutura espacial. Foi apresentado um algoritmo completo que permite a obtenção de mapas das composições respeitando as restrições do espaço paramétrico induzidas pelas composições. Além disso, os parâmetros do modelo trazem interpretações quanto a variabilidade decorrente do efeito espacial e do erro de medida, mensuram a correlação espúria e também a força da dependência espacial.

As dificuldades no procedimento de inferência em modelos espacialmente contínuos são comuns e agravados em estruturas multivariadas. A subestimação dos parâmetros de variância pelo método de máxima verossimilhança era esperado, e tende a diminuir quando o tamanho da amostra aumenta. A construção de intervalos de Wald, apesar de apresentar bons resultados no estudo de simulação, levou a construção de intervalos irreais no exemplo de aplicação. A obtenção de intervalos baseados em perfil de verossimilhança é um procedimento computacionalmente caro, o que dificulta a sua implementação em rotinas padrões em softwares estatísticos, para aplicação rotineira do modelo proposto. Avanços em computação pararela, devem acelerar muito a construção deste tipo de intervalo, uma vez que sua programação pode facilmente ser paralelizada.

Com relação ao algoritmo numérico utilizado na maximização da log-verossimilhança concentrada, a obtenção do gradiente analítico fazendo uso de técnicas de derivação matricial foi de suma importância para o sucesso da estimação. Uma extensão natural seria obter a matriz de derivadas segundas também analíticamente e fazer a otimização via o algoritmo de Newton-Raphson, esta abordagem foi avaliada, porém o tempo computacional para avaliar a matriz Hessiana é elevado uma vez que sua forma exige muitas operações matriciais.

Como perspectivas futuras de desenvolvimento, pretende-se a implementação deste modelo em um pacote *R* para a análise de dados composicionais espaciais. Além disso, a exploração de um modelo com distribuição Normal assimetrica também deve ser proposto e implementado.

Apêndice A. Algoritmo alternativo para estimação por máxima verossimilhança

O vetor de parâmetros do modelo é $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)'$ e sua função de verossimilhança é obtida a partir da função de densidade da distribuição normal multivariada:

$$L(\underline{\theta}; \underline{\mathrm{Y}}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\{-(1/2)(\underline{\mathrm{Y}} - \underline{\mu}_{\mathrm{Y}})^{\top} \Sigma^{-1} (\underline{\mathrm{Y}} - \underline{\mu}_{\mathrm{Y}})\},$$

em que o termo $\underline{\mu}_{\mathrm{Y}}$ é função dos parâmetros μ_1 e μ_2 e os demais parâmetros $(\sigma_1, \sigma_2, \tau_1, \tau_2, \phi, \rho)$ definem os termos em Σ.

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

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

em que $\mathbf{R}$ contém as correlações espaciais e $\mathbf{I}_b$ contém as correlações composicionais. A função de log-verossimilhança reparametrizada é dada por

$$l(\underline{\theta}; \underline{\mathrm{Y}}) = (-1/2)(n \ln(2\pi) + 2n \ln(\sigma_1) + \ln(|\mathbf{V}|) + Qe/\sigma_1^2). \qquad \text{(A.1)}$$

Considerando $\underline{\mu}_{\mathrm{Y}} = \mathbf{D}\underline{\mu}$, em que $\mathbf{D}$ é a matriz do delineamento de ordem $n \times 2$, tem-se que $Qe = (\underline{\mathrm{Y}} - \underline{\mu}_{\mathrm{Y}})' \mathbf{V}^{-1} (\underline{\mathrm{Y}} - \underline{\mu}_{\mathrm{Y}})$ pode ser reescrita como

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

Os estimadores de máxima verossimilhança de $\underline{\mu} = (\mu_1, \mu_2)'$ e σ_1 obtidos diferenciando a função (A.1) em relação aos respectivos parâmetros são dados por

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

Nota-se que $\hat{Q}e$ pode ser escrita como

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

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

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

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

O algoritmo de otimização empregados no processo de maximização foi "L-BFGS-B".

Do processo de maximização obtém-se $\hat{\underset{\sim}{\theta}}^* = (\hat{\eta}, \hat{\nu_1}, \hat{\nu_2}, \hat{\phi}, \hat{\rho})'$ e as respectivas variâncias através da matriz Hessiana numérica dada pela derivada segunda do logaritmo da função de verossimilhança em relação aos parâmetros em $\underset{\sim}{\theta}^*$. A matriz Informação de Fisher observada é definida como o negativo da matriz Hessiana e é dada por

$$\mathbf{I}_F(\hat{\underset{\sim}{\theta}}^*) = -\frac{\partial^2 l(\hat{\underset{\sim}{\theta}}^*)}{\partial \hat{\underset{\sim}{\theta}}^* \partial (\hat{\underset{\sim}{\theta}}^*)'}$$

Para se obter $\hat{\mu}_1$, $\hat{\mu}_2$, e $\hat{\sigma}_1$, basta substituir $\hat{\underset{\sim}{\theta}}^*$ nas Equações (A.2).

Como o interesse está na obtenção de $\hat{\underset{\sim}{\theta}} = (\hat{\mu_1}, \hat{\mu_2}, \hat{\sigma_1}, \hat{\sigma_2}, \hat{\tau_1}, \hat{\tau_2}, \hat{\phi}, \hat{\rho})'$ e suas respectivas variâncias, o método Delta é aplicado para obter uma aproximação da distribuição de $\hat{\underset{\sim}{\theta}}$. Maiores detalhes podem ser encontrados em [23]. Assintoticamente, a distribuição de $\hat{\underset{\sim}{\theta}}$ será aproximadamente multivariada gaussiana com vetor de médias $\hat{\underset{\sim}{\theta}} = g(\hat{\underset{\sim}{\theta}}^*)$ e variância

$$\mathrm{Var}(\hat{\underset{\sim}{\theta}}) \geq \nabla g(\hat{\underset{\sim}{\theta}}^*)' \, \mathbf{I}_{Fe}(\hat{\underset{\sim}{\theta}}^*)^{-1} \, \nabla g(\hat{\underset{\sim}{\theta}}^*),$$

em que $\mathbf{I}_{Fe}$ é a matriz Informação de Fisher esperada e

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

é a função escore $\mathsf{U}(\hat{\underset{\sim}{\theta}}^*)$. Assim, a matriz Informação de Fisher esperada para $\hat{\underset{\sim}{\theta}}^*$, baseada nos dados $\underset{\sim}{Y}$, é substituída pela matriz $\mathbf{I}_F(\hat{\underset{\sim}{\theta}}^*)$ que é assintoticamente equivalente, de modo que

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

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

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

de onde vem

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

Apêndice B. Implementação computacional

Neste apêndice apresentamos os principais passos para a análise geoestatística de dados composicionais, de acordo com a metodologia proposta no artigo. Todas as rotinas foram desenvolvidas na linguagem R e encontram-se disponíveis para uso público. Este apêndice foi escrito para ser auto suficiente no sentido de descrever uma análise completa.

Devido a complexidade do modelo, foram necessários alguns pacotes adicionais que são carregados no código abaixo.

```
> options(width = 65)
> require(bbmle)
> require(statmod)
> require(compositions)
> require(geoR)
> require(mvtnorm)
> require(bbmle)
```

O próximo passo é carregar o conjunto de dados, a fim de facilitar disponibilizamos o conjunto de dados utilizado no artigo em formato .RData. E um arquivo com todas as funções desenvolvidas para a análise.

```
> load("dados.RData")
> source("functions.R")
```

Como em todos os modelos espaciais, é necessário montar a estrutura que vai representar a dependência espacial através da matriz de distâncias.

```
> gride <- dados[[3]]
> Y = c(dados[[1]][, 1], dados[[1]][, 2])
> U <- dist(gride, diag = TRUE, upper = TRUE)
```

Para iniciar o processo de otimização precisamos de um guia inicial, seguindo a proposta do paper obtemos,

```
> ini <- inicial(dados[[1]], U)
```

O processo de otimização da função de log-verossimilhança, usando o escore obtido analiticamente e o algoritmo $L - BFGS - B$,

```
> modelo <- mle2(log.Vero, start = list(s1 = ini[1],
+     s2 = ini[2], t1 = ini[3], t2 = ini[4], phi = ini[5],
+     rho = ini[6]), method = "L-BFGS-B", control = list(factr = 1000),
+     gr = escore, lower = list(s1 = 1e-10, s2 = 1e-10,
+         t1 = 1e-05, t2 = 1e-05, phi = 1e-10, rho = -0.9999),
+     upper = list(s1 = Inf, s2 = Inf, t1 = Inf, t2 = Inf,
+         phi = Inf, rho = 0.9999), data = list(Y = Y,
+         U = U))
```

Como estamos trabalhando com a log-verossimilhança concentrada precisamos obter as estimativas pontuais para $\underline{\mu}$,

```
> medias <- estima.mu(modelo, U = U)
```

Feita a estimação dos parâmetros, seguimos para a predição espacial. Começamos criando a borda da área e um gride de predição.

```
> bor <- cbind(c(0, seq(0, 200, l = 100), 0), c(0,
+     sqrt(200^2 - seq(0, 200, l = 100)^2), 0))
> gr <- pred_grid(bor, by = 100)
```

O procedimento de cokrigagem, na escala transformada

```
> esti.par <- c(medias[, 1], coef(modelo))
> md.cov.ck <- cokrigagem(esti.par, loc = gr, dados.comp = dados)
```

A volta da cokrigagem para o espaço simplex por quadratura de Gauss-Hermite.

```
> preditos.gh <- volta.quad(md.cov.ck, n.pontos = 7,
+     Variancia = FALSE)
```

E por fim, a volta da cokrigagem por simulação.

```
> preditos.simu <- volta.cokri(md.cov.ck, num.simu = 500,
+     retorna.tudo = FALSE, int.conf = 0.95)
```

Referências

[1] G. Matheron, Principles of geostatistics, Economic Geology 58 (8) (1963) 1246–1266. doi:10.2113/gsecongeo.58.8.1246.

[2] N. Cressie, Statistics for Spatial Data, Wiley-Interscience; Revised Edition edition, 1993.

[3] T. Bailey, A. C. Gatrell, Interactive Spatial Data Analysis [Paperback], Prentice Hall; 1 edition, Harlow, 1996.

[4] S. Banerjee, B. Carlin, A. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, Chapman and Hall/CRC; 1 edition, Boca Raton, 2003.

[5] O. Schabenberger, C. A. Gotway, Statistical Methods for Spatial Data Analysis [Hardcover], Chapman and Hall/CRC; 1 edition, Boca Raton, 2004.

[6] H. Wackernagel, Multivariate Geostatistics [Hardcover], Springer-Verlag New York, Inc; 2 edition, 1998.

[7] P. Diggle, R. A. Moyeed, J. A. Tawn, Model-based geostatistics, Applied Statistics 47 (1998) 299–350.

[8] O. Schabenberger, F. J. Pierce, Contemporary Statistical Models for the Plant and Soil Sciences [Hardcover], CRC Press; 1 edition, Boca Raton, 2001.

[9] P. J. Diggle, P. J. Ribeiro, Model-based Geostatistics., Vol. 47 of Springer Series in Statistics, Springer, New York, 2007.

[10] A. O. Finley, S. Banerjee, B. P. Carlin, spBayes: Univariate and Multivariate Spatial Modeling, r package version 0.2-2 (2011).
URL `http://CRAN.R-project.org/package=spBayes`

[11] J. Aitchison, The statistical analysis of compositional data, Journal of the Royal Statistical Society. Series B (Methodological) 44 (2) (1982) pp. 139–177.

[12] J. Aitchson, The Statistical Analysis of Compositional Data, Springer; 1st edition, New Jersey, 1986.

[13] J. Aitchison, Logratios and Natural Laws in Compositional Data Analysis, Mathematical Geology 31 (5) (1999) 563–580.

[14] I. O. Odeh, A. J. Todd, J. Triantafilis, Spatial Prediction of Soil Particle-Size Fractions As Compositional Data, Soil Science 168 (7) (2003) 501–515.

[15] G. Pawlowsky, R. A. Olea, Geostatistical Analysis of Compositional Data, Oxford University Press, USA, New York, 2004.

[16] R. Tolosana-Delgado, N. Otero, V. Pawlowsky-Glahn, Some Basic Concepts of Compositional Geometry, Mathematical Geology 37 (7) (2005) 673–680.

[17] R. M. Lark, T. F. A. Bishop, Cokriging particle size fractions of the soil, European Journal of Soil Science 58 (3) (2007) 763–774.

[18] H. Tjelmeland, K. V. Lund, Bayesian modelling of spatial compositional data, Journal of Applied Statistics 30 (1) (2003) 87–100.

[19] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2009).
URL `http://www.R-project.org`

[20] A. B. T. Martins, P. J. R. Jr, W. H. Bonat, Um modelo geoestatístico para dados composicionais., Revista Brasileira de Biometria 27 (2009) 456–477.

[21] R. H. Byrd, A Limited Memory Algorithm for Bound Constrained Optimization, SIAM Journal on Scientific Computing 35 (5) (1995) 773.

[22] B. Bolker, R. D. C. Team, bbmle: Tools for general maximum likelihood estimation, r package version 0.9.7 (2011).
URL `http://CRAN.R-project.org/package=bbmle`

[23] M. H. DeGroot, M. J. Schervish, Probability and Statistics, 4th Edition, Addison Wesley; 4 edition, New York, 2011.

[24] M. Abramowits, I. A. Stegun, Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables, Dover Publications, Washington, 1965.

[25] D. Gamerman, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference [Paperback], Chapman and Hall/CRC; 1 edition, Londres, 1997.

[26] A. C. Gonçalves, Variabilidade espacial de propriedades físicas do solo para fins de manejo e irrigação., Master's thesis, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba (1997).

[27] F. Lindgren, H. v. Rue, J. Lindström, An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (4) (2011) 423–498.