Implementação de Modelos Geoestatísticos Robustos Baseados na Distribuição t-Student e Correlação Matérn

Rafael Colombo Miranda[1], Gabriel Wenningkamp[2], Rosangela Aparecida Botinha Assumpção[3]

[1]Universidade Tecnológica Federal do Paraná, Toledo, Brasil (rafaelcolombomiranda@gmail.com)
[2]Universidade Tecnológica Federal do Paraná, Toledo, Brasil
[3]Universidade Tecnológica Federal do Paraná, Toledo, Brasil

Resumo: Este trabalho apresenta a implementação de algoritmos para estimação de parâmetros em modelos geoestatísticos espaciais com distribuição t-Student e função de covariância da família Matérn. Para a estimação dos parâmetros do modelo foi utilizado o algoritmo EM (Expectation-Maximization), possibilitando o ajuste dos componentes espaciais e de dependência. Além disso, foram desenvolvidas, em Python, ferramentas para visualização de semivariogramas, análise de resíduos e validação cruzada, visando auxiliar a avaliação da qualidade do ajuste e do desempenho preditivo dos modelos. Os resultados demonstram a viabilidade da implementação proposta como uma ferramenta para análise e modelagem de dados espaciais com distribuições não gaussianas.

Palavras-chave: Geoestatística; Distribuição t-Student; Correlação Matérn; Algoritmo EM; Modelagem Espacial

INTRODUÇÃO

A geoestatística constitui uma importante ferramenta para a análise e modelagem de dados espacialmente referenciados, sendo amplamente aplicada em áreas como agricultura de precisão, geologia, ciências ambientais e recursos naturais. Os modelos geoestatísticos permitem descrever a dependência espacial entre observações e realizar predições em locais não amostrados, fornecendo suporte para a tomada de decisões em diferentes contextos.

Tradicionalmente, esses modelos são formulados sob a suposição de normalidade dos dados. No entanto, em aplicações práticas, é comum a ocorrência de observações discrepantes (outliers) ou distribuições com caudas mais pesadas do que a distribuição normal, situações que podem comprometer a estimação dos parâmetros espaciais e reduzir a qualidade das predições obtidas. Nesse contexto, a utilização da distribuição t-Student multivariada surge como uma alternativa robusta, por apresentar maior resistência à influência de valores extremos.

Além da especificação da distribuição probabilística, a escolha de uma função de correlação adequada é fundamental para representar a estrutura de dependência espacial dos dados. Entre as funções mais utilizadas destaca-se a família Matérn (1986), devido à sua flexibilidade para modelar diferentes níveis de suavidade do processo espacial.

A estimação dos parâmetros em modelos geoestatísticos com distribuição t-Student e correlação Matérn envolve procedimentos computacionais mais complexos do que aqueles empregados em modelos gaussianos, exigindo a utilização de métodos iterativos de otimização. Diante disso, este trabalho tem como objetivo apresentar uma implementação computacional, desenvolvida em Python, para a estimação dos parâmetros de modelos geoestatísticos espaciais com distribuição t-Student e função de correlação da família Matérn. Adicionalmente, a implementação incorpora ferramentas para análise residual, validação cruzada e visualização de semivariogramas, contribuindo para a avaliação do ajuste e do desempenho preditivo dos modelos. A implementação foi desenvolvida com base nos fundamentos teóricos e metodológicos apresentados por Assumpção (2010).

MATERIAL E MÉTODOS

A implementação foi desenvolvida utilizando a linguagem Python (VAN ROSSUM; DRAKE, 2009), com forte suporte de bibliotecas científicas para otimização de performance e manipulação matricial. O ecossistema baseou-se na biblioteca numpy (HARRIS et al., 2020) para operações de álgebra linear (como a decomposição de Cholesky), scipy.special (VIRTANEN et al., 2020) para o cálculo

de funções matemáticas complexas exigidas pelos modelos espaciais, e matplotlib (HUNTER, 2007) para a visualização dos semivariogramas e análises gráficas.

Considera-se $Y = (Y_{(s_i)}, \ldots, Y_{(s_n)})^T$ como o vetor de observações da variável resposta associadas às localizações espaciais observadas em s_i e s_j para $i \neq j = 1, 2, \ldots, n$. De forma matricial **Y** pode ser escrito como:

$$Y = X\beta + e \quad (1)$$

onde: **X** é a matriz n x p composta pelo vetor de um's e pelas (p -1) covariáveis, $\beta = (\beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros desconhecidos a serem estimados e e é um vetor estocástico n x 1.

A dependência espacial entre as observações é representada por uma matriz de covariância construída a partir dos parâmetros espaciais do modelo:

$$\Sigma = \phi_1 I + \phi_2 R(\phi_3), \quad (2)$$

em que: ***I*** é a matriz identidade n x n, ϕ_1 representa o efeito pepita puro (nugget), ϕ_2 corresponde à contribuição (sill), ϕ_3 define o alcance espacial (range) e $\mathbf{R}(\phi_3)$ é uma matriz n x n, simétrica, que representa a estrutura de correlação espacial em função de ϕ_3.

Para modelagem da correlação espacial foi utilizada a função Matérn, amplamente empregada em geoestatística devido à sua flexibilidade para representar diferentes níveis de suavidade espacial:

$$\rho(h) = \frac{1}{2^{k-1}\Gamma(k)} \left(\frac{h}{\phi_3}\right)^k K_k\left(\frac{h}{\phi_3}\right). \quad (3)$$

Nesta equação, h é a distância euclidiana entre as observações, κ é o parâmetro de suavidade, Γ é a função gama e K_k é a função de Bessel modificada de segundo tipo de ordem κ.

Para a estimação dos parâmetros do modelo espacial foi implementado o algoritmo Expectation-Maximization (EM) (ASSUMPÇÃO, 2014). Durante as iterações do algoritmo, o peso robusto associado à distribuição t-Student é atualizado segundo:

$$v = \frac{gl+n}{gl+u}, \quad (4)$$

em que gl representa os graus de liberdade da distribuição t-Student, n corresponde ao número de observações e u a distância de Mahalanobis associada aos resíduos. Esse mecanismo contribui para aumentar a robustez do modelo diante da presença de observações discrepantes (outliers).

O processo iterativo é encerrado quando o critério de convergência baseado no erro relativo entre parâmetros sucessivos atinge um valor suficientemente pequeno, definido pela norma máxima relativa:

$$erro = \frac{\|\theta_{old} - \theta_{new}\|}{\|\theta_{old}\|}, \quad (5)$$

onde: θ_{new} e θ_{old} representam os vetores de parâmetros antes e após a atualização iterativa. Esse critério permite verificar a estabilidade das estimativas produzidas pelo algoritmo.

Além da estimação dos parâmetros espaciais, foram implementadas ferramentas complementares para visualização de semivariogramas empíricos e teóricos e validação cruzada do tipo Leave-One-Out (LOOCV), permitindo avaliar a qualidade do ajuste obtido pelos modelos.

RESULTADOS E DISCUSSÃO

Para avaliar o desempenho da implementação, o algoritmo foi aplicado a um conjunto de dados experimentais composto n observações georreferenciadas, variável resposta de interesse e duas covariáveis.

A Figura 1 apresenta os semivariogramas empíricos ajustados pelo modelo teórico de correlação espacial da família Matérn com diferentes valores para o parâmetro de suavidade κ: κ = 0,5 (Exponencial), κ→∞ (Gaussiano) e κ = 0,7 (Matérn).

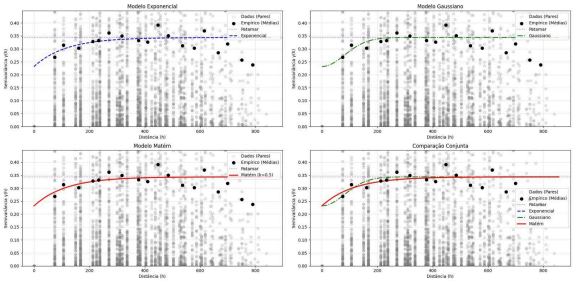

Figura 1. Comparação do ajuste do modelo da família Matérn ao semivariograma empírico.

Visualmente, constata-se que o modelo Exponencial cresce rapidamente no início e depois se estabiliza, enquanto o modelo Gaussiano apresenta crescimento mais suave próximo à origem, atingindo gradualmente o patamar. Em contrapartida, o modelo Matérn com κ = 0,7 proporcionou um ajuste superior, acompanhando adequadamente o crescimento da semivariância desde as menores distâncias até a estabilização no patamar.

Para atestar a qualidade preditiva dos modelos ajustados, realizou-se a Validação Cruzada Leave-One-Out (LOOCV). Para cada modelo obteve-se o Erro Médio (EM), o Erro médio reduzido (EMR), Desvio Padrão do Erro Médio (S_{EMR}) e Erro Absoluto (EA). Comparando esses resultados entre os três modelos verificou-se que o modelo Matérn com k=0,7 teve melhor ajuste. Esse resultado evidencia a flexibilidade da família Matérn para representar processos espaciais com níveis intermediários de suavidade. Para esse modelo os resultados da Validação Cruzada indicaram um Erro Médio (EM) muito próximo a zero (0,000986), confirmando que as predições geradas pelo modelo não são tendenciosas. Além disso, o Erro Médio Reduzido resultou em 0,000837, com um desvio padrão de 1,0108. Como o desvio padrão do erro médio é próximo de 1, conclui-se que o modelo estimou corretamente a variância do erro de predição. O Erro Absoluto (EA) consolidou-se em 0,4250, um valor satisfatório considerando a escala e a dispersão dos dados analisados.

Para atestar a robustez da implementação computacional e a estabilidade da distribuição t-Student frente a condições adversas, realizou-se um segundo experimento de sensibilidade (Figura 2). Neste cenário, os parâmetros espaciais iniciais foram definidos de forma intencionalmente distante de seus valores ótimos reais: o efeito pepita inicial foi superestimado para $\phi_1 = 2{,}0$ e o patamar parcial para $\phi_2 = 1{,}0$ e $\phi_3 = 100$. O parâmetro de suavidade da função Matérn foi considerado $\kappa = 0{,}7$.

Devido à grande distância entre as estimativas iniciais e o ótimo global, o algoritmo necessitou de um esforço computacional maior, atingindo o limite estipulado de 100 iterações sem acionar o critério de convergência estrita, estabilizando com um erro relativo residual na ordem de 10^{-4} (0,000760). No entanto, a implementação demonstrou robustez. Embora o critério formal de convergência não tenha sido atingido, as estimativas estabilizaram-se em uma região paramétrica equivalente à solução obtida no cenário inicial. O algoritmo EM estimou $\phi_1 = 0{,}2457$, $\phi_2 = 0{,}0940$ e $\phi_3 = 107{,}42$.

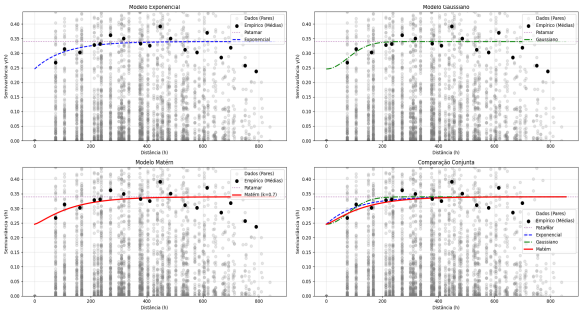

Figura 2. Ajuste dos semivariogramas empíricos aos modelos teóricos partindo de estimativas iniciais extremas e com $\kappa = 0{,}7$.

Essa estabilidade paramétrica é respaldada pelas métricas de validação cruzada (LOOCV), que permaneceram virtualmente inalteradas em comparação ao cenário com estimativas iniciais ideais. O Erro Absoluto (EA) consolidou-se em 0,4245 e o desvio padrão do erro médio manteve-se em 1,018. A manutenção da acurácia preditiva prova que a arquitetura do algoritmo implementado, amparada pelos pesos iterativos da distribuição t-Student, é altamente robusta, capaz de contornar entradas subótimas e garantir resultados confiáveis para a modelagem espacial.

CONCLUSÃO

Este trabalho alcançou com sucesso o objetivo de desenvolver uma implementação computacional em Python para a estimação robusta de modelos geoestatísticos. A combinação da distribuição t-Student multivariada com a função de correlação Matérn mostrou-se eficaz, garantindo flexibilidade para modelar a continuidade espacial dos dados e mitigar a influência de valores discrepantes por meio do ajuste iterativo de pesos.

Computacionalmente, a rotina de otimização desenvolvida demonstrou alta eficiência e estabilidade. Em cenários com estimativas iniciais adequadas, os algoritmos convergiram rapidamente, em aproximadamente 20 iterações. Além disso, testes de estresse comprovaram a estabilidade da implementação: mesmo sob condições adversas e estimativas iniciais intencionalmente distantes do ótimo, os algoritmos não divergiram e mantiveram a capacidade de localizar a região paramétrica correta.

As métricas de validação cruzada confirmaram a acurácia preditiva do modelo em todos os cenários testados, atestando a ausência de viés e a correta quantificação das incertezas. A disponibilização destas rotinas estruturadas utilizando bibliotecas científicas de alto desempenho contribui para que pesquisadores apliquem metodologias espaciais

robustas com maior acessibilidade, resiliência e rigor metodológico.

AGRADECIMENTOS

Os autores agradecem à Universidade Tecnológica Federal do Paraná (UTFPR) pelo suporte institucional, incentivo à pesquisa e infraestrutura fornecida, elementos essenciais para a execução e conclusão deste trabalho.

REFERÊNCIAS

ASSUMPÇÃO, R. A. B. Influência local em modelos geoestatísticos t-Student com aplicações a dados agrícolas. Tese (Doutorado) – Universidade Estadual do Oeste do Paraná, Cascavel, 2010.

ASSUMPÇÃO, R. A. B.; URIBE-OPAZO, M. A.; GALEA, M. Analysis of local influence in geostatistics using Student's t-distribution. Journal of Applied Statistics, v. 41, n. 11, p. 2323-2341, 2014.

HARRIS, C. R. et al. Array programming with NumPy. Nature, v. 585, n. 7825, p. 357-362, 2020.

HUNTER, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007.

MATÉRN, B. Spatial variation. Lectures notes in statistics. 2 ed. New York: Springer, 1986. 151p.

VAN ROSSUM, G.; DRAKE, F. L. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace, 2009.

VIRTANEN, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, v. 17, n. 3, p. 261-272, 2020.