Comparação de Krigagem Universal com um Modelo de Redes Neurais

Yuri Dessimon [1] Pedro Dresseno John [1] Márcia Helena Barbian [1,2]
[1]Departamento de Estatística, UFRGS [2]Programa de Pós-Graduação em Estatística, UFRGS
[1]yuridessimon@gmail.com [1]johnpd4@gmail.com [2]mhbarbian@gmail.com

2nd SouthStat Meeting

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Introdução

A modelagem de fenômenos ambientais desafia os métodos geoestatísticos tradicionais, que frequentemente assumem linearidade e normalidade nos dados. Tais suposições simplificam excessivamente a complexidade dos processos naturais, limitando a acurácia das predições. Para superar essas barreiras, este trabalho emprega o modelo híbrido NN-GLS desenvolvido por [1], o modelo utiliza Redes Neurais (NN) para capturar tendências não-lineares com Mínimos Quadrados Generalizados (GLS) para modelar a correlação espacial nos resíduos. A abordagem NN-GLS pode ser vista como uma Rede Neural de Grafo (GNN), em que as observações são nós em um grafo espacial, cujas dependências são modeladas localmente, esse tipo de metodologia tem sido aplicadas em dados ambientais como em [2]

Objetivo

O objetivo do trabalho é comparar a predição do modelo NN-GLS com o modelo que considera que os dados observados são oriundos de um campo aleatório gaussiano. Os dados utilizados indicam a quantidade média de precipitação anual no território Brasileiro durante o ano de 2024. Além disso, as covariáveis consideradas são: temperatura, amplitude e rajada médias anuais. Os dados foram coletados pelas estações automáticas do Instituto Nacional de Meteorologia (INMET) [3].

Campo Aleatório Gaussiano

A abordagem geoestatística modela a variável de interesse, $Y(s)$, em que s indica a localização do evento, representando a realização de um processo aleatório espacial definido sobre um domínio $D \subset \mathbb{R}^2$. É comum considerar que a variável $Y(s)$ pode ser representada por um campo aleatório gaussiano com uma estrutura linear para a média [4]:

$$Y(s) = \mathbf{X}(s)^T\beta + \omega(s) + \epsilon(s), \tag{1}$$

- $\mathbf{X}(s)^T\beta$ é a média gerada pelo vetor de covariáveis $\mathbf{X}(s)$ e o vetor de coeficientes de regressão β;
- $\omega(s)$ Resíduos que seguem um processo gaussiano de média 0 e função de covariância que depende da distância dos pontos de forma $\epsilon(\cdot) \sim GP(0, C(d))$.
- $\epsilon(s)$ É o resíduo não-espacial (efeito pepita), segue distribuição $\epsilon(s) \sim N(0, \tau^2)$, *IID*
- τ^2 o efeito pepita (nugget), variabilidade que não é espacialmente estruturada;

Comumente modela-se função a covariância $C(d)$ como uma função Matérn, devido à sua flexibilidade e interpretabilidade:

$$C(d) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{d}{\phi}\right)^{\nu} K_\nu \left(\frac{d}{\phi}\right), \tag{2}$$

- d distâncias dos pontos;
- ν Suavidade do campo aleatório;
- σ o patamar (sill), quantidade de variação que pode ser explicada pela localização dos pontos;
- ϕ o alcance (range), distância máxima na qual os valores de dois pontos são relacionados;
- K_ν Função de Bessel.

Podemos estimar os parâmetros desconhecidos desse modelo $(\beta, \sigma, \phi, \tau)$ através do método da máxima verossimilhança.
Com os parâmetros estimados, é possível fazer a predição, através do método de Krigagem universal [4]. Logo, busca-se estimar o valor da precipitação em locais onde não foi realizada a coleta de dados, a krigagem é definida pela equação:

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i), \tag{3}$$

- $\hat{Z}(s_0)$ é o valor predito no novo local s_0;
- $Z(s_i)$ representa o valor da amostra no local s_i;
- λ_i é o peso da amostra $Z(s_i)$ definido por $\boldsymbol{\lambda} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$

NN-GLS

O modelo da NN-GLS é definido por:

$$Y(s) = f(\mathbf{X}(s)) + \epsilon(\mathbf{s}), \quad \epsilon(\cdot) \sim GP(0, \Sigma(\cdot,\cdot)). \tag{4}$$

Para a estimação de dados dependentes o uso de estimarmos Mínimos Quadrados Generalizados (GLS) é mais eficiente que nínimos Quadrados Ordinários (OLS) [1], assim para estimar $f(\cdot)$ é usada a função perda:

$$\mathcal{L}_n(f) = \frac{1}{n}(\mathbf{Y} - f(\mathbf{X}))^{\mathsf{T}}\mathbf{Q}(\mathbf{Y} - f(\mathbf{X})), \tag{5}$$

Que leva em conta a dependência espacial pela matriz de precisão $Q = \Sigma^{-1}$.
Esse método em geral é computacionalmente inviável para mini-batching e retropropagação. Para mitigar isso, é usado OLS com a variável resposta decorrelacionada, é também aplicado a técnica de Nearest Neighbor Gaussian Process (NNGP) [5] para estimar o valor em um ponto $Y(s)$ com apenas um número fixo de vizinhos mais próximos e a NN-GLS é representada como uma GNN.

O método de krigagem da rede é descrito pela função:

$$\hat{Y}_0 = \hat{f}(\mathbf{X}_0) + \Sigma\left[s_0, N(0)\right]\Sigma\left[N(0), N(0)\right]^{-1}\left(\mathbf{Y}_{N(0)} - \hat{\mathbf{f}}_{N(0)}\right), \tag{6}$$

- $\hat{f}(\mathbf{X}_0)$ predição da tendência;
- $\left(\mathbf{Y}_{N(0)} - \hat{\mathbf{f}}_{N(0)}\right)$ vetor de resíduos observados nos locais vizinhos;
- $\Sigma\left[s_0, N(0)\right]$ é o vetor de covariâncias entre o novo local e cada um de seus vizinhos;
- $\Sigma\left[N(0), N(0)\right]^{-1}$ a inversa da matriz de covariância entre os locais vizinhos.

Resultados

Os resultados do modelo gaussiano foram obtidos no software **R** por meio do pacote *geoR* [6] e a rede neural foi estimada por meio do pacote do **Python** *geospaNN* [1]. A Figura 1 mostra a distribuição espacial estimada da precipitação e a Figura 2 o erro quadrático médio de predição dos dois modelos.

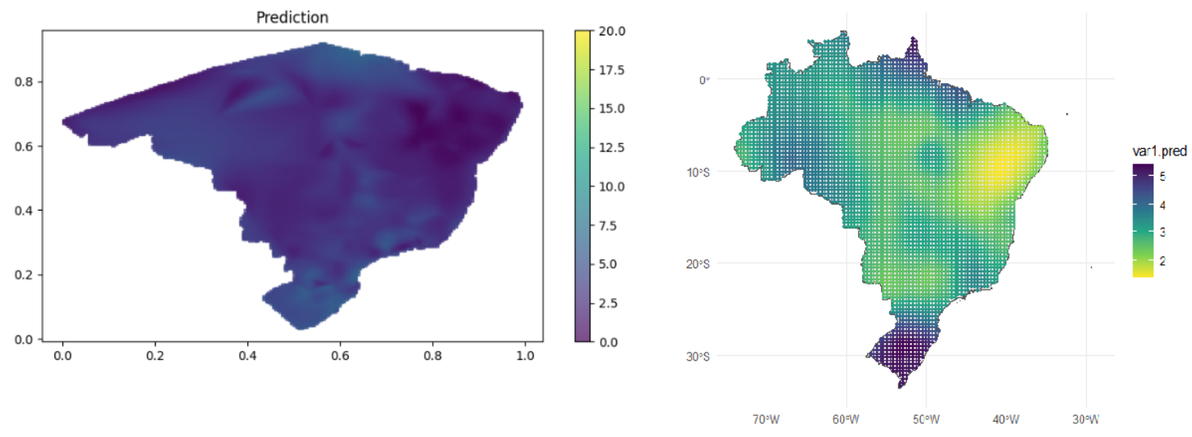

Figura 1: À esquerda resultado da predição da NN-GLS e à direita da Krigagem universal, Fonte: do Autor.

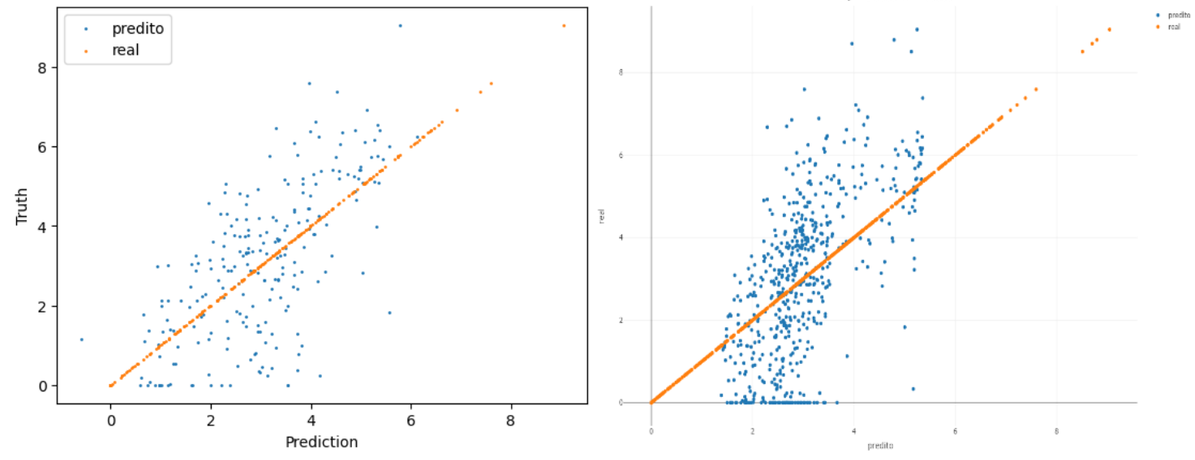

Figura 2: Gráfico de dispersão dos valores preditos e reais à direita o da NN-GLS e à esquerda o do processo gaussiano, Fonte: do Autor.

A validação dos modelos foi realizada comparando o Erro Quadrático Médio (EQM) das predições. Os resultados mostraram que o modelo NN-GLS alcançou um EQM de 2.128, enquanto o método de Krigagem Universal, usado como base de comparação, obteve um EQM de 2.308. Isso se dá pela não-linearidade das covariáveis que incluímos.

Conclusão

A fusão entre redes neurais e geoestatística, como implementada pelo NN-GLS, indica que que é um avanço grande avanço para a análise ambiental, permitindo modelar a complexidade dos fenômenos naturais de forma mais realista e computacionalmente viável. No exemplo de precipitação houve uma melhora do EQM indicando uma possível melhora em relação aos métodos geoestatísticos clássicos ao dispensar as suposições de linearidade e normalidade, resultando em predições mais acuradas.

Referências

[1] Wentao Zhan and Abhirup Datta.
Neural networks for geospatial data.
Journal of the American Statistical Association, 120(549):535–547, 2025.

[2] Ditsuhi Iskandaryan, Francisco Ramos, and Sergio Trilles.
Graph neural network for air quality prediction: A case study in madrid.
IEEE Access, 11:2729–2742, 2023.

[3] Instituto Nacional de Meteorologia (INMET).
Tabela de estações meteorológicas, 2025.
Acesso em: 16 mar. 2025.

[4] Noel Cressie.
Statistics for spatial data.
John Wiley & Sons, 2015.

[5] Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand.
Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets.
Journal of the American Statistical Association, 111(514):800–812, 2016.

[6] Paulo Justiniano Ribeiro Jr, Peter Diggle, Ole Christensen, Martin Schlather, Roger Bivand, and Brian Ripley.
geoR: Analysis of Geostatistical Data, 2024.
R package version 1.9-4.

Agradecimentos

Agradeçemos à FAPERGS por viabilizar este projeto (24/2551-0002361-5) através do Edital 06/2024 - Programa de Pesquisa e Desenvolvimento Voltado a Desastres Climáticos