ANÁLISE DE CONSISTÊNCIA DE DADOS PLUVIOGRÁFICOS E PREENCHIMENTO DE FALHAS NA BACIA HIDROGRÁFICA DO RIO MEIA PONTE

Natasha de Lima Dias Conceição [1] ;Kamila Almeida dos Santos [2] ; Ismael Torres Guedes [3]; Mel Martins Vaz [4]; Nicolle Silva Oliveira [5]; Julliana Gomes Martins[6]; Klebber Teodomiro Martins Formiga [7]

Abstract: This study aimed to evaluate and select the most suitable method for filling gaps in precipitation time series in the Meia Ponte river basin, located in the state of Goiás. The research was developed using rainfall data from 21 stations, obtained through the Agência Nacional de Águas e Saneamento Básico (ANA). In this sense, the daily precipitation series were initially processed to remove inconsistent values and later organized into monthly and annual averages. For the validation of the filing methods, a database without artificial gaps (Ground Truth) was generated and soon after, a 5% introduction of random gaps. The performance of eight imputation methods was evaluated: Simple Linear Regression with Available Records (SLR_VAR), Multiple Linear Regression with Fixed Predictors (MLR_FIX) K-Nearest Neighbors Regression (KNN_R), Regional Weighting (RW), Nonlinear Iterative Partial Least Squares (NIPALS), KNN Analog, Arithmetic Mean and Iterative PCA. The qualification was based on statistical metrics RMSE, MAE, R^2, NSE Spearman coefficient and Willmott's Concordance Index (d), calculated for each station and then a global average of each metric was performed (Wangwongchai et al.,2023). In this sense, the results showed that the MLR_FIX method presented the best average performance in practically all evaluated parameters, especially for the lowest RMSE (52.55 mm) and highest average NSE (0.7906), which motivated its choice for the final filling of the time series. The graphical analysis through the Taylor Diagram supported these results, highlighting the proximity between the filled data and the original values.

Resumo: Esse estudo teve como objetivo a avaliar e selecionar o método mais adequado para o preenchimento de falhas em séries temporais de precipitação na bacia hidrográfica do rio Meia Ponte, localizada no estado de Goiás. A pesquisa foi desenvolvida utilizando dados pluviométricos de 21 postos, obtidos por meio da Agência Nacional de Águas e Saneamento Básico (ANA). Nesse sentido,

1, 2, 3, 4, 5, 6 E 7 Universidade Federal de Goiás -UFG: Avenida Universitária, Quadra 86, Lote Área 1488 -Setor Leste Universitário, Goiânia -GO, 74605-220. Fone: (62) 3209-6086. e-mail autor correspondente: natashalimadias2005@gmail.com

as séries diárias de precipitação foram inicialmente tratadas para remoção de valores inconsistentes e posteriormente organizadas em médias mensais e anuais. Para a validação dos métodos de preenchimento, foi gerada uma base de dados sem falhas artificiais (*Grouth Truth*) e logo após, uma introdução de 5% de falhas aleatórias. O desempenho de oito métodos de imputação foi avaliado: *Simple Linear Regression with Available Records* (SLR_VAR), *Multiple Linear Regression with Fixed Predictors* (MLR_FIX), *K-Nearest Neighbors Regression* (KNN_R), Ponderação Regional (PR), *Nonlinear Iterative Partial Least Squares* (NIPALS), *KNN_Analog*, Média aritmética e PCA Iterativo. A qualificação foi com base em métricas estatísticas RMSE, MAE, R^2, NSE, coeficiente de Spearman e Índice de Concordância de Willmott(d), calculadas para cada estação e depois realizada uma média global de cada métrica (Wangwongchai *et al*.,2023). Nesse sentido, os resultados mostraram que o método MLR_FIX apresentou o melhor desempenho médio em praticamente todos os parâmetros avaliados, com destaque para o menor RMSE (52, 55 mm) e maior NSE médio (0,7906), o que motivou a sua escolha para o preenchimento final das séries temporais. A análise gráfica pelo Diagrama de Taylor apoiou esses resultados, evidenciando a proximidade entre os dados preenchidos e valores originais.

Palavras-Chave – Preenchimento de falhas, Dados diários, Validação

INTRODUÇÃO

Entre as variáveis do ciclo hidrológico, a precipitação é um dos fenômenos mais importantes dentro do sistema e por consequência disso, é necessário o fornecimento de dados para diversas análises hidrológicas para o monitoramento e planejamento dos recursos hídricos (Wangwongchai et al.,2023). Inúmeras áreas de pesquisa hidrológica, utilizam dados de séries de dados de chuvas, seja como a previsão de enchentes, avaliação de riscos, previsão de chuvas, análise de variabilidade climática e modelagem, entretanto, a análise de dados hidrogeológicos é problemática devido as falhas que os dados podem apresentar (Orian, 2020) (Chiu et al., 2021).

A análise e extração de dados hidrológicos possui uma importância sobre os estudos de fenômenos hidrológicos, a sua qualidade está associada a um padrão estável e livres de anomalias para garantir resultados eficazes (Zhao et al., 2024). As possíveis causas para as falhas encontradas estão relacionadas às imprecisões nas instalações, saltos anormais ou condições climáticas extremas provocando erros (Zhao et al., 2024).

Essas falhas consistem, por exemplo, na manutenção nos equipamentos de medição, mudanças na localização dos postos, erros sistemáticos e na transposição de dados, dessa forma,

obtendo-se valores ausentes (Filho e Lima, 2016) (Chiu et al.,2021). As redes de estações hidrometeorológicas têm como responsáveis entes públicos e privados, nacionais e subnacionais. A Agência Nacional de Águas e Saneamento (ANA), em que, somente 21% do total de estações são de responsabilidade da ANA, e 11% sob responsabilidade do CEMADEN (Centro Nacional de Monitoramento e Alertas de Desastres Naturais) (Sarmento, 2021).

Métodos tradicionais para a correção das lacunas tem sido utilizado na abordagem do vizinho mais próximo, em que se utiliza os dados daquelas estações vizinhas selecionadas geometricamente ou fazendo a ponderação das estações (Miró; Caselles, Estrela, 2017). A imputação múltipla é uma abordagem para a correção dos dados ausentes de precipitação, em que lida de uma maneira que produz uma inferência estatística, ao contrário de estimar os valores ausentes o mais próximo possível de valores observados (Hamzah, F; Hamzah, M; Razali, 2021).

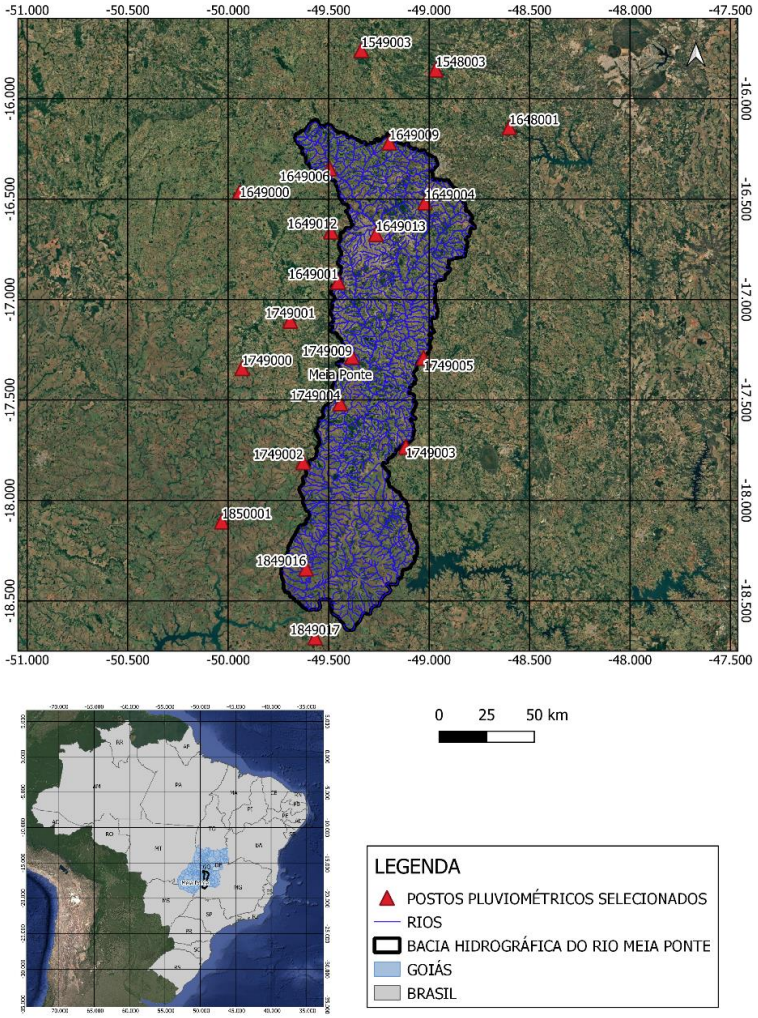
A bacia hidrográfica constitui uma unidade fundamental para o gerenciamento das atividades de uso e conservação dos recursos naturais (Li, 2019). Nesse contexto, o presente estudo utiliza dados pluviográficos coletados na bacia do rio Meia Ponte, cuja nascente está localizada na Serra dos Brandões, nos municípios de Itauçu e Taquaral de Goiás, desaguando posteriormente no rio Paranaíba (Calil et al., 2012). A região apresenta clima tropical úmido, com duas estações bem definidas-uma chuvosa e outra seca- conforme a classificação de Koppen (Calil et al., 2012). Além disso, destaca-se que o rio Meia Ponte é um dos principais responsáveis pelo abastecimento da cidade de Goiânia.

METODOLOGIA

1. ÁREA DE ESTUDO E COLETA DE DADOS PLUVIOMÉTRICOS.

A área de estudo está localizada na bacia hidrográfica do rio Meia Ponte, no estado de Goiás. A distribuição dos postos selecionados foi obtida a partir do banco de dados da ANA utilizando a ferramenta ANA Data Acquisition, disponível no software QGIS, conforme apresentado na figura 1. O preenchimento de falhas foi realizado para 21 estações pluviométricas localizadas na bacia e seu entorno, considerando exclusivamente métodos estatísticos e suas métricas de desempenho, uma vez que, o objetivo do estudo foi a comparação de métodos com base em séries temporais de precipitação.

FIGURA 1: Localização da área de estudo e a distribuição dos postos pluviométricos selecionados.



2. PREPARAÇÃO E ORGANIZAÇÃO DO DADOS.

As séries diárias de precipitação foram organizadas em um formato contendo dia, mês e ano. Nessa etapa, foi desenvolvido no *MATLAB*, a retirada de valores negativos e incoerentes, mantendo apenas dados válidos. Após isso, foram calculadas as médias mensais e anuais de precipitação para cada estação, considerando os valores válidos, para a obtenção da caracterização estatística das séries antes da etapa de validação.

3. VALIDAÇÃO DO MÉTODO DE PREENCHIMENTO.

Para a validação dos métodos de preenchimento, foi criado um *dataset* de referência sem falhas (*Ground Truth*), aplicando o método PCA iterativo sem falhas, por meio da função de preenchimento por KNN, nativo do *MATLAB*, preenchendo os dados faltantes antes da inclusão de

falhas artificiais. Em seguida, foi aplicada uma taxa de 5% de falhas artificiais de maneira aleatória nas séries completas, em que, os valores originais foram armazenados para posterior comparação com valores preenchidos.

Diante disso, foram aplicados oito métodos de preenchimento de falhas, nos quais, cada método foi implementado via funções próprias no *MATLAB*, considerando a configuração e parâmetros de cada abordagem, sendo eles: *Simple Linear Regression with Availab Records* (SLR_VAR); *Multiple Linear Regression with fixed Predictors* (MLR_FIX); *K-Nearest Neighbores Regression* (KNN_R); Ponderação Regional (PR); *Nonlinear Iterative Partial Least Squares* (NIPALS) e Média Aritmética.

Nesse sentido, o desempenho dos métodos foi avaliado por meio de seis métricas estatísticas, de modo que, os cálculos foram realizados para cada estação e cada método, usando os valores originais como referência, sendo: RSME (*Root Mean Square Error*) (Hodson, 2022); MAE (*Mean Absolute Error*) (Matsuura; Willmott, 2005); R^2 (*Coeficiente de Determinação*); NSE (*Nash-Sutcliffe Efficiency*); Spearman (*Coeficiente de correlação de Spearman*) e Índice de Concordância de Willmott (d).

4. ESCOLHA DO MÉTODO E PREENCHIMENTO FINAL DOS DADOS REAIS.

Para a escolha do método de preenchimento foi realizada com base em uma análise de médias de desempenho, considerando todas as estações do estudo (Wangwongchai et al.,2023). Nesse sentido, o Multiple Linear Regression with Fixed Predictors (MLR_FIX), foi o método selecionado para o preenchimento de falhas, por apresentar os melhores resultados médios das métricas (Hodson, 2022). Os dados finais foram salvos em formato (. mat) e também exportados para uma planilha excel com uma aba para cada estação.

RESULTADOS E DISCUSSÕES

Em relação a análise comparativa entre os métodos avaliados, a seleção do método definitivo para o preenchimento falhas foi determinada baseando-se nas médias das métricas estatísticas calculadas para todas as estações. Nesse sentido, o método *Multiple Linear Regression with Fixed Predictors* (MLR_FIX), mostrou um menor valor médio de RMSE (52, 55mm), além de obter os maiores valores médios de NSE (0,7906) e o Índice de Concordância de Willmott (0,94606), de acordo com a tabela 1.

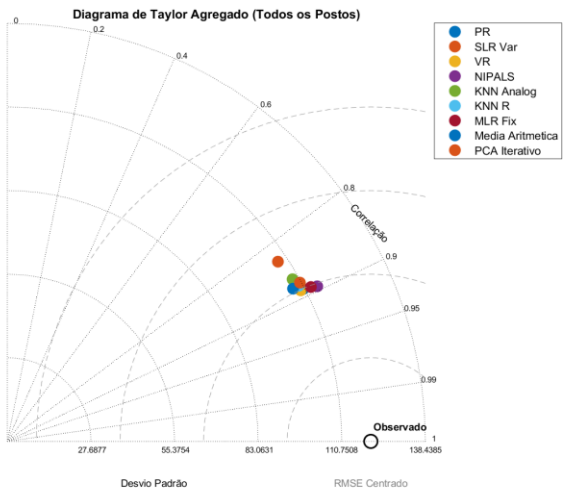
O estudo realizado por Wangwongchai *et al.*(2023), fizeram a comparação entre modelos, mostrando que o MLR atingiu os melhores parâmetros devido sua capacidade de fornecer bons resultados de estimativa, assim como nesse estudo.

Tabela 1: desempenho médio das métricas de análise.

MÉTODO	RMSE	MAE	R2	NSE	Spearman	Willmott_d
PR	52.69368532	33.29032854	0.805765885	0.783236471	0.928244823	0.938223569
SLR_Var	64.22262004	43.13738476	0.717905729	0.683068835	0.897528434	0.903599117
VR	52.21934344	33.89420539	0.809809862	0.786504124	0.928257263	0.937770098
NIPALS	52.58507021	34.32785193	0.813136949	0.78423947	0.930835175	0.940658844
KNN_Analog	57.19155558	36.539885	0.773008434	0.75328315	0.923122011	0.927553568
KNN_R	54.16914693	34.19165277	0.803223036	0.770054023	0.931846405	0.935093532
MLR_Fix	52.54608488	33.62704806	0.809664805	0.790686738	0.929178898	0.940616397
Media_Aritmetica	53.96143044	34.00581242	0.80628057	0.770530317	0.928505486	0.934376328
PCA_Iterativo	55.25502381	34.5222929	0.788474475	0.762634763	0.925061805	0.932351689

Para uma avaliação gráfica do desempenho dos métodos, foram gerados Diagramas de Taylor, que é uma ferramenta utilizada para comparação de modelos estatísticos, frequentemente utilizado em áreas como a climatologia, ciências ambientais e hidrologia, resumindo em um único gráfico, métricas estatísticas importantes como o coeficiente de correlação e erro médio centrado (HU *et al.*, 2021) (Iacobellis et al.,2024). O Diagrama de Taylor agregado pode ser observado na figura 2.

Figura 2: Diagrama de Taylor Agregado.



Após a escolha definitiva do método MLR_FIX, foi aplicado ao conjunto de dados reais contendo as falhas originais, de modo que, as séries preenchidas foram exportadas em formato Excel (.xlsx). De modo geral, o MLR superou as outras técnicas de preenchimento de lacunas em series diárias de precipitação, apresentando o menor RMSE e as maiores pontuações de habilidades (Elias et al., 2021).

Um estudo realizado no Rio Grande do Sul, evidenciou que o preenchimento de falhas de dados diários de precipitação apresentou maior coeficiente de determinação com o modelo de regressão linear múltipla, apresentando o R^2 de 0,697. No caso desse estudo, possivelmente por variação da região e análise temporal dos dados, apresentou um R^2 maior de 0.809 (Brubacher; Oliveira; Guasselli, 2020). Por fim, o uso do método de regressão linear com preditores fixos é uma ferramenta valiosa para preenchimento de falhas em regiões tropicais, desde que, haja seleção criteriosa de preditores e validação estatística dos resultados (Almeida et al.,2021).

O estudo de Pereira, Martínez, Agudelo (2018), observaram um bom desempenho do MLR em comparação com métodos baseados em interpolação espacial, embora as redes neurais artificiais tenham obtido resultados superiores, não houve diferença estatística relevante entre a rede neural, MLR e uma versão modificada do método de ponderação por distância inversa (IDWm), que incorporava a diferença de elevação entre as estações. Diante disso, é importante a consideração da variável altitude, especialmente em relevos mais acidentado. Assim, sugere-se para estudos futuros a avaliação do impacto da inclusão das variáveis topográficas, como a elevação.

CONCLUSÕES

Portanto, a análise realizada comprovou a eficácia do método *Multiple Linear Regression with Fixed Predictors* (MLR_Fix) na imputação de dados faltantes em séries temporais de precipitação na bacia hidrográfica do rio Meia Ponte. Diante disso, a escolha do método foi fundamentada em uma avaliação comparativa de desempenho, usando métricas estatísticas médias como RMSE, NSE e Índice de Concordância de Willmott, considerando os 21 postos observados.

O MLR_FIX apresentou os melhores resultados globais, tanto em termos numéricos quanto gráficos, como mostrado no Diagrama de Taylor gerados. Esse desempenho reforça a aplicabilidade de regressão linear múltipla com preditores fixos em cenários de dados diários faltantes. Com a aplicação do método, foi possível gerar séries de dados completas, preservando as qualidades hidrológicas das estações avaliadas, proporcionando dados adequados para futuros estudos hidrológicos.

REFERÊNCIAS

ALMEIDA, C. et al. Gap filling procedures of climatological series in the state of Pernambuco. *Irriga*, v. 1, n. 4, p. 754–764, 2021. DOI: https://doi.org/10.15809/irriga.2021v1n4p754-764. Acesso em: 23 jun. 2025.

BRUBACHER, João Paulo; OLIVEIRA, Guilherme Garcia de; GUASSELLI, Laurindo Antonio. Preenchimento de falhas em séries temporais de precipitação diária no Rio Grande do Sul. *Revista Brasileira de Meteorologia*, v. 35, n. 2, p. 261-274, 2020. DOI: https://doi.org/10.1590/0102-7786352035. Acesso em: 22 jun. 2025.

CALIL, P. M. et al. Caracterização geomorfométrica e do uso do solo da Bacia Hidrográfica do Alto Meia Ponte, Goiás. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 16, n. 4, p. 433–442, 2012. DOI: https://doi.org/10.1590/S1415-43662012000400014. Acesso em: 8 jun. 2025.

CHIU, Po Chan et al. Imputation of Rainfall Data Using the Sine Cosine Function Fitting Neural Network. *International Journal of Interactive Multimedia and Artificial Intelligence*, v. 6, n. 7, p. 39–48, 2021. DOI: https://doi.org/10.9781/ijimai.2021.08.013. Acesso em: 3 jun. 2025.

ELIAS, B. et al. Evaluation of seven gap-filling techniques for daily station-based rainfall datasets in South Ethiopia. *Advances in Meteorology*, 2021. DOI: https://doi.org/10.1155/2021/9657460. Acesso em: 23 jun. 2025.

FILHO, A. S. F.; LIMA, G. A. R. Gap filling of precipitation data by SSA – Singular Spectrum Analysis. *Journal of Physics: Conference Series*, v. 759, p. 012085, 2016. DOI: https://doi.org/10.1088/1742-6596/759/1/012085. Acesso em: 3 jun. 2025.

HAMZAH, F.; HAMZAH, M.; RAZALI, S. Multiple imputations by chained equations for recovering missing daily streamflow observations: a case study of Langat River basin in Malaysia. *Hydrological Sciences Journal*, v. 67, p. 137-149, 2021. DOI: https://doi.org/10.1080/02626667.2021.2001471. Acesso em: 23 jun. 2025.

HODSON, T. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, [S.l.], v. 15, n. 14, p. 5481–5489, 2022. Disponível em: https://doi.org/10.5194/gmd-15-5481-2022. Acesso em: 21 jun. 2025.

HU, Z. et al. Decompositions of Taylor diagram and DISO performance criteria. *International Journal of Climatology*, v. 41, p. 5726–5732, 2021. DOI: https://doi.org/10.1002/joc.7149. Acesso em: 23 jun. 2025.

IACOBELLIS, V. et al. A new diagram for performance evaluation of complex models. *Stochastic Environmental Research and Risk Assessment*, 2024. DOI: https://doi.org/10.1007/s00477-024-02678-3. Acesso em: 23 jun. 2025.

LI, L. Watershed reactive transport. *Reviews in Mineralogy and Geochemistry*, 2019. DOI: https://doi.org/10.2138/rmg.2018.85.13. Acesso em: 23 jun. 2025.

MATSUURA, K.; WILLMOTT, C. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, Oldendorf, v. 30, p. 79–82, 2005. Disponível em: https://doi.org/10.3354/CR030079. Acesso em: 21 jun. 2025.

MIRÓ, Juan Javier; CASELLES, Vicente; ESTRELA, María José. Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric Research*, v. 197, p. 313–330, 2017. DOI: https://doi.org/10.1016/j.atmosres.2017.07.016. Acesso em: 8 jun. 2025.

ORIAN, F. et al. Missing Data Imputation for Multisite Rainfall Networks: A Comparison between Geostatistical Interpolation and Pattern-Based Estimation on Different Terrain Types. *Journal of Hydrometeorology*, 2020. DOI: https://doi.org/10.1175/JHM-D-19-0220.1. Acesso em: 3 jun. 2025.

PEREIRA, Gabriel; MARTÍNEZ, Carlos; AGUDELO, Felipe. *Comparison of interpolation methods to estimate missing data in monthly precipitation series in south-central Chile*. Forest Ecosystems, v. 5, n. 1, p. 1–14, 2018. Disponível em: https://link.springer.com/article/10.1186/s40663-018-0147-x. Acesso em: 4 ago. 2025.

SARMENTO, L. Monitoramento hidrometeorológico no Brasil: uma análise sob a ótica da coordenação de políticas públicas. *Revista de Gestão de Água da América Latina*, v. 18, e3, 2021. DOI: https://doi.org/10.21168/rega.v18e3. Acesso em: 3 jun. 2025.

WANGWONGCHAI, Angkool et al. Imputation of missing daily rainfall data: a comparison between artificial intelligence and statistical techniques. *MethodsX*, v. 11, 102459, 2023. DOI: https://doi.org/10.1016/j.mex.2023.102459. Acesso em: 3 jun. 2025.

ZHAO, Q. et al. A novel online hydrological data quality control approach based on adaptive differential evolution. *Mathematics*, v. 12, n. 12, p. 1821, 2024. DOI: https://doi.org/10.3390/math12121821. Acesso em: 8 jun. 2025.

AGRADECIMENTOS

Os autores agradecem a Fundação de Apoio à Pesquisa – FUNAPE e ao Conselho Nacional de Desenvolvimento Cientifico e Tecnológico pelo apoio financeiro e a Universidade Federal de Goiás – UFG pelo programa de Iniciação Científica IC (2024 – 2025).