UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREENCHIMENTO DE FALHAS EM SÉRIES TELEMÉTRICAS DE NÍVEL

Josielton da Silva Santos [1] ; Milena Pereira Dantas[1]; Vitória Amélia Lemes Gonçalves[1]; Rafael Grinberg Chasles[1]; Helen Graciane Ruela Machado[1]; Brisa Maria Fregonesi[1]; Diego Monteiro[1]; Diego Freitas de Souza[1]; André Luis Navarro[1]; Anderson Barboza Esteves[1].

Abstract: Hydrological time series are essential for water resources management but often contain gaps that compromise their usability. This study evaluated the performance of machine learning (ML) models for gap-filling in river stage time series of the Tietê River, in São Paulo city, Brazil, using telemetric data from three streamflow monitoring stations located along a 29.5 km urbanized stretch. The “Ponte do Piqueri” station (296) presented approximately 13% missing data between 2011 and 2025, which were filled using records from the upstream “Ponte Dutra” station (232) and the downstream “Barragem Móvel Montante” station (1000847). The methodology combined linear regression models with stochastic gradient descent (SGDRegressor) and Random Forest, considering the criticality of water levels relative to warning thresholds. For remaining gaps, second-degree polynomial regression models were applied, adjusted based on the maximum temporal correlation between stations. The quality of the gap-filling process was evaluated using metrics such as R^2, MAE, MAPE, and RMSE. The results indicated that the use of machine learning is an efficient and robust approach for filling gaps in hydrological stage series, especially in complex urban environments prone to critical flood events.
Keywords: time series; machine learning; hydrological data.

Resumo: Séries temporais hidrológicas são essenciais para a gestão de recursos hídricos, mas frequentemente apresentam falhas que comprometem sua utilização. Este estudo avaliou o desempenho de modelos de aprendizado de máquina (ML) para o preenchimento de falhas em séries de nível do rio Tietê, na cidade de São Paulo - SP, utilizando dados telemétricos de três estações fluviométricas localizadas ao longo de um trecho urbanizado de 29,5 km. A estação "Ponte do Piqueri" (296) apresentou aproximadamente 13% de dados ausentes no período de 2011 a 2025, que foram preenchidos utilizando registros das estações a montante "Ponte Dutra" (232) e a jusante "Barragem Móvel Montante" (1000847). A metodologia combinou modelos de regressão linear por descida de gradiente estocástico (SGDRegressor) e Floresta Aleatória (Random Forest), considerando a criticidade dos níveis em relação às cotas de atenção. Para casos remanescentes, foram utilizados modelos de regressão polinomial de segundo grau, ajustados com base na máxima correlação temporal entre as estações. A qualidade do preenchimento foi avaliada por meio de métricas como R^2, MAE, MAPE e RMSE. Os resultados indicaram que o uso de aprendizado de máquina é uma abordagem eficiente e robusta para o preenchimento de falhas em séries hidrológicas de nível, especialmente em ambientes urbanos complexos e sujeitos a eventos críticos de cheia.
Palavras-Chave: séries temporais; aprendizado de máquina; dados hidrológicos.

[1]) SP-ÁGUAS – Agência de Águas do Estado de São Paulo. Rua Boa Vista – n° 175, 1° andar/ São Paulo – SP. jssantos@spaguas.sp.gov.br

INTRODUÇÃO

As séries temporais hidrológicas são fundamentais para a gestão de recursos hídricos e energéticos, bem como para a modelagem de impactos associados às mudanças climáticas (ARRIAGADA *et al*., 2021). No entanto, essas séries, quando disponíveis, frequentemente apresentam falhas e inconsistências, o que torna necessário o preenchimento dos valores ausentes (LEDRA *et al*., 2017).

A completude dos dados é essencial para aumentar a confiabilidade de diversos estudos meteorológicos e hidrológicos, incluindo análises de frequência de inundações, gestão de secas, abastecimento de água, avaliação de mudanças climáticas, previsão do tempo e planejamento da geração de energia hidrelétrica (KATIPOĞLU, 2023).

Nas últimas décadas, a evolução das técnicas de interpolação e de preenchimento de lacunas foi impulsionada pelos avanços na capacidade computacional e tecnológica. Esse progresso potencializou diversos métodos, ampliando o volume de dados que podem ser processados e utilizados nas análises (BRUBACHER *et al*., 2020).

Mais recentemente, técnicas baseadas em inteligência artificial vêm ganhando destaque como ferramentas promissoras para o preenchimento de falhas em dados meteorológicos (KATIPOĞLU, 2023). Em particular, algoritmos de aprendizado de máquina (*machine learning*) têm sido cada vez mais empregados para preencher séries de dados de vazão diária ausentes, utilizando informações de estações vizinhas, especialmente em regiões caracterizadas pela escassez de dados hidrológicos (ZHOU *et al*., 2023).

Neste contexto, o presente estudo tem como objetivo avaliar o desempenho de modelos de aprendizado de máquina no preenchimento de falhas em uma série temporal de dados de nível, utilizando registros provenientes de uma estação a montante e de outra a jusante, localizadas no mesmo curso d'água.

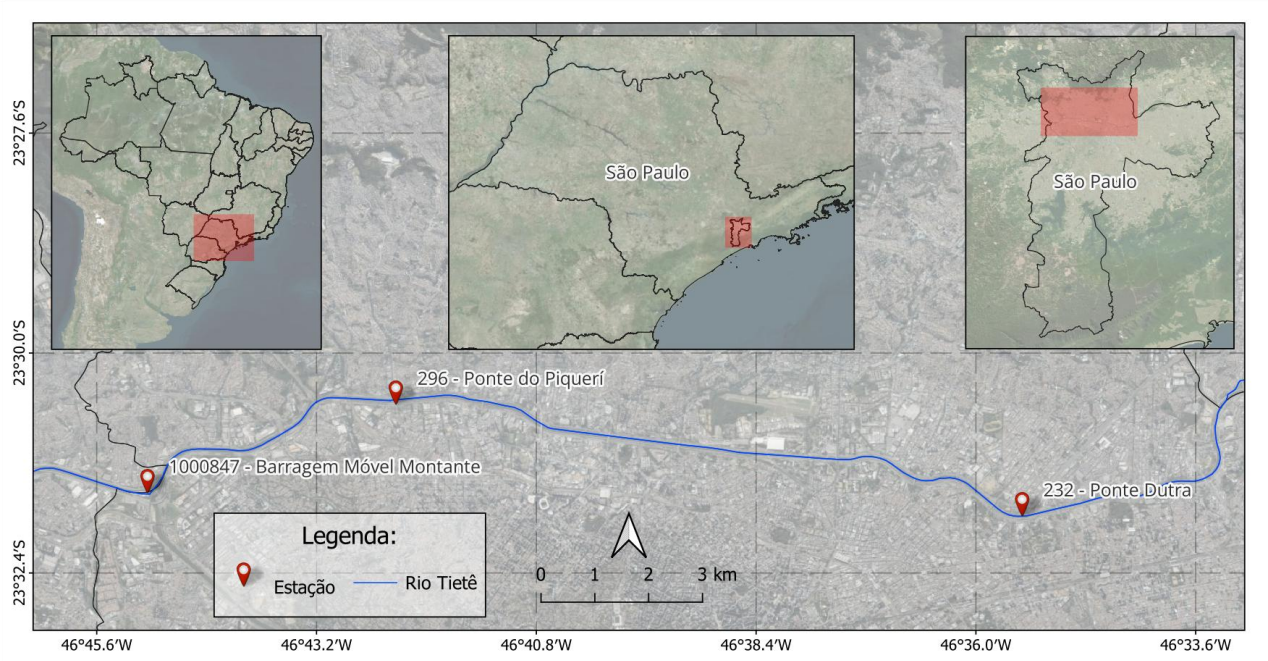
METODOLOGIA

Área de Estudo

A área em estudo do presente artigo corresponde ao trecho do rio Tietê que passa pelo município de São Paulo (SP). Trata-se de um trecho de 29,5 km em que o rio está totalmente urbanizado, com as margens concretadas, e sua área correspondente à Área de Preservação Permanente (APP) parcialmente tomada com vias marginais para o tráfego de veículos (MIRANDA *et al*., 2011).

A Figura 1 destaca três estações fluviométricas de medição de nível no rio, localizadas ao longo da Marginal Tietê. Essas estações são essenciais para o monitoramento contínuo do nível do rio e para a gestão de barramentos e reservatórios, especialmente em situações de cheias, a fim de evitar o transbordamento sobre a via. As cotas de atenção para eventos de cheia nessas estações estão apresentadas na Tabela 1.

Figura 1 – Área de Estudo



As cotas de atenção utilizadas neste estudo estão apresentadas na Tabela 1.

Tabela 1: Cota de atenção

Estação	Cota de Atenção (m)
296 - Rio Tietê - Ponte do Piqueri	717,980
232 - Rio Tietê - Ponte Dutra	719,409
1000847 - Rio Tietê Barragem Móvel Montante (BMM)	716,980

Fonte: Sistema Integrado de Bacias Hidrográficas (SIBH)

Dados

Os dados utilizados neste trabalho estão disponíveis no Sistema Integrado de Bacias Hidrográficas (SIBH) da Agência de Águas do Estado de São Paulo (SP-ÁGUAS), o qual disponibiliza produtos com informações de chuva, nível e vazão do estado de São Paulo, utilizando postos telemétricos e convencionais de diversas entidades. Estes dados estão disponíveis no endereço: https://cth.daee.sp.gov.br/sibh/.

As séries analisadas são provenientes de estações fluviométricas telemétricas com dados de nível para um período de 01/01/2011 até 02/05/2025 em intervalos de 10 minutos apresentadas na Tabela 2.

Os dados das estações Dutra (232) e BMM (1000847) foram utilizados para auxiliar no preenchimento das falhas da estação Piqueri (296).

Tabela 2: Estações utilizadas

Código	Nome	Descrição
296	Ponte do Piqueri	Estação em que os dados foram preenchidos
232	Rio Tietê - Ponte Dutra	Estação à montante
1000847	Rio Tietê Barragem Móvel Montante (BMM)	Estação à jusante

Desenvolvimento

A metodologia adotada para o preenchimento de falhas teve início com a identificação das falhas. A série da estação 296 foi carregada e processada em ambiente *Python*, onde os dados espúrios foram removidos. Durante a análise, foram identificados 95.572 valores ausentes (NaN), correspondendo a aproximadamente 13% do total da série.

O preenchimento dessas falhas foi realizado em duas etapas. Primeiro optou-se pelo desenvolvimento de modelo baseado em algoritmo de *Machine Learning* (ML) para preencher os dados ausentes nos períodos em que havia registros simultâneos nas estações Dutra e BMM. Posteriormente foram aplicados modelos de regressão polinomial para os casos em que os dados estavam disponíveis ou apenas na estação Dutra ou na BMM.

Desenvolvimento de modelo de *Machine Learning*

Inicialmente, todas as amostras com valores ausentes em pelo menos uma das três séries foram removidas. A variável de saída considerada foi o nível da estação Piqueri, enquanto as variáveis de entrada corresponderam às medições das estações Dutra e Barragem Móvel Montante (BMM). O conjunto de dados foi dividido em dois subconjuntos: 70% para treinamento e 30% para teste, assegurando uma avaliação adequada dos modelos. Antes do treinamento, todas as variáveis foram normalizadas, com o objetivo de melhorar o desempenho e a estabilidade dos algoritmos.

Dois modelos de aprendizado de máquina foram empregados, considerando a criticidade associada aos níveis observados. O primeiro modelo foi uma regressão linear utilizando o algoritmo de Regressão Linear por Descida de Gradiente Estocástico (SGDRegressor). É um algoritmo de otimização que atualiza os parâmetros de um modelo linear de forma incremental, utilizando gradientes aproximados calculados a partir de subconjuntos dos dados de treinamento (TSURUOKA et al,. 2009). Após a validação do modelo, os valores ausentes na estação Piqueri foram preenchidos, desde que as estações Dutra e BMM apresentassem dados disponíveis no mesmo instante. Adicionalmente, as falhas de curta duração (inferiores ou iguais a 1 hora) nas três séries foram preenchidas com a média dos valores imediatamente anteriores e subsequentes à ausência.

De forma complementar, para melhorar o desempenho do preenchimento em situações críticas, foi implementado um segundo modelo, baseado no algoritmo Floresta Aleatória – RF (*Random Forest)*. Trata-se de um algoritmo que combina múltiplas árvores de decisão, cada uma construída a partir de amostras aleatórias dos dados, para melhorar a precisão e reduzir o erro de generalização à medida que o número de árvores aumenta (BREIMAN, 2001). Este modelo foi treinado especificamente para os casos em que o nível da estação Piqueri era maior ou igual à respectiva cota de atenção (Tabela 1). Assim, o modelo RF foi aplicado sempre que pelo menos uma das estações auxiliares (Dutra ou BMM) apresentava nível igual ou superior à sua cota de

atenção. Após a aplicação dos dois modelos (SGD e RF), os dados que ainda não haviam sido preenchidos foram completados utilizando o modelo de Regressão Polinomial.

Desenvolvimento de modelo de *Regressão Polinomial*

Após o preenchimento inicial da série da estação Piqueri com modelos de aprendizado de máquina, identificou-se o maior período contínuo sem falhas para os pares (Dutra - Piqueri) e (BMM - Piqueri). Esse procedimento visou determinar o melhor deslocamento temporal (*lag*) entre as variáveis, por meio da correlação de Pearson (PEARSON, 1895).

Nos maiores intervalos contínuos sem falhas, foi realizada uma análise de *lag* de até 180 minutos, com incrementos de 10 minutos, para identificar o deslocamento que maximizasse a correlação entre as estações. Com os *lags* definidos, foram ajustados dois modelos de regressão polinomial de segundo grau: um entre Dutra e Piqueri e outro entre BMM e Piqueri. Esses modelos permitiram o preenchimento de Piqueri nos períodos em que apenas uma das estações auxiliares possuía dados disponíveis. Por fim, as lacunas foram preenchidas com a média dos valores imediatamente anterior e posterior. Com a série toda preenchida, valores espúrios foram removidos e substituídos pela média da série.

Estas modelagens foram construídas com a biblioteca *Scikit-learn* (Pedregosa et al., 2011) em ambiente de programação *python*.

Para avaliação dos modelos utilizou-se das métricas Coeficiente de Determinação (R^2), Erro Absoluto Médio (MAE), Erro Percentual Médio Absoluto (MAPE) e Raiz do Erro Quadrático Médio (RMSE) (HYNDMAN e KOEHLER, 2006). Calculados com as Equações 1, 2, 3 e 4, respectivamente.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$MAE\,(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$MAPE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{max(\epsilon, |y_i|)} \quad (3)$$

$$RMSE\,(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

Onde:

y_i: dado real;

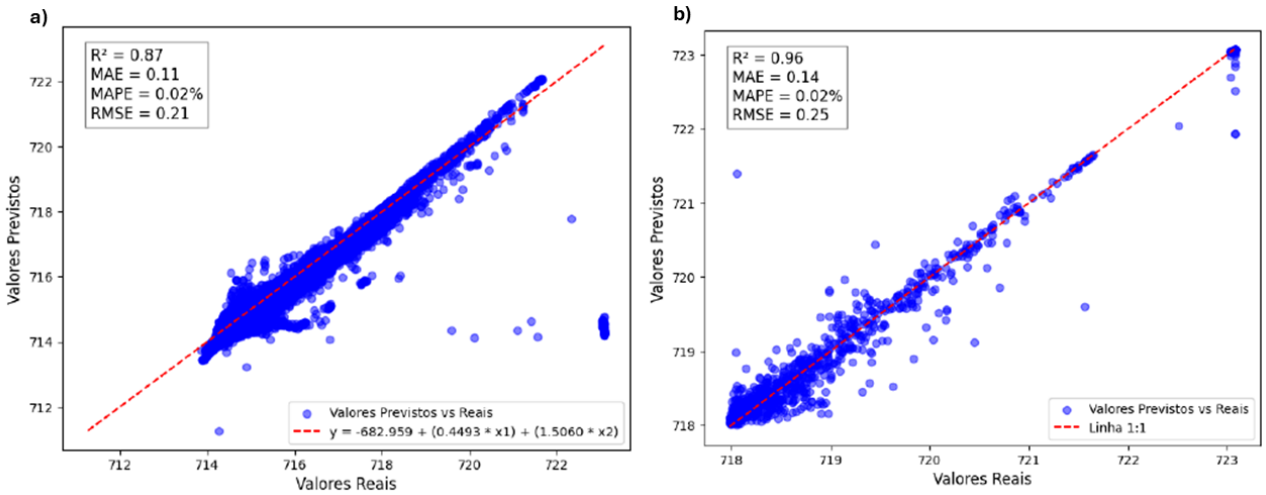
$\hat{y}_i$: dado estimado pelo modelo e

$\bar{y}$: média dos dados reais.

RESULTADOS E DISCUSSÃO

Para avaliar o desempenho dos modelos, foram elaborados dois gráficos de dispersão, que permitiram a comparação dos valores reais com os previstos para os dois modelos utilizados: *SGDRegressor* (Figura 2a) e *Random Forest* (Figura 2b), acompanhados da reta 1:1 como referência. Além disso, as métricas de desempenho e a equação ajustada foram adicionadas.

Figura 2 – a) resultados do modelo *SGDRegressor* e b) resultados do modelo *Random Forest*



Observa-se que, no modelo SGD, houve uma tendência à subestimação, onde os valores previstos concentraram-se próximos de 714 cm, enquanto os valores reais superavam 720 cm. Em contrapartida, esse comportamento não foi observado na aplicação do modelo *Random Forest*. Ambos os modelos apresentaram resultados satisfatórios nas métricas avaliadas; contudo, apesar do *Random Forest* apresentar um MAE maior, a amplitude do erro foi menor, conforme ilustrado na Figura 3. Esse comportamento indica maior aderência do modelo RF para estimativa de valores em eventos de cheia.

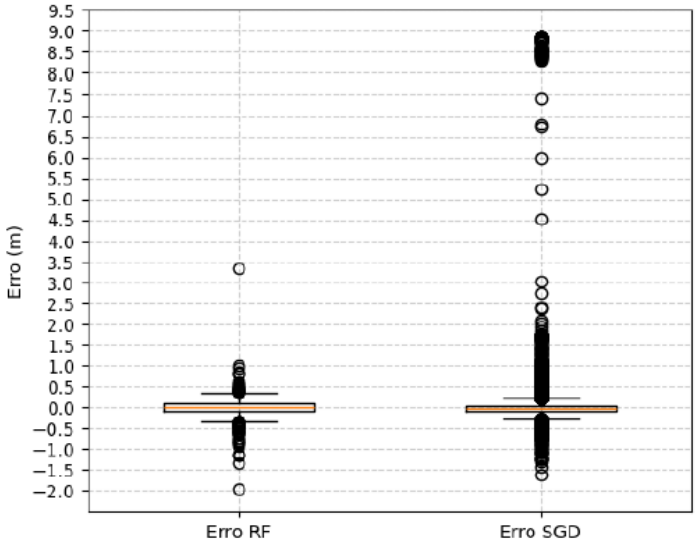
A limitação observada na utilização do SGD é corroborada pelo estudo de Moura et al. (2017), os autores concluíram que o gradiente descendente é eficaz para preencher falhas em séries de vazão, porém com melhor desempenho em séries mensais. Ou seja, em uma escala temporal onde os picos de cheia são suavizados.

Para a construção dos modelos de regressão polinomial, os resultados obtidos na definição do melhor tempo de resposta (lag) são apresentados na Tabela 2. Para ambas as estações, as correlações foram fortes e próximas de 1 (correlação perfeita), com um tempo de resposta de 10 minutos em relação à estação à Dutra e resposta imediata em relação à BMM.

Tabela 2: Melhores *lags* e correlações

Estação	Período mais longo sem falhas	Melhor *lag*	Correlação
Dutra	2012-11-28 15:50:00 2013-06-17 15:00:00	10 min	0,94
BMM	2011-12-26 13:50:00 2012-09-27 13:10:00	0 min	0,95

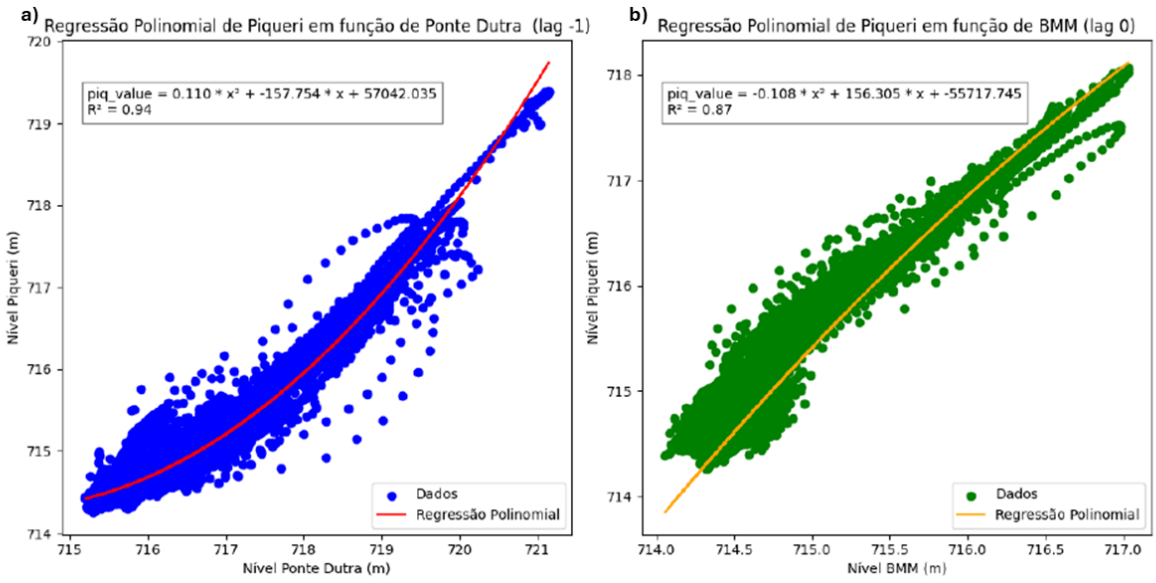
Figura 3 – Comparação entre erros



Para visualizar os resultados, foram gerados dois gráficos de dispersão (Figura 4), nos quais os pontos representam os dados originais e as curvas ajustadas correspondem às previsões dos modelos polinomiais. As equações e os valores de R^2 estão incluídos nos gráficos.

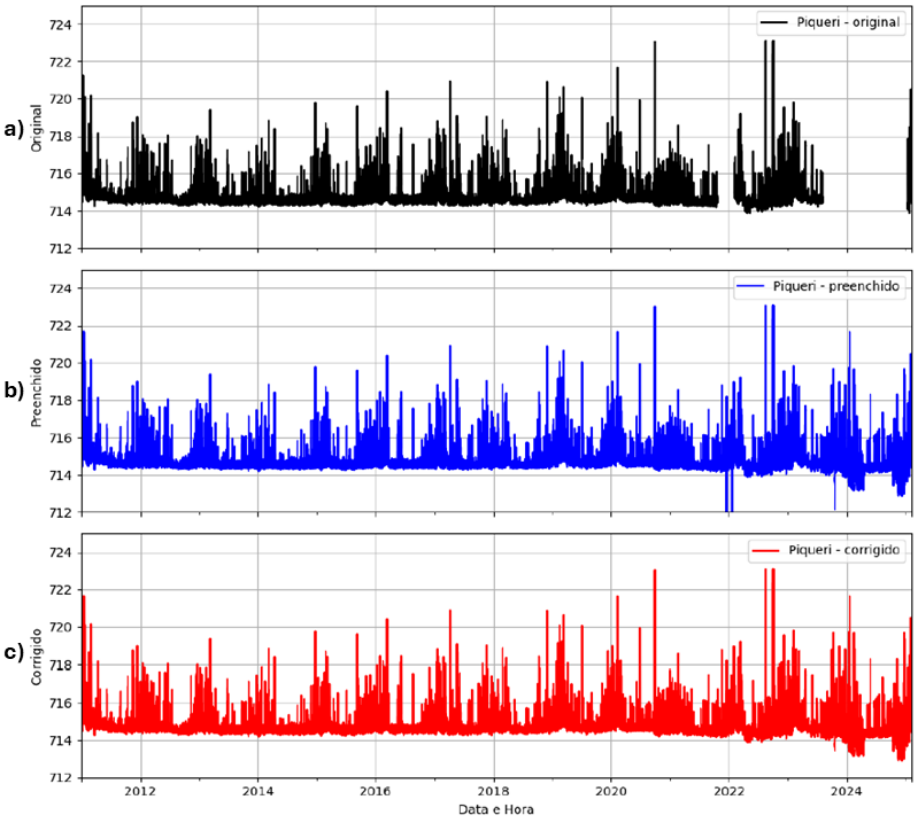
Nota-se que em ambos os modelos o valor de R^2 é relativamente alto: 0,94 para a estação Dutra e 0,87 para estação BMM. Sugerindo bom desempenho na utilização. Mesmo após validação da metodologia e redução de suas limitações, de acordo com Tamiosso et al. (2013) recomenda-se cautela em sua utilização, uma vez que esta é uma equação matemática que busca descrever um fenômeno aleatório e aplicada a uma estação específica.

Figura 4 – Modelos de regressão polinomial



A comparação entre as séries temporais de Piqueri: “Original” (com as falhas), “Preenchido” (após preenchimento gerado pelos modelos) e “Corrigido” (após a remoção dos dados espúrios) é apresentada na Figura 5.

5 – Comparação entre as séries temporais



Os valores atipicamente baixos observados na série em 2024 podem ser justificados pela operação de bombeamento realizada pelo Governo do Estado de São Paulo, naquele ano. A ação promoveu a circulação de água do canal inferior para o canal superior do rio Pinheiros, com o objetivo de reduzir a proliferação de algas, em um contexto de forte estiagem (São Paulo, 2024a, 2024b). Esse processo alterou a dinâmica hidrológica do rio, especialmente no trecho próximo à confluência com o Tietê, impactando diretamente os níveis registrados na estação Piqueri. O estudo de Zhou et al. (2023) constatou que o desempenho do modelo de *machine learning* utilizado para o preenchimento de lacunas em dados diários de vazão foi degradado devido à influência da operação do reservatório e da implementação de um projeto de desvio de água.

A Tabela 3 apresenta uma estatística descritiva básica da série temporal de Piqueri para todas as versões dos dados. Na série temporal preenchida, os dados ausentes foram completados, resultando em um menor valor mínimo (711,34 m). Já na série corrigida, o valor mínimo foi elevado para 712,87 m, chegando mais próximo do valor mínimo original. A média e o desvio padrão permaneceram praticamente inalterados entre as versões.

Tabela 3: Estatísticas básicas da série temporal de Piqueri nas versões original, preenchida e corrigida.

	Média (m)	Desvio Padrão (m)	Mínimo (m)	Máximo (m)
Original	714,75	0,56	713,87	723,08
Preenchido	714,72	0,59	711,34	723,08
Corrigido	714,72	0,59	712,87	723,08

CONCLUSÃO

O método empregado possibilitou a utilização de algoritmos de aprendizado de máquina (*machine learning*) para a realização do preenchimento das falhas na série temporal da estação Piqueri, preservando suas características estatísticas e tornando os dados adequados para análises subsequentes.

Por outro lado, algumas particularidades precisam ser observadas em pesquisas futuras. É o caso da interferência operacional, na medida que um conjunto de valores modelados a partir do ano de 2024 apresentou níveis significativamente abaixo da média histórica. Esse comportamento está relacionado à operação do sistema de bombas próximos a estação 1000847 – Rio Tietê Barragem Móvel Montante, que foi utilizada como auxiliar para o preenchimento de falhas.

O uso de modelagem com apoio de técnicas de *machine learning* é uma ferramenta viável para lidar com grandes volumes de dados de nível, tornando o preenchimento de falhas essencial para a elaboração de curvas-chave e o desenvolvimento de sistemas de monitoramento e previsão de cheias e secas, contribuindo significativamente para a gestão de recursos hídricos.

REFERÊNCIAS BIBLIOGRÁFICAS

ARRIAGADA, P.; KARELOVIC, B.; LINK, O. Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. Journal of Hydrology, v. 598, 2021, 126454. Disponível em: https://doi.org/10.1016/j.jhydrol.2021.126454. Acesso em: 27 maio 2025.

BREIMAN, L. Random Forests. Machine Learning, v. 45, p. 5–32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: 27 maio 2025.

BRUBACHER, J. P.; OLIVEIRA, G. G. de; GUASSELLI, L. A. Preenchimento de Falhas e Espacialização de Dados Pluviométricos: desafios e perspectivas. Revista Brasileira de Meteorologia, v. 35, n. 4, p. 615-629, dez. 2020. Disponível em: http://dx.doi.org/10.1590/0102-77863540067. Acesso em: 27 maio 2025.

HYNDMAN, R. J.; KOEHLER, A. B.. Another look at measures of forecast accuracy. International Journal of Forecasting, [s. l.], n. 22, issue 4, p. 679-688, Oct.–Dec. 2006. Disponível em: https://doi.org/10.1016/j.ijforecast.2006.03.001. Acesso em: 09 maio 2025.

KATIPOĞLU, O. M. Evaluation of the performance of data-driven approaches for filling monthly precipitation gaps in a semi-arid climate conditions. Acta Geophysica, v. 71, p. 2265–2285, 2023. Disponível em: https://doi.org/10.1007/s11600-022-00963-9. Acesso em: 27 maio 2025.

LAIS, L.; BARROS, V. G.; HENNING, E. Preenchimento de falhas de dados diários de precipitação utilizando redes neurais artificiais e método das correlações. In: XXII Simpósio

Brasileiro de Recursos Hídricos. 2017. Disponível em: https://anais.abrhidro.org.br/job.php?Job=3045. Acesso em: 27 maio 2025.

MIRANDA, R. B. de; SCARPINELLA, G. D. de A.; GOUVEA, T. H.; MAUAD, F. F. Rio Tietê: Iniciativas governamentais para revitalização do trecho urbano no município de São Paulo (SP). In: XIX Simpósio Brasileiro de Recursos Hídricos. 2011. SBRH04108. Disponível em: https://anais.abrhidro.org.br/job.php?Job=11525. Acesso em: 28 maio 2025.

MOURA, C. N. de; SÁ, É. A. S.; PADILHA, V. L.; NETO, S. L. R. Desempenho do machine learning para o preenchimento de falhas em séries de vazões diárias e mensais. In: XXII Simpósio Brasileiro de Recursos Hídricos. Florianópolis - SC, 2017. Disponível em: https://files.abrhidro.org.br/Eventos/Trabalhos/60/PAP022405.pdf. Acesso em: 27 maio 2025.

PEARSON, K. (1895). Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58, 240–242.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011. Disponível em: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Acesso em: 27 maio 2025.

SÃO PAULO. Agência de Notícias do Governo do Estado de São Paulo. Governo de SP e Emae realizam operação para aumentar a circulação da água no Rio Pinheiros. 2024b. Disponível em: https://www.agenciasp.sp.gov.br/governo-de-sp-e-emae-realizam-operacao-para-aumentar-a-circulacao-da-agua-no-rio-pinheiros/. Acesso em: 27 maio 2025.

SÃO PAULO. Secretaria de Meio Ambiente, Infraestrutura e Logística. Governo de SP e Emae farão nova operação de bombeamento no rio Pinheiros. 2024a. Disponível em: https://semil.sp.gov.br/2024/09/governo-de-sp-e-emae-farao-nova-operacao-de-bombeamento-no-rio-pinheiros/. Acesso em: 27 maio 2025.

TAMIOSSO, M. F.; TAMIOSSO, C. F.; ARAÚJO, R. K. de et al. Preenchimento de falhas de dados observados de vazão utilizando a equação de manning. In: XX Simpósio Brasileiro de Recursos Hídricos. Bento Gonçalves – RS, 2013. Disponível em: https://anais.abrhidro.org.br/job.php?Job=1802. Acesso em: 28 maio 2025.

TSURUOKA, Y.; TSUJII, J.; ANANIADOU, S. Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Suntec, Singapore, 2009. p. 477–485. Disponível em: https://aclanthology.org/P09-1054.pdf. Acesso em: 27 maio 2025.

ZHOU, Y.; TANG, Q.; ZHAO, G. Gap infilling of daily streamflow data using a machine learning algorithm (MissForest) for impact assessment of human activities. Journal of Hydrology, v. 627, parte A, 2023, 130404. Disponível em: https://doi.org/10.1016/j.jhydrol.2023.130404. Acesso em: 27 maio 2025.

AGRADECIMENTOS

Os autores agradecem à SP-ÁGUAS pelo fornecimento das diretrizes que orientaram esta análise e ao consórcio HCC3 pelo apoio ao longo do desenvolvimento desta pesquisa.