UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA CIVIL

PEDRO AUGUSTO TOLEDO RIOS

MODELOS DE PREENCHIMENTO DE FALHAS DE DADOS PLUVIOMÉTRICOS DIÁRIOS USANDO APRENDIZADO DE MÁQUINA

Uberlândia

2024

PEDRO AUGUSTO TOLEDO RIOS

MODELOS DE PREENCHIMENTO DE FALHAS DE DADOS PLUVIOMÉTRICOS DIÁRIOS USANDO APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado à Universidade Federal de Uberlândia, curso de Engenharia Civil, como requisito parcial para a obtenção do título de Bacharel em Engenharia Civil.

Orientador: Prof. Dr. Carlos Eugênio Pereira
Coorientador: Prof. Dr. Rodrigo Sanches Miani

Uberlândia
2024

RESUMO

A reconstrução de séries temporais pluviométricas incompletas é crucial tanto para a Climatologia e Hidrologia, quanto para a Engenharia Civil. A qualidade e integralidade desses dados impactam diretamente a acurácia de previsões meteorológicas e investigações hidrológicas, bem como o dimensionamento adequado de infraestruturas hídricas, como reservatórios, barragens e sistemas de drenagem urbana, cuja precisão depende da confiabilidade dos dados pluviométricos. Neste trabalho, investiga-se a aplicação de modelos de aprendizado de máquina, como KNN Imputer e Iterative Imputer, para a imputação de dados pluviométricos faltantes, comparando-os com métodos tradicionais, utilizando dados que abrangem o período de 2002 a 2019, coletados na estação de Itirapina-CRHEA, e o intervalo de 1979 a 2023, fornecidos pelo Instituto Nacional de Meteorologia (INMET) de São Carlos. Para tanto, foram empregadas três abordagens distintas: uma análise global, considerando a totalidade do conjunto de dados, e análises separadas para os períodos chuvoso e seco, a fim de avaliar o desempenho dos métodos sob diferentes regimes sazonais, para validar os modelos, retirou-se porcentagens dos dados (2,5%, 5,0%, 7,5% e 10,0%), simulando falhas, e os valores imputados foram comparados aos originais, além de ser realizada uma comparação detalhada entre os modelos de aprendizado de máquina e técnicas tradicionais. Os resultados demonstram que os modelos de aprendizado de máquina apresentaram desempenho superior aos métodos convencionais em termos de precisão e manutenção da consistência dos dados, mostrando-se eficazes na melhoria da qualidade dos dados e no aumento da confiabilidade das previsões, e evidenciando seu potencial para aplicação em estudos climáticos mais amplos. Para pesquisas futuras, sugere-se a aplicação dos métodos em diferentes regiões geográficas, biomas e regimes climáticos, o que permitirá a avaliação da adaptabilidade e eficiência dos métodos propostos em diversos contextos, contribuindo para a gestão de recursos hídricos e o planejamento climático de longo prazo no Brasil.

Palavras-chave: Séries Temporais Pluviométricas; Imputação de Dados; Aprendizado de Máquina; KNN Imputer; Iterative Imputer; Engenharia Civil; Recursos Hídricos; Climatologia; Hidrologia.

ABSTRACT

The reconstruction of incomplete rainfall time series is crucial for both Climatology/Hydrology and Civil Engineering. Data quality and integrity directly impact the accuracy of weather forecasts and hydrological investigations, as well as the proper design of water infrastructure such as reservoirs, dams, and urban drainage systems, whose precision relies on the reliability of rainfall data. This study investigates the application of machine learning models, namely KNN Imputer and Iterative Imputer, for imputing missing rainfall data, comparing them to traditional methods. Datasets covering the periods from 2002 to 2019 (collected at the Itirapina-CRHEA station) and 1979 to 2023 (provided by the Brazilian National Institute of Meteorology (INMET) in São Carlos) were used. Three distinct approaches were employed: a global analysis considering the entire dataset, and separate analyses for the wet and dry seasons to evaluate the methods' performance under different seasonal regimes. To validate the models, percentages of the data (2.5%, 5.0%, 7.5%, and 10.0%) were removed, simulating missing values, and the imputed values were compared to the originals. A detailed comparison between the machine learning models and traditional techniques was also performed. The results demonstrate that the machine learning models outperformed conventional methods in terms of accuracy and maintaining data consistency, proving effective in improving data quality and increasing the reliability of predictions, and highlighting their potential for application in broader climate studies. For future research, the application of these methods in different geographical regions, biomes, and climate regimes is suggested, which will allow for the evaluation of the adaptability and efficiency of the proposed methods in diverse contexts, contributing to water resource management and long-term climate planning in Brazil.

Keywords: Rainfall Time Series; Data Imputation; Machine Learning; KNN Imputer; Iterative Imputer; Civil Engineering; Water Resources; Climatology; Hydrology.

Lista de Ilustrações

Lista de Tabelas

SUMÁRIO

1. INTRODUÇÃO

A precipitação, componente intrínseco ao ciclo hidrológico, desempenha um papel fundamental em uma ampla gama de setores, influenciando não apenas a agricultura e o abastecimento hídrico, mas também a gestão de riscos associados a desastres naturais e as estratégias de planejamento climático. A precisão e a completude dos dados pluviométricos são, portanto, essenciais para a construção de modelos hidrológicos robustos, para previsões meteorológicas confiáveis e para a realização de estudos aprofundados sobre as mudanças climáticas. Entretanto, a ocorrência de falhas nesses dados, decorrentes de problemas técnicos em estações de medição, erros de registro ou limitações inerentes à própria coleta, configura um desafio persistente, conforme destacado por Sanches et al. (2020). Tais falhas comprometem a análise de séries temporais, introduzindo dificuldades na identificação de padrões sazonais e interanuais, além de reduzirem a confiabilidade das análises e a precisão das previsões, especialmente no que tange a eventos extremos, como secas e inundações.

Visando mitigar as dificuldades impostas pela incompletude dos dados pluviométricos, diversas estratégias têm sido desenvolvidas, abrangendo desde métodos tradicionais, como discutido por Horta et al. (2021), até técnicas mais avançadas, baseadas em aprendizado de máquina. Métodos tradicionais, como a interpolação linear e o cálculo de médias móveis, são frequentemente empregados devido à sua relativa simplicidade e eficiência computacional. No entanto, a aplicação desses métodos frequentemente se mostra insuficiente para capturar a complexidade e a não linearidade que caracterizam os padrões de precipitação, resultando em estimativas potencialmente imprecisas, particularmente em situações de eventos extremos (Blake, 2011; Latif et al., 2023).

Nesse cenário, o aprendizado de máquina surge como uma alternativa promissora, disponibilizando algoritmos com capacidade para modelar padrões complexos, lidar com grandes volumes de dados e capturar relações não lineares. Algoritmos como KNN Imputer, Iterative Imputer se destacam na imputação de falhas em séries temporais, exibindo, em estudos como os de Li et al. (2023) e Saad et al. (2020), maior precisão em comparação com métodos tradicionais. A escolha do algoritmo mais adequado, contudo, deve ser realizada considerando as características específicas dos dados, as particularidades da região investigada, e as discussões presentes em trabalhos como os de Stephenson (2002), Rusticuci e Tencer (2008), Tucci (2001), e Barry e Chorley (1998). Adicionalmente, a relevância dos

dados pluviométricos para estudos ambientais, como o monitoramento de queimadas em regiões tropicais, é demonstrada por Bosch et al. (2003).

Diante disso, este trabalho tem como objetivo investigar a aplicação de modelos de aprendizado de máquina, como. KNN Imputer e Iterative Imputer, para o preenchimento de falhas em dados pluviométricos da estação P16 (Itirapina-CRHEA, 2002-2019) e do INMET de São Carlos (1979-2023). A análise compreenderá diferentes cenários, incluindo um modelo global e modelos específicos para os períodos chuvoso e seco, buscando avaliar o desempenho dos métodos em diferentes condições sazonais e compará-los com métodos tradicionais de imputação. Os resultados obtidos contribuirão para o aprimoramento da qualidade dos dados pluviométricos, com implicações significativas para a gestão eficiente de recursos hídricos e para o desenvolvimento de estratégias de planejamento climático mais robustas no Brasil.

2. OBJETIVOS

Neste capítulo, serão apresentados os objetivos gerais e específicos do trabalho, que norteiam a pesquisa e definem o escopo das análises realizadas. O objetivo geral representa o direcionamento principal da investigação, enquanto os objetivos específicos descrevem as etapas e abordagens adotadas para atingir o resultado final esperado.

2.1 Objetivo Geral

O objetivo geral desta pesquisa é preencher dados faltantes em estações pluviométricas por meio da aplicação e comparação de diferentes modelos de aprendizado de máquina, utilizando dados coletados na estação de Itirapina/SP (2002-2019) e pelo INMET de São Carlos (1979-2023).

2.2 Objetivos Específicos

Os objetivos específicos desta pesquisa são:

- Identificar e avaliar a eficácia de métodos tradicionais e de aprendizado de máquina na imputação de falhas em séries temporais de dados pluviométricos.
- Aplicar o uso de algoritmos como KNN Imputer, Iterative Imputer e outras abordagens de aprendizado de máquina no preenchimento de dados meteorológicos incompletos.
- Comparar os resultados obtidos pelos métodos de aprendizado de máquina com os métodos tradicionais, como interpolação e regressão simples, para verificar a robustez dos modelos propostos.

3. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma revisão da literatura sobre os modelos de preenchimento de falhas em dados pluviométricos, com ênfase no uso de algoritmos de aprendizado de máquina aplicados a séries temporais. A fundamentação teórica abrange tanto as técnicas clássicas quanto os modelos mais recentes de imputação de dados, buscando fornecer uma base sólida para o desenvolvimento deste trabalho. Esses conceitos são fundamentais para a análise comparativa dos métodos de preenchimento de falhas propostos e para demonstrar a relevância das abordagens de aprendizado de máquina no contexto dos dados pluviométricos.

3.1 Impacto das Falhas nos Dados

A presença de falhas nos dados pluviométricos compromete diretamente a análise de padrões sazonais e interanuais, impactando também a performance de modelos preditivos. Essas falhas ocorrem, principalmente, devido a problemas técnicos nas estações meteorológicas, que podem apresentar períodos de inatividade por questões de manutenção ou falhas de comunicação (SANCHES et al., 2020). Quando dados ausentes não são adequadamente preenchidos, modelos preditivos baseados em séries temporais têm seu desempenho prejudicado, resultando em previsões imprecisas ou distorcidas, especialmente para eventos climáticos extremos, como enchentes ou secas (BLAKE, 2011). Bier e Ferraz (2017) compararam diversas metodologias de preenchimento de falhas em dados meteorológicos no sul do Brasil, destacando que a escolha do método adequado pode influenciar diretamente a qualidade das previsões climáticas.

Estudos anteriores propõem diferentes soluções para o preenchimento de falhas de dados. BáRdossy e Pegram (2014) desenvolveram um método baseado em copula para preencher registros de precipitação ausentes, comparando-o com outras técnicas e constatando sua superioridade em termos de precisão. El Hachimi et al. (2023), em seu framework *ClimateFiller*, aplicaram inteligência artificial e dados de reanálise para diagnosticar e preencher lacunas em séries temporais climáticas, mostrando-se uma abordagem promissora para a correção de falhas. Da mesma forma, Afrifa-Yamoah et al. (2020) investigaram a imputação de dados ausentes em séries temporais climáticas de alta resolução, demonstrando que métodos avançados são capazes de aumentar a precisão das previsões meteorológicas.

Outros trabalhos enfatizam abordagens baseadas em modelos preditivos e métodos tradicionais. Barrios Alonso et al. (2018) exploraram abordagens alternativas para estimar dados climáticos ausentes, aplicando-as a registros mensais de precipitação no centro-sul do Chile, o que demonstrou a eficiência de métodos específicos para regiões com características climáticas distintas. Hasanpour Kashani e Dinpashoh (2012) avaliaram a eficiência de diferentes métodos de estimativa para dados climatológicos ausentes, concluindo que técnicas adequadas de imputação podem melhorar significativamente a acurácia de modelos preditivos.

No contexto do Brasil, Collischonn et al. (2007) analisaram o desempenho do satélite TRMM para a estimativa de precipitação sobre a bacia do Paraguai Superior, destacando os desafios de precisão em regiões específicas. Além disso, Dornelles, Goldenfum e Pedrollo (2013) demonstraram a eficácia do uso de redes neurais artificiais na previsão de níveis de rios, aplicando-as em previsões hidrológicas.

A literatura evidencia a importância de utilizar técnicas robustas e específicas para o preenchimento de falhas em séries temporais, de modo a evitar distorções que possam comprometer as previsões climáticas. A interpolação ou imputação de dados torna-se essencial para manter a continuidade dos dados e garantir a qualidade das previsões, especialmente em eventos extremos, como secas e enchentes (MACHADO; ASSIS, 2018).

3.2 Métodos de Preenchimento de Falhas em Séries Temporais

Os métodos tradicionais utilizados para lidar com falhas em séries temporais, como interpolação linear e médias móveis, são amplamente empregados devido à sua simplicidade. No entanto, esses métodos apresentam limitações significativas, especialmente em dados complexos e não lineares, como os pluviométricos (LATIF et al., 2023). A interpolação linear, por exemplo, assume que os dados ausentes seguem um comportamento linear entre os pontos conhecidos, o que muitas vezes não reflete a realidade dos fenômenos climáticos (BLAKE, 2011).

Esses métodos tradicionais falham em capturar variações abruptas e extremos climáticos, resultando em previsões que podem ser imprecisas para eventos como precipitação intensa ou períodos prolongados de seca. Como argumentado por Bier e Ferraz (2017), a interpolação ou uso de médias tende a suavizar os dados, o que pode eliminar informações

importantes sobre variações extremas, essenciais para a previsão de fenômenos críticos. Para superar essas limitações, métodos mais avançados de imputação, baseados em aprendizado de máquina, começaram a ser utilizados, fornecendo uma maneira mais robusta de preencher falhas em séries temporais (SANCHES et al., 2020).

3.3 Trabalhos Correlatos

Essa seção da fundamentação teórica apresenta trabalhos que abordam métodos de preenchimento de falhas em séries temporais de dados meteorológicos, com foco em técnicas de imputação baseadas em métodos estatísticos e de aprendizado de máquina, sem o uso de redes neurais. Posteriormente, serão mostradas as similaridades dos trabalhos apresentados com o trabalho proposto.

Li, Ren e Zhao (2023) investigam a imputação de falhas em dados meteorológicos utilizando um framework multidimensional baseado em aprendizado de máquina, aplicado a várias estações meteorológicas. Entre os 20 métodos testados, o Random Forest se destacou como o mais eficaz para dados meteorológicos, principalmente por sua capacidade de lidar com dados complexos e ausentes, fornecendo resultados robustos para a imputação de dados em séries temporais. Esse estudo corrobora o uso de métodos avançados de machine learning, como Random Forest, que também foi explorado no presente trabalho.

Saad et al. (2020) comparam diferentes métodos de imputação, incluindo métodos tradicionais como interpolação e métodos mais sofisticados, como regressão linear e vetores de suporte. Os autores mostram que, para séries temporais com falhas, métodos baseados em aprendizado de máquina, como o Linear Regression e o Support Vector Regression (SVR), obtêm melhores resultados do que métodos simples como preenchimento por média ou interpolação linear, destacando a importância de escolher métodos adequados para a natureza dos dados.

Horta et al. (2021) realizam uma revisão sistemática dos métodos de preenchimento de falhas em dados de precipitação, destacando que métodos estatísticos clássicos, como a interpolação e o preenchimento por médias móveis, são amplamente utilizados para dados com padrões diários ou horários. Esses métodos, embora simples, podem ser eficazes em contextos onde o volume de dados ausentes não é muito elevado, mas apresentam limitações ao lidar com padrões não lineares.

Ou et al. (2024) exploram a combinação de Random Forest com métodos de imputação baseados em redes generativas adversariais (GAN), sugerindo que a combinação de diferentes métodos pode resultar em uma maior precisão na imputação de dados ausentes. Embora o presente trabalho não faça uso de redes neurais, a ideia de combinar múltiplos métodos de imputação, como diferentes formas de interpolação, médias móveis e algoritmos de aprendizado supervisionado, também foi aplicada para melhorar a qualidade da imputação de dados.

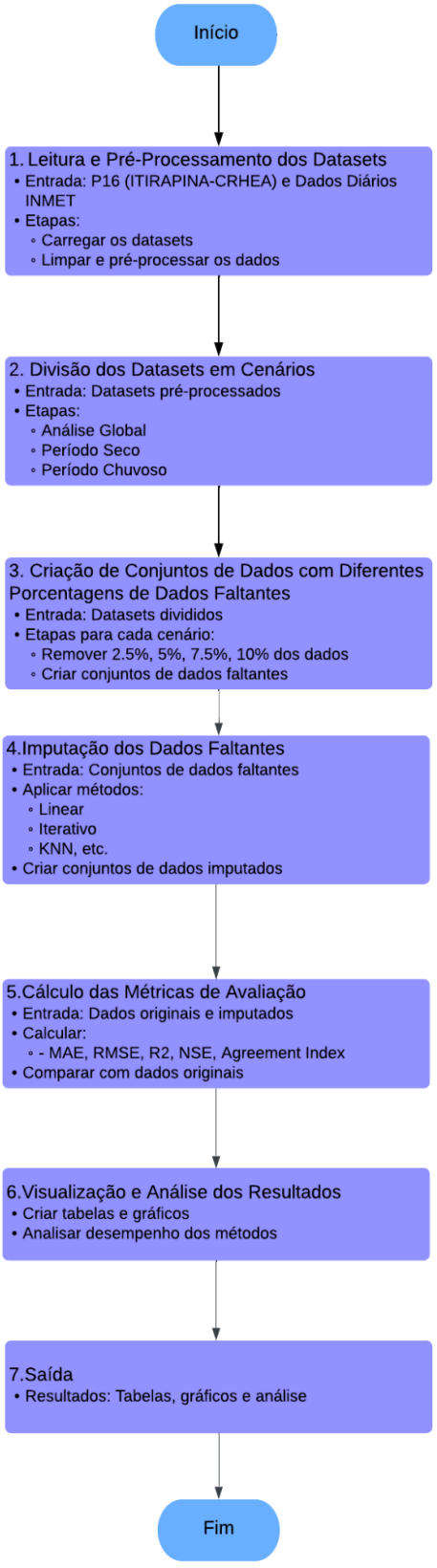
Embora Li et al. (2023), Saad et al. (2020), Horta et al. (2021) e Ou et al. (2024) explorem diferentes abordagens para a imputação de dados meteorológicos, observa-se uma diversidade de técnicas aplicadas. O presente trabalho segue a linha de imputação utilizando métodos estatísticos e algoritmos de aprendizado de máquina, como interpolação linear, KNN Imputer e Iterative Imputer, que foram selecionados por sua capacidade de lidar com séries temporais e garantir uma imputação eficiente dos dados pluviométricos ausentes. Essas abordagens não utilizam redes neurais, mas se concentram em métodos de aprendizado de máquina e estatística, em conformidade com as técnicas utilizadas no código implementado.

4. MATERIAIS E MÉTODOS

4.1 Visão geral

Este capítulo descreve os conjuntos de dados utilizados, detalhando a coleta e as etapas metodológicas para o preenchimento de falhas nos dados pluviométricos. A Figura 1 apresenta o fluxograma da metodologia adotada.

Figura 1. Etapas da metodologia



Fonte: Autor (2024).

A metodologia de imputação de dados faltantes nos conjuntos de dados meteorológicos compreende as seguintes etapas, resumidas na Figura 1:

1. Preparação dos conjunto de dados – Nesta etapa, os dados coletados foram carregados, limpos e pré-processados, abrangendo a remoção de valores inválidos e a conversão de tipos de dados. Os dados utilizados provêm dos conjunto de dados P16 (ITIRAPINA-CRHEA) e dados diários do INMET (1979 a 2023). A preparação adequada dos dados é essencial para garantir a confiabilidade das análises subsequentes.
2. Divisão dos conjunto de dados em cenários – Os conjunto de dados pré-processados foram segmentados em três cenários principais: Análise Global, Período Seco e Período Chuvoso. Essa divisão permite uma avaliação mais direcionada do desempenho dos modelos de imputação em diferentes condições climáticas, sendo importante definir critérios claros para a classificação dos períodos seco e chuvoso, baseados em índices de precipitação.
3. Criação de conjuntos de dados com diferentes porcentagens de dados faltantes – Após a divisão dos conjunto de dados, foram criados conjuntos de dados com 2,5%, 5%, 7,5% e 10% de dados faltantes, de forma aleatória, para cada cenário. Essa estratégia visa simular diferentes níveis de ausência de dados e testar a robustez dos métodos de imputação.
4. Imputação dos dados faltantes – Nesta etapa, foram aplicados diversos métodos de imputação, como imputação linear, iterativa e KNN, em cada conjunto de dados com dados faltantes. O objetivo é preencher as lacunas de forma precisa, mantendo a integridade dos dados originais.
5. Cálculo das métricas de avaliação – Para avaliar o desempenho dos métodos de imputação, foram utilizadas as métricas de MAE, RMSE, R^2 e o Índice de Concordância (Agreement Index). Cada métrica foi calculada comparando os dados imputados com os dados originais, proporcionando uma visão clara sobre a precisão e eficácia de cada método de imputação em diferentes cenários e porcentagens de dados faltantes.
6. Visualização e análise dos resultados – Foram criadas tabelas e gráficos para visualizar as métricas de avaliação de cada método de imputação. A análise visual permite uma comparação direta do desempenho dos métodos, destacando aqueles que apresentaram melhor acurácia em diferentes condições.

7. Saída dos resultados – Por fim, os resultados finais incluem tabelas, gráficos e uma análise comparativa dos métodos de imputação, considerando os diferentes cenários e porcentagens de dados faltantes. Essa análise proporciona uma base sólida para a escolha do método de imputação mais adequado para cada situação.

4.2 Conjunto de dados

Este estudo utiliza dados pluviométricos diários e estimativas por satélite para imputar falhas em séries temporais de precipitação, avaliando a eficácia de modelos de aprendizado de máquina nessa tarefa. As fontes de dados incluem medições in situ, além de produtos de sensoriamento remoto como o CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data) e o GPM (Global Precipitation Measurement), sucessor do TRMM (Tropical Rainfall Measuring Mission), lançado em 1997 com o objetivo de monitorar e estudar a precipitação tropical. A transição para o GPM, iniciada em 2014, ampliou a cobertura para além dos trópicos e aprimorou a precisão das estimativas de precipitação, resolvendo limitações de bases de dados climáticas mais antigas que dependiam, muitas vezes, de métodos de coleta e registro manuais. Esta modernização, impulsionada por avanços na tecnologia de satélites geoespaciais, possibilitou a criação de séries temporais de dados pluviométricos mais completas e consistentes, cruciais para análises climáticas robustas e para o desenvolvimento de modelos hidrológicos mais precisos.

Além disso, o estudo incorpora dados diários de precipitação (1961-2022) provenientes do National Tibetan Plateau Data Center (https://data.tpdc.ac.cn/) e nove Índices de Circulação Climática em Larga Escala (LCCIs), selecionados por sua influência reconhecida no clima da região de estudo. Entre os LCCIs utilizados, destacam-se o Índice de Fluxo Solar e o Índice de Intensidade da Alta Subtropical do Pacífico Ocidental, conhecidos por sua associação com eventos extremos de precipitação no norte da China, conforme demonstrado em pesquisas prévias (Wei et al., 2023; Wang et al., 2022). A inclusão desses índices visa capturar as complexas dinâmicas atmosféricas que modulam a precipitação na região e aprimorar a capacidade dos modelos de aprendizado de máquina em lidar com a imputação de dados faltantes.

4.2.1 Estações Meteorológicas e Fontes de Dados

O principal conjunto de dados utilizado nesta pesquisa foi obtido a partir de duas fontes principais:

- CHIRPS (Climate Hazards Group InfraRed Precipitation with Station Data), que fornece estimativas de precipitação diárias com base em sensores de satélite e dados de estações meteorológicas.
- Prcp_GPM, dados de precipitação diária medidos pelo satélite GPM (Global Precipitation Measurement).

Este estudo utilizou dados de precipitação diários de duas fontes distintas. A estação meteorológica de Itirapina/CRHEA/USP, localizada nas coordenadas 22°10'13.04"S e 47°53'55.66"W, forneceu dados para o período de 2002 a 2019, conforme representado pela Figura 2. Foram utilizados também os dados diários do INMET (Instituto Nacional de Meteorologia) de São Carlos-SP, representadas pela Figura 3, com coordenadas aproximadas de 21.98°S e 47.88°W, referentes ao período de 1979 a 2023. As duas fontes de dados, cuja localizações estão representadas pela Figura 4, permitiram uma análise mais abrangente do desempenho dos modelos de imputação, considerando diferentes contextos temporais.

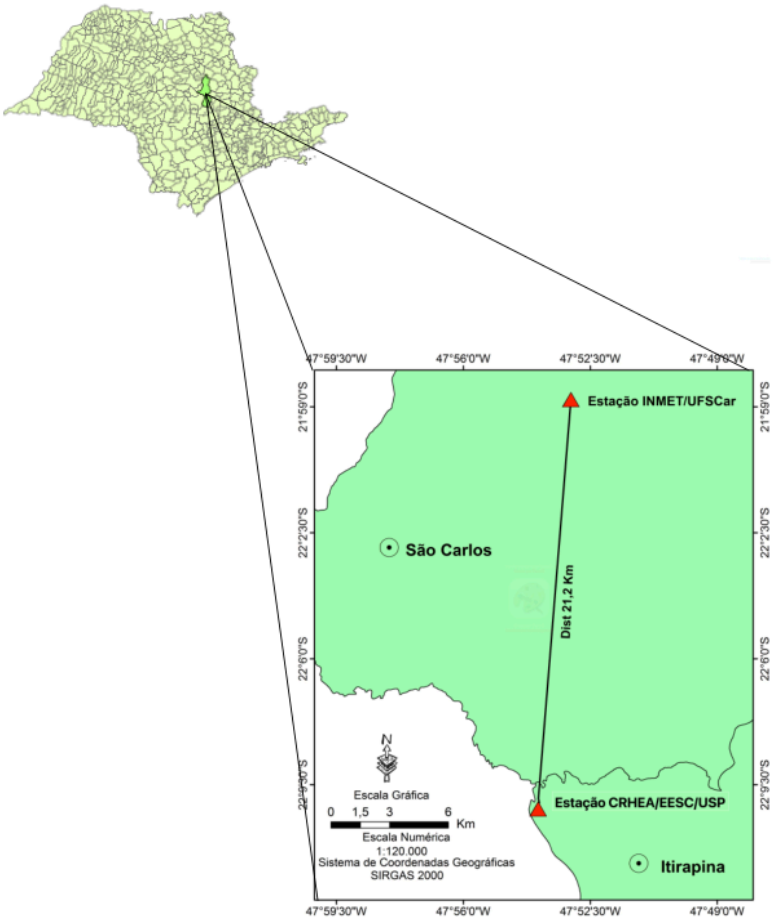
Figura 2. A estação meteorológica Itirapina/CRHEA/USP.

Fonte: Disponível em: http://nh.eesc.usp.br/crhea/. Acesso em: 08 nov. 2024.

Figura 3. Estação meteorológica do INMET de São Carlos - SP.

Fonte: Autor (2024).

Figura 4. Localização das Estações Meteorológicas.



Fonte: Autor (2024).

4.2.2 Variáveis Climáticas Utilizadas

As principais variáveis climáticas consideradas no estudo são descritas na Tabela 1 abaixo:

Tabela 1. Variáveis Utilizadas para Análise de Precipitação e Imputação de Dados.

Variável	Descrição	Tipo de Dado
CHIRPS	Precipitação diária estimada pelo conjunto de dados CHIRPS	Numérico
Prcp_GPM	Precipitação diária medida pelo satélite GPM	Numérico
DIA	Data da observação	Data
Período	Classificação do período como 'Chuvoso' ou 'Seco' com base na data	Categórico
É outlier?	Indicador binário se o valor é um outlier (0 = não, 1 = sim)	Binário
Precipitação - Itirapina-CRHEA	Precipitação medida na estação Itirapina-CRHEA (2002 a 2019)	Numérico
Precipitação -INMET São Carlos-SP	Dados do INMET (Instituto Nacional de Meteorologia) de São Carlos-SP (1979 a 2023)	Numérico

Fonte: Autor (2024).

Essas variáveis foram essenciais para a modelagem e análise da imputação dos dados ausentes, permitindo uma análise robusta dos períodos chuvosos e secos.

4.2.3 Classificação dos Períodos Climáticos

Para este trabalho, o ano foi dividido em dois períodos climáticos distintos: chuvoso e seco. O período chuvoso foi definido como os meses entre outubro e março, enquanto o período seco compreende os meses entre abril e setembro. Essa classificação permitiu a segmentação dos dados para avaliar o comportamento dos modelos de imputação em diferentes condições sazonais, garantindo que o impacto das variações sazonais fosse adequadamente considerado na análise.

4.2.4 Pré-processamento dos Dados e Simulação de Dados Faltantes

Antes da aplicação dos modelos de imputação, foi realizado um processo de pré-processamento dos dados que incluiu a classificação de outliers e a simulação de dados faltantes. O pré-processamento consistiu nas seguintes etapas:

- Classificação de outliers: Para evitar que valores extremos distorcerem os resultados da imputação, foram identificados e classificados outliers com base no cálculo dos quartis e do Intervalo Interquartil (IQR). O IQR foi calculado como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1):

$$IQR = Q3 - Q1 \text{ (Equação 10)}$$

  Os limites inferior e superior para a remoção de outliers foram definidos da seguinte forma:

$$Limite\ inferior = Q1 - 1,5 * IQR \text{ (Equação 11)}$$

$$Limite\ superior = Q3 + 1,5 * IQR \text{ (Equação 12)}$$

- Para testar a eficácia dos modelos de imputação, foram removidos intencionalmente 2,5%, 5%, 7,5% e 10% dos dados de precipitação medida, em diferentes cenários. Essa abordagem permitiu avaliar o desempenho dos métodos de imputação em preencher as lacunas de dados reais sob diferentes proporções de dados faltantes.

4.2.5 Imputação de Dados Ausentes

Após o pré-processamento e a simulação de dados faltantes, foram aplicados diversos métodos de imputação de dados para preencher as lacunas observadas. Entre os métodos testados, os principais e mais utilizados foram:

- KNNImputer: Este método preenche os valores faltantes com base nos k vizinhos mais próximos, utilizando a similaridade entre os dados observados para estimar os valores ausentes.
- Iterative Imputer: Realiza imputações iterativas, onde os valores ausentes são previstos repetidamente em um processo que leva em conta todas as variáveis disponíveis.

Esses métodos foram escolhidos por sua robustez na captura de padrões subjacentes nos dados pluviométricos, proporcionando uma imputação precisa e consistente. Embora outros métodos tenham sido testados, o KNNImputer e o Iterative Imputer se mostraram os mais eficazes para o preenchimento de falhas nos dados.

4.2.6 Período de Tempo Abrangido

Os dados utilizados cobrem um período extenso, permitindo a análise de tendências sazonais e anuais. O conjunto de dados da estação Itirapina-CRHEA abrange o período de 2002 a 2019, enquanto os dados do INMET de São Carlos fornecem informações de 1979 a 2023. Essa ampla janela temporal é fundamental para garantir a robustez dos modelos de imputação e para capturar variações significativas nos padrões de precipitação ao longo dos anos.

4.3 Métodos tradicionais de preenchimento de dados

A análise de séries temporais frequentemente enfrenta o desafio de lidar com falhas nos dados, resultantes de problemas técnicos ou ambientais que afetam a coleta contínua de informações. No caso dos dados pluviométricos, essas falhas podem comprometer a precisão de previsões e análises climáticas. Métodos estatísticos simples têm sido amplamente utilizados para preencher essas lacunas, proporcionando soluções rápidas e de fácil implementação. Nesta seção, são apresentados os principais métodos tradicionais de preenchimento de falhas, suas limitações e suas respectivas fórmulas.

4.3.1 Interpolação Linear

A interpolação linear é um dos métodos mais simples para preenchimento de falhas em séries temporais. Ela assume que a variação entre dois pontos conhecidos segue uma tendência linear e calcula o valor ausente com base nessa relação. A fórmula para a interpolação linear é dada por:

$$x_t = x_{t-1} + \frac{x_{t+1} - x_{t-1}}{(t+1)-(t-1)} * (t - (t-1)) \quad \text{(Equação 1)}$$

Onde:

- x_t é o valor interpolado no tempo t,
- x_{t-1} e x_{t+1} são os valores conhecidos imediatamente antes e depois do tempo t.

Embora simples, a interpolação linear não é adequada para séries temporais com alta variabilidade ou eventos extremos, como os observados em dados pluviométricos, onde as mudanças de precipitação podem ser abruptas e não lineares (BLAKE, 2011).

4.3.2 Médias Móveis

A média móvel é uma técnica que calcula o valor de um ponto ausente com base na média dos valores adjacentes a ele, em uma janela de tempo especificada. A fórmula da média móvel simples para uma janela de n períodos é:

$$x_t = \frac{1}{n} \sum_{i=t-k}^{t+k} x_i \quad \text{(Equação 2)}$$

Onde:

- x_t é o valor preenchido no tempo t,
- n é o número total de pontos dentro da janela (incluindo k períodos antes e depois de t),
- x_i são os valores observados dentro da janela de tempo.

Apesar de sua simplicidade, esse método tende a suavizar os dados, o que pode ser problemático em séries temporais de dados pluviométricos, onde picos e vales extremos são importantes para a análise (SANCHES et al., 2020).

4.3.3 Média Simples

A média simples de toda a série temporal pode ser utilizada para substituir os valores ausentes. Ela é calculada pela seguinte fórmula:

$$x_t = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{(Equação 3)}$$

Onde:

- x_t é o valor preenchido no tempo t,
- n é o número total de observações disponíveis,
- x_i são os valores observados.

Embora fácil de implementar, a média simples é inadequada para séries temporais que apresentam variações sazonais ou padrões específicos, uma vez que ela ignora as variações de curto prazo nos dados (LATIF et al., 2023).

4.4 Limitações dos métodos tradicionais

Métodos tradicionais de preenchimento de falhas, como interpolação linear e média móvel, embora frequentemente empregados devido à sua simplicidade computacional e facilidade de implementação (LY et al., 2013), apresentam limitações inerentes que comprometem sua eficácia, especialmente quando aplicados a séries temporais de precipitação, caracterizadas por alta variabilidade e complexidade. A interpolação linear, por exemplo, assume uma relação linear entre os pontos amostrais, uma premissa raramente válida em séries pluviométricas. Essa simplificação excessiva ignora a natureza não-linear dos processos hidrológicos e atmosféricos que regem a precipitação, resultando em estimativas imprecisas, particularmente em situações com eventos extremos. A suavização artificial introduzida pela interpolação linear mascara a verdadeira magnitude e frequência de picos de chuva e períodos de seca, dificultando a análise de eventos críticos como inundações e estiagens (DAS, 2021a). Além disso, a interpolação linear não considera a influência de variáveis externas, como a sazonalidade, que pode modular a relação entre os pontos amostrais.

A média móvel, por sua vez, atua como um filtro passa-baixa, suavizando flutuações de curto prazo e removendo ruído da série temporal. Embora útil para identificar tendências de longo prazo, esse processo de suavização pode obliterar informações relevantes sobre a variabilidade da precipitação. Eventos extremos, como chuvas torrenciais ou secas prolongadas, são atenuados, resultando em uma representação distorcida da dinâmica pluviométrica e comprometendo a análise de riscos associados a esses eventos (WAGNER et al., 2012). Em séries temporais com sazonalidade marcante, como as observadas em regimes

de monções (DAS & ISLAM, 2021), a aplicação da média móvel pode distorcer os padrões sazonais, subestimando a intensidade das chuvas no período úmido e superestimando a precipitação na estação seca.

A principal deficiência dos métodos tradicionais reside na sua incapacidade de incorporar informações auxiliares e modelar a complexa interação entre a precipitação e outras variáveis meteorológicas e geofísicas (AYDIN & RAJA, 2016; OLGU & RAJA, 2016). Variáveis como temperatura, umidade, pressão atmosférica, velocidade e direção do vento, radiação solar, e características topográficas, como altitude, declividade e exposição, desempenham um papel crucial na modulação dos padrões de precipitação (ADHIKARY et al., 2017; LUO et al., 2011; MAIR & FARES, 2011, 2012). A omissão dessas variáveis na modelagem, inerente aos métodos tradicionais, limita sua capacidade preditiva, especialmente em regiões com alta variabilidade espacial, como áreas montanhosas (BAJAT et al., 2013; KATIPOĞLU, 2021; PELLICONE et al., 2018, 2019, 2020). Diante dessas limitações, a utilização de métodos mais sofisticados, como os baseados em aprendizado de máquina, torna-se imperativa para uma modelagem mais precisa e robusta da precipitação.

4.5 Métodos baseados em aprendizado de máquina

Com o avanço das técnicas de aprendizado de máquina, surgiram métodos mais sofisticados para lidar com o preenchimento de falhas em séries temporais. Diferentemente dos métodos tradicionais, que muitas vezes falham em capturar a complexidade dos dados climáticos, os algoritmos de aprendizado de máquina têm a capacidade de modelar padrões não lineares e lidar com grandes volumes de dados. Essas características tornam os modelos de aprendizado de máquina especialmente úteis em dados pluviométricos, onde a variabilidade e a irregularidade são frequentes (LATIF et al., 2023).

Nesta seção, são apresentadas as principais técnicas de aprendizado de máquina aplicadas ao preenchimento de falhas, com destaque para o K-Nearest Neighbors Imputer (KNN Imputer), o Iterative Imputer e o Random Forest, que foram selecionados neste trabalho por sua eficiência e precisão em lidar com séries temporais complexas e de alta variabilidade.

4.5.1 K-Nearest Neighbors Imputer (KNN Imputer)

O K-Nearest Neighbors Imputer (KNN Imputer) é uma técnica baseada na identificação dos vizinhos mais próximos de uma observação com falha, preenchendo o valor ausente com base nos valores das observações mais similares. Esse método utiliza a distância entre os pontos no espaço de variáveis para determinar quais são os dados mais próximos.

A fórmula básica para a imputação por KNN pode ser expressa da seguinte forma:

$$x_t = \frac{1}{k}\sum_{i=1}^{k} w_i x_i \quad \text{(Equação 4)}$$

Onde:

- x_t é o valor imputado no tempo t,
- k é o número de vizinhos mais próximos,
- w_i são os pesos baseados na distância entre os vizinhos,
- x_i são os valores observados dos k vizinhos mais próximos.

O KNN Imputer é eficaz em situações onde existem padrões similares entre os dados e onde a correlação entre as variáveis é forte. Em séries temporais de precipitação, o método é útil porque permite imputar valores com base nas características de pontos adjacentes e correlacionados, preservando a integridade dos padrões temporais (TROYANSKY et al., 2020).

4.5.2 Iterative Imputer

O Iterative Imputer é uma técnica avançada que preenche os valores ausentes de forma iterativa, ajustando os valores com base em previsões feitas por um modelo preditivo. O processo envolve a inicialização dos valores ausentes com uma imputação simples, seguida de várias iterações, onde as variáveis ausentes são previstas e ajustadas em função das outras variáveis da série temporal.

O processo de imputação iterativa é formalizado da seguinte maneira:

1. Inicialize os valores ausentes com uma técnica simples (por exemplo, média ou interpolação linear).

2. Para cada valor ausente, ajuste-o usando uma regressão múltipla com base nas outras variáveis disponíveis.
3. Repita o processo até que os valores imputados converjam, ou seja, até que os ajustes subsequentes sejam mínimos.

Esse método é particularmente útil quando há correlação entre várias variáveis da série temporal, como temperatura, umidade e precipitação. O Iterative Imputer é capaz de capturar a interdependência entre essas variáveis, resultando em imputações mais precisas e consistentes ao longo do tempo (LATIF et al., 2023).

4.6 Métricas de avaliação

Para avaliar a eficácia dos modelos de preenchimento de falhas de dados pluviométricos utilizando aprendizado de máquina, foram empregadas várias métricas que medem a precisão e a qualidade da imputação dos dados faltantes. Cada métrica desempenha um papel importante na identificação de como os modelos conseguem prever os valores de precipitação ausentes, comparando-os com os dados reais disponíveis. As métricas escolhidas analisam desde a magnitude dos erros até a concordância global dos valores imputados.

4.6.1 Erro Médio Absoluto (MAE - Mean Absolute Error)

O Erro Médio Absoluto (MAE) mede a média das diferenças absolutas entre os valores imputados e os valores reais de precipitação. No contexto da imputação de dados pluviométricos, o MAE oferece uma medida clara de quão perto os valores preenchidos estão dos valores observados, sem considerar a direção do erro (se é para mais ou para menos).

A fórmula do MAE é:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \quad \text{(Equação 5)}$$

Onde:

- n é o número de amostras (dias com dados pluviométricos),
- y_i são os valores reais de precipitação,
- $\hat{y}_i$ são os valores imputados pelo modelo.

No preenchimento de dados pluviométricos, um MAE menor indica que o modelo de aprendizado de máquina preencheu as falhas de dados de maneira mais precisa, aproximando os valores imputados dos dados reais.

4.6.2 Erro Quadrático Médio (RMSE - Root Mean Square Error)

O Erro Quadrático Médio (RMSE) mede o desvio quadrático médio entre os valores imputados e os valores reais de precipitação. Ele penaliza erros maiores com mais severidade do que o MAE, o que é relevante no preenchimento de dados pluviométricos, pois grandes desvios podem distorcer as análises climáticas e hidrológicas.

A fórmula do RMSE é:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \text{ (Equação 6)}$$

Onde:

- n é o número de amostras (dias com dados pluviométricos),
- y_i são os valores reais de precipitação,
- $\hat{y}_i$ são os valores imputados pelo modelo

No contexto da imputação de dados pluviométricos, valores mais baixos de RMSE indicam uma maior precisão, pois grandes erros no preenchimento dos dados são menos frequentes.

4.6.3 Coeficiente de Determinação (R^2)

O Coeficiente de Determinação (R^2) indica a proporção da variabilidade nos dados pluviométricos que é explicada pelos valores imputados. Um valor de R^2 próximo de 1 indica que o modelo explica bem a variabilidade dos dados, enquanto valores mais baixos indicam que os dados imputados não seguem a tendência dos dados reais.

A fórmula do R^2 é:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \text{ (Equação 7)}$$

Onde:

- $\bar{y}$ é a média dos valores reais de precipitação.

Essa métrica é importante, pois ajuda a avaliar se o modelo está capturando corretamente a variabilidade natural dos dados pluviométricos ao imputar os valores ausentes.

4.6.4 Índice de Concordância (d)

O Índice de Concordância (d) mede o grau de concordância entre os valores imputados e os valores reais de precipitação. Esse índice varia de 0 a 1, onde 1 representa concordância perfeita. No contexto da imputação de dados, o Índice de Concordância avalia a capacidade do modelo de prever corretamente os valores ausentes, comparando-os com os valores reais.

A fórmula do Índice de Concordância é:

$$d = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} \quad \text{(Equação 8)}$$

Esse índice é particularmente útil na análise de dados climáticos, onde é importante garantir que os valores imputados sigam as mesmas tendências e padrões dos dados observados.

5. RESULTADOS

Os resultados obtidos com os modelos de aprendizado de máquina serão discutidos neste capítulo. Serão apresentadas as tabelas e gráficos que mostram o desempenho dos modelos em termos das métricas de avaliação. Comparações entre os resultados dos diferentes métodos de imputação e modelos de previsão foram realizadas, mostrando qual abordagem foi mais eficaz em termos de precisão e robustez das previsões de chuvas.

5.1 Visão geral dos experimentos

A análise foi realizada considerando dois conjuntos de dados distintos: Conjunto de dados A, correspondente aos dados da estação P16 (Itirapina-CRHEA), e Conjunto de dados B, referente aos dados do INMET. Três cenários principais foram avaliados: Análise Global, Período Seco e Período Chuvoso. Para cada conjunto de dados e cenário, foram aplicados os métodos de imputação KNN Imputer e Iterative Imputer, bem como os métodos tradicionais (interpolação linear, médias móveis e média simples). O desempenho foi avaliado utilizando as seguintes métricas: Erro Médio Absoluto (MAE), Erro Quadrático Médio (RMSE), Coeficiente de Determinação (R^2) e Índice de Concordância (d).

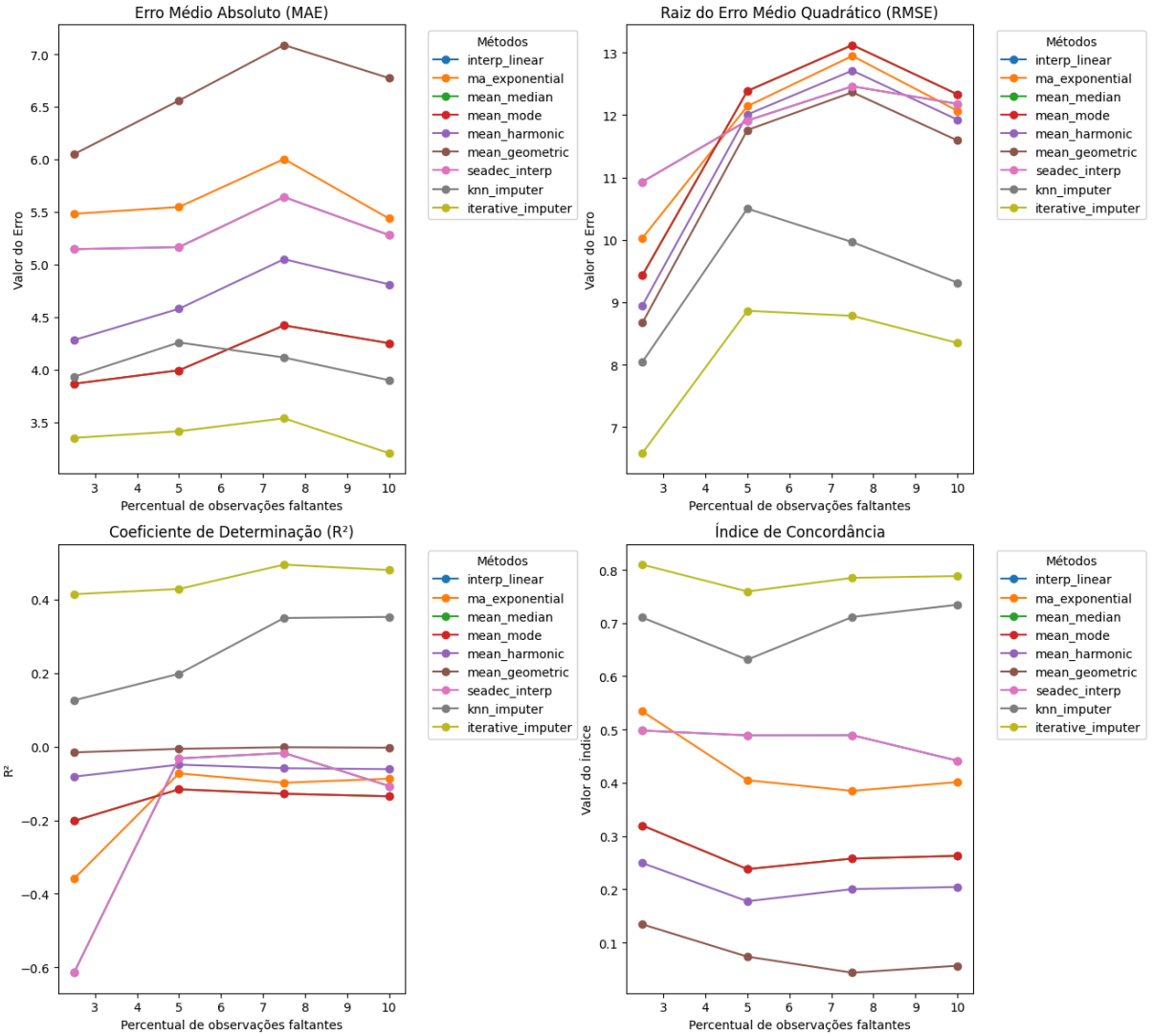
5.2 Análise Global

5.2.1 Conjunto de dados A

Os resultados da análise global do Conjunto de Dados A, considerando diferentes métodos de imputação e percentuais de observações faltantes (2.5%, 5%, 7% e 10%), são apresentados na Figura 5. Os métodos avaliados incluem: Interpolação Linear (interp_linear), Média Móvel Exponencial (ma_exponential), Mediana das Médias (mean_median), Moda das Médias (mean_mode), Média Harmônica (mean_harmonic), Média Geométrica (mean_geometric), Interpolação Sedecc (seadec_interp), Imputação por K-Vizinhos Mais Próximos (knn_imputer) e Imputação Iterativa (iterative_imputer). Para cada combinação de método e percentual de dados faltantes, foram calculadas as seguintes métricas de desempenho: Erro Médio Absoluto (MAE), Raiz do Erro Médio Quadrático (RMSE), Coeficiente de Determinação (R^2) e Índice de Concordância. A análise comparativa dessas métricas permite avaliar a eficácia de cada método na reconstrução dos dados faltantes e

identificar o mais adequado para o Conjunto de Dados A, considerando diferentes níveis de perda de informação.

Figura 5. Desempenho dos Métodos de Imputação no Conjunto de Dados A - Análise Global.



Fonte: Autor (2024).

A Tabela 2 apresenta os valores numéricos das métricas de desempenho (MAE, RMSE, R² e Índice de Concordância) para cada método de imputação e para diferentes percentuais de dados faltantes considerados na análise global do Conjunto de Dados A. Essa

tabela complementa a Figura 5, fornecendo os valores exatos, o que possibilita uma análise mais detalhada dos resultados.

Tabela 2. Métricas de Desempenho dos Métodos de Imputação para o Conjunto de Dados A - Análise Global.

Método	Percentual Faltante (%)	MAE	RMSE	R2	Índice de Concordância
interp_linear	2,5	5,146	10,930	-0,614	0,498
ma_exponential	2,5	5,482	10,029	-0,359	0,534
mean_median	2,5	3,866	9,431	-0,202	0,320
mean_mode	2,5	3,866	9,431	-0,202	0,320
mean_harmonic	2,5	4,282	8,947	-0,082	0,249
mean_geometric	2,5	6,049	8,670	-0,016	0,134
seadec_interp	2,5	5,146	10,930	-0,614	0,498
knn_imputer	2,5	3,933	8,043	0,126	0,710
iterative_imputer	2,5	3,352	6,584	0,414	0,810
interp_linear	5	5,166	11,911	-0,032	0,489
ma_exponential	5	5,547	12,142	-0,073	0,405
mean_median	5	3,995	12,386	-0,116	0,238
mean_mode	5	3,995	12,386	-0,116	0,238
mean_harmonic	5	4,579	12,008	-0,049	0,178
mean_geometric	5	6,559	11,760	-0,006	0,074
seadec_interp	5	5,166	11,911	-0,032	0,489
knn_imputer	5	4,259	10,501	0,198	0,631
iterative_imputer	5	3,415	8,865	0,428	0,759
interp_linear	7,5	5,642	12,461	-0,017	0,489
ma_exponential	7,5	6,003	12,947	-0,098	0,385
mean_median	7,5	4,421	13,122	-0,128	0,258
mean_mode	7,5	4,421	13,122	-0,128	0,258
mean_harmonic	7,5	5,051	12,713	-0,059	0,200
mean_geometric	7,5	7,087	12,365	-0,002	0,044
seadec_interp	7,5	5,642	12,461	-0,017	0,489
knn_imputer	7,5	4,116	9,964	0,350	0,711
iterative_imputer	7,5	3,538	8,783	0,495	0,785
interp_linear	10	5,280	12,180	-0,107	0,442
ma_exponential	10	5,437	12,070	-0,087	0,401
mean_median	10	4,252	12,333	-0,135	0,263
mean_mode	10	4,252	12,333	-0,135	0,263
mean_harmonic	10	4,812	11,926	-0,061	0,205
mean_geometric	10	6,773	11,593	-0,003	0,057
seadec_interp	10	5,280	12,180	-0,107	0,442
knn_imputer	10	3,900	9,316	0,352	0,734
iterative_imputer	10	3,207	8,350	0,480	0,788

Fonte: Autor (2024).

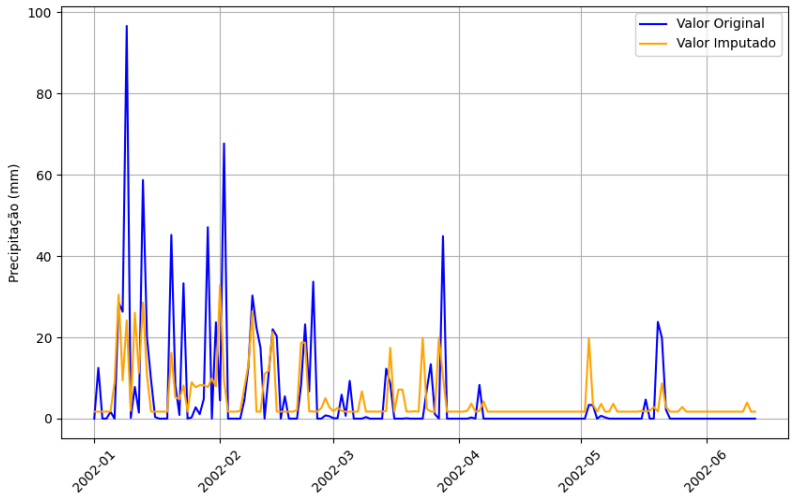
Nas Figuras 6 a 9, são apresentados gráficos que comparam os valores imputados com os valores originais para diferentes percentuais de dados faltantes: 2,5%, 5%, 7,5% e 10%. Esses gráficos ilustram como a precisão dos métodos de imputação varia à medida que a porcentagem de dados faltantes aumenta, permitindo observar as tendências de erro em função da quantidade de dados ausentes.

Os resultados mostram que os erros tendem a aumentar quando o percentual de dados faltantes é maior. Isso ocorre porque, à medida que o conjunto de dados faltantes cresce, o modelo tem menos informações para treinar, o que compromete sua capacidade de prever com precisão os valores ausentes. Esse efeito é particularmente evidente nos percentuais mais

altos, como 7,5% e 10%, onde a distorção entre os valores imputados e os valores reais se torna mais significativa.

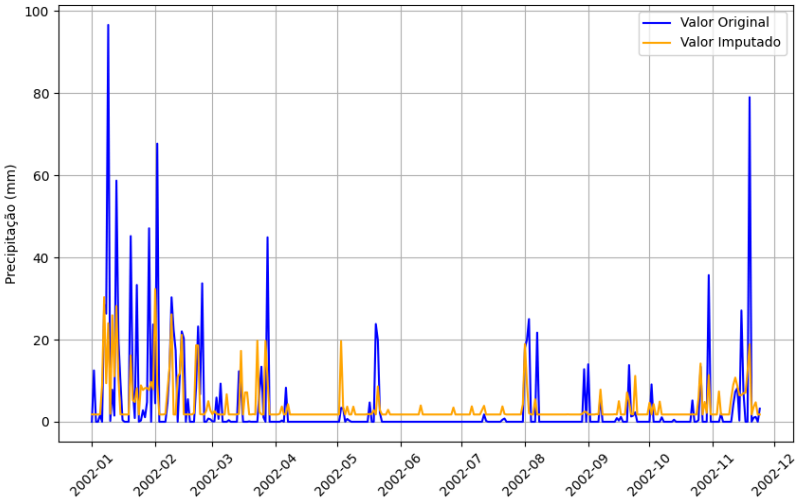
Além disso, o modelo apresenta uma tendência a cometer erros maiores quando os dados contêm outliers, como no caso de chuvas extremas. Nesses períodos, os valores originais podem ser significativamente diferentes da tendência geral dos dados, o que dificulta a tarefa de imputação. Como resultado, o modelo pode falhar ao estimar esses valores extremos com precisão, aumentando o erro na imputação. Esse comportamento é especialmente notável quando as chuvas extremas ocorrem em períodos com grande volume de dados faltantes, uma vez que o modelo tem poucas informações confiáveis para ajustar suas previsões de forma adequada.

Figura 6. Análise Global, Conjunto de Dados A, 2.5% de dados faltantes - Precipitação Observada vs. Imputada



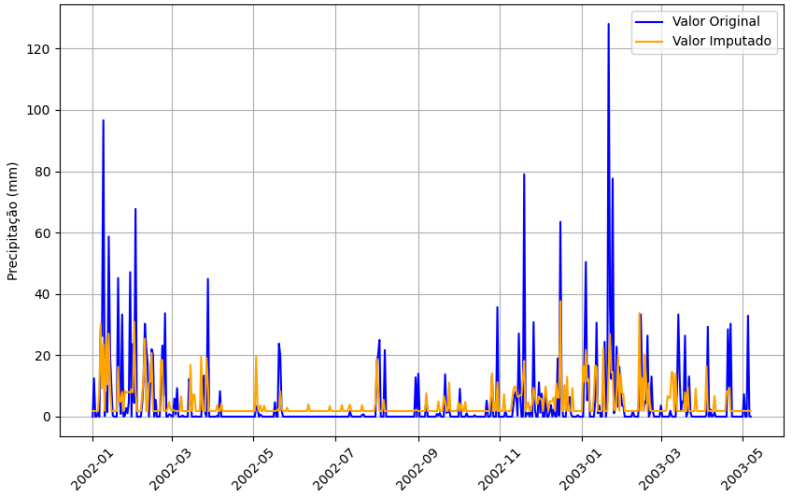
Fonte: Autor (2024).

Figura 7. Análise Global, Conjunto de Dados A, 5.0% de dados faltantes - Precipitação Observada vs. Imputada



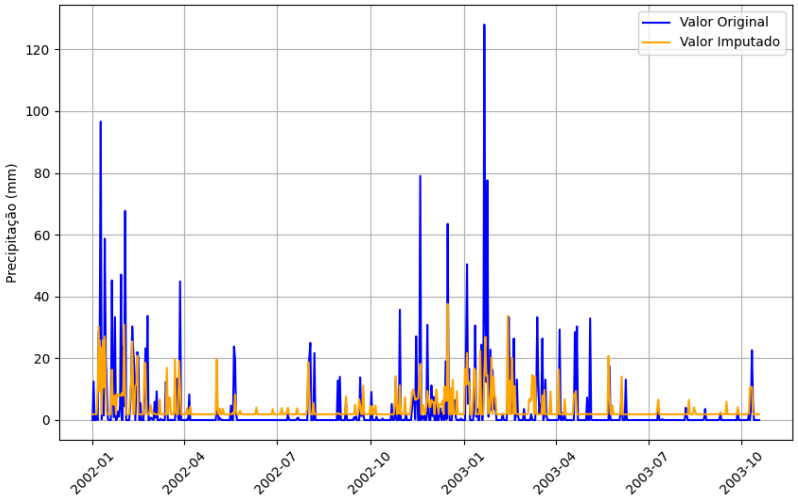
Fonte: Autor (2024).

Figura 8. Análise Global, Conjunto de Dados A, 7.5% de dados faltantes - Precipitação Observada vs. Imputada



Fonte: Autor (2024).

Figura 9. Análise Global, Conjunto de Dados A, 10.0% de dados faltantes - Precipitação Observada vs. Imputada

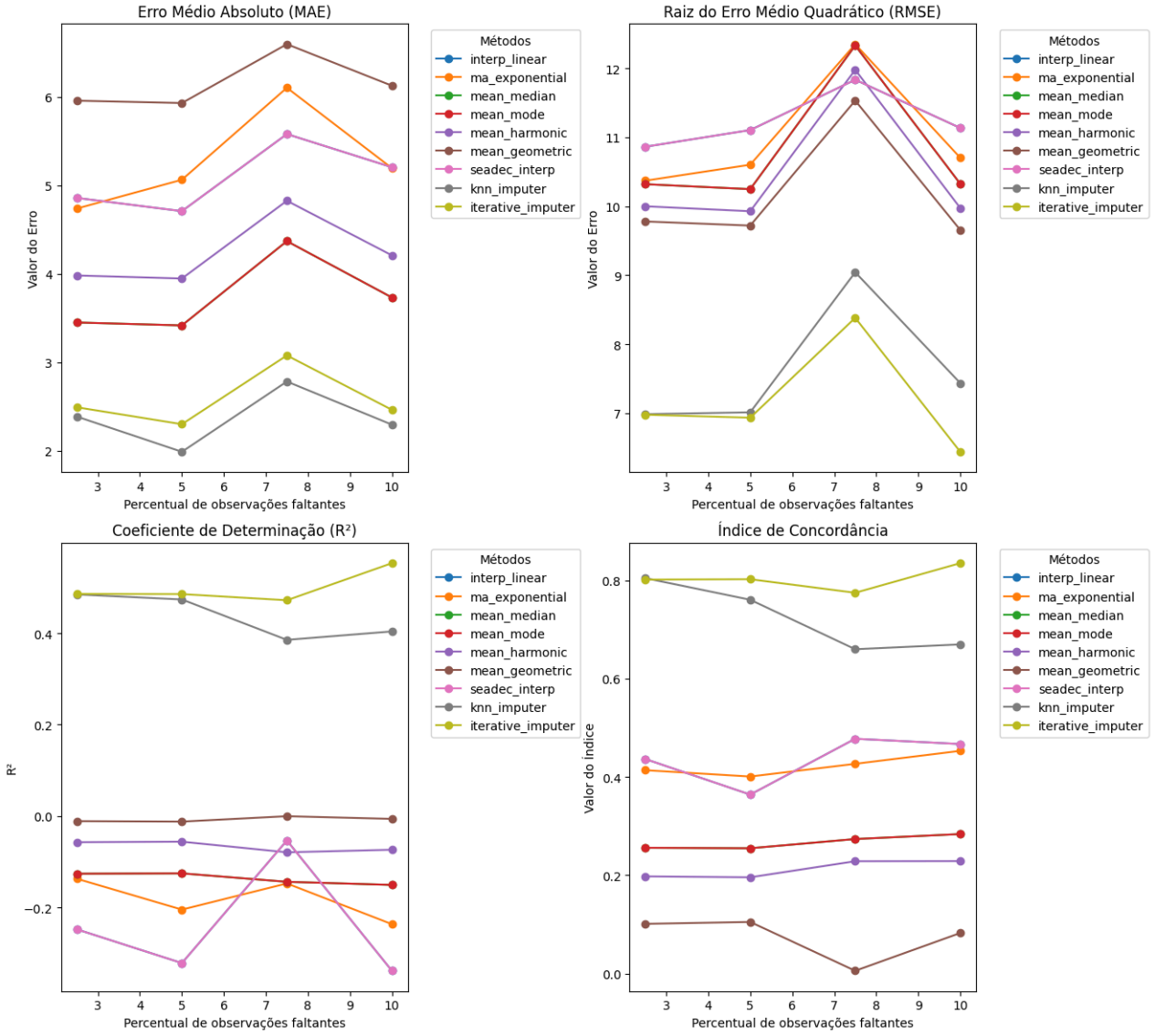


Fonte: Autor (2024).

5.2.2 Conjunto de dados B

A Figura 10 apresenta os resultados da análise global do Conjunto de Dados B, seguindo a mesma metodologia e métricas de desempenho descritas para o Conjunto de Dados A.

Figura 10. Desempenho dos Métodos de Imputação no Conjunto de Dados B - Análise Global.



Fonte: Autor (2024).

De forma semelhante, a Tabela 3 apresenta os valores numéricos das métricas de desempenho para o Conjunto de Dados B, seguindo a mesma estrutura da Tabela 2.

Tabela 3. Métricas de Desempenho dos Métodos de Imputação para o Conjunto de Dados B - Análise Global.

Method	Percentual Faltante (%)	MAE	RMSE	R2	Índice de Concordância
interp_linear	2,5	4,856	10,864	-0,247	0,436
ma_exponential	2,5	4,736	10,371	-0,137	0,414
mean_median	2,5	3,449	10,321	-0,126	0,256
mean_mode	2,5	3,449	10,321	-0,126	0,256
mean_harmonic	2,5	3,980	10,000	-0,057	0,198
mean_geometric	2,5	5,955	9,780	-0,011	0,101
seadec_interp	2,5	4,856	10,864	-0,247	0,436
knn_imputer	2,5	2,386	6,983	0,485	0,805
iterative_imputer	2,5	2,491	6,975	0,486	0,802
interp_linear	5,0	4,707	11,106	-0,321	0,364
ma_exponential	5,0	5,062	10,603	-0,204	0,401
mean_median	5,0	3,416	10,249	-0,125	0,255
mean_mode	5,0	3,416	10,249	-0,125	0,255
mean_harmonic	5,0	3,946	9,928	-0,056	0,196
mean_geometric	5,0	5,929	9,720	-0,012	0,105
seadec_interp	5,0	4,707	11,106	-0,321	0,364
knn_imputer	5,0	1,988	7,010	0,474	0,761
iterative_imputer	5,0	2,300	6,932	0,485	0,802
interp_linear	7,5	5,578	11,836	-0,053	0,478
ma_exponential	7,5	6,101	12,352	-0,147	0,427
mean_median	7,5	4,368	12,333	-0,143	0,274
mean_mode	7,5	4,368	12,333	-0,143	0,274
mean_harmonic	7,5	4,824	11,980	-0,079	0,228
mean_geometric	7,5	6,594	11,534	0,000	0,006
seadec_interp	7,5	5,578	11,836	-0,053	0,478
knn_imputer	7,5	2,782	9,042	0,385	0,660
iterative_imputer	7,5	3,077	8,381	0,472	0,775
interp_linear	10,0	5,203	11,136	-0,338	0,467
ma_exponential	10,0	5,193	10,704	-0,236	0,453
mean_median	10,0	3,731	10,326	-0,150	0,284
mean_mode	10,0	3,731	10,326	-0,150	0,284
mean_harmonic	10,0	4,207	9,975	-0,073	0,229
mean_geometric	10,0	6,125	9,656	-0,006	0,082
seadec_interp	10,0	5,203	11,136	-0,338	0,467
knn_imputer	10,0	2,295	7,435	0,404	0,670
iterative_imputer	10,0	2,462	6,437	0,553	0,835

Fonte: Autor (2024).

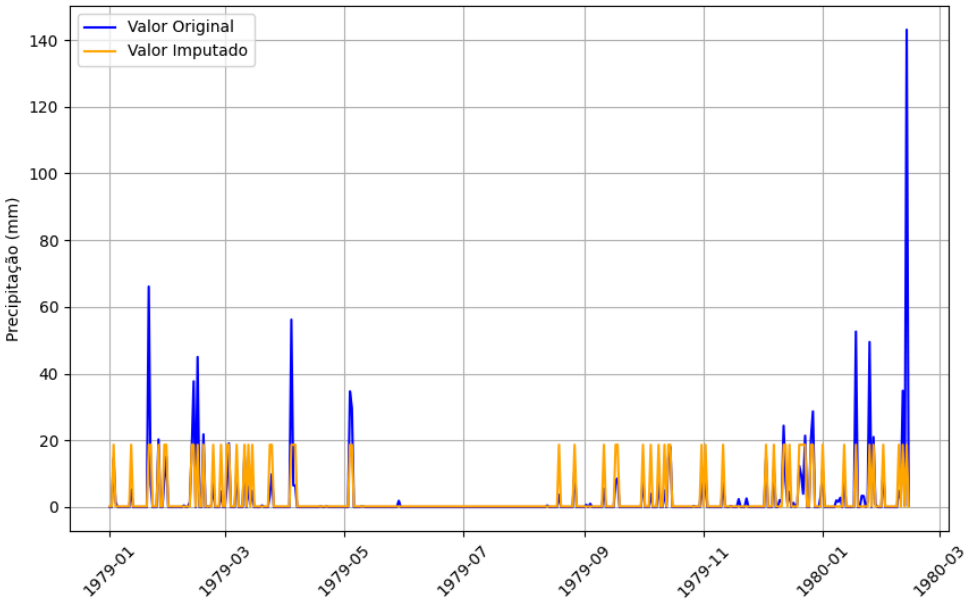
Nas Figuras 11 a 14, são apresentados gráficos semelhantes aos das Figuras 5 a 8, desta vez comparando os valores imputados com os valores originais do Conjunto de Dados B para diferentes percentuais de dados faltantes: 2,5%, 5%, 7,5% e 10%. Esses gráficos seguem a mesma estrutura e metodologia, permitindo analisar a variação da precisão dos métodos de imputação conforme o aumento da porcentagem de dados ausentes.

De maneira consistente com os resultados obtidos para o Conjunto de Dados A, observa-se que os erros de imputação aumentam com o crescimento do percentual de dados faltantes. Esse comportamento reflete a dificuldade do modelo em aprender e fazer previsões

precisas quando a quantidade de dados disponíveis para treinamento diminui, o que compromete a exatidão das imputações.

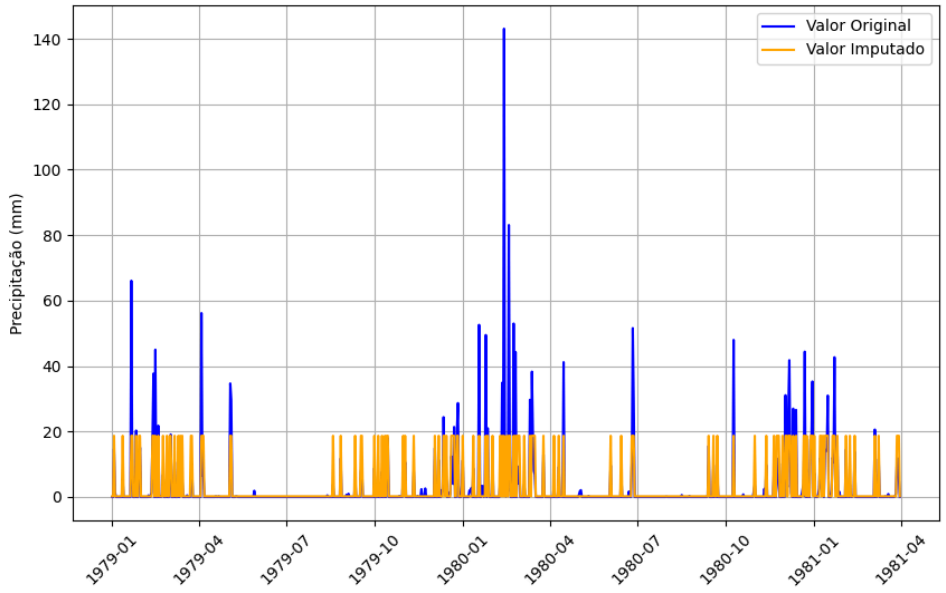
Além disso, a presença de outliers, como valores extremos de chuvas, continua a impactar negativamente a performance do modelo. Da mesma forma que no Conjunto de Dados A, em períodos com chuvas extremas e grandes lacunas de dados, o modelo encontra dificuldades em ajustar suas previsões, resultando em erros mais significativos na imputação dos valores extremos.

Figura 11. Análise Global, Conjunto de Dados B, 2.5% de dados faltantes - Precipitação Observada vs. Imputada



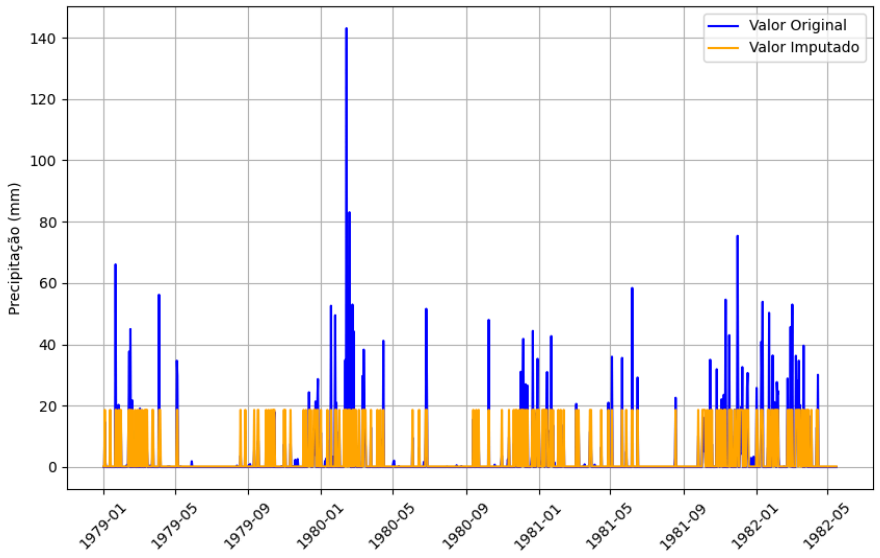
Fonte: Autor (2024).

Figura 12. Análise Global, Conjunto de Dados B, 5.0% de dados faltantes - Precipitação Observada vs. Imputada



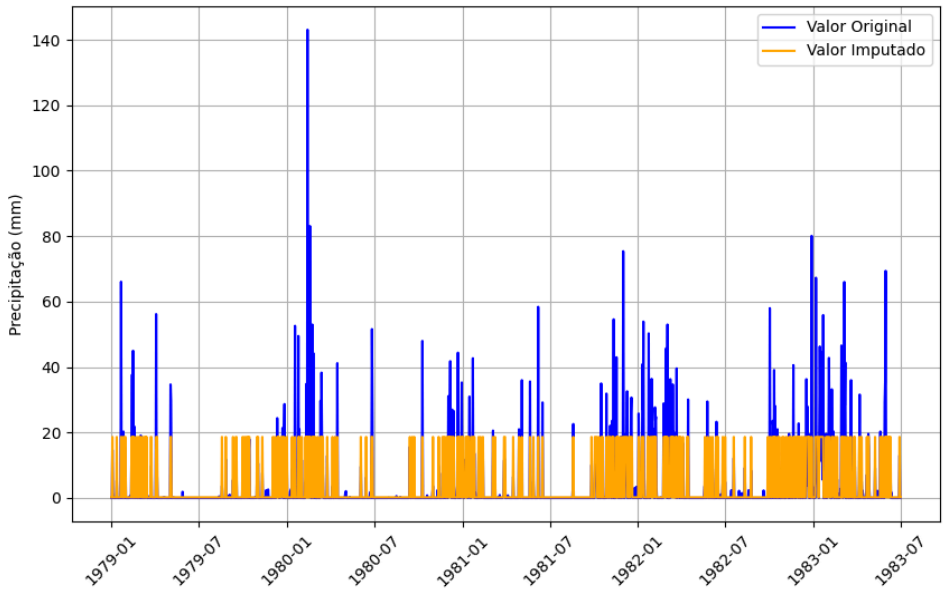
Fonte: Autor (2024).

Figura 13. Análise Global, Conjunto de Dados B, 7.5% de dados faltantes - Precipitação Observada vs. Imputada



Fonte: Autor (2024).

Figura 14. Análise Global, Conjunto de Dados B, 10.0% de dados faltantes - Precipitação Observada vs. Imputada



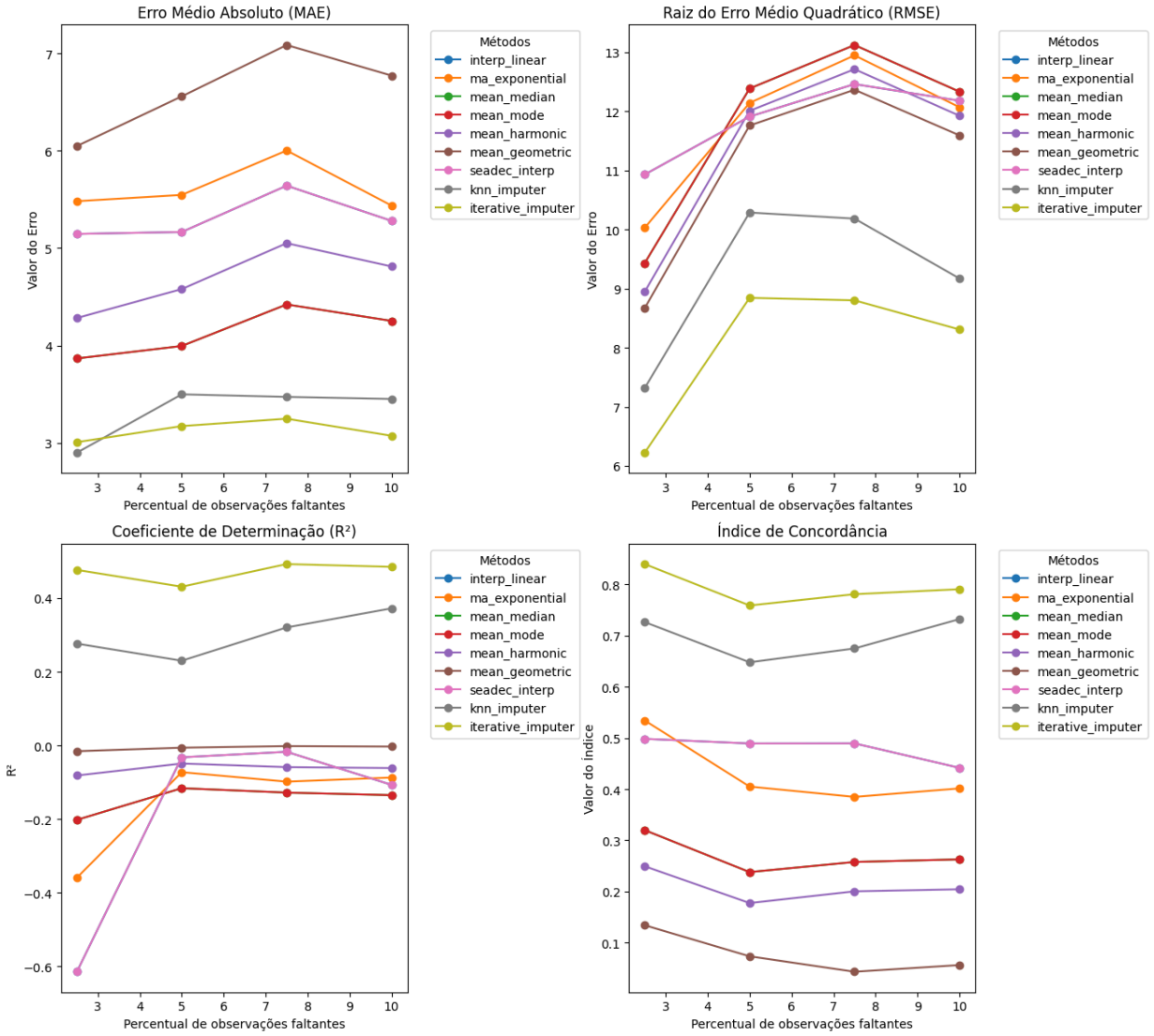
Fonte: Autor (2024).

5.3 Análise - Período Seco

5.3.1 Conjunto de dados A

Nos resultados para o período seco do Conjunto de Dados A, os gráficos apresentados na Figura 15 comparam os valores imputados com os valores originais para diferentes percentuais de dados faltantes (2,5%, 5%, 7,5% e 10%). Para cada combinação de método de imputação e percentual de dados faltantes, as métricas de desempenho, como MAE, RMSE, R^2 e Índice de Concordância, foram calculadas.

Figura 15. Desempenho dos Métodos de Imputação no Conjunto de Dados A - Período Seco.



Fonte: Autor (2024).

A Tabela 4 apresenta os valores numéricos das métricas de desempenho (MAE, RMSE, R² e Índice de Concordância) para cada método de imputação no Conjunto de Dados A, considerando diferentes percentuais de dados faltantes (2,5%, 5%, 7,5% e 10%) durante o período seco. Esses valores complementam os resultados apresentados na Figura 14, permitindo uma análise mais detalhada e precisa do desempenho dos métodos de imputação nas condições do período seco.

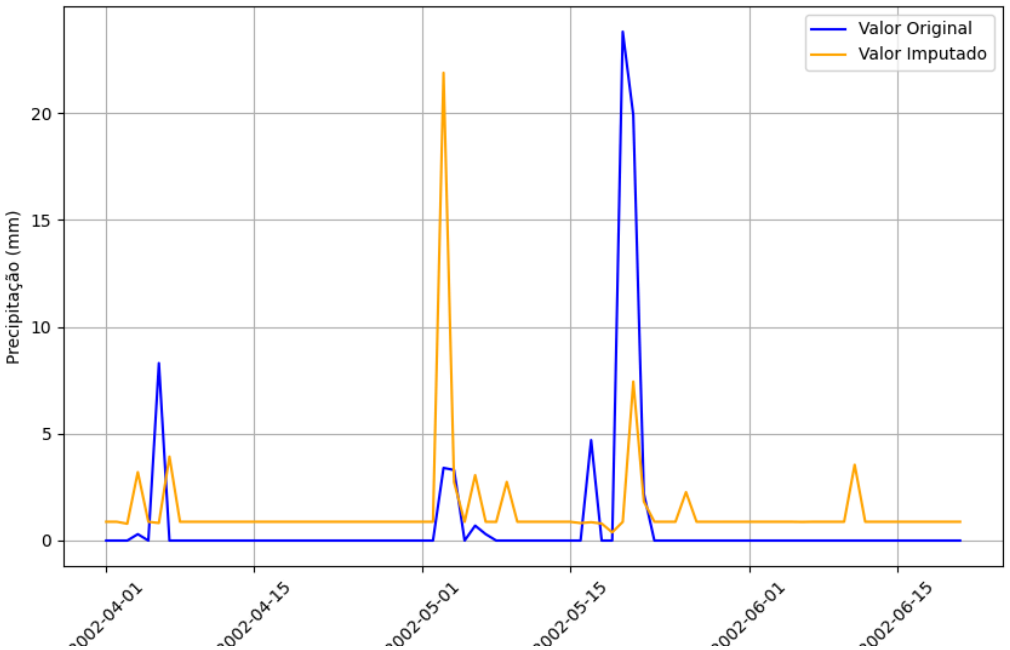
Tabela 4. Métricas de Desempenho dos Métodos de Imputação para o Conjunto de Dados A - Período Seco.

Método	Percentual Faltante (%)	MAE	RMSE	R2	Índice de Concordância
interp_linear	2,5	5,146	10,930	-0,614	0,498
ma_exponential	2,5	5,482	10,029	-0,359	0,534
mean_median	2,5	3,866	9,431	-0,202	0,320
mean_mode	2,5	3,866	9,431	-0,202	0,320
mean_harmonic	2,5	4,282	8,947	-0,082	0,249
mean_geometric	2,5	6,049	8,670	-0,016	0,134
seadec_interp	2,5	5,146	10,930	-0,614	0,498
knn_imputer	2,5	2,898	7,317	0,277	0,726
iterative_imputer	2,5	3,007	6,223	0,477	0,839
interp_linear	5,0	5,166	11,911	-0,032	0,489
ma_exponential	5,0	5,547	12,142	-0,073	0,405
mean_median	5,0	3,995	12,386	-0,116	0,238
mean_mode	5,0	3,995	12,386	-0,116	0,238
mean_harmonic	5,0	4,579	12,008	-0,049	0,178
mean_geometric	5,0	6,559	11,760	-0,006	0,074
seadec_interp	5,0	5,166	11,911	-0,032	0,489
knn_imputer	5,0	3,498	10,287	0,230	0,648
iterative_imputer	5,0	3,172	8,844	0,431	0,759
interp_linear	7,5	5,642	12,461	-0,017	0,489
ma_exponential	7,5	6,003	12,947	-0,098	0,385
mean_median	7,5	4,421	13,122	-0,128	0,258
mean_mode	7,5	4,421	13,122	-0,128	0,258
mean_harmonic	7,5	5,051	12,713	-0,059	0,200
mean_geometric	7,5	7,087	12,365	-0,002	0,044
seadec_interp	7,5	5,642	12,461	-0,017	0,489
knn_imputer	7,5	3,471	10,184	0,321	0,675
iterative_imputer	7,5	3,248	8,802	0,492	0,781
interp_linear	10,0	5,280	12,180	-0,107	0,442
ma_exponential	10,0	5,437	12,070	-0,087	0,401
mean_median	10,0	4,252	12,333	-0,135	0,263
mean_mode	10,0	4,252	12,333	-0,135	0,263
mean_harmonic	10,0	4,812	11,926	-0,061	0,205
mean_geometric	10,0	6,773	11,593	-0,003	0,057
seadec_interp	10,0	5,280	12,180	-0,107	0,442
knn_imputer	10,0	3,450	9,172	0,372	0,732
iterative_imputer	10,0	3,072	8,308	0,485	0,790

Fonte: Autor (2024).

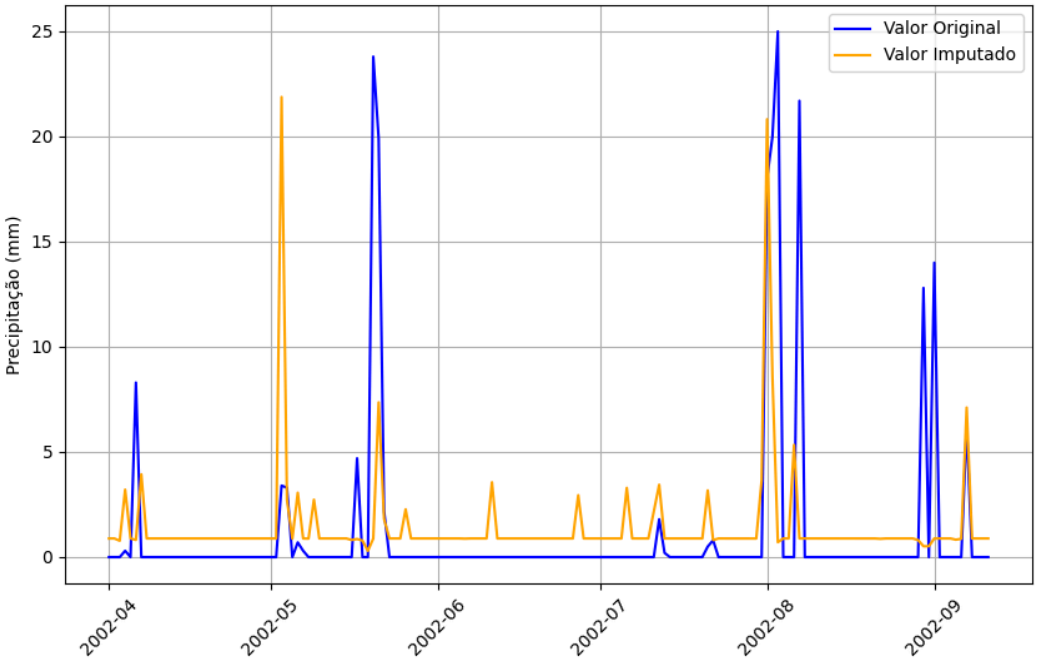
No período seco, os métodos de imputação apresentaram um desempenho mais estável, com um aumento gradual nos erros à medida que o percentual de dados faltantes crescia. A menor variabilidade dos dados nesse período contribuiu para uma imputação mais eficaz. As Figuras 16 a 19 ilustram os valores imputados pelo método Iterative Imputer, comparados aos valores originais, destacando a precisão da imputação em diferentes cenários de dados faltantes.

Figura 16. Período Seco, Conjunto de Dados A, 2.5% de dados faltantes - Precipitação Observada vs. Imputada



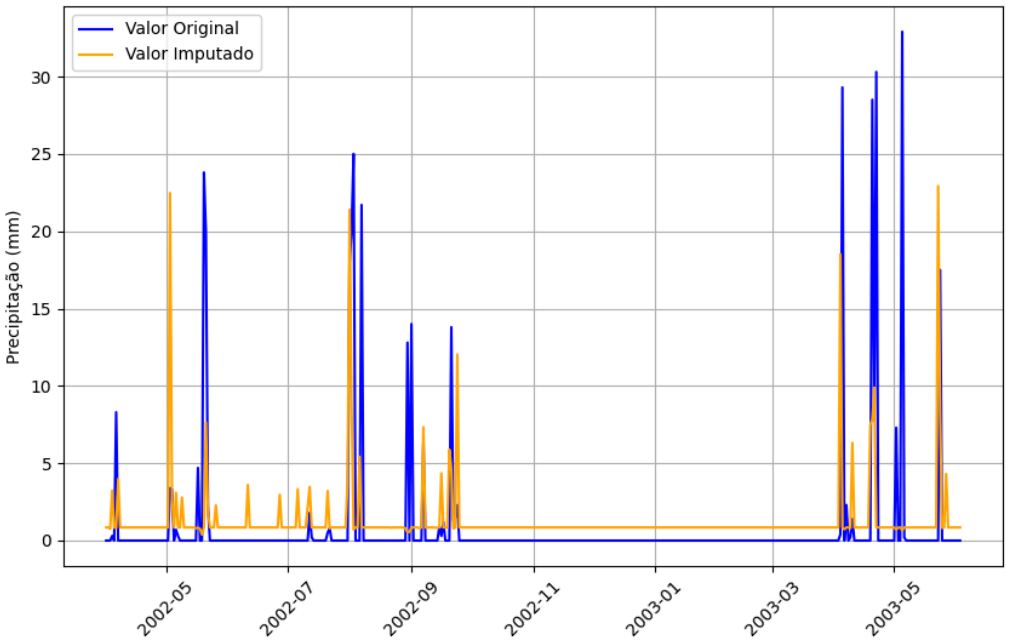
Fonte: Autor (2024).

Figura 17. Período Seco, Conjunto de Dados A, 5.0% de dados faltantes - Precipitação Observada vs. Imputada.



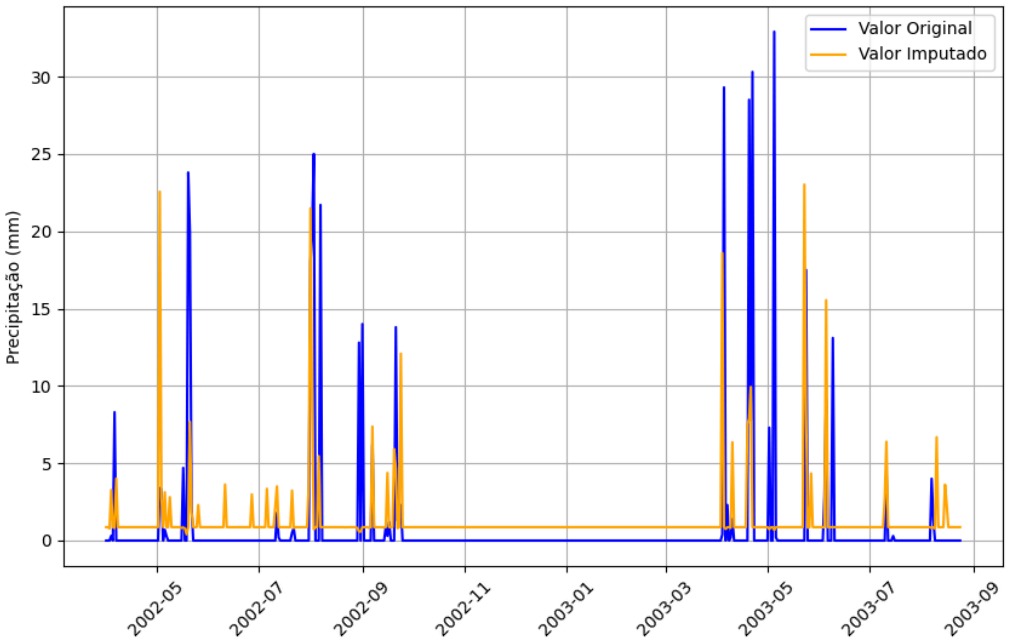
Fonte: Autor (2024).

Figura 18. Período Seco, Conjunto de Dados A, 7.5% de dados faltantes - Precipitação Observada vs. Imputada



Fonte: Autor (2024).

Figura 19. Período Seco, Conjunto de Dados A, 10.0% de dados faltantes - Precipitação Observada vs. Imputada

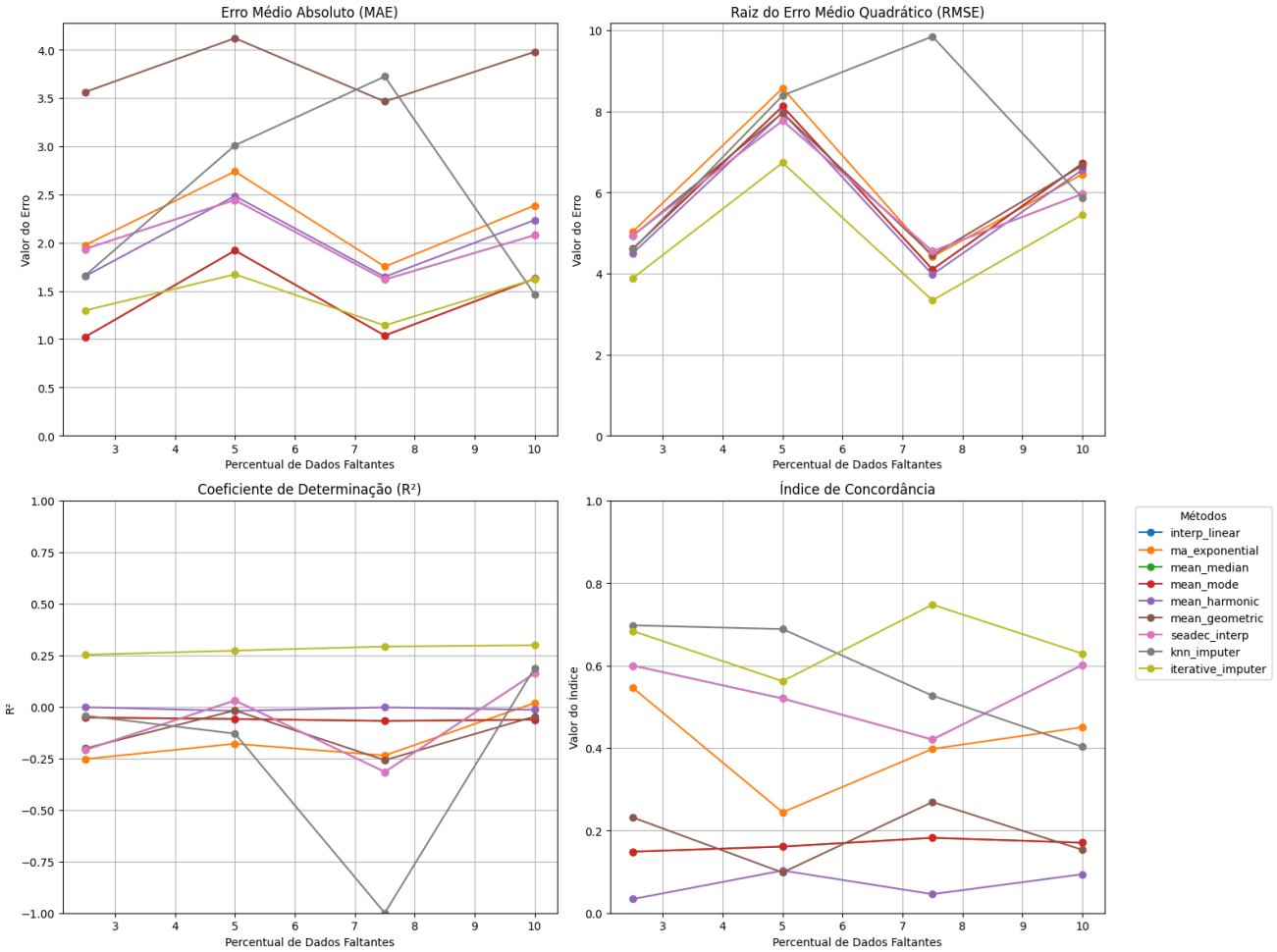


Fonte: Autor (2024).

5.3.2 Conjunto de dados B

Para o período seco no Conjunto de Dados B, os resultados, apresentados na Figura 20, mostram o comportamento das diferentes técnicas de imputação com os percentuais de dados faltantes (2,5%, 5%, 7,5% e 10%). Como no Conjunto de Dados A, foram calculadas as métricas de desempenho, incluindo MAE, RMSE, R^2 e Índice de Concordância.

Figura 20. Desempenho dos Métodos de Imputação no Conjunto de Dados B - Período Seco.



Fonte: Autor (2024).

A Tabela 5 apresenta os valores das métricas de desempenho (MAE, RMSE, R^2 e Índice de Concordância) para os métodos de imputação no Conjunto de Dados B, considerando diferentes percentuais de dados faltantes (2,5%, 5%, 7,5% e 10%) no período seco. Esses valores complementam os resultados visualizados na Figura 18, permitindo uma análise detalhada da precisão dos métodos nas condições do conjunto B.

Tabela 5. Métricas de Desempenho dos Métodos de Imputação para o Conjunto de Dados B - Período Seco.

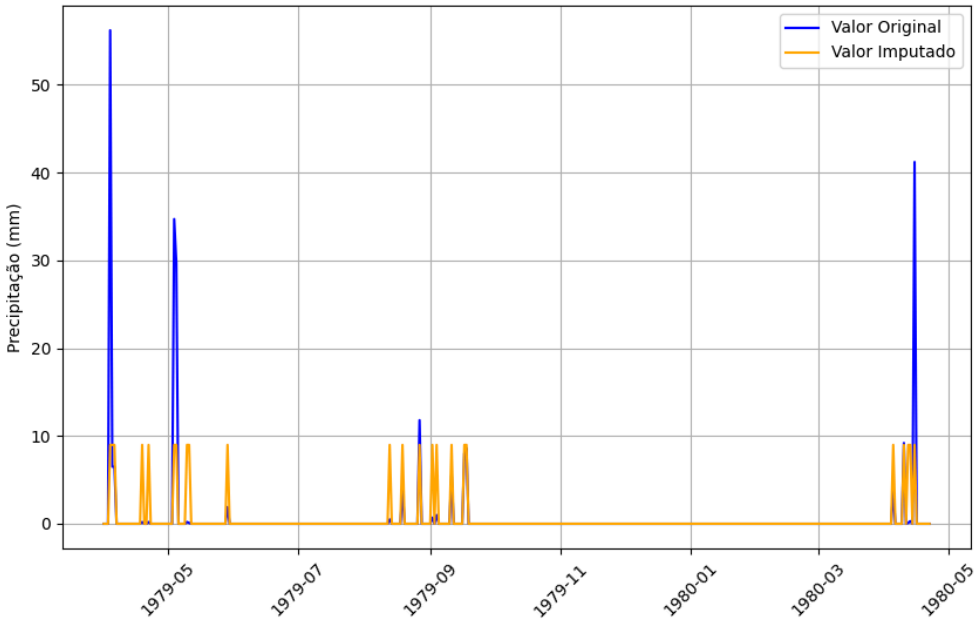
Método	Percentual Faltante (%)	MAE	RMSE	R2	Índice de Concordância
interp_linear	2,5	1,934	4,940	-0,209	0,600
ma_exponential	2,5	1,974	5,029	-0,253	0,546
mean_median	2,5	1,023	4,607	-0,052	0,149
mean_mode	2,5	1,023	4,607	-0,052	0,149
mean_harmonic	2,5	1,653	4,496	-0,002	0,034
mean_geometric	2,5	3,561	4,925	-0,202	0,232
seadec_interp	2,5	1,934	4,940	-0,209	0,600
knn_imputer	2,5	1,659	4,587	-0,043	0,698
iterative_imputer	2,5	1,297	3,885	0,252	0,684
interp_linear	5,0	2,443	7,767	0,031	0,520
ma_exponential	5,0	2,738	8,567	-0,179	0,244
mean_median	5,0	1,918	8,120	-0,059	0,161
mean_mode	5,0	1,918	8,120	-0,059	0,161
mean_harmonic	5,0	2,483	7,967	-0,020	0,103
mean_geometric	5,0	4,119	7,959	-0,018	0,098
seadec_interp	5,0	2,443	7,767	0,031	0,520
knn_imputer	5,0	3,008	8,387	-0,130	0,688
iterative_imputer	5,0	1,672	6,730	0,272	0,562
interp_linear	7,5	1,619	4,550	-0,316	0,421
ma_exponential	7,5	1,751	4,411	-0,236	0,398
mean_median	7,5	1,038	4,101	-0,068	0,182
mean_mode	7,5	1,038	4,101	-0,068	0,182
mean_harmonic	7,5	1,646	3,972	-0,003	0,046
mean_geometric	7,5	3,462	4,451	-0,259	0,269
seadec_interp	7,5	1,619	4,550	-0,316	0,421
knn_imputer	7,5	3,721	9,843	-5,157	0,527
iterative_imputer	7,5	1,142	3,337	0,292	0,748
interp_linear	10,0	2,077	5,959	0,162	0,601
ma_exponential	10,0	2,385	6,447	0,019	0,450
mean_median	10,0	1,627	6,709	-0,062	0,170
mean_mode	10,0	1,627	6,709	-0,062	0,170
mean_harmonic	10,0	2,234	6,556	-0,015	0,094
mean_geometric	10,0	3,977	6,661	-0,047	0,153
seadec_interp	10,0	2,077	5,959	0,162	0,601
knn_imputer	10,0	1,465	5,868	0,187	0,403
iterative_imputer	10,0	1,622	5,451	0,299	0,629

Fonte: Autor (2024).

No conjunto de dados B, o desempenho dos métodos de imputação no período seco foi semelhante ao observado no conjunto de dados A, com um aumento gradual nos erros à

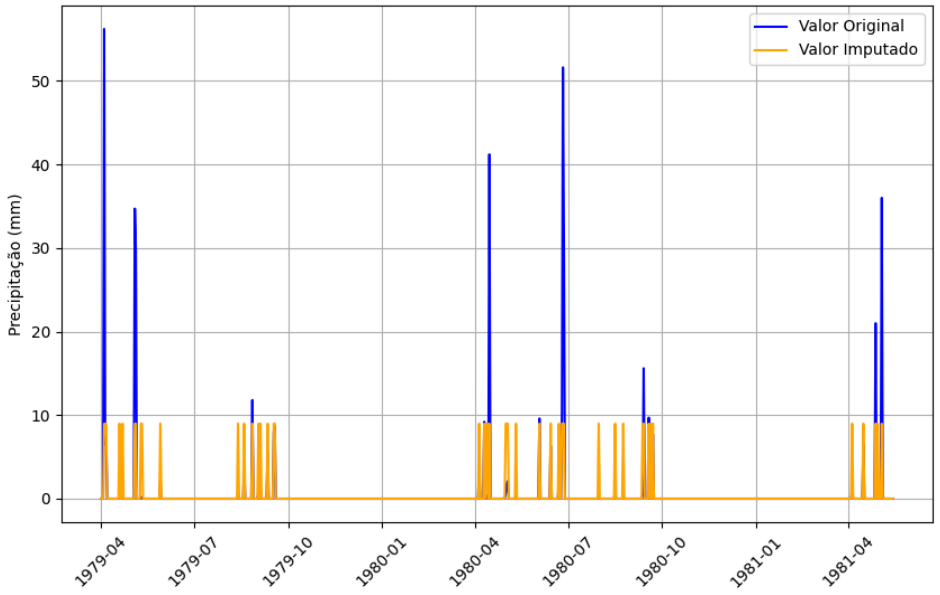
medida que o percentual de dados faltantes crescia. A baixa variabilidade dos dados nesse período também contribuiu para uma imputação mais eficaz. As Figuras 21 a 24, análogas às Figuras 16 a 19 do conjunto A, ilustram os valores imputados pelo método Iterative Imputer em comparação com os valores originais, evidenciando a precisão da imputação em diferentes cenários de dados ausentes.

Figura 21. Período Seco, Conjunto de Dados B, 2.5% de dados faltantes - Precipitação Observada vs. Imputada



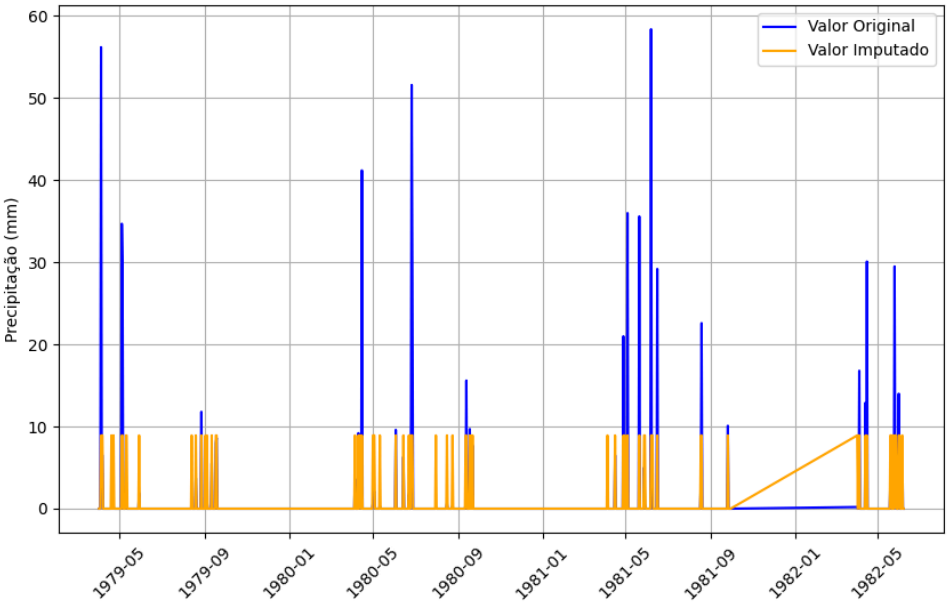
Fonte: Autor (2024).

Figura 22. Período Seco, Conjunto de Dados B, 5.0% de dados faltantes - Precipitação Observada vs. Imputada



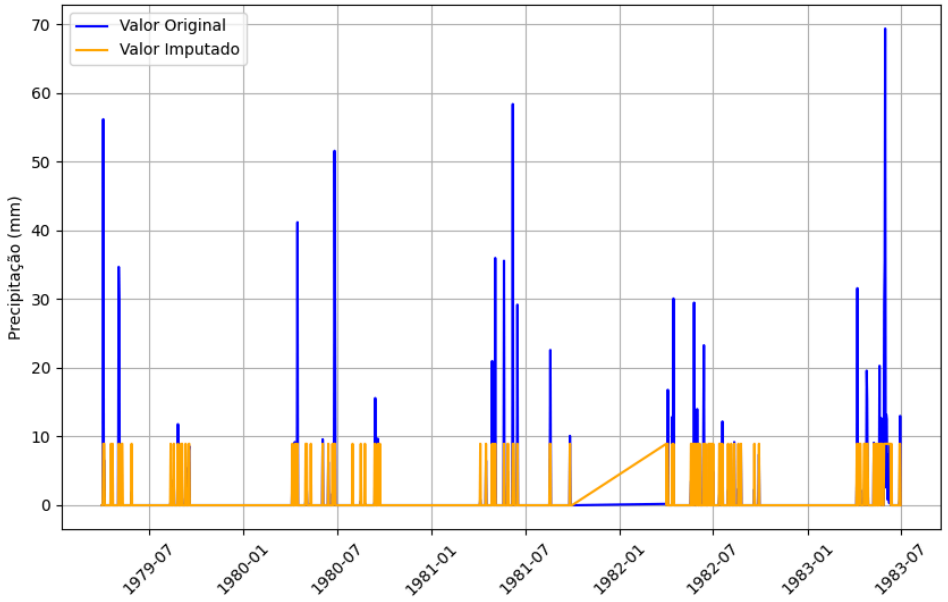
Fonte: Autor (2024).

Figura 23. Período Seco, Conjunto de Dados B, 7.5% de dados faltantes - Precipitação Observada vs. Imputada



Fonte: Autor (2024).

Figura 24. Período Seco, Conjunto de Dados B, 10.0% de dados faltantes - Precipitação Observada vs. Imputada



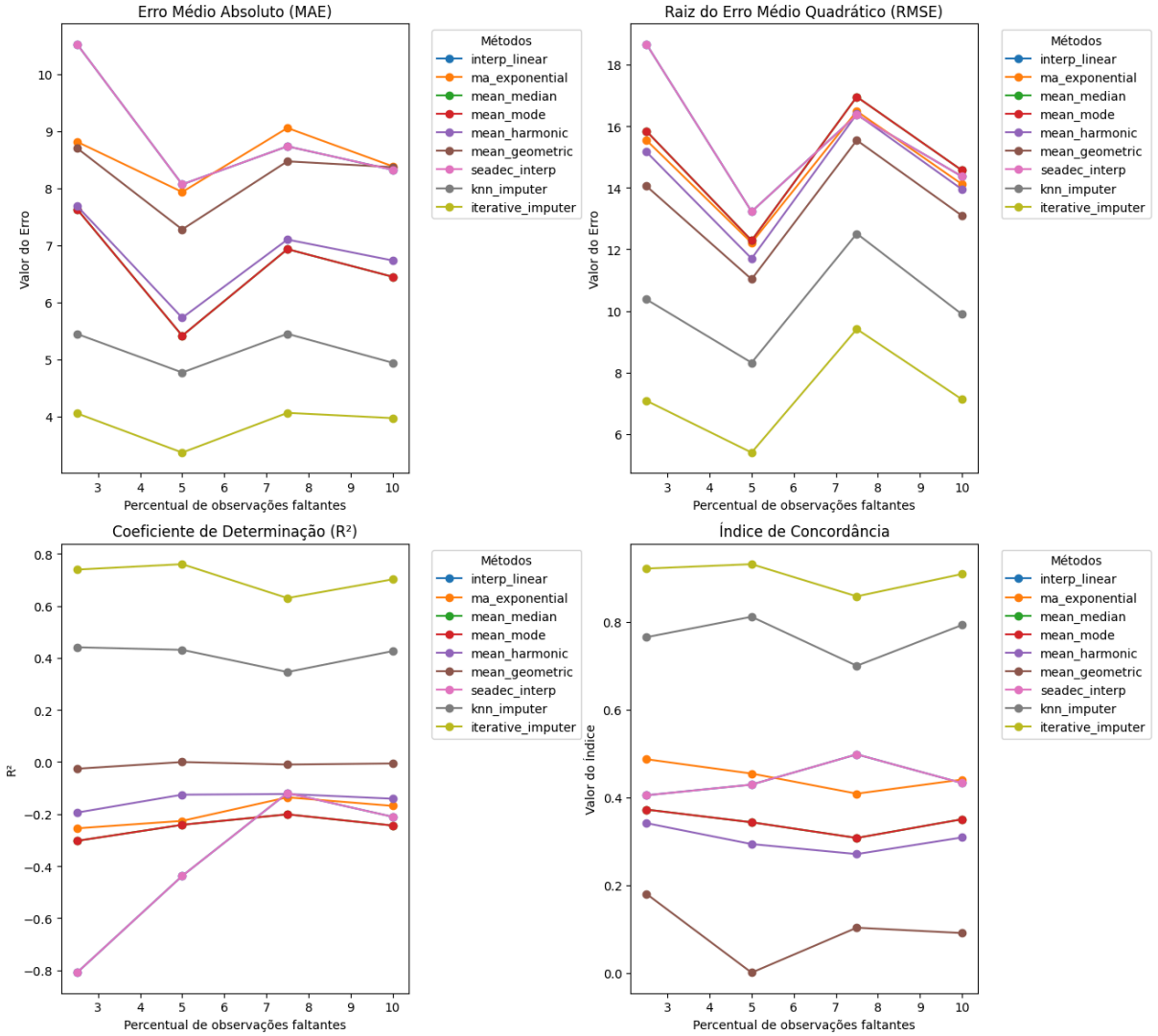
Fonte: Autor (2024).

5.4 Análise - Período Chuvoso

5.4.1 Conjunto de dados A

Para o período chuvoso do Conjunto de Dados A, os gráficos na Figura 25 revelam o comportamento dos métodos de imputação em face da maior variabilidade dos dados e da presença de outliers, como chuvas extremas. Como no período seco e análise global, foram calculadas as métricas MAE, RMSE, R^2 e Índice de Concordância para cada combinação de método e percentual de dados faltantes.

Figura 25. Desempenho dos Métodos de Imputação no Conjunto de Dados A -Período Chuvoso.



Fonte: Autor (2024).

A Tabela 6 apresenta os valores numéricos das métricas de desempenho (MAE, RMSE, R^2 e Índice de Concordância) para cada método de imputação no Conjunto de Dados A, considerando diferentes percentuais de dados faltantes (2,5%, 5%, 7,5% e 10%) durante o período chuvoso. Esses valores complementam os resultados apresentados na Figura 24, permitindo uma análise mais detalhada e precisa do desempenho dos métodos de imputação nas condições do período chuvoso.

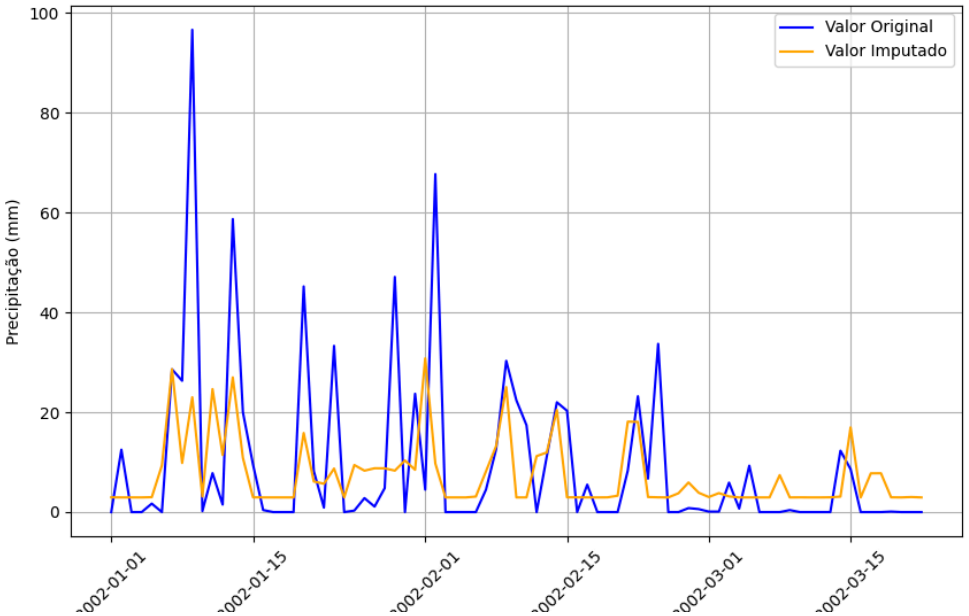
Tabela 6. Métricas de Desempenho dos Métodos de Imputação para o Dados A -Período Chuvoso.

Método	Percentual Faltante (%)	MAE	RMSE	R2	Índice de Concordância
interp_linear	2,5	10,528	18,662	-0,809	0,405
ma_exponential	2,5	8,816	15,542	-0,255	0,487
mean_median	2,5	7,636	15,837	-0,303	0,372
mean_mode	2,5	7,636	15,837	-0,303	0,372
mean_harmonic	2,5	7,691	15,163	-0,194	0,341
mean_geometric	2,5	8,705	14,053	-0,026	0,180
seadec_interp	2,5	10,528	18,662	-0,809	0,405
knn_imputer	2,5	5,447	10,375	0,441	0,765
iterative_imputer	2,5	4,056	7,086	0,739	0,921
interp_linear	5,0	8,069	13,224	-0,438	0,429
ma_exponential	5,0	7,935	12,213	-0,226	0,454
mean_median	5,0	5,417	12,287	-0,241	0,343
mean_mode	5,0	5,417	12,287	-0,241	0,343
mean_harmonic	5,0	5,731	11,701	-0,125	0,294
mean_geometric	5,0	7,279	11,029	-3,87E-07	0,001
seadec_interp	5,0	8,069	13,224	-0,438	0,429
knn_imputer	5,0	4,772	8,320	0,431	0,811
iterative_imputer	5,0	3,368	5,403	0,760	0,931
interp_linear	7,5	8,736	16,363	-0,119	0,498
ma_exponential	7,5	9,058	16,483	-0,136	0,408
mean_median	7,5	6,934	16,950	-0,201	0,308
mean_mode	7,5	6,934	16,950	-0,201	0,308
mean_harmonic	7,5	7,102	16,389	-0,123	0,271
mean_geometric	7,5	8,473	15,540	-0,010	0,103
seadec_interp	7,5	8,736	16,363	-0,119	0,498
knn_imputer	7,5	5,450	12,510	0,346	0,700
iterative_imputer	7,5	4,064	9,411	0,630	0,858
interp_linear	10,0	8,324	14,364	-0,210	0,432
ma_exponential	10,0	8,387	14,114	-0,169	0,440
mean_median	10,0	6,450	14,563	-0,244	0,350
mean_mode	10,0	6,450	14,563	-0,244	0,350
mean_harmonic	10,0	6,736	13,947	-0,141	0,309
mean_geometric	10,0	8,371	13,092	-0,006	0,091
seadec_interp	10,0	8,324	14,364	-0,210	0,432
knn_imputer	10,0	4,942	9,892	0,426	0,792
iterative_imputer	10,0	3,972	7,130	0,702	0,909

Fonte: Autor (2024).

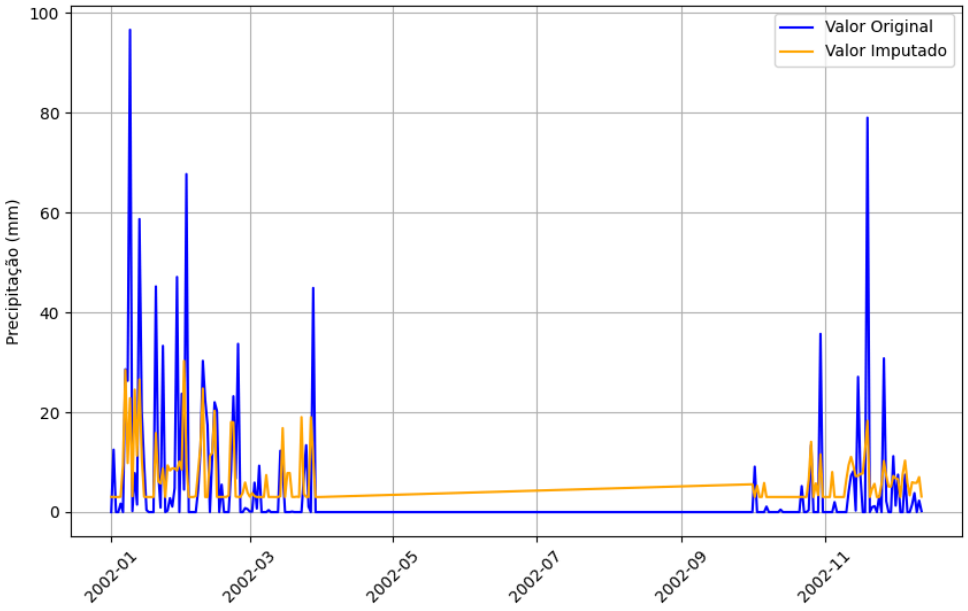
Durante o período chuvoso, os métodos de imputação mostraram um desempenho mais variável, com os erros aumentando de forma mais acentuada à medida que o percentual de dados faltantes aumentava. A maior variabilidade dos dados nesse período contribuiu para uma maior dificuldade na imputação. As Figuras 26 a 29 ilustram os valores imputados pelo método Iterative Imputer, comparados aos valores originais, destacando a precisão da imputação em diferentes cenários de dados faltantes.

Figura 26. Período Chuvoso, Conjunto de Dados A, 2.5% de dados faltantes - Precipitação Observada vs. Imputada



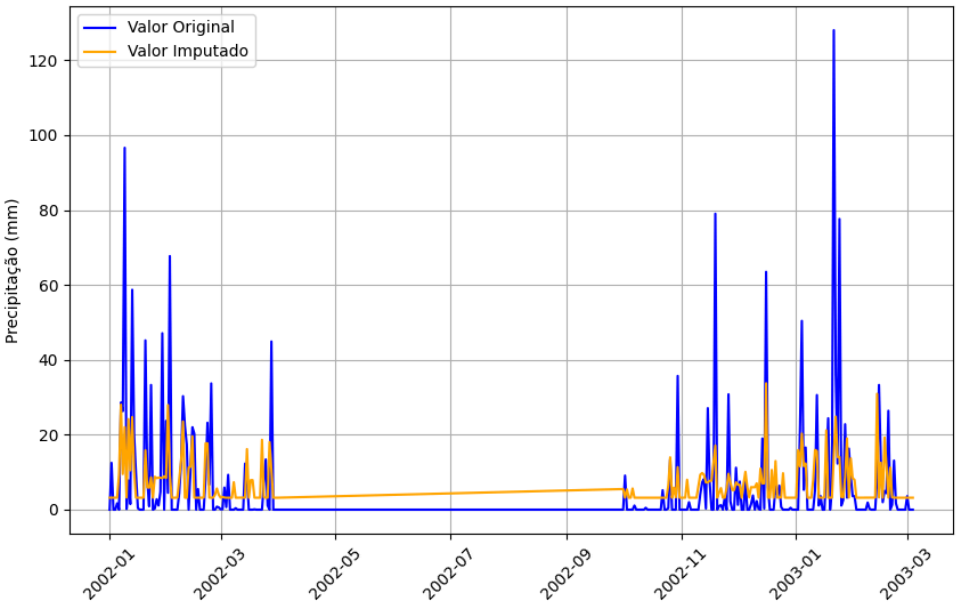
Fonte: Autor (2024).

Figura 27. Período Chuvoso, Conjunto de Dados A, 5.0% de dados faltantes - Precipitação Observada vs. Imputada



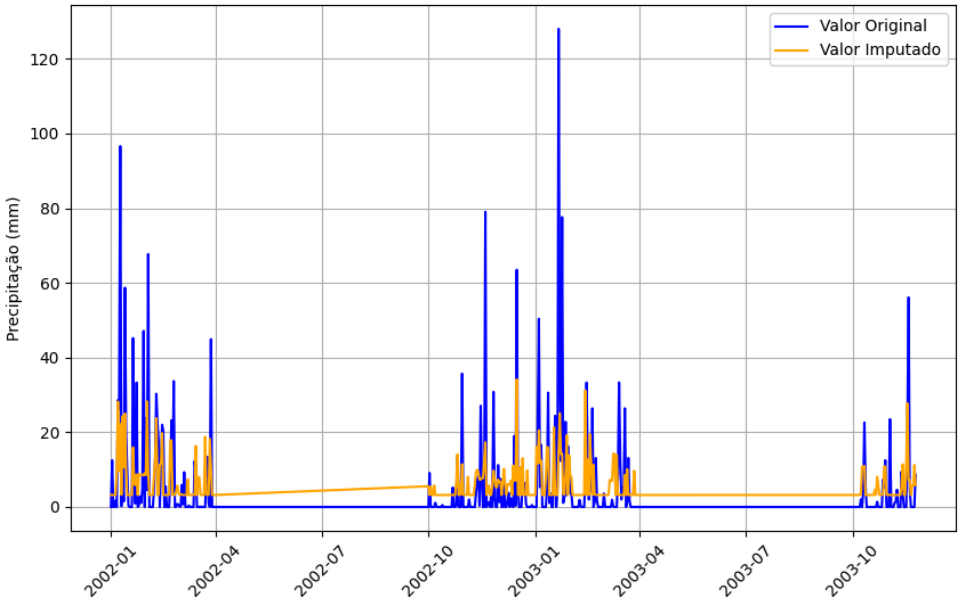
Fonte: Autor (2024).

Figura 28. Período Chuvoso, Conjunto de Dados A, 7.5% de dados faltantes - Precipitação Observada vs. Imputada



Fonte: Autor (2024).

Figura 29. Período Chuvoso, Conjunto de Dados A, 10.0% de dados faltantes - Precipitação Observada vs. Imputada

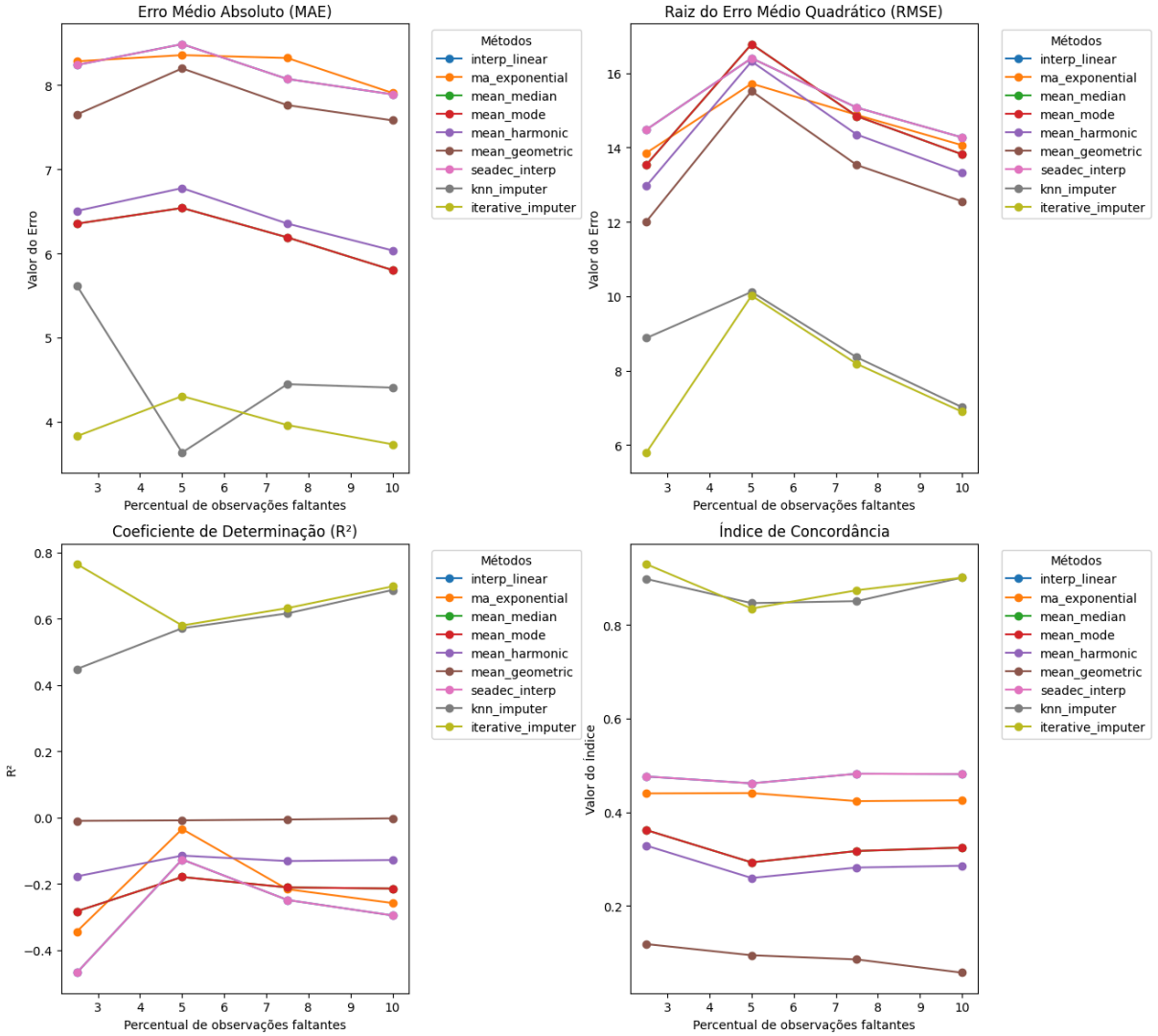


Fonte: Autor (2024).

5.4.2 Conjunto de dados B

Nos resultados para o período chuvoso do Conjunto de Dados B, os gráficos apresentados na Figura 30 comparam os valores imputados com os valores originais para diferentes percentuais de dados faltantes (2,5%, 5%, 7,5% e 10%). Para cada combinação de método de imputação e percentual de dados faltantes, as métricas de desempenho, como MAE, RMSE, R^2 e Índice de Concordância, foram calculadas.

Figura 30. Desempenho dos Métodos de Imputação no Conjunto de Dados B -Período Chuvoso.



Fonte: Autor (2024).

A Tabela 7 apresenta os valores numéricos das métricas de desempenho (MAE, RMSE, R² e Índice de Concordância) para cada método de imputação no Conjunto de Dados B, considerando diferentes percentuais de dados faltantes (2,5%, 5%, 7,5% e 10%) durante o período chuvoso. Esses valores complementam os resultados apresentados na Figura 29, permitindo uma análise mais detalhada e precisa do desempenho dos métodos de imputação nas condições do período chuvoso.

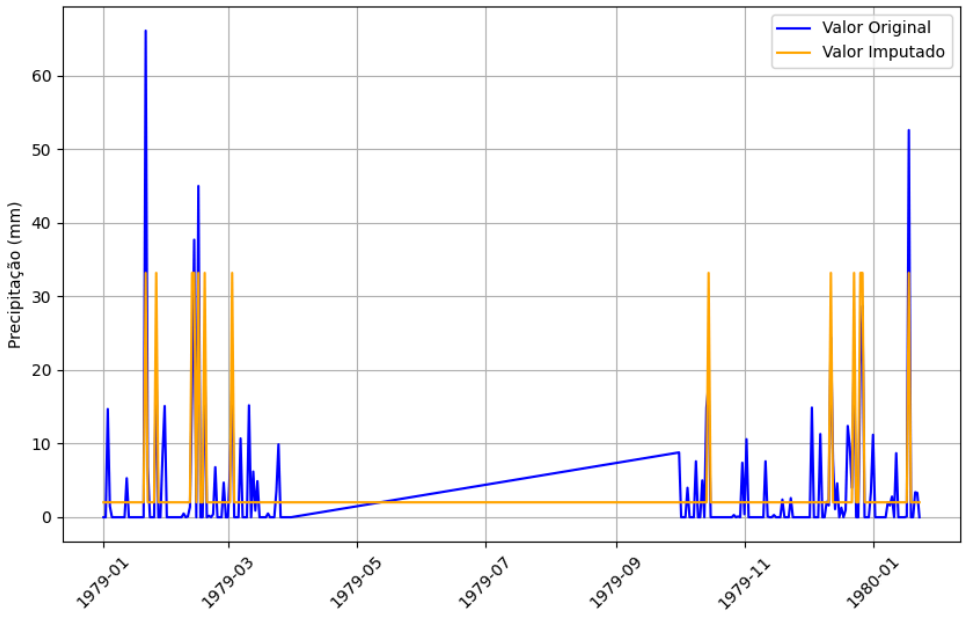
Tabela 7. Métricas de Desempenho dos Métodos de Imputação para o Dados B -Período Chuvoso.

Método	Percentual Faltante (%)	MAE	RMSE	R2	Índice de Concordância
interp_linear	2,5	8,237	14,478	-0,468	0,476
ma_exponential	2,5	8,281	13,851	-0,344	0,440
mean_median	2,5	6,353	13,534	-0,283	0,362
mean_mode	2,5	6,353	13,534	-0,283	0,362
mean_harmonic	2,5	6,504	12,966	-0,177	0,328
mean_geometric	2,5	7,649	12,008	-0,010	0,118
seadec_interp	2,5	8,237	14,478	-0,468	0,476
knn_imputer	2,5	5,616	8,875	0,448	0,898
iterative_imputer	2,5	3,828	5,797	0,765	0,930
interp_linear	5,0	8,486	16,397	-0,126	0,461
ma_exponential	5,0	8,355	15,714	-0,035	0,441
mean_median	5,0	6,540	16,776	-0,179	0,293
mean_mode	5,0	6,540	16,776	-0,179	0,293
mean_harmonic	5,0	6,775	16,311	-0,115	0,259
mean_geometric	5,0	8,197	15,513	-0,008	0,094
seadec_interp	5,0	8,486	16,397	-0,126	0,461
knn_imputer	5,0	3,634	10,115	0,571	0,846
iterative_imputer	5,0	4,305	10,015	0,580	0,835
interp_linear	7,5	8,072	15,070	-0,248	0,482
ma_exponential	7,5	8,320	14,873	-0,216	0,423
mean_median	7,5	6,190	14,840	-0,211	0,317
mean_mode	7,5	6,190	14,840	-0,211	0,317
mean_harmonic	7,5	6,354	14,344	-0,131	0,282
mean_geometric	7,5	7,761	13,527	-0,006	0,085
seadec_interp	7,5	8,072	15,070	-0,248	0,482
knn_imputer	7,5	4,447	8,357	0,616	0,851
iterative_imputer	7,5	3,961	8,181	0,632	0,874
interp_linear	10,0	7,888	14,268	-0,295	0,481
ma_exponential	10,0	7,905	14,060	-0,258	0,425
mean_median	10,0	5,801	13,814	-0,214	0,324
mean_mode	10,0	5,801	13,814	-0,214	0,324
mean_harmonic	10,0	6,034	13,315	-0,128	0,285
mean_geometric	10,0	7,580	12,551	-0,002	0,057
seadec_interp	10,0	7,888	14,268	-0,295	0,481
knn_imputer	10,0	4,405	7,016	0,687	0,901
iterative_imputer	10,0	3,733	6,894	0,698	0,901

Fonte: Autor (2024).

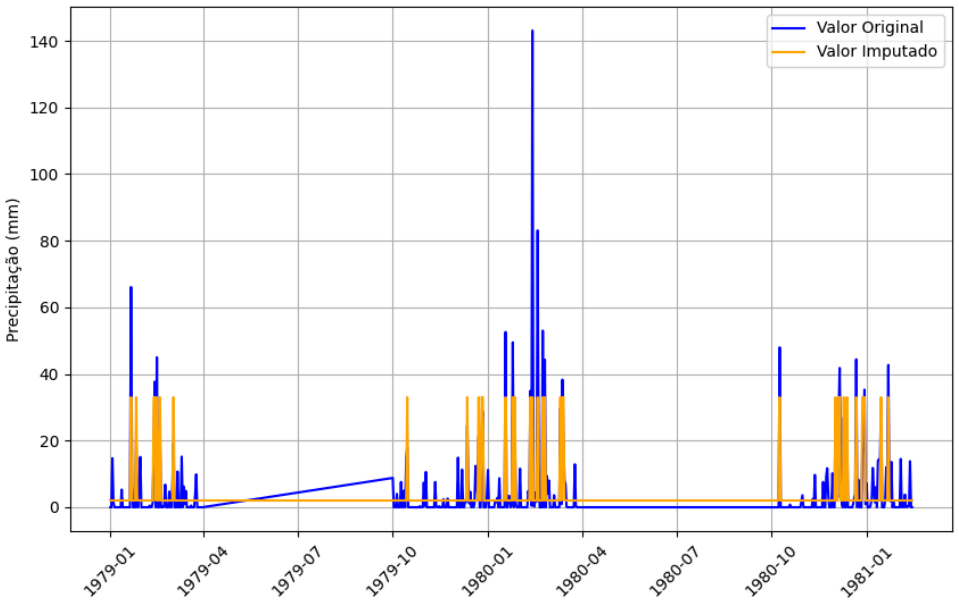
Durante o período chuvoso, os métodos de imputação apresentaram maior variação no desempenho, com os erros aumentando de forma mais significativa à medida que o percentual de dados faltantes se elevava. A maior variabilidade dos dados nesse período gerou desafios adicionais para a imputação. As Figuras 31 a 34 ilustram os valores imputados pelo método Iterative Imputer, comparados aos valores originais, destacando a precisão da imputação em diferentes cenários de dados faltantes.

Figura 31. Período Chuvoso, Conjunto de Dados A, 2.5% de dados faltantes - Precipitação Observada vs. Imputada



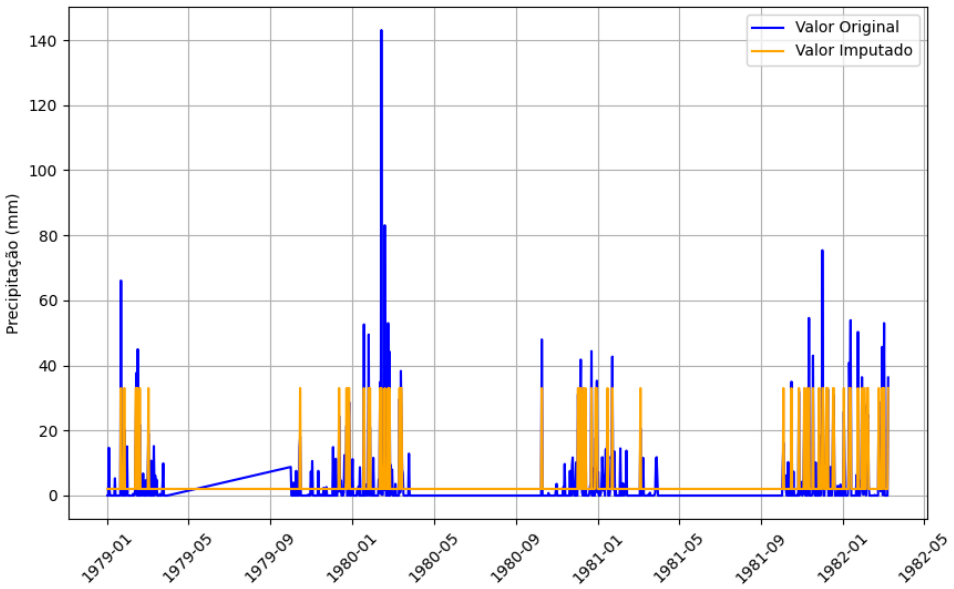
Fonte: Autor (2024).

Figura 32. Período Chuvoso, Conjunto de Dados A, 5.0% de dados faltantes - Precipitação Observada vs. Imputada



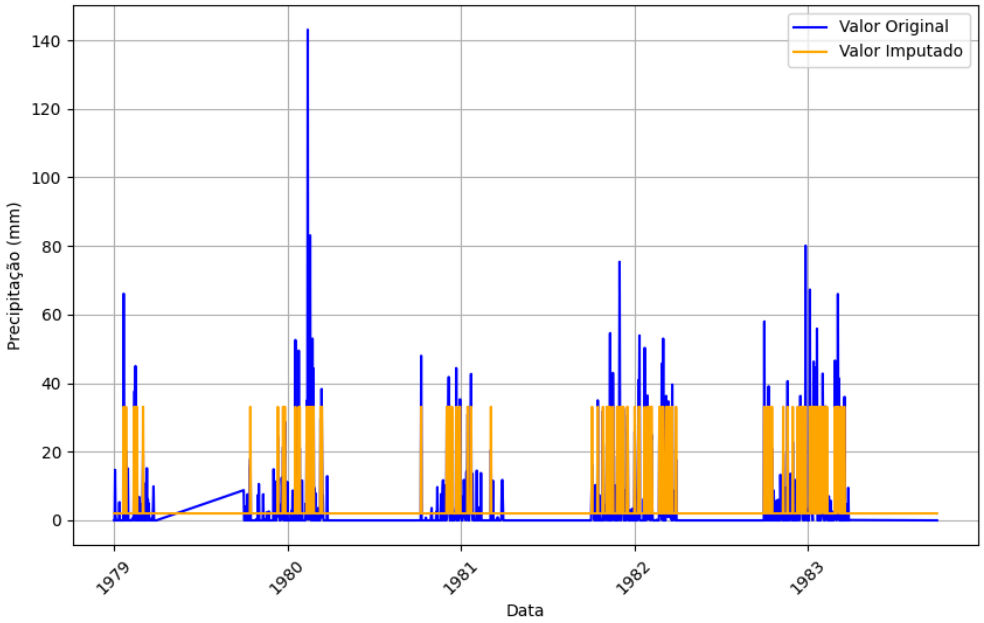
Fonte: Autor (2024).

Figura 33. Período Chuvoso, Conjunto de Dados A, 7.5% de dados faltantes - Precipitação Observada vs. Imputada



Fonte: Autor (2024).

Figura 34. Período Chuvoso, Conjunto de Dados A, 10.0% de dados faltantes - Precipitação Observada vs. Imputada



Fonte: Autor (2024).

5.5 Discussão

5.5.1 Melhores algoritmos

Esta seção discute os resultados da avaliação dos algoritmos de imputação utilizados neste trabalho, com ênfase nos métodos que apresentaram o melhor desempenho. Conforme mostrado nas seções anteriores, os algoritmos baseados em aprendizado de máquina, em particular o KNNImputer e o IterativeImputer, destacaram-se como as abordagens mais eficazes para lidar com os dados faltantes no conjunto de dados analisado.

A superioridade desses métodos é evidenciada por diversas análises. Primeiramente, as figuras e gráficos apresentados ao longo deste trabalho ilustram a capacidade superior do KNNImputer e do IterativeImputer em preencher as lacunas nos dados de forma mais precisa e consistente, em comparação com métodos mais simples, como a imputação pela média ou mediana. Observa-se que as séries temporais imputadas por esses métodos mantêm uma aderência considerável aos dados originais, minimizando distorções e preservando padrões importantes nas séries temporais.

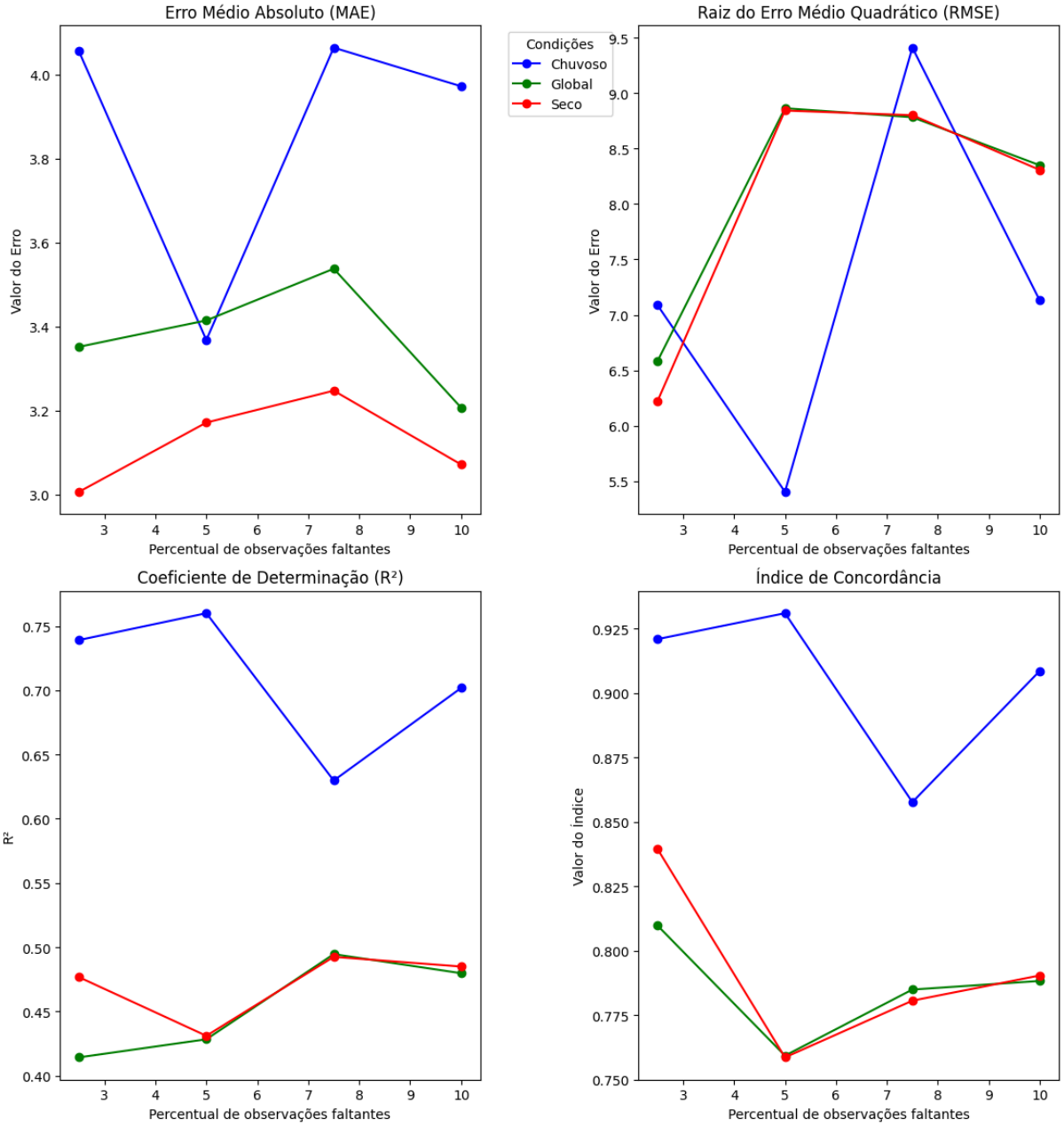
Além da análise visual, as métricas quantitativas de avaliação reforçam a superioridade do KNNImputer e do IterativeImputer. Os resultados indicam que esses algoritmos apresentaram os menores valores em diversas métricas de desempenho, como RMSE, MAE e MAPE. Por exemplo, o KNNImputer obteve um RMSE significativamente inferior ao dos métodos univariados, enquanto o IterativeImputer alcançou um desempenho igualmente superior em outras métricas.

A eficácia do KNNImputer pode ser atribuída à sua capacidade de aproveitar as informações dos vizinhos mais próximos para estimar os valores faltantes, levando em consideração a correlação entre as variáveis. O IterativeImputer, por sua vez, utiliza um modelo de regressão para prever os valores ausentes, aprimorando as estimativas de maneira iterativa. Ambos os métodos demonstraram uma capacidade superior de capturar relações complexas e não lineares nos dados, justificando seu desempenho superior em relação aos métodos mais simples.

É importante observar que a escolha entre KNNImputer e IterativeImputer pode depender das características específicas do conjunto de dados e dos objetivos da análise. No entanto, os resultados obtidos neste estudo indicam que ambos os métodos são robustos e eficazes para a imputação de dados faltantes, superando as abordagens tradicionais e proporcionando uma análise mais precisa e confiável dos dados. Futuros trabalhos poderão explorar combinações desses métodos ou investigar a utilização de outros algoritmos de aprendizado de máquina para aprimorar ainda mais a qualidade da imputação.

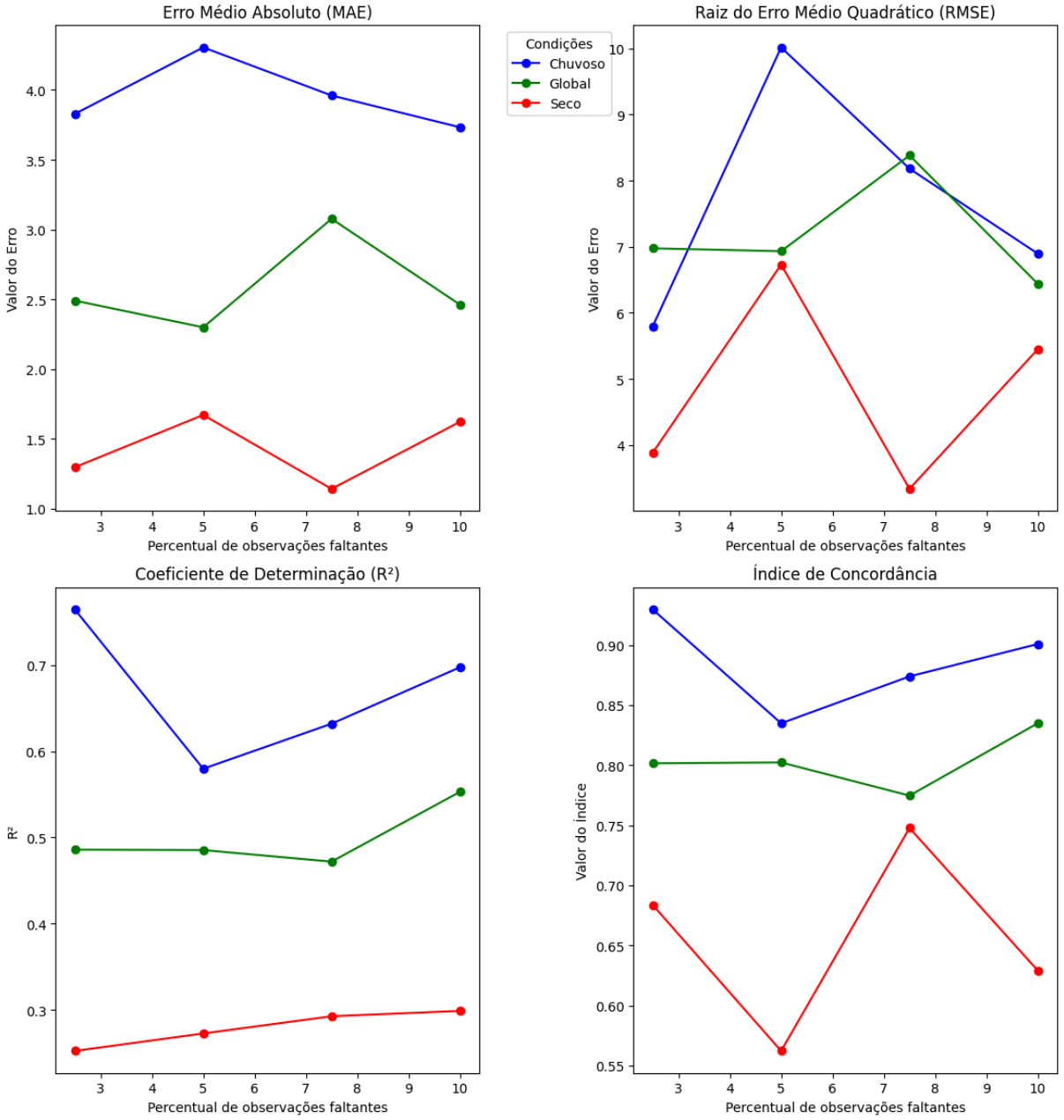
5.5.2 Global x Seco x Chuvoso

Figura 35. Condições Global, Seco e Chuvoso, Conjunto de Dados A - Desempenho do Iterative Imputer para Diferentes Percentuais de Dados Faltantes



Fonte: Autor (2024).

Figura 36. Condições Global, Seco e Chuvoso, Conjunto de Dados B - Desempenho do Iterative Imputer para Diferentes Percentuais de Dados Faltantes.



Fonte: Autor (2024).

A análise do desempenho do método de imputação iterative imputer nas condições Global, Seco e Chuvoso para os conjuntos de dados A e B revela padrões específicos para cada métrica: Erro Médio Absoluto (MAE), Raiz do Erro Médio Quadrático (RMSE),

Coeficiente de Determinação (R^2) e Índice de Concordância, conforme ilustram as Figuras 34 e 35.

5.5.2.1 Erro Médio Absoluto (MAE)

No conjunto de dados A, observa-se que o período seco apresenta os menores valores de MAE, indicando maior precisão na imputação. Em contraste, a condição Chuvoso possui os maiores valores de MAE, refletindo maior dificuldade do iterative imputer em lidar com a variabilidade dessa condição. A condição Global ocupa uma posição intermediária entre Seco e Chuvoso.

No conjunto B, a tendência é semelhante, com a condição Seco apresentando os menores valores de MAE e a condição Chuvoso os maiores. No entanto, os valores de MAE são ligeiramente mais baixos no conjunto B em comparação ao conjunto A, indicando uma imputação mais precisa para esse conjunto.

5.5.2.2 Raiz do Erro Médio Quadrático (RMSE)

A análise do RMSE confirma a tendência observada no MAE. No conjunto A, a condição Chuvoso apresenta valores de RMSE mais altos e variáveis, especialmente em percentuais maiores de dados faltantes. Já no conjunto B, o RMSE na condição Chuvoso é menor e menos variável do que no conjunto A, o que sugere que o iterative imputer teve mais facilidade em lidar com essa condição no conjunto B.

5.5.2.3 Coeficiente de Determinação (R^2)

O Coeficiente de Determinação (R^2) é maior no conjunto A, principalmente na condição Chuvoso para percentuais mais baixos de dados faltantes, o que indica que a imputação captura bem a variabilidade dos dados reais nessa condição. No conjunto B, o R^2 é mais baixo na condição Chuvoso, o que sugere uma abordagem mais conservadora e menos suscetível a variações.

5.5.2.4 Índice de Concordância

No conjunto A, o Índice de Concordância é mais elevado para a condição Chuvoso em percentuais baixos de dados faltantes, mas apresenta queda acentuada com o aumento do percentual de dados ausentes. No conjunto B, embora o Índice de Concordância também seja mais elevado na condição Chuvoso para percentuais mais baixos, seus valores são mais altos e consistentes, indicando maior estabilidade.

Conclui-se que o método iterative imputer apresenta um desempenho mais robusto na condição seca, onde as métricas indicam menor erro e maior estabilidade. Em contrapartida, a condição chuvosa representa um desafio maior para a imputação, com valores mais elevados de erro e sensibilidade ao aumento de dados faltantes, o que é possivelmente atribuído à maior variabilidade associada ao período chuvoso. A condição Global, como esperado, reflete uma média dos comportamentos observados nas condições Seco e Chuvoso, oferecendo um desempenho intermediário em todas as métricas.

6. CONCLUSÃO

Este trabalho investigou o uso de modelos de aprendizado de máquina, como o KNN Imputer e o Iterative Imputer, para preenchimento de falhas em séries de dados pluviométricos. A análise foi conduzida com dados da estação P16 (Itirapina-CRHEA, 2002-2019) e do INMET em São Carlos (1979-2023), aplicados a três cenários distintos: uma visão global, o período seco e o período chuvoso. Esse enfoque permitiu observar como esses modelos respondem a diferentes regimes de chuva e variações sazonais, proporcionando uma visão mais detalhada sobre a eficácia dos métodos de imputação.

Os resultados, expressos pelas métricas de avaliação MAE, RMSE, R^2 e Índice de Concordância (d), indicaram que os algoritmos de aprendizado de máquina superam métodos tradicionais de preenchimento, como a interpolação linear e médias móveis, especialmente em séries temporais complexas e com alta variabilidade, características comuns em dados pluviométricos (TROYANSKY et al., 2020). A capacidade desses modelos de capturar padrões não lineares e lidar com grandes volumes de dados foi essencial para alcançar resultados mais precisos e robustos, considerando também a interdependência entre variáveis climáticas.

Embora esses métodos apresentem desafios computacionais — como a necessidade de grandes volumes de dados para treinamento e o aumento do tempo de processamento, especialmente em séries temporais extensas (LATIF et al., 2023) —, o desempenho superior justifica sua aplicação no contexto deste estudo. A comparação com métodos tradicionais permitiu quantificar os ganhos do aprendizado de máquina, evidenciando seu potencial para aprimorar a qualidade dos dados pluviométricos disponíveis para pesquisas.

Assim, o presente estudo oferece contribuições importantes para o desenvolvimento de métodos mais robustos e eficazes no tratamento de dados incompletos, com potencial para aprimorar a gestão de recursos hídricos e o planejamento climático de longo prazo. A análise comparativa realizada fornece resultados valiosos que podem auxiliar pesquisadores e profissionais na escolha do método de imputação mais adequado para diferentes cenários e conjuntos de dados, melhorando a confiabilidade de estudos em climatologia e hidrologia no Brasil.

6.1 Trabalhos futuros

A pesquisa realizada neste trabalho abre caminho para diversas investigações futuras, com o objetivo de aprofundar a compreensão e o desenvolvimento de métodos para o preenchimento de falhas em dados pluviométricos. Algumas das principais direções para trabalhos futuros incluem:

1. Abordagens Híbridas: Explorar a combinação de diferentes métodos de imputação, criando abordagens híbridas que possam combinar as vantagens de cada técnica e, assim, potencialmente aprimorar os resultados. Um exemplo promissor seria a integração do Random Forest com métodos baseados em Redes Generativas Adversariais (GANs), como explorado por Ou et al. (2024).
2. Análise de Robustez em Diferentes Contextos: Avaliar o desempenho dos modelos propostos (KNN Imputer e Iterative Imputer) em diferentes regiões geográficas, biomas e condições climáticas, a fim de testar sua adaptabilidade e eficiência em contextos variados. Essa análise contribuirá para uma melhor compreensão da generalização dos resultados obtidos neste estudo.
3. Investigação de Outros Algoritmos: Investigar a aplicação de outros algoritmos de aprendizado de máquina, como o Random Forest, para a imputação de falhas em dados pluviométricos. Considerando o seu reconhecido potencial para lidar com dados complexos e ausentes (LI et al., 2023), essa linha de pesquisa pode revelar novas abordagens promissoras para o preenchimento de falhas.
4. Impacto na Previsão de Eventos Extremos: Analisar o impacto da imputação de falhas na previsão de eventos extremos, como secas e enchentes. Essa investigação deve considerar a importância de preservar a magnitude e a frequência desses eventos nas séries temporais, garantindo que a imputação não distorça as características essenciais dos eventos extremos.
5. Reconstrução de Séries Temporais de Longo Prazo: Investigar a aplicação dos métodos estudados (KNN Imputer, Iterative Imputer e Random Forest) na reconstrução de séries temporais de precipitação de longo prazo, considerando a presença de lacunas extensas e a variabilidade climática interanual e interdecadal. A reconstrução de séries históricas é crucial para estudos de mudanças climáticas e para a calibração de modelos hidrológicos.

6. Previsão de Séries Temporais: Explorar o potencial dos algoritmos de aprendizado de máquina, como Redes Neurais Recorrentes (RNNs), incluindo variantes como Long Short-Term Memory (LSTM) (DJERBOUAI, 2022) e Gated Recurrent Unit (GRU), para prever valores futuros de precipitação e, assim, complementar a imputação de falhas com previsões em tempo real.

A continuidade da pesquisa nessas áreas contribuirá significativamente para o avanço do conhecimento e para o desenvolvimento de técnicas mais robustas e eficazes no tratamento de dados pluviométricos incompletos, gerando impactos positivos na gestão de recursos hídricos, no planejamento climático e em outras áreas que dependem da disponibilidade de dados pluviométricos confiáveis.

REFERÊNCIAS

AFRIFA-YAMOAH, E. et al. Imputation of missing data in high-resolution climate conjunto de dados: A comparison of methods. **International Journal of Climatology**, v. 40, n. 9, p. 4075-4093, 2020.

BÁR DOSSY, A.; PEGRAM, G. Copula based multisite rainfall generation using atmospheric circulation patterns for filling gaps in daily precipitation measurements. **Hydrology and Earth System Sciences**, v. 18, n. 4, p. 1463–1479, 2014.

DJERBOUAI, Salim. Missing precipitation data estimation using long short-term memory deep neural networks. **Journal of Ecological Engineering**, v. 23, n. 5, 2022.

BARRY, R. G.; CHORLEY, R. J. Atmosphere, weather and climate. 7. ed. Londres: Routledge, 1998.

BIER, A. A.; FERRAZ, S. E. T. Comparação de metodologias de preenchimento de falhas em dados meteorológicos para estações no Sul do Brasil. **Revista Brasileira de Meteorologia**, v. 32, n. 2, p. 215–226, jun. 2017.

BLAKE, S. Meteorological data imputation. Disponível em: http://www.maths.lancs.ac.uk/~blake/Stats4/MissingDataImputation.pdf. Acesso em: 15 nov. 2011.

BOSCH, E.T. et al. The use of Meteosat and GMS to detect burned areas in tropical environments. **Remote Sensing of Environment**, v. 85, p. 329-340, 2003.

COLLISCHONN, W. et al. Evaluation of TRMM rainfall estimates over the Upper Paraguay River basin. **Journal of Hydrology**, v. 331, n. 1-2, p. 218–232, 2007.

DORNELLES, F.; GOLDENFUM, J. A.; PEDRO LLO, O. C. Desempenho das redes neurais artificiais (RNAs) para simulação hidrológica mensal. **Revista Brasileira de Recursos Hídricos**, v. 19, n. 2, p. 251-265, 2014.

EL HACHIMI, A.; RIBES, A.; CAREAS, S. ClimateFiller: A machine learning framework to fill gaps and diagnose inconsistencies in climate data. **Geoscientific Model Development**, v. 16, n. 5, p. 1483-1504, 2023.

HORTA, I. T. L. G. et al. Preenchimento de falhas de dados de precipitação por redes neurais artificiais: uma revisão sistemática. In: **SIMPÓSIO BRASILEIRO DE**

CLIMATOLOGIA GEOGRÁFICA, 14., 2021, João Pessoa. Anais... João Pessoa: UFPB, 2021. p. 1768-1780.

LATIF, S. D. et al. Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches. **Alexandria Engineering Journal**, v. 82, p. 16–25, nov. 2023.

LI, C.; REN, X.; ZHAO, G. Machine-learning-based imputation method for filling missing values in ground meteorological observation data. **Algorithms**, v. 16, n. 9, p. 422, set. 2023.

MACHADO, L. A.; ASSIS, W. L. Comparação entre métodos de preenchimento de falhas em séries de dados meteorológicos da bacia hidrográfica do Rio das Velhas (MG). **Revista Geografias**, v. 26, n. 1, p. 73–90, 2018.

OU, H. et al. Missing data imputation method combining random forest and generative adversarial imputation network. **Sensors**, v. 24, n. 4, p. 1112, fev. 2024.

RUSTICUCCI, M.; TENCER, B. Observed changes in return values of annual temperature extremes over Argentina. **Journal of Climate**, v. 21, p. 5455-5467, 2008.

SAAD, M. et al. Machine learning based approaches for imputation in time series data and their impact on forecasting. In: **IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN, AND CYBERNETICS (SMC)**, 2020, Toronto. **Proceedings**... Toronto: IEEE, 2020. p. 2621-2627.

SANCHES, R. G.; BOLLELI, T. M.; SANTOS, B. C. Previsão das chuvas: revisão bibliográfica das técnicas para as predições climáticas na América do Sul. **Revista Brasileira de Climatologia**, v. 27, 2020.

STEPHENSON, A. G. Extreme value distributions. **R News**, v. 2, n. 2, p. 31-32, 2002.

TUCCI, C. E. M. Hidrologia: ciência e aplicação. Porto Alegre: Ed. Universidade/UFRGS, 2001.