Avaliação de Técnicas de Imputação em Séries Temporais Curtas no Contexto da Atenção Primária à Saúde

Lucas Emanoel Silva Barros[1], Laissa Maria Gama Silva[1], Gabriel Silva de Aquino[1], Caio Guilherme Moreira da Rocha[1], Fábio Guerra Medeiros[1], Veruska Borges Santos[3], Tiago Brasileiro Araújo[2], Dimas Cassimiro Nascimento Filho[1]

[1] Universidade Federal do Agreste de Pernambuco (UFAPE)
Garanhuns, PE – Brasil

[2]Instituto Federal da Paraíba (IFPB)
Soledade, PB – Brasil

[3]Centro de Engenharia Elétrica e Informática – Universidade Federal de Campina Grande (UFCG) – Campina Grande, PB – Brasil.
{lucas2014.barros, jlaestudos, fabio.guerramedeiros, dimascnf}@gmail.com,
{gabriel.silvaa, caio.guilherme}@ufape.edu.br,
tiago.brasileiro@ifpb.edu.br, veruska@copin.ufcg.edu.br

Abstract. *Data quality is fundamental to the effectiveness of solutions using Machine Learning, with the absence of values being an even more critical problem in short time series. In Primary Health Care, this problem is still recurrent due to the lack of uniformity in data sources, the non-mandatory nature of data registration (depending on the region), and the lack of adequate user training in the use of the tools. Thus, this study compares seven imputation methods applied to data from Primary Health Care units in two Brazilian cities, highlighting lower error rates in the MICE and Random Forest models, respectively.*

Resumo. *A qualidade dos dados é fundamental para a eficácia de soluções com a aplicação de Aprendizado de Máquina, sendo a ausência de valores um problema ainda mais crítico em séries temporais pequenas. Nesse sentido, na Atenção Primária à Saúde, este problema ainda é recorrente devido à falta de uniformidade das fontes de dados, à não obrigatoriedade de certas informações (a depender da região) e à falta de treinamento dos usuários no uso das ferramentas. Portanto, este estudo compara sete métodos de imputação aplicados a dados de atendimentos das unidades da Atenção Primária à Saúde das cidades de Tacaratu e Bom Conselho (Pernambuco), destacando menor erro nos modelos MICE e Random Forest, respectivamente.*

1. Introdução

A análise de dados tem desempenhado um papel cada vez mais relevante no apoio à tomada de decisão em serviços públicos de saúde, especialmente no contexto da Atenção Primária à Saúde (APS). No âmbito do Sistema Único de Saúde (SUS), gestores dependem de informações confiáveis para planejar recursos, dimensionar equipes e antecipar demandas por atendimentos [Santos et al. 2024, de Almeida et al. 2025]. Nesse cenário,

sistemas de informação que integram dados provenientes de diferentes fontes e incorporam técnicas analíticas têm sido utilizados para apoiar decisões gerenciais e aprimorar o planejamento de serviços de saúde [Pereira et al. 2025, Nascimento et al. 2026].

Apesar dos avanços recentes no uso de técnicas de análise de dados em saúde, bases de dados administrativas frequentemente apresentam problemas de qualidade, sendo a presença de valores ausentes (*missing data*) uma das limitações mais recorrentes [Lima 2025]. Em bases clínicas e epidemiológicas tais lacunas podem surgir devido a falhas de coleta, integração de múltiplas fontes ou inconsistências no preenchimento dos sistemas de informação [Ganem et al. 2025]. Quando não tratadas adequadamente, essas ausências podem introduzir vieses analíticos ou comprometer a confiabilidade de modelos preditivos. Esse desafio torna-se ainda mais crítico quando os dados estão organizados na forma de séries temporais curtas, situação comum em dados administrativos da APS, nos quais o número reduzido de observações limita a aplicação de técnicas mais sofisticadas e dificulta a captura de padrões temporais estáveis [Massuda et al. 2022].

Diante desse contexto, este trabalho investiga o desempenho de diferentes técnicas de imputação de dados aplicadas a séries temporais curtas provenientes de indicadores da Atenção Primária à Saúde. O objetivo é comparar sete métodos de imputação amplamente utilizados na literatura, avaliando seu comportamento por meio de métricas quantitativas de erro. Ao analisar o impacto dessas técnicas em séries temporais curtas derivadas de dados reais da APS, este estudo busca fornecer evidências empíricas que auxiliem pesquisadores e gestores na escolha de estratégias mais adequadas para o tratamento de dados faltantes em sistemas de informação em saúde.

2. Fundamentação Teórica e Trabalhos Relacionados

Dados ausentes podem ocorrer sob diferentes mecanismos, usualmente classificados como *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) e *Missing Not at Random* (MNAR), os quais possuem implicações distintas para a análise e para a confiabilidade dos modelos preditivos. Em saúde, esse problema é particularmente relevante, pois lacunas nos registros podem introduzir vieses e comprometer a interpretação dos resultados [Ganem et al. 2025].

Para lidar com esse problema, diversas técnicas de imputação têm sido propostas na literatura, variando desde métodos estatísticos simples até abordagens mais sofisticadas baseadas em aprendizado de máquina. Entre os métodos amplamente utilizados estão estratégias baseadas em média, abordagens fundamentadas em vizinhança, como o *k-Nearest Neighbors* (KNN) [Alnowaiser 2024], e técnicas mais avançadas como o *Multivariate Imputation by Chained Equations* (MICE) [Hegde et al. 2019] e o *MissForest* [Hu et al. 2024]. Estudos recentes como [Ganem et al. 2025] e [Joel et al. 2025] têm demonstrado que métodos mais complexos podem alcançar maior precisão na reconstrução de valores ausentes, embora frequentemente apresentem custo computacional elevado e maior complexidade de implementação.

Apesar dos avanços observados na literatura, muitos trabalhos avaliam técnicas de imputação em cenários com grandes volumes de dados ou bases clínicas altamente estruturadas. Entretanto, em diversos contextos aplicados, os dados disponíveis estão organizados em séries temporais curtas, nas quais o número reduzido de observações limita a capacidade de métodos mais complexos capturarem padrões consistentes. Nes-

sas situações, torna-se relevante investigar empiricamente o comportamento de diferentes técnicas de imputação, analisando em que medida abordagens simples e métodos mais sofisticados se comportam quando aplicados a séries temporais curtas.

3. Metodologia

Este trabalho compara estratégias de imputação em séries temporais curtas oriundas de registros administrativos de saúde, cenário em que o número limitado de observações dificulta a captura de padrões temporais estáveis por métodos mais complexos [Joel et al. 2025]. O conjunto de dados analisado consiste em séries temporais mensais que representam a evolução de uma variável de interesse ao longo do tempo. Cada série possui um número relativamente pequeno de observações, característica comum em dados administrativos utilizados para monitoramento de serviços públicos. Para permitir a aplicação de métodos baseados em aprendizado de máquina, foram geradas variáveis auxiliares derivadas da estrutura temporal da série, incluindo codificações de ano e unidade observacional, transformações sazonais baseadas em seno e cosseno do mês[1], variáveis de defasagem temporal e estatísticas móveis calculadas em janelas deslizantes, calculadas *após* a remoção simulada (MCAR), exclusivamente a partir da série corrompida, evitando vazamento temporal.

O experimento[2] compara sete técnicas de imputação pertencentes a diferentes categorias metodológicas. Foram avaliados dois métodos de interpolação temporal (**linear** e **polinomial**), dois métodos estatísticos simples (**média** e **mediana**), um método baseado em proximidade (**k-Nearest Neighbors**), e duas abordagens baseadas em aprendizado de máquina, utilizando modelos de **Random Forest** e imputação iterativa por cadeias condicionais (**Multivariate Imputation by Chained Equations - MICE**). Essa seleção permite analisar o comportamento de métodos com diferentes níveis de complexidade quando aplicados a séries temporais curtas [Ganem et al. 2025].

Para avaliar o desempenho das técnicas de imputação, foi adotada uma estratégia de simulação controlada de dados ausentes. Inicialmente, parte dos valores originais das séries foi removida aleatoriamente (MCAR), reconhecendo que ausências em registros administrativos tendem a seguir mecanismos MAR ou MNAR, essa foi a abordagem inicial do presente trabalho para simular a ocorrência de *missing data*. Em seguida, cada método de imputação foi aplicado para reconstruir os valores removidos. Essa abordagem permite comparar diretamente os valores imputados com os valores reais originalmente observados.

A fim de garantir robustez estatística, o processo de imputação foi repetido em múltiplas execuções independentes. Em cada iteração uma semente (*seed*) aleatória distinta é utilizada para remover aproximadamente 20% dos valores da série, garantindo variações no padrão de dados ausentes. No total, foram realizadas 50 iterações do experimento, nas quais todos os métodos de imputação são aplicados e avaliados.

O desempenho das técnicas é medido por meio de três métricas de erro amplamente utilizadas: *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE) e *Mean Absolute Error* (MAE) [Pereira et al. 2025]. Enquanto o MSE penaliza mais fortemente erros de grande magnitude, o RMSE permite interpretar o erro na mesma escala da

[1]Essas transformações representam a natureza cíclica e sazonal dos meses [Taylor and Letham 2018].

[2]Os códigos dos experimentos estão disponíveis em: https://zenodo.org/records/20497605.

Cidade	Min	Q1	Mediana	Q3	Max	Total de Zeros	Total de Registros
Tacaratu	0	596	1405	2697	8015	4 (1%)	397
Bom Conselho	0	144	404	621	1511	173 (18%)	948

Tabela 1. Estatísticas das séries temporais de Atendimentos da APS

variável original e o MAE fornece uma medida robusta menos sensível a *outliers*. Além da análise quantitativa das métricas, também são realizadas inspeções visuais das séries imputadas para comparar o comportamento dos métodos ao longo do tempo.

4. Experimentos e Resultados

Nesta seção, apresentamos os resultados obtidos a partir da aplicação da metodologia descrita na seção anterior, cujas análises foram divididas da seguinte forma: (i) análise dos dados e (ii) avaliação comparativa dos 7 métodos de imputação.

4.1. Característica dos Dados

Para avaliar o desempenho dos diferentes modelos na imputação de dados ausentes, foram utilizadas séries temporais das cidades de Tacaratu e Bom Conselho, ambas brasileiras, com registros de 01/01/2019 a 01/07/2024 e de 01/01/2022 a 01/12/2025, respectivamente. A diferença na quantidade de registros de cada município está relacionada à quantidade de unidades APS da localidade, sendo o município de Bom Conselho o que possui mais unidades, portanto, mais valores, embora em um período menor. Os dados são oriundos do sistema de gestão das unidades APS, com acesso restrito aos gestores das unidades.

A variável imputada (total de atendimentos por unidades APS) possui valores entre 0 e 8015, conforme as estatísticas descritas na Tabela 1, que serve como referência para avaliar os erros introduzidos por cada método. É importante observar a proporção de valores nulos, o que indica uma possível inconsistência nas séries temporais de múltiplas unidades APS do mesmo município: em um conjunto extremamente pequeno, com 397 ocorrências, 1% dos valores já são zeros, enquanto em um conjunto relativamente maior, a proporção de dados zerados aumenta progressivamente para 18%. Tais exemplos demonstram a necessidade de uma adequada preparação dos dados antes da aplicação de técnicas mais robustas e avançadas, como Aprendizado de Máquina, para a predição do total de atendimentos por unidade APS.

4.2. Comparação entre Métodos de Imputação

A Tabela 2 apresenta os valores médios de MSE, RMSE e MAE para cada modelo utilizado. Os resultados mostram maior efetividade dos métodos MICE ($max\_iter = 10$) e Random Forest ($n\_estimators = 100$) em relação aos demais; o kNN ($k = 3$) apresentou desempenho intermediário, enquanto Média e Mediana registraram os maiores erros, conforme esperado, devido à simplicidade das abordagens.

Com base nos experimentos, o modelo MICE apresenta-se como a solução mais apropriada para a maioria dos cenários, embora o Random Forest tenha superado seu desempenho no MAE em Bom Conselho (72,01 vs. 77,06). O MICE obteve erro absoluto médio (MAE) de 182,78 em Tacaratu e 77,06 em Bom Conselho. Sua acurácia geral, demonstrada por erros significativamente menores na maioria das métricas, permite que

Método	MSE		RMSE		MAE	
	Tacaratu	Bom Conselho	Tacaratu	Bom Conselho	Tacaratu	Bom Conselho
Iterative Imputer (MICE)	65.712,02	12.647,02	255,35	112,03	182,78	77,06
Random Forest	105.238,9	13.275,85	321,80	114,01	226,45	72,01
KNN Imputer	181.848,2	20.348,54	424,05	142,19	315,13	98,61
Linear Interpolation	247.942,9	35.164,84	492,90	186,70	333,70	127,78
Polynomial Interpolation	387.939,1	48.086,69	618,43	218,31	418,93	155,01
Mean Imputation	2.578.212	100.653,57	1.600,86	316,99	1.261,33	259,20
Median Imputation	2.745.512	100.759,15	1.650,50	317,20	1.224,05	259,27

Tabela 2. Comparação da eficácia dos métodos de imputação

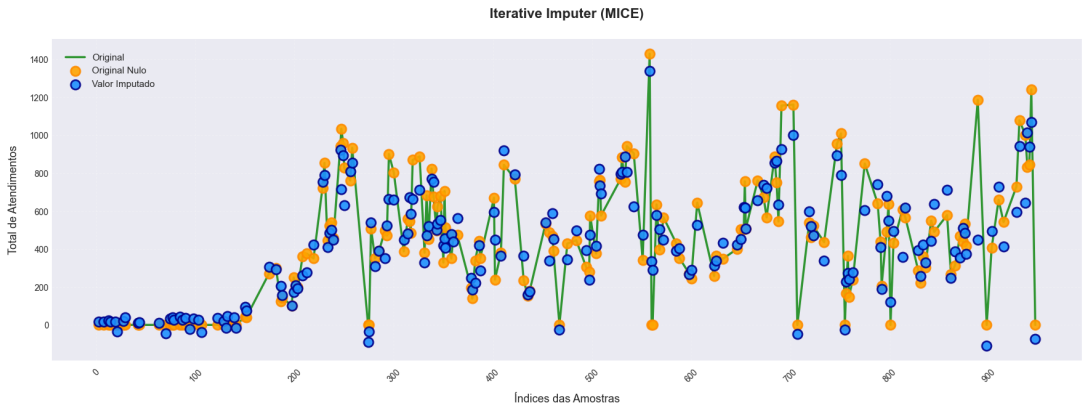

Figura 1. Comparação da eficácia do MICE na série completa do município de Bom Conselho.

os valores imputados sejam os mais próximos possíveis da realidade, conforme exibido na Figura 1. Na figura, são exibidos os 948 registros avaliados de Bom Conselho, nos quais o eixo x representa o índice das observações analisadas, enquanto os pontos destacam os valores reais removidos e os valores imputados pelo modelo.

Conforme descrito na Metodologia, 20% dos valores originais foram removidos para verificar a eficácia dos métodos de imputação. Assim, nas Figuras 1 e 2, os valores removidos estão representados na cor laranja e os valores estimados pelo algoritmo MICE estão na cor azul. Com isso, é possível validar a proximidade dos valores estimados com os valores reais, especialmente quando há pouca variação nos dados, como é o caso do início da série (Figura 1). Quando há variação, a estimativa tende a apresentar maior erro, como é possível observar nas Figuras 2a e 2b. Ainda assim, mesmo com maior variância nos valores, a estimativa ainda segue o padrão da série temporal, demonstrando a flexibilidade da abordagem para estimar dados em qualquer posição da série, independente de sua variância. Nas Figuras 2a e 2b, têm-se as séries temporais exclusivas de duas unidades APS do município de Bom Conselho, onde é possível observar a estabilidade na estimativa cuja série temporal tem maior variância nos dados.

O melhor desempenho do MICE pode estar associado ao uso de atributos derivados da estrutura temporal da série, incluindo componentes sazonais, defasagens e estatísticas móveis. Esses atributos permitem incorporar padrões de curto e longo prazo sem exigir grandes volumes de dados, o que é particularmente relevante em séries temporais curtas. Na sequência, o Random Forest também apresentou resultados promissores, embora inferiores ao MICE na maioria das métricas avaliadas. Sua capacidade de modelar relações não lineares e lidar com variações nos dados sugere que o método pode ser útil

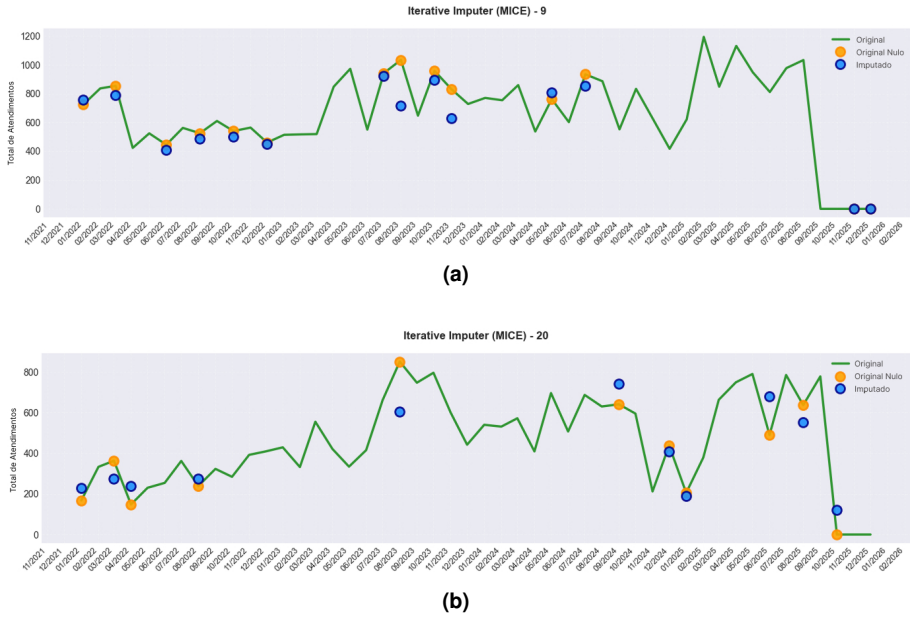

Figura 2. Séries temporais de atendimentos em duas unidades de Bom Conselho (2022–2025): tendência da série em verde, dados removidos em laranja e estimativas de imputação em azul.

em cenários com maior volume de amostras ou com padrões temporais mais complexos.

5. Conclusão

Este trabalho apresentou uma avaliação comparativa de técnicas de imputação de dados aplicadas a séries temporais curtas no contexto da Atenção Primária à Saúde. Os resultados indicam que métodos baseados em aprendizado de máquina, especialmente o *MICE*, apresentaram melhor desempenho na reconstrução de valores ausentes, com menores taxas de erro que abordagens estatísticas simples e métodos de interpolação. O Random Forest também obteve resultados promissores, enquanto média e mediana apresentaram os maiores erros, devido à limitação em capturar padrões temporais mais complexos.

De forma geral, os resultados indicam que a utilização de atributos derivados da série temporal, como componentes sazonais e variáveis de defasagem, contribui significativamente para o desempenho de métodos mais robustos, permitindo capturar padrões não lineares e variações temporais presentes nos dados. No entanto, observou-se que a qualidade da imputação é influenciada pela variabilidade da série, sendo mais precisa em cenários com menor flutuação e mais desafiadora em séries com maior variância, conforme observado nos experimentos realizados. Como trabalho futuro, pretende-se investigar o impacto da parametrização dos modelos e explorar abordagens adicionais de imputação, incluindo métodos híbridos e técnicas baseadas em aprendizado profundo, com o intuito de aprimorar a qualidade das estimativas em cenários com dados limitados.

Agradecimentos

Agradecemos o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq e financiamento do Departamento de Ciência e Tecnologia da Secretaria de Ciência, Tecnologia, Inovação e Complexo da Saúde do Ministério da Saúde - Decit/SECTICS/MS no projeto de pesquisa (processo 445259/2023-0) em execução.

Referências

Alnowaiser, K. (2024). Improving healthcare prediction of diabetic patients using knn imputed features and tri-ensemble model. *IEEE Access*, 12:16783–16793.

de Almeida, P. F., Giovanella, L., Schenkman, S., Franco, C. M., Duarte, P. O., Houghton, N., Báscolo, E., and Bousquat, A. (2025). Health surveillance as a strategic pillar of a comprehensive primary health care. *Ciencia & saude coletiva*, 30(5):e04572025.

Ganem, I. M., Bianco, G. D., Serufo Filho, J. C., Lima, L., Rocha, L., and Gonçalves, M. A. (2025). Avaliando as limitações e potenciais do algoritmo k-vizinhos mais próximos (knn) na imputação de dados clínicos. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 809–815. SBC.

Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., and Acharya, A. (2019). Mice vs ppca: Missing data imputation in healthcare. *Informatics in medicine unlocked*, 17:100275.

Hu, Y.-H., Wu, R.-Y., Lin, Y.-C., and Lin, T.-Y. (2024). A novel missforest-based missing values imputation approach with recursive feature elimination in medical applications. *BMC Medical Research Methodology*, 24(1):269.

Joel, L. O., Doorsamy, W., and Paul, B. S. (2025). A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *International Journal of Data Science and Analytics*, 20(7):6357–6373.

Lima, A. M. (2025). Assessing the impact of missing value mechanisms on anomaly detection in healthcare wearable data. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 781–787. SBC.

Massuda, A., Malik, A., Lotta, G., Siqueira, M., Tasca, R., and Rocha, R. (2022). Brazil's primary health care financing: case study. *Lancet Global Health Commission on Financing Primary Health Care*.

Nascimento, I. Y. N., Santos, V. B., Pereira, L. F. A., Araújo, T. B., da Rocha, C. G. M., Medeiros Filho, F. G., de Andrade, R. C. A., Vanderlei, I. M., and Nascimento Filho, D. C. (2026). Primary healthcare efficiency indicators for analysis and decision-making support for managers. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 556–575. SBC.

Pereira, L. F. A., Ferreira, L. B., Nascimento, D. C., Vanderlei, I. M., Silva, D., Santos, V. B., Gomes, M. S., and Melo, I. A. (2025). Design, integration, and evaluation of a demand forecasting service in the context of primary healthcare. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 506–514. SBC.

Santos, A. M. d., Giovanella, L., Fausto, M. C. R., Cabral, L. M. d. S., and Almeida, P. F. d. (2024). Dynamics of regionalization and repercussions of gaps in care on health marketing in remote rural municipalities. *Cadernos de Saude Publica*, 40:e00194523.

Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.