

# Imputação Multivariada de Séries Temporais Meteorológicas do INMET com Reanálise ERA5-Land e Self-Attention

Davi Aquila de Souza Dias<sup>1</sup>, Maria José Pereira Dantas<sup>1</sup>

<sup>1</sup>Pontifícia Universidade Católica de Goiás (PUC Goiás)  
Goiânia – GO – Brasil

daviaquila64489@gmail.com, mjdantas@pucgoias.edu.br

**Abstract.** *Missing values in meteorological time series compromise climate analyses and predictive models. This work investigates a multivariate imputation pipeline for observational data from the Brazilian National Institute of Meteorology (INMET), using ERA5-Land as a source of auxiliary features. SAITS was adopted as the main model and compared with Median, LOCF, MissForest, XGBoost, Transformer and ImputeFormer. MCAR scenarios ranging from 10% to 80% artificial missingness were generated and evaluated using MAE, RMSE, and MASE. The results show that SAITS achieves the best performance under low missingness, while XGBoost obtains the lowest global errors from 40% to 80% missingness, indicating greater robustness under severe data degradation.*

**Resumo.** *Valores ausentes em séries temporais meteorológicas comprometem análises climáticas e modelos preditivos. Este trabalho investiga um pipeline de imputação multivariada para dados observacionais do INMET, utilizando o ERA5-Land como fonte de atributos auxiliares. O SAITS foi adotado como modelo principal e comparado com Median, LOCF, MissForest, XGBoost, Transformer e ImputeFormer. Foram criados cenários MCAR de 10% a 80% de ausência artificial, com avaliação por MAE, RMSE e MASE. Os resultados mostram que o SAITS apresenta melhor desempenho em cenários de baixa ausência, enquanto o XGBoost obtém os menores erros globais de 40% a 80% de mascaramento, indicando maior robustez sob degradação severa dos dados.*

## 1. Introdução

Séries temporais meteorológicas são essenciais para análises climáticas e aplicações operacionais, mas frequentemente apresentam lacunas decorrentes de falhas de sensores, interrupções de comunicação e descontinuidades operacionais. Quando não tratadas adequadamente, essas ausências comprometem as análises e podem introduzir vieses em modelos preditivos. No Brasil, os dados observacionais do Instituto Nacional de Meteorologia constituem uma das principais fontes de séries históricas utilizadas em pesquisas e aplicações meteorológicas [Instituto Nacional de Meteorologia 2024].

Para mitigar esses efeitos, reanálises globais de superfície, como o ERA5-Land, têm sido utilizadas como fontes auxiliares na reconstrução de séries incompletas, com resultados promissores quando consideradas as particularidades locais [Araújo et al. 2022, Varotsos et al. 2023]. Além disso, diferentes mecanismos de ausência, como *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) e *Missing Not At Random* (MNAR), afetam diretamente o desempenho e a confiabilidade dos métodos de imputação [Alabadla et al. 2022, Lien et al. 2023].

Embora a literatura apresente avanços na imputação multivariada de séries temporais, ainda é relevante avaliar comparativamente diferentes classes de métodos em séries meteorológicas observacionais brasileiras, especialmente em cenários controlados de ausência e com apoio de variáveis auxiliares de reanálise [Alabadla et al. 2022, Alejo-Sanchez et al. 2025].

Neste trabalho, avalia-se um *pipeline* de imputação multivariada para séries meteorológicas do INMET com apoio do ERA5-Land. O SAITS foi adotado como modelo principal e comparado com métodos estatísticos, modelos baseados em árvores e arquiteturas baseadas em atenção, sob diferentes níveis de ausência artificial.

## 2. Trabalhos Relacionados

A imputação de dados ausentes em séries temporais multivariadas tem sido abordada por métodos estatísticos, técnicas de aprendizado de máquina e modelos de aprendizado profundo. Métodos simples, como média, mediana e LOCF, continuam sendo utilizados como *baselines* devido ao baixo custo computacional, embora apresentem limitações para capturar dependências temporais e correlações entre variáveis [Alabadla et al. 2022, Zainuddin et al. 2022].

Métodos de aprendizado de máquina, como MissForest e XGBoost, têm se mostrado competitivos em cenários de imputação multivariada por explorarem relações não lineares entre atributos [Troyanskaya et al. 2001, Stekhoven and Bühlmann 2012, Chen and Guestrin 2016]. Em paralelo, arquiteturas baseadas em Transformers passaram a ser aplicadas à imputação de séries temporais por sua capacidade de modelar dependências complexas. Nesse contexto, destacam-se o SAITS, baseado em *self-attention* para séries temporais multivariadas, e o ImputeFormer, voltado à imputação espaço-temporal com uso de *low-rankness* [Vaswani et al. 2017, Du et al. 2023, Nie et al. 2024].

No domínio meteorológico, reanálises como o ERA5-Land têm sido utilizadas como suporte auxiliar à reconstrução de séries observacionais incompletas [Muñoz-Sabater et al. 2021, Araújo et al. 2022, Varotsos et al. 2023]. Assim, este estudo compara diferentes famílias de métodos nesse contexto.

## 3. Materiais e Métodos

### 3.1. Fontes de Dados

Os dados observacionais foram obtidos de 96 estações meteorológicas automáticas do INMET na região Centro-Oeste, cobrindo o período de 2015 a 2024, com frequência horária e aproximadamente 8,14 milhões de amostras [Instituto Nacional de Meteorologia 2024]. As variáveis de interesse compreendem a temperatura do ar, ponto de orvalho, umidade relativa, pressão atmosférica, radiação global, precipitação e velocidade do vento.

Como fonte auxiliar, utilizou-se a reanálise horária ERA5-Land, com resolução espacial de  $0,1^\circ \times 0,1^\circ$  ( $\approx 9$  km), pareada a cada estação do INMET [Muñoz-Sabater et al. 2021]. Seus atributos foram convertidos para unidades compatíveis com o padrão INMET: temperatura a 2 m ( $t_{2m}$ ) e ponto de orvalho ( $d_{2m}$ ) de Kelvin para  $^\circ\text{C}$ ; pressão de superfície ( $s_p$ ) de Pascal para hPa; radiação de onda curta ( $s_{srd}$ ) de  $\text{J/m}^2$  para  $\text{kJ/m}^2$ ; precipitação total ( $t_p$ ) de metros para mm; e as componentes do vento zonal ( $u_{10}$ ) e meridional ( $v_{10}$ ) mantidas em m/s. Essas variáveis atuam estritamente como atributos complementares, sem substituir as observações originais de superfície.

### 3.2. Pré-processamento e Integração dos Dados

Após alinhamento temporal e pareamento estação–grade, foram incluídos três atributos estáticos (latitude, longitude, altitude) e quatro atributos temporais cíclicos (seno/cosseno da hora do dia e do dia do ano). O conjunto final possui 21 atributos por instante, frequência horária, janelas temporais de 72 horas e aproximadamente 8,14 milhões de amostras para 96 estações. A taxa global de ausência natural no INMET foi de 21,61%, enquanto os atributos auxiliares do ERA5-Land pareados não apresentaram lacunas no recorte utilizado. Os dados foram normalizados estatisticamente antes do treinamento.

### 3.3. Métodos de imputação avaliados

Os métodos foram organizados em três grupos. Nos *baselines* estatísticos, utilizaram-se mediana simples e LOCF com propagação bidirecional e preenchimento residual por mediana [Lachin 2016, Engels 2003]. No grupo de aprendizado de máquina, avaliaram-se XGBoost [Chen and Guestrin 2016], treinado por variável-alvo com atributos observados e auxiliares, e MissForestLike [Stekhoven and Bühlmann 2012], adaptação da imputação iterativa por árvores sem equivalência estrita à implementação canônica. Para aprendizado profundo, foram avaliados SAITS [Du et al. 2023], adotado como modelo principal por empregar *self-attention* com tarefas ORT e MIT, Transformer [Vaswani et al. 2017] e ImputeFormer [Nie et al. 2024], este último incorporando *low-rankness* para imputação espaço-temporal.

### 3.4. Protocolo Experimental e Métricas

Foram gerados cenários MCAR com oito níveis de ausência artificial, de 10% a 80%, substituindo posições selecionadas aleatoriamente por NaN e preservando os valores originais apenas para avaliação. A divisão dos dados foi realizada temporalmente: os 80% mais antigos foram destinados ao treinamento, sendo os 20% finais desse bloco usados para validação, enquanto os 20% mais recentes foram reservados para teste.

Os hiperparâmetros foram ajustados com Optuna, minimizando o RMSE no conjunto de validação. As métricas foram calculadas exclusivamente sobre os valores artificialmente mascarados. A avaliação utilizou MAE, RMSE e MASE, permitindo analisar erro médio, penalização de erros maiores e comparação entre séries com diferentes escalas.

Os experimentos foram executados em servidor com cinco GPUs NVIDIA RTX A4000, cada uma com 16 GB de VRAM, e processador AMD Ryzen Threadripper Pro 5965WX.

## 4. Resultados e Discussão

### 4.1. Desempenho global sob diferentes níveis de ausência artificial

A Tabela 1 resume o desempenho global dos modelos em cenários representativos de mascaramento MCAR. Em 20% de ausência artificial, o SAITS obteve os menores erros globais, indicando que arquiteturas baseadas em *self-attention* podem explorar de forma eficaz a estrutura temporal quando ainda há maior disponibilidade de observações. A partir de 40%, o XGBoost apresentou os menores erros e manteve esse comportamento até 80%, evidenciando maior robustez em cenários de degradação severa.

Essa robustez pode estar associada ao fato de que, na configuração adotada, o XGBoost não depende explicitamente da continuidade sequencial da série para realizar

estimativas. Como modelo baseado em árvores, ele explora relações não lineares entre atributos meteorológicos observados, variáveis auxiliares do ERA5-Land e atributos temporais e espaciais, tornando-se menos sensível à perda de contexto sequencial em cenários mais esparsos.

**Tabela 1. Desempenho global em cenários representativos de mascaramento MCAR. Em cada célula: MAE / RMSE / MASE.**

| Modelo         | 20%                           | 40%                           | 60%                           | 80%                           |
|----------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| Median         | 123.22 / 504.69 / 0.274       | 123.34 / 505.05 / 0.267       | 123.35 / 505.13 / 0.271       | 123.32 / 505.01 / 0.281       |
| LOCF+Med       | 48.88 / 216.82 / 0.109        | 60.50 / 269.34 / 0.131        | 81.15 / 351.01 / 0.178        | 121.61 / 479.72 / 0.277       |
| MissForestLike | 38.62 / 161.80 / 0.086        | 39.05 / 164.67 / 0.085        | 39.57 / 168.33 / 0.090        | 40.19 / 172.80 / 0.092        |
| XGBoost        | 28.51 / 136.68 / 0.064        | <b>30.35 / 143.48 / 0.066</b> | <b>32.10 / 149.83 / 0.070</b> | <b>33.90 / 157.33 / 0.077</b> |
| Transformer    | 54.35 / 190.20 / 0.120        | 73.57 / 277.59 / 0.160        | 78.69 / 271.62 / 0.180        | 111.32 / 393.38 / 0.250       |
| ImputeFormer   | 48.01 / 221.25 / 0.107        | 49.11 / 205.63 / 0.107        | 64.37 / 254.94 / 0.142        | 95.85 / 352.85 / 0.220        |
| SAITS          | <b>28.14 / 125.91 / 0.062</b> | 36.68 / 173.02 / 0.079        | 41.28 / 171.21 / 0.090        | 40.16 / 170.54 / 0.091        |

## 4.2. Análise por variável meteorológica

Não houve um único método dominante em todos os atributos. O desempenho depende da natureza estatística e física de cada variável, conforme detalhado na Tabela 2.

Os modelos baseados em Transformers destacaram-se em ponto de orvalho, umidade relativa e temperatura do ar em baixos níveis de ausência. Para radiação global, o XGBoost foi consistentemente superior, evidenciando a eficácia da exploração de relações não lineares com as variáveis auxiliares. Na pressão atmosférica, o LOCF+Median obteve os menores erros em todos os cenários, sugerindo que a alta persistência temporal dessa variável favorece abordagens simples. Para precipitação, os resultados foram próximos entre os métodos, reflexo da natureza intermitente e assimétrica dessa variável.

**Tabela 2. Desempenho por variável (MAE / RMSE / MASE)**

| Var.         | Modelo   | 20%                   | 50%                   | 80%                   |
|--------------|----------|-----------------------|-----------------------|-----------------------|
| Pt. Orvalho  | Median   | 3.98/5.20/3.39        | 3.98/5.20/4.87        | 3.97/5.20/6.06        |
|              | LOCF+Med | 0.64/0.97/0.54        | 0.79/1.20/0.96        | 1.11/1.66/1.70        |
|              | MissFor. | 1.25/1.68/1.07        | 1.31/1.76/1.60        | 1.36/1.84/2.07        |
|              | XGBoost  | 1.36/1.76/1.16        | 1.49/1.92/1.83        | 1.62/2.07/2.47        |
|              | Transf.  | <b>0.28/0.47/0.24</b> | 0.52/0.76/0.63        | 0.90/1.23/1.40        |
|              | ImpFor.  | 0.32/0.50/0.27        | <b>0.51/0.76/0.62</b> | <b>0.80/1.13/1.22</b> |
| Rad. Global  | SAITS    | 0.49/0.71/0.42        | 0.78/1.08/0.96        | 1.10/1.47/1.68        |
|              | Median   | 785/1314/0.94         | 785/1315/1.65         | 785/1315/2.36         |
|              | LOCF+Med | 323/565/0.39          | 457/795/0.96          | 806/1249/2.43         |
|              | MissFor. | 248/421/0.30          | 252/434/0.53          | 258/450/0.78          |
|              | XGBoost  | <b>179/356/0.21</b>   | <b>194/384/0.41</b>   | <b>208/409/0.63</b>   |
|              | Transf.  | 364/496/4.33          | 636/868/1.33          | 744/1025/2.24         |
| Umidade Rel. | ImpFor.  | 322/576/0.39          | 465/736/0.97          | 639/919/1.92          |
|              | SAITS    | 185/328/0.22          | 224/400/0.47          | 260/444/0.78          |
|              | Median   | 18.6/22.6/1.44        | 18.6/22.7/2.51        | 18.7/22.7/3.55        |
|              | LOCF+Med | 5.13/7.66/0.40        | 7.11/10.7/0.96        | 12.2/17.3/2.33        |
|              | MissFor. | 6.66/8.63/0.52        | 6.92/9.02/0.93        | 7.19/9.42/1.37        |
|              | XGBoost  | 6.85/9.88/0.53        | 10.2/13.4/1.38        | 13.2/16.2/2.50        |
| Precipitação | Transf.  | <b>1.48/2.39/0.11</b> | <b>2.82/4.20/0.40</b> | <b>6.20/8.20/1.20</b> |
|              | ImpFor.  | 1.56/2.69/0.12        | 2.99/4.74/0.40        | 6.49/9.08/1.23        |
|              | SAITS    | 2.28/3.33/0.18        | 4.59/6.20/0.62        | 6.84/8.95/1.30        |
|              | Median   | <b>0.15/1.45/0.57</b> | <b>0.15/1.44/0.64</b> | <b>0.15/1.46/0.69</b> |
|              | LOCF+Med | 0.22/1.76/0.84        | 0.24/1.85/1.01        | 0.27/1.96/1.22        |
|              | MissFor. | 0.24/1.38/0.89        | 0.24/1.39/0.98        | 0.23/1.42/1.04        |
| Pressão Atm. | XGBoost  | 0.21/1.34/0.80        | 0.20/1.37/0.83        | 0.18/1.43/0.80        |
|              | Transf.  | <b>0.15/1.50/0.60</b> | 0.16/1.44/0.64        | 0.16/1.50/0.70        |
|              | ImpFor.  | <b>0.15/1.24/0.54</b> | 0.20/1.35/0.82        | 0.20/1.45/0.89        |
|              | SAITS    | <b>0.15/1.45/0.57</b> | <b>0.15/1.44/0.64</b> | <b>0.15/1.46/0.69</b> |
|              | Median   | 4.18/5.36/1.23        | 4.18/5.36/2.21        | 4.18/5.36/3.19        |
|              | LOCF+Med | 1.27/1.88/0.38        | 1.80/2.72/0.95        | 3.20/4.55/2.45        |
| Temp. Ar     | MissFor. | 1.43/1.91/0.42        | 1.48/1.99/0.79        | 1.55/2.09/1.18        |
|              | XGBoost  | 1.12/1.56/0.33        | 1.42/1.88/0.75        | 1.55/2.01/1.19        |
|              | Transf.  | <b>0.40/0.65/0.12</b> | <b>0.76/1.10/0.40</b> | 1.62/2.12/1.24        |
|              | ImpFor.  | 0.47/0.84/0.14        | 0.86/1.33/0.45        | 1.89/2.66/1.45        |
|              | SAITS    | 0.62/0.87/0.18        | 1.10/1.46/0.58        | <b>1.50/1.96/1.15</b> |
|              | Median   | 1.10/1.43/1.13        | 1.10/1.43/1.45        | 1.10/1.43/1.67        |
| Vel. Vento   | LOCF+Med | 0.65/0.97/0.67        | 0.74/1.10/0.98        | 0.95/1.36/1.45        |
|              | MissFor. | 0.89/1.15/0.91        | 0.89/1.15/1.18        | 0.90/1.16/1.37        |
|              | XGBoost  | 0.77/1.02/0.80        | 0.78/1.03/1.04        | 0.80/1.05/1.22        |
|              | Transf.  | <b>0.51/0.76/0.52</b> | 0.60/0.84/0.80        | 0.73/1.00/1.11        |
|              | ImpFor.  | <b>0.51/0.76/0.53</b> | <b>0.57/0.82/0.75</b> | 0.73/1.00/1.11        |
|              | SAITS    | 0.56/0.79/0.57        | 0.61/0.85/0.80        | <b>0.72/0.97/1.10</b> |

Como síntese metodológica, os resultados sugerem que a escolha do método de imputação deve considerar a variável analisada e a severidade da ausência. Modelos baseados em atenção mostraram-se competitivos em variáveis com dependência temporal bem definida e em cenários de menor degradação; métodos baseados em árvores, como o XGBoost, foram mais robustos em cenários severos com atributos auxiliares disponíveis; e abordagens simples, como LOCF+Median, permaneceram adequadas para variáveis altamente persistentes, como pressão atmosférica.<sup>1</sup>

### 4.3. Custo computacional e Limitações experimentais

Além da precisão, o custo computacional destacou-se como um aspecto limitante na comparação. O ImputeFormer apresentou um tempo de execução substancialmente superior ao dos demais modelos, o que impactou a viabilidade operacional dos testes e restringiu a exploração em diferentes cenários.

A etapa de ajuste de hiperparâmetros foi estritamente condicionada à infraestrutura disponível de memória das GPUs. Enquanto a otimização do SAITS favoreceu o uso de arquiteturas mais enxutas e eficientes, a busca por configurações mais profundas no ImputeFormer resultou em diversas interrupções (trials) por esgotamento de memória. Essa assimetria na exploração do espaço de busca é uma limitação importante e deve ser considerada na interpretação comparativa dos resultados.

## 5. Conclusão

Este trabalho avaliou métodos de imputação multivariada em séries meteorológicas do INMET com apoio de variáveis auxiliares do ERA5-Land, considerando cenários MCAR entre 10% e 80% de ausência artificial. Os resultados mostram que o desempenho depende da taxa de ausência e da variável analisada: o SAITS foi mais competitivo em cenários de baixa ausência, enquanto o XGBoost apresentou maior robustez entre 40% e 80% de mascaramento. A análise por variável reforça que não há método universalmente superior, pois diferentes atributos meteorológicos favorecem estratégias distintas de imputação.

As variáveis auxiliares do ERA5-Land contribuíram para enriquecer a representação do contexto meteorológico sem substituir as observações do INMET. Como limitações, destacam-se o uso exclusivo de cenários MCAR, a ausência de testes estatísticos de significância e as restrições computacionais na otimização dos modelos profundos. Trabalhos futuros devem investigar mecanismos MAR/MNAR, falhas em blocos, novas regiões do Brasil e estratégias mais eficientes de otimização para modelos baseados em Transformers.

## Referências

- Alabadla, M., Sidi, F., Ishak, I., Ibrahim, H., Affendey, L. S., Che Ani, Z., Jabar, M. A., Bukar, U. A., Devaraj, N. K., Muda, A. S., Tharek, A., Omar, N., and Jaya, M. I. M. (2022). Systematic Review of Using Machine Learning in Imputing Missing Values. *IEEE Access*, 10:44483–44502.
- Alejo-Sanchez, L. E., Márquez-Grajales, A., Salas-Martínez, F., Franco-Arcega, A., López-Morales, V., Acevedo-Sandoval, O. A., González-Ramírez, C. A., and Villegas-Vega, R. (2025). Missing data imputation of climate time series: A review. *MethodsX*, 15:103455.

---

<sup>1</sup>Código-fonte disponível em: <https://github.com/DaviRGB/pipeline-imputation>

- Araújo, C. S. P. D., Silva, I. A. C. E., Ippolito, M., and Almeida, C. D. G. C. D. (2022). Evaluation of air temperature estimated by ERA5-Land reanalysis using surface data in Pernambuco, Brazil. *Environmental Monitoring and Assessment*, 194(5):381.
- Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.
- Du, W., Côté, D., and Liu, Y. (2023). SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619.
- Engels, J. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, 56(10):968–976.
- Instituto Nacional de Meteorologia (2024). Sobre o instituto nacional de meteorologia. Acesso em: 28 nov. 2025.
- Lachin, J. M. (2016). Fallacies of last observation carried forward analyses. *Clinical Trials*, 13(2):161–168.
- Lien, P. L., Do, T. T., and Nguyen, T. (2023). Data Imputation for Multivariate Time-series Data. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6. ISSN: 2694-4804.
- Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383. Publisher: Copernicus GmbH.
- Nie, T., Qin, G., Ma, W., Mei, Y., and Sun, J. (2024). ImputeFormer: Low Rankness-Induced Transformers for Generalizable Spatiotemporal Imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 2260–2271, New York, NY, USA. Association for Computing Machinery.
- Stekhoven, D. J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.
- Varotsos, K. V., Katavoutas, G., and Giannakopoulos, C. (2023). On the Use of Reanalysis Data to Reconstruct Missing Observed Daily Temperatures in Europe over a Lengthy Period of Time. *Sustainability*, 15(9):7081.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need.
- Zainuddin, A., Hairuddin, M. A., Yassin, A. I. M., Latiff, Z. I. A., and Azhar, A. (2022). Time Series Data and Recent Imputation Techniques for Missing Data: A Review. In *2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*, pages 346–350.