Seleção de Variáveis para Imputação de Dados Ausentes em Séries Temporais de Qualidade do Ar com XGBoost

Taiane Campos Braga de Assis[a*], Édler Lins de Albuquerque[b], Cristiano Hora de Oliveira Fontes[c]
[a,b,c] *Programa de Engenharia Industrial, Universidade Federal da Bahia, Salvador-BA Brasil.*
**taianecbraga@gmail.com*

RESUMO

A presença de dados ausentes em séries temporais ambientais representa um desafio relevante para análises estatísticas e modelagem preditiva. Este estudo avaliou o desempenho de quatro cenários de imputação aplicados a dados horários de qualidade do ar coletados entre 2011 e 2016, considerando variações nas variáveis utilizadas como entrada. O modelo foi avaliado por meio de métricas quantitativas, como RMSE e MAE, e análise estatística comparativa. Os resultados indicaram que o Cenário 4 apresentou desempenho semelhante ao cenário completo, com valores médios de RMSE de aproximadamente [0,117] e MAE de [0,147], sem diferenças estatisticamente relevantes ($p > 0,05$) quando comparado ao cenário completo. A análise visual por meio de boxplots corroborou a estabilidade das estimativas, reforçando a viabilidade do Cenário 4 como alternativa operacional para o tratamento de dados ausentes em séries ambientais.

Palavras-chave: imputação de dados; qualidade do ar; dados ausentes; séries temporais; análise estatística.

1. Motivação

O monitoramento da qualidade do ar é essencial para a gestão ambiental, pois fornece informações necessárias para o controle da poluição e o suporte a políticas públicas. Reconhecendo essa importância, a Organização das Nações Unidas (ONU) recomenda que cidades com mais de 500 mil habitantes realizem o monitoramento sistemático da qualidade do ar. No entanto, estudos recentes indicam que menos da metade dos municípios brasileiros com esse quantitativo populacional possuem estações automáticas de monitoramento, evidenciando limitações estruturais associadas principalmente à restrição de recursos financeiros para implantação e manutenção dessas redes (IEMA, 2024). Além disso, desafios operacionais, como falhas na operação contínua das estações e ausência de monitoramento de variáveis relevantes, contribuem para a ocorrência de períodos ausentes nas séries temporais ambientais (IEMA, 2022).

A presença de dados ausentes representa um dos principais desafios no tratamento de séries temporais, pois compromete a representatividade das informações e dificulta a análise da variabilidade temporal dos fenômenos monitorados, podendo prejudicar avaliações sobre as condições reais do ar ambiente e a eficácia das redes de monitoramento (IEMA, 2022).

Nesse contexto, técnicas de imputação de dados tornam-se ferramentas fundamentais para a reconstrução de informações faltantes e preservação de padrões temporais relevantes. No âmbito da Engenharia de Sistemas em Processos (Process Systems Engineering – PSE), a aplicação de modelos computacionais para tratamento de dados constitui abordagem estratégica para o desenvolvimento de sistemas mais robustos e confiáveis. Entretanto, o desempenho desses modelos depende diretamente do conjunto de variáveis utilizadas como preditoras, sendo necessária a definição de conjuntos que equilibrem qualidade das estimativas e complexidade computacional.

Diante desse cenário, este trabalho tem como objetivo avaliar o desempenho de modelos de imputação em diferentes cenários de seleção de variáveis, com foco na identificação de um conjunto ótimo de preditores capaz de manter a qualidade das estimativas com menor complexidade do modelo. Para isso, foi conduzido um estudo de caso utilizando dados históricos de monitoramento da qualidade do ar da cidade de Salvador, Bahia, Brasil. A abordagem proposta permite analisar o impacto da seleção de variáveis no desempenho dos modelos e apresenta potencial de aplicação em outras redes de monitoramento ambiental e sistemas baseados em séries temporais.

2. Metodologia

2.1. Área de estudo e base de dados

Os dados utilizados neste estudo foram disponibilizados pela CETREL – Environmental Protection Company Inc.,

responsável pela operação da rede de monitoramento da qualidade do ar de Salvador, Bahia, Brasil, no período de 2011 a 2016. A rede era composta por oito estações fixas distribuídas em diferentes regiões urbanas, fornecendo registros horários de concentrações de poluentes atmosféricos e variáveis meteorológicas. O conjunto de dados inclui medições horárias de monóxido de carbono (CO), ozônio (O_3), dióxido de nitrogênio (NO_2) e óxido nítrico (NO), além de variáveis meteorológicas como velocidade do vento, temperatura e umidade relativa do ar.

Neste trabalho, o monóxido de carbono (CO) foi selecionado como variável-alvo devido à sua associação com emissões veiculares e à presença de padrões temporais bem definidos em ambientes urbanos. Para a aplicação e avaliação do modelo, adotou-se como referência a estação do Centro Administrativo da Bahia (CAB), localizada próxima à Avenida Paralela, selecionada em função da maior completude relativa de suas séries temporais em comparação às demais estações da rede.

2.2. Preparação dos dados

O pré-processamento da base de dados foi realizado a partir de uma *dataframe*, contendo registros horários de qualidade do ar e variáveis meteorológicas coletados por oito estações automáticas de monitoramento localizadas na cidade de Salvador (BA), no período de 2011 a 2016. Essa etapa teve como objetivo padronizar as variáveis e estruturar o conjunto de dados para aplicação dos modelos de imputação. Inicialmente, as colunas correspondentes à data e à hora das medições foram convertidas para formatos temporais apropriados e combinadas em uma única marcação temporal (*timestamp*). A partir dessa estrutura, foi criado um índice temporal contínuo com frequência horária, garantindo a regularidade da série temporal e permitindo a identificação explícita de lacunas nos registros.

Em seguida, foi realizada a padronização dos tipos de dados das variáveis monitoradas, com a conversão de colunas originalmente armazenadas como texto para formato numérico, adotando-se a substituição de valores inválidos por valores ausentes (*NaN*). Posteriormente, foi conduzida uma análise da disponibilidade e completude dos dados para todas as estações e variáveis, considerando as datas oficiais de início de operação de cada estação, de modo a evitar a superestimação de períodos ausentes. Com base nessa análise, a estação CAB foi selecionada como estação-alvo do estudo, em função da maior disponibilidade relativa de dados e da maior extensão temporal de suas séries. Após essa definição, foi construído um subconjunto específico correspondente à estação selecionada, com os dados recortados a partir do início de sua operação e organizados para a identificação e caracterização dos períodos ausentes.

2.3. Modelagem com XGBoost e protocolo de avaliação

Para a imputação das lacunas em séries temporais ambientais foi utilizado o modelo *Extreme Gradient Boosting* (XGBoost), uma técnica baseada em árvores de decisão amplamente empregada em problemas de regressão devido à sua capacidade de modelar relações não lineares entre variáveis. O modelo foi implementado em linguagem *Python*, utilizando as bibliotecas *NumPy* e Pandas para manipulação dos dados, *Scikit-learn* para padronização e cálculo de métricas e a biblioteca XGBoost para treinamento do modelo.

A avaliação do desempenho foi realizada por meio da geração de lacunas artificiais, obtidas pela remoção controlada de observações do banco de dados original, permitindo a comparação direta entre valores imputados e valores observados.

No cenário inicial de modelagem, foram consideradas como variáveis exógenas as concentrações de dióxido de nitrogênio (NO_2), monóxido de nitrogênio (NO), ozônio (O_3), temperatura do ar (TEMP), umidade relativa (UMID) e velocidade do vento (VEL.VENTO), registradas no mesmo instante temporal da variável alvo, correspondente à concentração de monóxido de carbono (CO).Considerando que o modelo XGBoost não incorpora dependência temporal de forma explícita, foram incluídas variáveis defasadas da concentração de CO, correspondentes aos atrasos de 12, 24, 36 e 48 horas. Dessa forma, cada instância de entrada do modelo foi composta pelas quatro defasagens da variável alvo e pelas seis variáveis exógenas no instante corrente, totalizando dez atributos preditores no cenário completo.

O modelo foi configurado utilizando parâmetros típicos para problemas de regressão baseados em árvores de decisão, incluindo múltiplos estimadores, profundidade controlada das árvores e mecanismos de amostragem parcial das observações e variáveis. Foi adotada função objetivo baseada no erro quadrático, com controle da variabilidade estocástica do treinamento e paralelização computacional. Embora o modelo tenha sido treinado para prever valores em um único passo temporal, o preenchimento das lacunas foi realizado de forma sequencial, estimando sucessivamente os valores ausentes até a reconstrução completa de cada intervalo.

2.4. Análise de Feature Importance (seleção de variáveis e construção dos cenários)

A análise de importância das variáveis (*feature importance*) foi conduzida a partir do modelo XGBoost treinado utilizando o cenário completo de preditores. Esse procedimento teve como objetivo identificar a contribuição relativa de cada variável para o desempenho do modelo e subsidiar a definição de diferentes configurações de entrada para avaliação comparativa. A estimativa da importância das variáveis foi obtida diretamente a partir das métricas internas do modelo XGBoost, que quantificam a contribuição de cada preditor para a redução do erro ao longo do processo de construção das árvores de decisão. A partir desse procedimento, foi gerado um ranqueamento das variáveis em ordem decrescente de importância relativa. O cenário inicial do estudo foi definido como o conjunto completo de variáveis disponíveis, incluindo poluentes atmosféricos, variáveis meteorológicas e defasagens temporais da concentração de CO, sendo adotado como referência para comparação com as demais configurações.

Com base na análise de importância das variáveis e na natureza física das relações entre os preditores e a variável alvo, foram definidos quatro cenários distintos de entrada, com o objetivo de avaliar o impacto da composição do conjunto de variáveis sobre o desempenho do modelo:

- **Cenário 1 – Base completa:** inclui poluentes atmosféricos (NO_2, NO e O_3), variáveis meteorológicas (TEMP, UMID e VEL.VENTO) e defasagens temporais da concentração de CO, constituindo o cenário de referência para comparação com as demais configurações;
- **Cenário 2 – Apenas poluentes atmosféricos:** inclui as concentrações de NO_2, NO e O_3, juntamente com as defasagens temporais da concentração de CO, permitindo avaliar se variáveis diretamente associadas às fontes de emissão são suficientes para imputar a concentração da variável alvo;
- **Cenário 3 – Apenas variáveis meteorológicas:** inclui as variáveis TEMP, UMID e VEL.VENTO, juntamente com as defasagens temporais da concentração de CO, permitindo examinar o grau de influência das condições atmosféricas sobre os padrões temporais da concentração de CO;
- **Cenário 4 – Poluentes nitrogenados e meteorologia:** inclui as concentrações de NO_2 e NO, combinadas com variáveis meteorológicas (TEMP, UMID e VEL.VENTO) e defasagens temporais da concentração de CO, permitindo avaliar o equilíbrio entre complexidade do modelo e desempenho preditivo com um conjunto reduzido de variáveis.

Todos os cenários foram avaliados utilizando o mesmo protocolo de geração de lacunas artificiais descrito anteriormente, garantindo a comparabilidade direta entre os resultados obtidos para diferentes configurações de entrada. Para cada cenário, o modelo XGBoost foi treinado novamente e aplicado ao preenchimento das lacunas artificiais, sendo o desempenho avaliado por meio de métricas quantitativas de erro.

2.5. Avaliação das métricas (tabelas resumo)

O desempenho dos modelos foi avaliado por meio de métricas quantitativas de erro, permitindo quantificar a discrepância entre os valores imputados e os valores observados nas lacunas artificiais. As métricas utilizadas foram o Erro Absoluto Médio (MAE) e a Raiz do Erro Quadrático Médio (RMSE), definidas conforme as equações 1 e 2, respectivamente. O MAE representa a magnitude média dos erros absolutos, enquanto o RMSE atribui maior peso a discrepâncias de maior magnitude, sendo sensível à presença de erros elevados.

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

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

onde y_i representa o valor observado, $\hat{y}_i$ representa o valor estimado pelo modelo, e n corresponde ao número total de observações avaliadas nas lacunas artificiais.

As métricas foram calculadas exclusivamente sobre os intervalos correspondentes às lacunas artificiais, garantindo que a avaliação refletisse diretamente o desempenho do modelo no processo de imputação.

Para verificar a existência de diferenças estatisticamente significativas entre os cenários avaliados, foi aplicado o teste não paramétrico de *Kruskal–Wallis*, adequado para comparação entre múltiplos grupos independentes sem a necessidade de assumir distribuição normal dos dados. As hipóteses testadas foram:

- H_0: não há diferença significativa entre os cenários avaliados;
- H_1: pelo menos um cenário apresenta diferença significativa em relação aos demais.

Nos casos em que foram identificadas diferenças estatisticamente significativas no teste global, foi aplicado o teste pós-hoc de Dunn, com correção de *Bonferroni*, a fim de identificar quais pares de cenários apresentaram diferenças significativas. Essa correção foi adotada para controlar o erro do tipo I decorrente da realização de múltiplas comparações. O nível de significância adotado em todos os testes foi de 5% ($\alpha = 0,05$).

3. Resultados e discussão

3.1. *Avaliação da importância das variáveis*

A importância relativa das variáveis foi avaliada com base no modelo XGBoost treinado, utilizando o conjunto completo de variáveis preditoras (Cenário 1). A figura 1 apresenta o ranqueamento das variáveis em ordem decrescente de importância relativa. Adicionalmente, a tabela 1 apresenta os valores numéricos correspondentes obtidos a partir da métrica *Gain*.

Figura 1. Feature Importance – XGBoost (Target: CO_CAB).

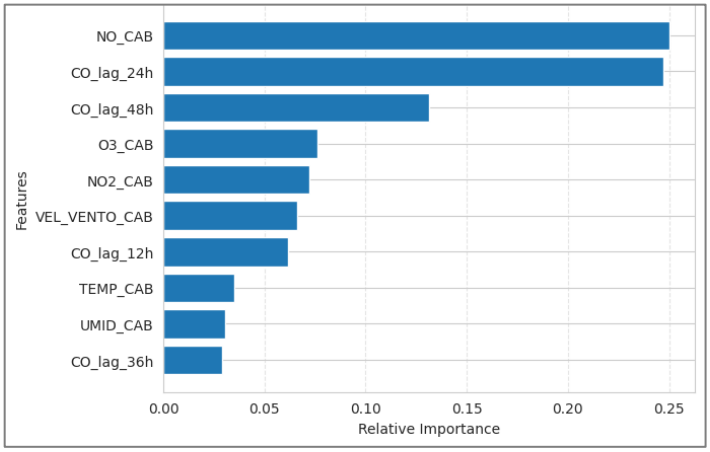


Tabela 1. Importance_gain.

Feature	Importance_gain
NO_CAB	0.25027314
CO_lag_24h	0.24717821
CO_lag_48h	0.1312309
O3_CAB	0.076375246
NO2_CAB	0.07222089
VEL_VENTO_CAB	0.06615129
CO_lag_12h	0.061597463
TEMP_CAB	0.035180546
UMID_CAB	0.030770546
CO_lag_36h	0.029021736

Assim, observa-se que a variável NO apresentou a maior contribuição para o desempenho do modelo, seguida pela defasagem temporal CO_lag_24h. A elevada relevância da concentração de NO indica forte associação entre a dinâmica do CO e poluentes primários relacionados a fontes de combustão, especialmente em ambientes urbanos com predominância de emissões veiculares.

As defasagens temporais da concentração de CO, particularmente CO_lag_24h e CO_lag_48h, também apresentaram elevada importância relativa, evidenciando a presença de dependência temporal significativa na variável alvo. Esse comportamento é esperado em séries temporais ambientais, nas quais valores passados contribuem diretamente para a estimativa de valores futuros.

Outros poluentes, como O_3 e NO_2, tiveram contribuição reduzida, sugerindo que interações físico-químicas foram pouco relevantes para a imputação. Da mesma forma, variáveis meteorológicas (velocidade do vento, temperatura e umidade) apresentaram baixo impacto, indicando que, para este conjunto de dados, a variabilidade das emissões decorrente do tráfego prevaleceu a dispersão e o transporte atmosférico.

3.2. Desempenho dos cenários avaliados

O desempenho dos diferentes cenários foi avaliado por meio das métricas MAE, RMSE, calculados a partir das estimativas realizadas nas lacunas artificiais. A tabela 2 apresenta os valores médios obtidos para cada cenário analisado.

Tabela 2. Desempenho médio dos cenários.

Cenários	MAE	RMSE	R^2
Cenário 1	**0.112**	**0.140**	**0.293**
Cenário 2	0.125	0.157	0.126
Cenário 3	0.137	0.173	0.017
Cenário 4	0.117	0.147	0.283

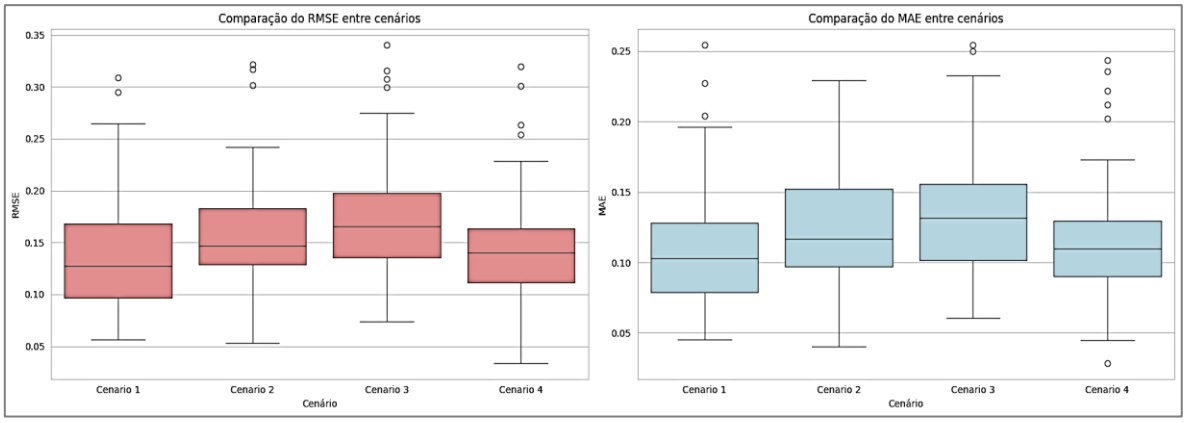
A partir disso, é possível observar que o **Cenário 1** apresentou os menores valores médios de MAE (0.112) e RMSE (0.140), indicando o melhor desempenho preditivo entre os cenários avaliados. Em contrapartida, o **Cenário 3** apresentou os maiores valores de erro (MAE = 0.137; RMSE = 0.173), evidenciando desempenho inferior quando comparado aos demais cenários.

O **Cenário 4** apresentou desempenho intermediário, com valores de erro próximos aos observados no Cenário 1, sugerindo que a redução do conjunto de variáveis não comprometeu significativamente a capacidade preditiva do modelo. Já o **Cenário 2** apresentou desempenho inferior ao Cenário 1 e ao Cenário 4, indicando que a utilização isolada de variáveis associadas a poluentes não foi suficiente para alcançar o melhor desempenho observado.

Além dos valores médios, a Figura 2 apresenta os boxplots das métricas RMSE e MAE para os cenários avaliados, permitindo a análise da distribuição e variabilidade dos erros. De modo geral, o Cenário 1 apresenta medianas inferiores e menor dispersão relativa, indicando maior estabilidade nas estimativas, enquanto o Cenário 3 apresenta maiores medianas e maior variabilidade, evidenciando desempenho inferior, comportamento consistente com seus maiores valores médios.

O Cenário 4 apresenta distribuição semelhante à observada no Cenário 1, com medianas próximas e intervalos interquartis parcialmente sobrepostos, sugerindo desempenho comparável entre esses cenários. O Cenário 2 apresenta comportamento intermediário, com dispersão superior à dos Cenários 1 e 4, porém inferior à do Cenário 3. Observa-se ainda a presença de valores discrepantes em todos os cenários, comportamento esperado em processos de imputação aplicados a séries ambientais.

Figura 2. Distribuição das métricas MAE e RMSE entre os cenários avaliados.



3.3. Análise estatística dos cenários

A comparação estatística entre os cenários foi realizada por meio do teste não paramétrico de Kruskal–Wallis, seguido pelo teste pós-hoc de Dunn com correção de Bonferroni, considerando nível de significância de

5%. Os resultados do teste de Kruskal–Wallis indicaram diferença estatisticamente significativa entre os cenários para ambas as métricas avaliadas (MAE: p = 0.021; RMSE: p = 0.014), evidenciando que pelo menos um dos cenários apresentou desempenho distinto em relação aos demais.

A análise pós-hoc realizada por meio do teste de Dunn revelou que a diferença estatisticamente significativa ocorreu entre o **Cenário 1** e o **Cenário 3**, para ambas as métricas (MAE e RMSE). Não foram observadas diferenças estatisticamente significativas entre os demais pares de cenários. Esses resultados indicam que o **Cenário 3** apresentou desempenho significativamente inferior ao **Cenário 1**, enquanto os demais cenários apresentaram desempenhos estatisticamente semelhantes entre si. Esse comportamento sugere que a exclusão de determinados grupos de variáveis pode comprometer significativamente a capacidade preditiva do modelo, enquanto a utilização de conjuntos reduzidos, mas representativos, pode manter desempenho comparável ao cenário completo.

Adicionalmente, observa-se que o **Cenário 4** apresentou desempenho estatisticamente equivalente ao **Cenário 1**, uma vez que não foram identificadas diferenças significativas entre esses cenários nos testes pós-hoc realizados. Considerando que o Cenário 4 utiliza um conjunto reduzido de variáveis e apresentou métricas de erro próximas às observadas no cenário completo, esse resultado sugere que tal configuração representa uma alternativa eficiente para o processo de imputação, mantendo desempenho comparável ao modelo completo com menor complexidade de entrada.

4. Conclusão

Os resultados obtidos demonstraram que o desempenho do modelo de imputação é influenciado pela composição do conjunto de variáveis utilizadas como preditoras. A análise de importância indicou maior contribuição das variáveis associadas à dependência temporal e à presença de poluentes atmosféricos, enquanto as variáveis meteorológicas apresentaram contribuição complementar.

A avaliação dos cenários evidenciou desempenho inferior no Cenário 3, caracterizado por maior variabilidade e maiores valores de erro. Em contrapartida, os Cenários 1 e 4 apresentaram distribuições semelhantes dos erros, conforme observado nos boxplots e confirmado pelos testes estatísticos, que não indicaram diferenças significativas entre essas configurações.

Destaca-se que o Cenário 4 apresentou desempenho estatisticamente equivalente ao cenário completo, mesmo utilizando um conjunto reduzido de variáveis. Esse resultado evidencia que a redução criteriosa do número de preditores pode ser realizada sem comprometer significativamente a capacidade preditiva do modelo, caracterizando o Cenário 4 como a configuração mais eficiente entre as avaliadas.

De modo geral, os resultados reforçam a importância da seleção adequada de variáveis no desenvolvimento de modelos de imputação, contribuindo para o equilíbrio entre desempenho preditivo e eficiência computacional em aplicações envolvendo séries temporais ambientais.

Agradecimentos: Ao Instituto do Meio Ambiente e Recursos Hídricos da Bahia (INEMA) e à CETREL pelo fornecimento dos dados de monitoramento da qualidade do ar utilizados neste estudo.

Referências

Betancourt, C., et al. (2023). Graph machine learning for improved imputation of missing tropospheric ozone data. Environmental Science & Technology, 57, 18246–18258. https://doi.org/10.1021/acs.est.3c05104

Instituto de Energia e Meio Ambiente (IEMA). (2022). Qualidade do ar no Brasil: Relatório técnico. IEMA.

Instituto de Energia e Meio Ambiente (IEMA). (2024). Dimensionamento da rede básica de monitoramento da qualidade do ar no Brasil: Cenários iniciais. IEMA.

Thomas, T., & Rajabi, E. (2021). A systematic review of machine learning-based missing value imputation techniques. Data Technologies and Applications, 55(4), 558–585. https://doi.org/10.1108/DTA-12-2020-0298