

UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA – PPGCAP

MIRACELI BONJARDIM WALDEMAR

**PREVISÃO INTELIGENTE DE OZÔNIO TROPOSFÉRICO EM ESTAÇÕES DE
QUALIDADE DO AR COM MODELOS PREDITIVOS RESILIENTES A DADOS
FALTANTES**

JOINVILLE

2026

MIRACELI BONJARDIM WALDEMAR

**PREVISÃO INTELIGENTE DE OZÔNIO TROPOSFÉRICO EM ESTAÇÕES DE
QUALIDADE DO AR COM MODELOS PREDITIVOS RESILIENTES A DADOS
FALTANTES**

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada.

Orientador: Dr. Yuri Kaszubowski Lopes

JOINVILLE

2026

Ficha catalográfica elaborada pelo(a) autor(a), com
auxílio do programa de geração automática da
Biblioteca Setorial do CCT/UDESC

Waldemar, Miraceli Bonjardim

Previsão Inteligente de Ozônio Troposférico em Estações de Qualidade
do Ar com Modelos Preditivos Resilientes a Dados Faltantes / Miraceli
Bonjardim Waldemar. – Joinville, 2026.

78 p. : il.

Orientador: Yuri Kaszubowski Lopes.

Dissertação (Mestrado) – Universidade do Estado de Santa Catarina,
Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Compu-
tação Aplicada, Joinville, 2026.

1. Ozônio troposférico. 2. Aprendizado de máquina. 3. Imputação de
dados. 4. Qualidade do ar. I. Lopes, Yuri Kaszubowski. II. Universidade do
Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de
Pós-Graduação em Computação Aplicada. III. Título.

MIRACELI BONJARDIM WALDEMAR

**PREVISÃO INTELIGENTE DE OZÔNIO TROPOSFÉRICO EM ESTAÇÕES DE
QUALIDADE DO AR COM MODELOS PREDITIVOS RESILIENTES A DADOS
FALTANTES**

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada.

Orientador: Yuri Kaszubowski Lopes

BANCA EXAMINADORA:

Prof. Dr. Yuri Kaszubowski Lopes - Presidente
Universidade do Estado de Santa Catarina

Membros:

Prof. Dr. Fernando Augusto Silveira Armani
Universidade Federal do Paraná

Prof. Dr. Douglas Wildgrube Bertol
Universidade do Estado de Santa Catarina

Prof. Dr. Roberto Silvio Ubertino Rosso Junior - Suplente
Universidade do Estado de Santa Catarina

Prof. Dr. André Roberto Ortoncelli - Suplente
Universidade Tecnológica Federal do Paraná

Joinville, 26 de fevereiro de 2026

AGRADECIMENTOS

Agradeço à minha mãe Wanusa e ao meu pai Sanclé por sempre valorizarem a minha educação e por me proporcionarem as condições que tornaram possível o meu caminho até aqui.

Ao Professor Yuri, meu orientador, sou grata pela disponibilidade, pela paciência e por me guiar em cada etapa deste caminho. Sua orientação foi muito importante para o amadurecimento deste trabalho e para a minha formação acadêmica.

À minha irmã Anita e à minha avó Iracema pelo carinho, acolhimento e presença nos momentos em que mais precisei. Ao meu amigo Eduardo, agradeço pelo apoio, pela escuta e pelo incentivo em momentos importantes ao longo dessa jornada. Aos demais familiares e amigos, agradeço por sempre acreditarem em mim e torcerem pelas minhas conquistas.

Ao Bruno, colega de projeto e aluno de graduação, agradeço pelo comprometimento, dedicação e colaboração ao longo do desenvolvimento deste trabalho. Também sou grata às instituições envolvidas por disponibilizarem a base de dados que tornou esta pesquisa possível.

“Life’s too short to not give it a try. True brew
will always be the aim.” (Millencolin)

RESUMO

O ozônio troposférico é um poluente atmosférico com impactos negativos à saúde humana e ao meio ambiente. A previsão de sua concentração enfrenta desafios devido à complexidade dos processos fotoquímicos e à presença frequente de dados faltantes em estações de monitoramento. Isso é especialmente relevante no Brasil, onde a cobertura é limitada e os dados são muitas vezes incompletos. Este trabalho desenvolve e avalia modelos preditivos baseados em aprendizado de máquina para prever a concentração de ozônio troposférico, utilizando dados horários coletados em uma estação de qualidade do ar em Ponta Grossa, Paraná, no período de outubro de 2016 a dezembro de 2020. A série apresenta proporção média de dados faltantes superior a 55%, com lacunas concentradas em blocos (a mais extensa estende-se de março de 2017 a junho de 2018). O problema de dados ausentes é tratado por meio de sete métodos de imputação: média, mediana, LOCF, interpolação linear, média móvel, MICE e dados meteorológicos externos. Para este último, realizou-se uma compatibilização estatística em três etapas (normalização Z-score, transformações específicas por variável, logarítmica para chuva e raiz quadrada para radiação solar global, e normalização quantílica), garantindo que os dados externos preservassem a média, o desvio padrão e a distribuição das medições originais da estação. Os dados imputados foram então utilizados na modelagem preditiva com três algoritmos de aprendizado de máquina: XGBoost, Random Forest e kNN. Os resultados mostram que o XGBoost apresentou melhor desempenho ($R^2 = 0.858$), com elevada estabilidade entre os métodos de imputação. Os métodos clássicos (interpolação linear, LOCF) apresentaram desempenho ligeiramente superior, enquanto os dados externos obtiveram resultados comparáveis, com diferença inferior a 1% em relação aos melhores métodos e sem significância estatística quando considerados múltiplos algoritmos. Dados externos utilizam medições reais de estações independentes, operam sem dependência de histórico local e podem ser aplicados imediatamente após falhas de sensor, o que os torna especialmente adequados a contextos com infraestrutura de monitoramento limitada. O estudo demonstra a viabilidade dessa abordagem para superar limitações do monitoramento tradicional, contribuindo para estratégias mais robustas de previsão da qualidade do ar.

Palavras-chave: Ozônio troposférico. Aprendizado de máquina. Imputação de dados. Qualidade do ar.

ABSTRACT

Tropospheric ozone is an atmospheric pollutant with negative impacts on human health and the environment. Predicting its concentration faces challenges due to the complexity of photochemical processes and the frequent presence of missing data in monitoring stations. This is especially relevant in Brazil, where coverage is limited and data are often incomplete. This work develops and evaluates predictive models based on machine learning to forecast tropospheric ozone concentration, using hourly data collected at an air quality monitoring station in Ponta Grossa, Paraná, from October 2016 to December 2020. The series presents an average missing data rate exceeding 55%, with gaps concentrated in blocks (the most extensive spanning from March 2017 to June 2018). The missing data problem is addressed through seven imputation methods: mean, median, LOCF, linear interpolation, moving average, MICE, and external meteorological data. For the latter, a three-step statistical harmonization was carried out (Z-score normalization, variable-specific transformations, logarithmic for rainfall and square root for global solar radiation, and quantile normalization), ensuring that the external data preserved the mean, standard deviation, and distribution of the original station measurements. The imputed data were then used in predictive modeling with three machine learning algorithms: XGBoost, Random Forest, and kNN. The results show that XGBoost achieved the best performance ($R^2 = 0.858$), with high stability across imputation methods. Classical methods (linear interpolation, LOCF) showed slightly superior performance, while external data obtained comparable results, with differences below 1% relative to the best methods and no statistical significance when multiple algorithms were considered. External data rely on real measurements from independent stations, operate without dependence on local historical records, and can be applied immediately after sensor failures, making them particularly suitable for contexts with limited monitoring infrastructure. The study demonstrates the feasibility of this approach to overcome limitations of traditional monitoring, contributing to more robust air quality forecasting strategies.

Keywords: Tropospheric ozone. Machine learning. Data imputation. Air quality.

LISTA DE ILUSTRAÇÕES

Figura 2.1 – Variação da temperatura com a altitude nas camadas da atmosfera	20
Figura 2.2 – Ciclo do ozônio troposférico com e sem a presença de COVs	22
Figura 2.3 – Fatores que influenciam a formação do ozônio troposférico	23
Figura 3.1 – Fluxograma PRISMA (Page <i>et al.</i> , 2021) do processo de seleção de artigos .	31
Figura 3.2 – Distribuição anual dos estudos selecionados na revisão sistemática	32
Figura 3.3 – Evolução do uso de modelos de aprendizado de máquina nos estudos selecio- nados	33
Figura 3.4 – Evolução do uso das métricas de avaliação nos estudos selecionado	34
Figura 3.5 – Evolução do desempenho (R^2) dos modelos por ano de publicação com agrupamento por desempenho	36
Figura 4.1 – Classificações climáticas de Ponta Grossa	38
Figura 4.2 – Proporção de dados faltantes por variável (%)	40
Figura 4.3 – Temperatura média diária (2018 – 2020)	41
Figura 4.4 – Radiação solar global média diária (2018 – 2020)	42
Figura 4.5 – Umidade relativa média diária (2018 – 2020)	42
Figura 4.6 – Concentração média diária de ozônio (2018 – 2020)	43
Figura 4.7 – Séries temporais após imputação por média	45
Figura 4.8 – Análise de resíduos e erro acumulado para imputação por média	46
Figura 4.9 – Séries temporais após imputação por mediana	48
Figura 4.10–Séries temporais após interpolação linear	49
Figura 4.11–Análise de resíduos e erro acumulado para interpolação linear	50
Figura 4.12–Séries temporais após imputação por média móvel	51
Figura 4.13–Séries temporais após LOCF	53
Figura 4.14–Análise de resíduos e erro acumulado para LOCF	53
Figura 4.15–Séries temporais após imputação com MICE para variáveis meteorológicas .	55
Figura 4.16–Séries temporais após imputação com MICE para variáveis precursoras . .	55
Figura 5.1 – Dados meteorológicos externos (2018-2021)	56
Figura 5.2 – Compatibilização estatística de variáveis meteorológicas	57
Figura 5.3 – Transformações específicas para Chuva e RSG	58
Figura 5.4 – Séries temporais após imputação com dados externos	59
Figura 5.5 – Séries temporais das variáveis precursoras após imputação com dados exter- nos e MICE	59
Figura 5.6 – Análise de resíduos e erro acumulado para imputação por dados externos . .	60
Figura A.1 – Sensor portátil Aeroqual S500 Starter Kit com capacidade de medição de ozônio	73
Figura A.2 – Sensor fixo Serinus® 10, voltado à análise contínua de ozônio em ambientes	73

Figura A.3 – Sensor pessoal POM TM , desenvolvido para monitoramento individual de ozônio.	73
Figura B.1 – Distribuições de variáveis meteorológicas e químicas atmosféricas	76
Figura B.2 – Séries temporais de variáveis meteorológicas e químicas atmosféricas	77
Figura B.3 – Diagramas de dispersão entre ozônio e variáveis meteorológicas e química atmosféricas	78

LISTA DE TABELAS

Tabela 3.1 – Quantidade de estudos coletados por base de dados	30
Tabela 3.2 – Critérios de inclusão e exclusão dos estudos	30
Tabela 3.3 – Síntese dos algoritmos de aprendizado de máquina e métricas de desempenho utilizados nos estudos selecionados	33
Tabela 3.4 – Resumo dos estudos organizados por desempenho (R^2)	34
Tabela 3.5 – Métodos de imputação utilizados nos estudos selecionados	35
Tabela 3.6 – Desempenho (R^2) de diferentes estudos com classificação KMeans	36
Tabela 4.1 – Convenção para o período sazonal	39
Tabela 4.2 – Estatísticas descritivas das variáveis medidas	39
Tabela 4.3 – Estatísticas descritivas sem valores atípicos antes da imputação.	44
Tabela 4.4 – Estatísticas descritivas depois da imputação por média.	45
Tabela 4.5 – Resultados da imputação por média ordenados por R^2	46
Tabela 4.6 – Estatísticas descritivas depois da imputação por mediana.	47
Tabela 4.7 – Resultados da imputação por mediana ordenados por R^2	48
Tabela 4.8 – Estatísticas descritivas depois da imputação por interpolação linear.	49
Tabela 4.9 – Resultados da imputação por interpolação linear ordenados por R^2	49
Tabela 4.10–Estatísticas descritivas depois da imputação por média móvel.	51
Tabela 4.11–Resultados da imputação por média móvel ordenados por R^2	52
Tabela 4.12–Estatísticas descritivas depois da imputação por LOCF.	52
Tabela 4.13–Resultados da imputação por LOCF ordenados por R^2	53
Tabela 4.14–Estatísticas descritivas depois da imputação por Mice.	54
Tabela 4.15–Resultados da imputação por MICE ordenados por R^2	55
Tabela 5.1 – Comparação de unidades entre os conjuntos de dados	56
Tabela 5.2 – Resultados da imputação por dados externos ordenados por R^2	60
Tabela 5.3 – Desempenho médio por algoritmo de aprendizado.	61
Tabela 5.4 – Ordenação dos métodos de imputação no XGBoost por desempenho.	61
Tabela 5.5 – Diferença de desempenho entre Interpolação Linear e Dados Externos no XGBoost.	61
Tabela 5.6 – Diferença de desempenho entre LOCF e Dados Externos considerando três algoritmos.	62
Tabela A.1 – Especificações gerais de sensores portáteis para monitoramento de ozônio .	74

LISTA DE ABREVIATURAS E SIGLAS

COVs	Compostos Orgânicos Voláteis
kNN	Algoritmo dos k-Vizinhos Mais Próximos (do inglês <i>k-Nearest Neighbors</i>)
MAE	Erro Absoluto Médio (do inglês <i>Mean Absolute Error</i>)
MICE	Imputação Múltipla por Equações Encadeadas (do inglês <i>Multiple Imputation by Chained Equations</i>)
MNAR	Dados Ausentes Não ao Acaso (do inglês <i>Missing Not at Random</i>)
MAR	Dados Ausentes ao Acaso Dependente (do inglês <i>Missing At Random</i>)
MCAR	Dados Completamente Ausentes ao Acaso (do inglês <i>Missing Completely At Random</i>)
MSE	Erro Quadrático Médio (do inglês <i>Mean Square Error</i>)
RMSE	Raiz do Erro Quadrático Médio (do inglês <i>Root Mean Square Error</i>)
XGBoost	<i>eXtreme Gradient Boosting</i>

LISTA DE SÍMBOLOS

CO	Monóxido de carbono
H ₂ O ₂	Peróxido de hidrogênio
NO	Óxido nítrico
NO ₂	Dióxido de nitrogênio
NO _x	Óxidos de nitrogênio
O	Oxigênio atômico
O ₂	Oxigênio molecular
O ₃	Ozônio troposférico
OH	Radical hidroxila
°C	Graus Celsius
$\bar{y}$	Média dos valores reais observados
R^2	Coeficiente de determinação
y_i	Valor observado da i -ésima amostra
$\hat{y}_i$	Valor previsto da i -ésima amostra

SUMÁRIO

1	INTRODUÇÃO	14
1.1	MOTIVAÇÃO	15
1.2	OBJETIVOS	16
1.3	METODOLOGIA	17
1.4	ESTRUTURA DO TEXTO	19
2	FUNDAMENTAÇÃO TEÓRICA	20
2.1	OZÔNIO TROPOSFÉRICO	20
2.1.1	Formação e fontes de precursores do ozônio troposférico	21
2.1.2	Variáveis meteorológicas que afetam a concentração de ozônio	22
2.2	APRENDIZADO DE MÁQUINA PARA MODELAGEM PREDITIVA . . .	23
2.2.1	Séries temporais	23
2.2.2	Modelos de regressão	24
2.2.3	Métricas de avaliação	25
2.3	DADOS FALTANTES	26
2.3.1	Técnicas de imputação	27
2.3.1.1	<i>MICE</i>	28
3	MAPEAMENTO SISTEMÁTICO	29
4	ANÁLISE EXPLORATÓRIA E MÉTODOS DE IMPUTAÇÃO	38
4.1	DADOS DA ESTAÇÃO DE QUALIDADE DO AR	39
4.2	IMPUTAÇÃO COM MÉDIA	44
4.3	IMPUTAÇÃO COM MEDIANA	47
4.4	IMPUTAÇÃO COM INTERPOLAÇÃO	48
4.5	IMPUTAÇÃO COM MÉDIA MÓVEL	50
4.6	IMPUTAÇÃO COM LOCF	52
4.7	IMPUTAÇÃO COM MICE	54
5	IMPUTAÇÃO COM DADOS EXTERNOS	56
5.1	VIABILIDADE DO USO DE DADOS EXTERNOS PARA IMPUTAÇÃO .	60
5.2	DESEMPENHO DOS MODELOS PREDITIVOS	63
6	CONSIDERAÇÕES FINAIS	64
	REFERÊNCIAS	66
	APÊNDICE A – ILUSTRAÇÕES E CARACTERÍSTICAS TÉCNICAS DE SENSORES PORTÁTEIS PARA DETECÇÃO DE OZÔNIO	73
	APÊNDICE B – DISTRIBUIÇÕES E RELAÇÕES ENTRE VARIÁVEIS	76

1 INTRODUÇÃO

O ozônio troposférico (O_3) é um poluente atmosférico e sua presença pode ter efeitos distintos dependendo da altitude onde se encontra. Na estratosfera, forma a camada de ozônio, essencial para a proteção contra a radiação ultravioleta. Já na troposfera, o ozônio é um poluente secundário, formado por reações fotoquímicas entre hidrocarbonetos e óxidos de nitrogênio ($NO + NO_2$) na presença de luz solar. Seu acúmulo prejudica a qualidade do ar e a saúde humana, e é um dos principais componentes do *smog* fotoquímico (Manahan, 2013).

O *smog* fotoquímico aparece como uma névoa amarronzada sobre áreas urbanas densamente povoadas e causa efeitos negativos na biosfera, na tecnosfera e na saúde humana (Sher, 1998; Trashko *et al.*, 2022). Mesmo longe dos grandes centros urbanos, a exposição ao ozônio ainda representa riscos à saúde. Estudos mostram que níveis mais altos de ozônio, mesmo em curtos períodos, aumentam o número de internações por doenças respiratórias, como as que compõem a Doença Pulmonar Obstrutiva Crônica (DPOC), principalmente nos dias mais quentes (Gao *et al.*, 2020).

Além disso, evidências mostram que o ozônio contribui para o aumento da mortalidade por causas gerais, cardiovasculares e respiratórias, mesmo após a implementação de políticas ambientais, como observado em diversas cidades chinesas (Zhang *et al.*, 2022). Além dos impactos na saúde e os efeitos negativos sobre o meio ambiente, em que o ozônio pode alterar o desenvolvimento das plantas, ele também compromete as interações com polinizadores e reduz o sucesso reprodutivo de algumas espécies (Duque; Poelman; Steffan-Dewenter, 2021).

Por causa desses impactos à saúde e ao meio ambiente, o monitoramento contínuo da qualidade do ar é fundamental. De acordo com Vormittag *et al.* (2021), esse monitoramento permite acompanhar as concentrações dos poluentes atmosféricos, entender as condições atuais, construir séries históricas de dados e apoiar decisões para a proteção da população e do ambiente.

No entanto, essa atividade ainda enfrenta diversos desafios no Brasil. A rede nacional de monitoramento tem limitações e está concentrada principalmente na Região Sudeste e, em muitos estados, quase não existe. Além disso, as estações em operação, sejam públicas ou privadas, nem sempre fornecem dados continuamente. Problemas como falhas em sensores, quedas de energia, manutenção precária ou falhas na comunicação comprometem a qualidade dos dados e a análise das séries temporais (Pereira *et al.*, 2024; Correa *et al.*, 2021).

Apesar desses problemas operacionais, a previsão da concentração de ozônio continua sendo um desafio que vai além dos dados, e precisa de métodos para antecipar episódios de alta poluição e apoiar decisões. Isso se deve, em grande parte, à complexidade envolvida na própria formação do poluente, altamente dependente de condições meteorológicas, topográficas e de fontes específicas de emissão de precursores, como veículos automotores e atividades industriais (Netto *et al.*, 2017).

Estudos como o de Alves e Alves (2019) mostram a variação temporal e espacial de concentrações de ozônio, especialmente em áreas industriais, reforçando a necessidade de um

monitoramento contínuo. Soma-se a isso o desafio do alto custo das estações de monitoramento da qualidade do ar, uma vez que, segundo Sokolova, Yurgenson e Shakhov (2025), implantar e manter essas estações pode ser muito caro.

No Brasil, segundo levantamento conduzido por Vormittag *et al.* (2021) com 26 Secretarias Estaduais/Distrital de Meio Ambiente, 12 estados afirmaram não realizar o monitoramento da qualidade do ar. As principais justificativas foram a falta de recursos para obtenção e manutenção de equipamentos, além da falta de pessoal qualificado.

Nesse contexto, ganha importância a promulgação da Lei nº 14.850/2024 (BRASIL, 2024), que institui a Política Nacional de Qualidade do Ar, com diretrizes para a padronização de indicadores, à ampliação da cobertura do monitoramento, à integração nacional dos dados e à garantia de acesso público às informações. Apesar de representar um avanço importante, a nova legislação ainda não resolve os problemas estruturais existentes, como a inconsistência dos registros históricos e a limitada cobertura de monitoramento em várias regiões do país.

Diante dessas limitações, a análise de séries temporais tem se destacado na predição da concentração de ozônio troposférico, pois consegue identificar padrões e sazonalidades ao longo do tempo. Com o avanço das técnicas de computação e em especial da inteligência artificial, modelos baseados em aprendizado de máquina têm apresentado bons resultados no apoio à gestão da qualidade do ar e na mitigação de impactos ambientais e à saúde pública.

Estudos recentes conduzidos em países como Índia (Venuka *et al.*, 2024), Malásia (Latif *et al.*, 2024), México (Domínguez-García; Arellano-Vázquez, 2024) e Suíça (Räss; Leuenberger, 2025) evidenciam o interesse internacional pelo tema, reforçando sua relevância no contexto atual das pesquisas em qualidade do ar. No Brasil, a importância de expandir o monitoramento da qualidade do ar também é evidente, e, por isso, este estudo foca na análise de dados da cidade de Ponta Grossa, no Paraná, que representa uma realidade urbana fora dos grandes centros, buscando explorar soluções aplicáveis à realidade local.

Embora existam dados públicos disponíveis, em muitos casos eles são limitados em cobertura, granularidade ou frequência de atualização. Por isso, este trabalho recorre a dados privados, que oferecem maior detalhamento e possibilitam análises mais robustas e alinhadas aos objetivos propostos.

1.1 MOTIVAÇÃO

O ozônio troposférico é um dos principais poluentes atmosféricos e é prejudicial à saúde humana, estando associado a sintomas como dor de cabeça, dor no peito, tosse, dor de garganta e congestão respiratória (Chu *et al.*, 2024).

No contexto brasileiro, estudos como o de Santana, Moreira e Armani (2020), realizado em Paranaguá, Paraná, associam a exposição a poluentes atmosféricos ao aumento de internações por doenças respiratórias. O autor também mostra que, no período de maio de 2016 a fevereiro de 2017, Ponta Grossa registrou um índice de internações em relação à população mais elevado

que o de Cubatão, São Paulo, conhecida historicamente por seus altos níveis de poluição.

Apesar do avanço representado pela nova legislação, sensores específicos para O_3 são caros (ver Apêndice A), precisam de manutenção especializada e boa infraestrutura, o que limita a quantidade de estações que realizam esse monitoramento no Brasil. Além disso, falhas operacionais e condições ambientais adversas frequentemente resultam em dados faltantes, e a presença de lacunas nos dados segue sendo um obstáculo crítico, afetando diretamente a análise contínua e confiável desse poluente.

Durante o estudo, Santana, Moreira e Armani (2020) enfrentou interrupções nas medições da estação de qualidade do ar em Paranaguá devido a quedas de energia, manutenção e presença de valores atípicos (identificados pelos valores superiores ao limite de detecção dos sensores), removendo estes últimos e avaliando apenas dias com mais de 80% das horas medidas, o que resultou em 217 dias válidos de 10 meses monitorados.

Portanto, torna-se necessário buscar alternativas que permitam estimar e prever a concentração de ozônio com base em variáveis meteorológicas e de outros poluentes primários. A análise de séries temporais, combinada com técnicas de imputação de dados, aparece como uma boa solução para lidar com a ausência de medições diretas e aprimorar modelos preditivos.

Este trabalho busca contribuir nesse sentido, explorando estratégias para o tratamento de dados faltantes em séries temporais, permitindo estimativas mais precisas e confiáveis, e avaliando diferentes algoritmos de aprendizado de máquina para a previsão de concentração de ozônio.

1.2 OBJETIVOS

O objetivo geral deste trabalho é desenvolver e avaliar modelos inteligentes resilientes a dados faltantes, voltados à previsão da concentração de ozônio na troposfera, com o objetivo de contribuir para o monitoramento da qualidade do ar e a mitigação de impactos ambientais e à saúde. Para alcançar este objetivo, são definidos os seguintes Objetivos Específicos (OE):

OE 1. Realizar uma revisão da literatura sobre métodos de predição da concentração de ozônio troposférico com algoritmos de aprendizado de máquina e métodos de imputação de dados;

OE 2. Tratar os dados da base utilizada, aplicando técnicas de imputação para lidar com valores ausentes e garantir a qualidade das informações;

OE 3. Implementar modelos preditivos baseados em aprendizado de máquina para previsão do ozônio troposférico;

OE 4. Explorar o impacto de variáveis meteorológicas e poluentes precursores na concentração de ozônio, visando melhorar o desempenho dos modelos; e

OE 5. Avaliar a eficácia dos modelos por meio de métricas estatísticas como MAE, RMSE e R^2 .

1.3 METODOLOGIA

Esta pesquisa se insere na área da Ciência da Computação, entendida aqui como um campo diverso que combina métodos formais e empíricos, conforme a reflexão de Wazlawick (2020). O estudo propõe o desenvolvimento e a avaliação de modelos preditivos com aprendizado de máquina para prever a concentração de ozônio troposférico, utilizando dados reais de uma estação local de monitoramento da qualidade do ar. Técnicas como análise de séries temporais, imputação de dados faltantes e o uso de modelos computacionais caracterizam a pesquisa em Ciência da Computação.

Quanto ao tipo de ciência, este trabalho pode ser classificado como uma combinação de diferentes dimensões segundo Wazlawick (2020). Trata-se de uma ciência exata, pelo uso de modelos computacionais e métricas estatísticas como RMSE, MAE e R^2 ; empírica, por depender da observação e análise de dados ambientais reais; aplicada, já que visa resolver um problema prático de interesse público, como a previsão da concentração de poluentes; dura, devido ao rigor científico e o uso de métodos estatísticos para comprovar os resultados; e nomotética, pois busca encontrar correlações no comportamento do ozônio a partir de variáveis meteorológicas e precursoras.

O paradigma adotado neste trabalho é o tecnocrático, conforme definido por Eden (2007), que caracteriza uma abordagem focada na engenharia de soluções práticas. Diferentemente da busca por provas formais ou desenvolvimento de novas teorias, este estudo concentra-se no desenvolvimento e avaliação empírica de modelos computacionais para previsão da concentração de ozônio em cenários reais, visando apoiar políticas públicas e decisões ambientais.

Do ponto de vista da maturidade da pesquisa, ela se enquadra no Estilo 3, segundo Wazlawick (2020), caracterizado pela “apresentação de algo presumivelmente melhor”. Isso porque o trabalho propõe e valida modelos de aprendizado de máquina para prever a concentração de ozônio, enfrentando desafios operacionais de estações brasileiras, como dados faltantes, utilizando conjuntos de dados específicos que não fazem parte de uma base de dados internacional consolidada. Assim, a pesquisa busca superar limitações identificadas na literatura por meio de abordagens computacionais robustas, deixando claro o contexto e as condições em que as melhorias são obtidas.

Quanto ao objetivo geral, classifica-se inicialmente como uma pesquisa exploratória, de acordo com a definição de Gil (2022), pois busca compreender e levantar hipóteses sobre a influência de variáveis meteorológicas e poluentes precursores na previsão de ozônio com dados ambientais incompletos. No entanto, o trabalho também apresenta aspectos descritivos e explicativos ao identificar padrões e testar modelos que permitem entender relações entre variáveis e melhorar a capacidade preditiva dos modelos estudados.

Em relação ao raciocínio lógico, esta pesquisa adota uma abordagem hipotético-dedutiva, conforme descrita por Lakatos e Marconi (2021). Parte-se de fundamentos teóricos sobre a formação do ozônio e sobre os algoritmos de aprendizado de máquina para formular hipóteses aplicáveis a um caso específico, envolvendo dados reais da estação de monitoramento de uma cidade brasileira. A hipótese central é que, com base no conhecimento existente e nos dados disponíveis, é possível construir modelos computacionais capazes de prever, com eficácia, os níveis de ozônio na região estudada. Essas hipóteses são então testadas empiricamente.

Quanto à abordagem do problema, ou natureza das variáveis, esta pesquisa é de natureza quantitativa, conforme caracterização de Gil (2019). Trabalha-se com dados numéricos, como temperatura, umidade relativa do ar e concentração de poluentes, analisados com base em métricas estatísticas que permitem mensurar o desempenho dos modelos.

Quanto ao procedimento de coleta de dados, ainda que os dados não tenham sido coletados diretamente pela pesquisadora, sua origem e natureza caracterizam-nos como uma pesquisa de campo, segundo Lakatos e Marconi (2021). Os dados foram originalmente coletados por sensores instalados em uma estação de monitoramento da qualidade do ar, diretamente no ambiente em que os fenômenos ocorrem. Isso diferencia esta pesquisa de uma pesquisa puramente documental. Os dados ambientais históricos, como temperatura, umidade e concentração de poluentes, refletem medições contínuas realizadas em campo e exigem tratamento para garantir sua qualidade e validade, especialmente diante da presença de lacunas causados por valores ausentes.

Quanto ao grau de controle das variáveis, a pesquisa adota a distinção entre variáveis independentes e dependentes, conforme Lakatos e Marconi (2021). A concentração de ozônio é considerada a variável dependente, cuja variação busca ser explicada por meio de variáveis independentes como condições meteorológicas e concentrações de poluentes precursores. A estrutura dos modelos de aprendizado de máquina estudados utiliza essa distinção para identificar padrões e construir correlações entre os dados, permitindo a predição do ozônio troposférico neste contexto.

Do ponto de vista experimental, os modelos serão implementados e avaliados em ambiente computacional utilizando a biblioteca *scikit-learn* (Pedregosa *et al.*, 2011), que oferece ferramentas integradas para criação de *pipelines*, pré-processamento de dados, cálculo de métricas estatísticas e separação dos dados em conjuntos de treino e teste. A validação será conduzida com base em uma abordagem de *train-test split*, assegurando que os modelos sejam avaliados em dados não vistos durante o treinamento, o que reforça a confiabilidade dos resultados.

Embora os dados utilizados sejam provenientes de uma única estação de monitoramento, localizada em uma cidade brasileira do estado do Paraná, reconhece-se que essa limitação geográfica pode implicar em restrições na generalização dos resultados. Os padrões identificados podem não se aplicar diretamente a outras localidades com características climáticas, geográficas ou de poluição diferentes.

Por fim, quanto ao procedimento técnico, esta pesquisa combina a bibliográfica, pela revisão de literatura existente sobre o tema, e experimental, pela implementação e avaliação de

modelos computacionais em ambiente controlado. Wazlawick (2020) destaca que a pesquisa bibliográfica, embora essencial para fundamentar teoricamente o estudo e ampliar o conhecimento do pesquisador, por si só não é suficiente para gerar novos conhecimentos. Por isso, ela é complementada com um procedimento experimental, que permite aplicar, testar e validar, de forma prática, os conceitos estudados.

Essa caracterização tem como objetivo situar o trabalho dentro das diversas abordagens possíveis na Ciência da Computação, justificando as escolhas adotadas. Ao identificar o tipo de ciência, o paradigma epistemológico, o estilo de maturidade, os métodos de coleta e análise de dados, entre outros aspectos, pode-se comprovar seu rigor e relevância científica.

1.4 ESTRUTURA DO TEXTO

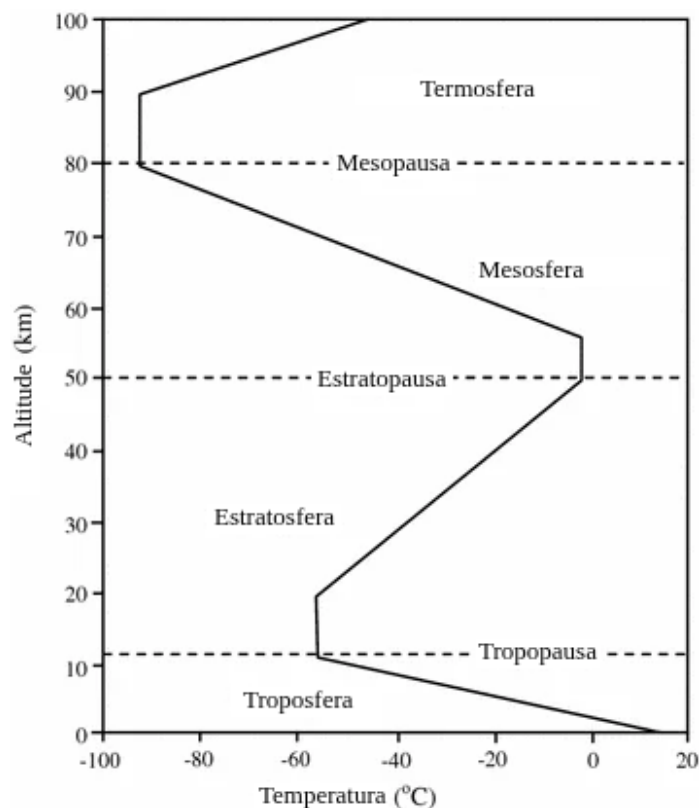
Este trabalho está estruturado da seguinte forma: O Capítulo 2 apresenta a fundamentação teórica sobre ozônio troposférico, variáveis meteorológicas associadas, aprendizado de máquina para modelagem preditiva, bem como conceitos de dados faltantes e técnicas de imputação. O Capítulo 3 descreve um mapeamento sistemático da literatura sobre aprendizado de máquina aplicado à predição de ozônio e tratamento de dados faltantes. O Capítulo 4 apresenta a análise exploratória dos dados das estações de qualidade do ar e a avaliação de métodos para a imputação dos dados faltantes. O Capítulo 5 apresenta um método de imputação com dados externos oriundos de estação meteorológica, abordando a viabilidade dessa estratégia e o desempenho dos modelos preditivos nesse cenário. Por fim, o Capítulo 6 apresenta as considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 OZÔNIO TROPOSFÉRICO

O ozônio troposférico é um gás composto por três átomos de oxigênio O_3 presente na camada mais baixa da atmosfera, conhecida como troposfera (Figura 2.1). Essa camada se estende do nível do solo até aproximadamente 8 a 16 km de altitude, variando conforme a latitude e as condições climáticas (Barry; Chorley, 2009). É na troposfera que ocorrem os fenômenos meteorológicos e onde se concentram as principais atividades humanas (Atkinson, 2000). O ozônio nessa região não é emitido diretamente por fontes naturais ou antrópicas, sendo classificado como um poluente secundário, formado por reações químicas entre precursores na presença da radiação solar (Avino; Manigrasso, 2015).

Figura 2.1 – Variação da temperatura com a altitude nas camadas da atmosfera. O gráfico mostra a relação entre temperatura ($^{\circ}C$) e altitude (km), identificando as principais camadas atmosféricas e suas regiões de transição.



Fonte: Adaptado de Chandrappa e Kulshrestha (2016).

A luz solar atua como catalisadora de reações fotoquímicas que envolvem misturas complexas de compostos orgânicos e inorgânicos, resultando na formação de gases altamente reativos como o próprio ozônio O_3 , o dióxido de nitrogênio NO_2 , o peróxido de hidrogênio H_2O_2 e o peroxiacetilnitrato (PAN) (Marshall; Fairbridge, 1999). Essa dinâmica faz com que suas concentrações variem ao longo do dia e dependam diretamente das quantidades relativas de seus precursores atmosféricos (Ahmad *et al.*, 2023). Além disso, fatores meteorológicos como a

radiação solar, a temperatura e a inversão térmica (fenômeno atmosférico onde a temperatura do ar aumenta com a altitude, em vez de diminuir) também apresentam correlação significativa com a variação das concentrações de ozônio ao longo do dia (Ahmad *et al.*, 2022) e durante as diferentes estações do ano (Qu *et al.*, 2021).

2.1.1 Formação e fontes de precursores do ozônio troposférico

A formação do ozônio troposférico ocorre por meio de reações fotoquímicas iniciadas pela decomposição do dióxido de nitrogênio (NO_2) na presença de luz solar, envolvendo compostos orgânicos e inorgânicos já presentes na atmosfera, que então libera monóxido de nitrogênio (NO) e oxigênio atômico (O), conforme a Equação 1 (Atkinson, 2000; Seinfeld; Pandis, 2016):



Em seguida, esse oxigênio atômico reage rapidamente com o oxigênio molecular (O_2) presente na atmosfera, formando ozônio (O_3), como mostra a Equação 2:



Embora essa segunda etapa seja mais eficiente durante o dia, devido à maior disponibilidade de radiação solar para iniciar a fotólise do NO_2 , a formação de ozônio pode continuar à noite. Isso ocorre por meio de reações envolvendo oxigênio atômico remanescente, que ainda pode colidir com O_2 e formar O_3 , mesmo em condições de baixa luminosidade (Marshall; Fairbridge, 1999).

No entanto, o ozônio pode ser consumido novamente se reagir com o NO, regenerando NO_2 e O_2 , conforme a Equação 3:

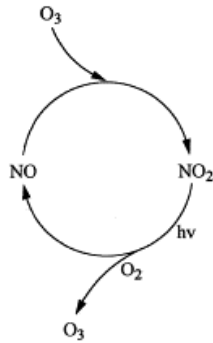


Segundo Ahmad *et al.* (2023), NO, NO_2 e O_3 atingem rapidamente um estado fotoestacionário, formando um sistema fechado e mantendo constante a relação entre substâncias precursoras e compostos oxidantes no sistema fechado representado na Figura 2.2a. No entanto, esse equilíbrio muda quando há presença de Compostos Orgânicos Voláteis (COVs) (Figura 2.2b), que ao reagirem com outras substâncias presentes no ar, como os radicais hidroxila (OH), acabam consumindo parte do óxido nítrico (Seinfeld; Pandis, 2016). Como resultado, há uma diminuição da disponibilidade de NO para atuar na destruição do O_3 , favorecendo seu acúmulo na atmosfera.

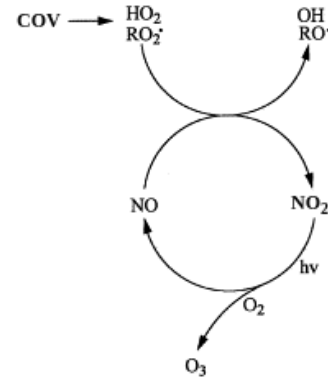
Esse acúmulo de O_3 é acelerado pela radiação solar e pela disponibilidade de precursores. Estudos em áreas urbanas mostram que a velocidade de formação do ozônio atinge seu máximo no período da tarde, quando altas concentrações de NO_x (provenientes do tráfego) e intensa radiação UV favorecem as reações fotoquímicas (Ahmad *et al.*, 2023). No entanto, a relação entre a concentração de ozônio e a presença de dióxido de nitrogênio não é linear e, em baixas concentrações de NO_2 , o ozônio pode ser degradado, enquanto em concentrações mais altas sua formação tende a ser intensificada (Benchirifa *et al.*, 2024).

Figura 2.2 – Ciclo do ozônio troposférico com e sem a presença de COVs. Destaca-se o papel dos COVs na quebra do equilíbrio natural e no aumento da concentração de O_3 .

(a) O ciclo entre NO , NO_2 e O_3 ocorre sem interferência externa.



(b) Os COVs introduzem radicais que intensificam a formação de O_3 .



Fonte: Adaptado de Atkinson (2000).

2.1.2 Variáveis meteorológicas que afetam a concentração de ozônio

A formação do ozônio depende, além dos precursores, de fatores como as emissões locais, que podem ser de origem antrópica, como aquelas provenientes de veículos e indústrias, ou biogênica, como os compostos liberados pela vegetação. Segundo Li *et al.* (2024), outros dois fatores que também influenciam esse processo são o transporte atmosférico, que envolve tanto o deslocamento de precursores quanto o transporte direto do próprio ozônio entre regiões, e as condições meteorológicas.

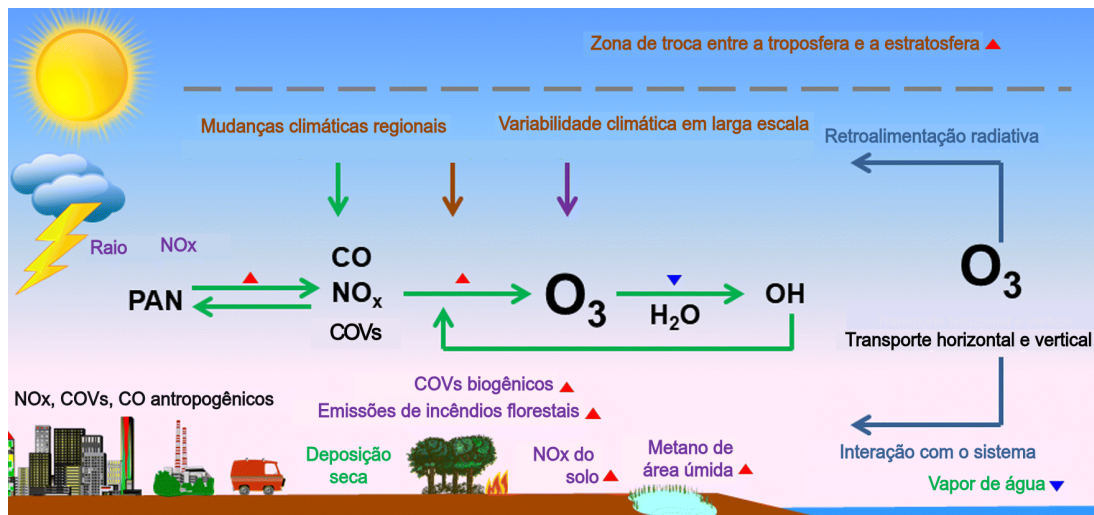
A temperatura é um dos fatores meteorológicos de maior impacto. Temperaturas mais elevadas tendem a aumentar a concentração de ozônio, particularmente em regiões onde seus precursores ainda não foram totalmente consumidos pelas reações atmosféricas (Aw; Kleeman, 2003). Essa relação se explica pela aceleração das taxas de reação fotoquímica que geram o ozônio troposférico, fazendo com que o calor intensifique os picos de poluição quando as condições são favoráveis.

Além da temperatura, Lu, Zhang e Shen (2019) destacam que a vida média do ozônio e de seus precursores na troposfera livre pode chegar a meses, tornando-os altamente suscetíveis ao transporte por ventos. Situações de ventilação intensa (passagem de frentes, ciclones ou brisas marítimas) tendem a diluir o poluente, ao passo que períodos de estagnação ou sistemas de bloqueio favorecem a acumulação local de NO_x , COVs e, consequentemente, de O_3 .

A Figura 2.3 ilustra as interações entre emissões naturais e antrópicas, processos químicos, transporte atmosférico e retroalimentações que afetam a concentração de ozônio (O_3) na troposfera. Os triângulos vermelhos indicam processos que, no cenário de aquecimento, tendem a aumentar o O_3 , enquanto os azuis apontam os que o reduzem. Entre os processos de redução está a umidade, pois o vapor de água favorece rotas de remoção do poluente na troposfera ao estimular a formação de radicais que consomem ozônio (Jacob; Winner, 2009). As cores dos títulos na figura indicam as categorias dos processos envolvidos: emissões naturais (roxo),

química e deposição (verde), transporte atmosférico (marrom) e retroalimentações climáticas (azul).

Figura 2.3 – Fatores que influenciam a formação do ozônio troposférico. Elementos como NO_x, COVs e CO reagem sob a influência da radiação solar e mudanças climáticas, contribuindo para o aumento ou redução do O₃. A figura destaca também a importância das condições meteorológicas e do papel do vapor de água.



Fonte: Adaptado de Lu, Zhang e Shen (2019).

2.2 APRENDIZADO DE MÁQUINA PARA MODELAGEM PREDITIVA

Os modelos de aprendizado de máquina para modelagem preditiva são técnicas não paramétricas que utilizam dados históricos para aprender padrões e relações complexas entre variáveis. Esses modelos, como regressão linear, Support Vector Regression (SVR), Gaussian Processes Regression (GPR), Multilayer Perceptron (MLP) e métodos de *ensemble* (Jafarzadeh *et al.*, 2021), são aplicados em diversas áreas, incluindo a previsão de poluição por ozônio, e demonstram flexibilidade e capacidade de extrair características relevantes de dados multivariados (Pan; Harrou; Sun, 2023).

Além disso, modelos não lineares, especialmente aqueles que incorporam interações entre variáveis como radiação, temperatura e poluentes, podem superar redes neurais complexas em precisão (Räss; Leuenberger, 2025). Dentre esses métodos, destaca-se o XGBoost (*eXtreme Gradient Boosting*) (Chen; Guestrin, 2016), um algoritmo de aprendizado de máquina não linear baseado em árvores de decisão.

2.2.1 Séries temporais

Uma série temporal é uma sequência de observações de uma mesma variável, registradas ao longo do tempo em intervalos regulares (Vishwas; Patel, 2020; Weiß, 2021), como medições horárias da concentração de ozônio. Diferentemente de conjuntos de dados em que a ordem

das observações é irrelevante, em séries temporais a posição no tempo é fundamental, pois valores próximos tendem a ser mais parecidos entre si do que valores distantes. Essa dependência temporal é o que permite utilizar o comportamento passado para prever valores futuros da série.

No contexto da qualidade do ar, as séries de ozônio apresentam padrões bem definidos ao longo do dia e do ano. Em geral, observam-se ciclos diários associados à variação da radiação solar e da temperatura, além de padrões sazonais relacionados a mudanças meteorológicas e de emissões. Esses efeitos fazem com que a série de ozônio seja, em muitos casos, não estacionária, isto é, suas médias e variâncias podem mudar ao longo do tempo em função de condições meteorológicas e antrópicas.

A função de autocorrelação permite quantificar a dependência temporal, medindo o quanto a série em um instante t se relaciona com a mesma série em instantes futuros $t + k$, sendo k a defasagem.

Na modelagem preditiva com aprendizado de máquina, a estrutura temporal pode ser incorporada ao transformar observações passadas da própria série e de variáveis associadas em variáveis explicativas (*lags*). Essas observações defasadas são então utilizadas como entradas em modelos de regressão, permitindo que o algoritmo aprenda relações entre o passado e o futuro da série. Além disso, a avaliação dos modelos deve respeitar a ordem temporal dos dados, utilizando esquemas de validação que evitam o uso de informações futuras no treinamento, prevenindo vazamento de dados e fornecendo estimativas mais realistas de desempenho.

2.2.2 Modelos de regressão

Os algoritmos de regressão estimam uma variável alvo contínua a partir de variáveis de entrada, aprendendo uma relação funcional entre elas com base em dados históricos (Lorena *et al.*, 2021). Em séries temporais, esses modelos aprendem padrões entre variáveis observadas em instantes anteriores (variáveis preditoras) e o valor futuro ou atual da variável de interesse (variável alvo). Os modelos podem ser lineares ou não lineares e incluem métodos baseados em instâncias, árvores de decisão e técnicas de combinação de modelos (*ensemble*) (Jafarzadeh *et al.*, 2021).

A escolha do algoritmo depende das características do problema. Métodos lineares (regressão linear, ElasticNet, Lasso) assumem relações aditivas entre variáveis e são interpretáveis, mas limitados em capturar interações complexas (Manoj; Manjunath; Kotagi, 2025). Modelos baseados em instâncias (kNN) operam por similaridade local entre observações. Árvores de decisão (Decision Tree, CART) dividem os dados em grupos cada vez menores baseados em regras sequenciais, mas tendem a sobreajustar (Lorena *et al.*, 2021). Técnicas de combinação de modelos combinam múltiplos modelos para reduzir viés ou variância. Métodos de agregação por amostragem (*bagging*), como Random Forest, agregam árvores independentes treinadas em amostras aleatórias, enquanto métodos de impulsionamento (*boosting*) (Jafarzadeh *et al.*, 2021), como XGBoost, AdaBoost e LightGBM, constroem modelos sequencialmente, corrigindo erros das iterações anteriores.

O *k-Nearest Neighbors* (kNN) estima o valor de saída para uma nova observação a partir dos valores dos k vizinhos mais próximos no espaço de características (Lorena *et al.*, 2021). A previsão é obtida pela média ou mediana dos alvos vizinhos, definidos por uma medida de distância (euclidiana, Manhattan, Minkowski). Não há treinamento explícito de parâmetros. O modelo apenas armazena os dados de treino e faz buscas no momento da predição. O comportamento é controlado por k (número de vizinhos) e pela métrica de distância escolhida. Valores baixos de k tornam o modelo sensível a ruídos, enquanto valores altos suavizam as predições mas podem perder padrões locais. O kNN funciona bem quando as relações entre variáveis são localmente consistentes e não seguem padrões complexos. Em séries temporais, o kNN pode capturar padrões recorrentes ao identificar instantes históricos similares ao momento atual.

O *Random Forest* (RF) é um método de combinação de modelos baseado em múltiplas árvores de decisão treinadas de forma independente (Lorena *et al.*, 2021). Cada árvore é construída a partir de uma amostragem com reposição (*bootstrap*) dos dados de treino e, em cada divisão da árvore, apenas um subconjunto aleatório de variáveis preditoras é considerado. A previsão final é calculada pela média das previsões individuais de todas as árvores. Essa aleatoriedade dupla (nas observações e nas variáveis) força as árvores a aprenderem aspectos diferentes dos dados, reduzindo a correlação entre elas. Como resultado, a agregação das previsões reduz a variância em relação a uma única árvore, tornando o modelo mais robusto a ruídos e menos propenso ao sobreajuste. O RF também fornece medidas de importância de variáveis, indicando quais preditores mais contribuem para as predições, o que auxilia na interpretação do modelo.

O XGBoost é um algoritmo de combinação de modelos baseado em impulsionamento por gradiente (*gradient boosting*), que constrói múltiplas árvores de decisão de forma sequencial. Cada nova árvore é treinada para corrigir os erros residuais das árvores anteriores, minimizando uma função de perda através de otimização por gradiente. O algoritmo incorpora regularização L1 e L2 nos pesos das folhas, penalizando modelos complexos e controlando o sobreajuste. Outros mecanismos de regularização incluem limitação de profundidade das árvores, taxa de aprendizado reduzida (que controla a contribuição de cada nova árvore) e subamostragem de dados e variáveis a cada iteração. O XGBoost também implementa tratamento nativo de valores faltantes e otimizações computacionais (paralelização, uso eficiente de memória cache, estruturas de dados esparsas), permitindo o treinamento em grandes volumes de dados com recursos limitados. O algoritmo é flexível e pode ser aplicado em problemas de classificação, regressão e ordenação (*ranking*). Em séries temporais, o XGBoost captura relações complexas e não-lineares entre múltiplas variáveis, sendo especialmente eficaz quando há interações heterogêneas entre preditores ao longo do tempo.

2.2.3 Métricas de avaliação

A avaliação de modelos preditivos é realizada com base em métricas estatísticas que quantificam o erro entre os valores previstos e os observados. Uma das métricas mais comuns é

o Erro Absoluto Médio (MAE, do inglês *Mean Absolute Error*), definido por:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

O MAE mede o valor médio das diferenças absolutas entre os valores reais (y_i) e os valores previstos ($\hat{y}_i$), dentre as n amostras. Quanto menor o MAE, melhor o desempenho do modelo, sendo que ele não penaliza tanto grandes erros isolados.

O Erro Quadrático Médio (MSE, do inglês *Mean Squared Error*) é outra métrica importante, calculada como:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

O MSE representa a média dos erros ao quadrado. Por exemplo, um MSE de 4 indica que, em média, os erros quadráticos têm magnitude equivalente a 4 vezes a unidade original ao quadrado. Como o erro é elevado ao quadrado, discrepâncias maiores têm impacto maior no valor final, o que torna o MSE sensível a casos extremos (*outliers*).

Outra métrica amplamente utilizada é a Raiz do Erro Quadrático Médio (RMSE, do inglês *Root Mean Square Error*), definida pela Equação (6). Diferente do MAE, o RMSE penaliza mais fortemente os casos extremos pois eleva as diferenças ao quadrado antes de calcular a média. Assim como no MAE, um RMSE baixo indica melhor desempenho preditivo.

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

O Coeficiente de Determinação (R^2) é uma métrica que avalia a proporção da variância dos dados que foi explicada pelo modelo. O valor de R^2 varia entre 0 e 1 e indica quanto da variabilidade dos dados o modelo consegue explicar. Quanto mais próximo de 1, indica que o modelo explica bem os dados observados, o que reflete uma alta proporção da variância dos valores reais sendo explicada pelas previsões do modelo. Quando próximo de 0, significa que o modelo não consegue explicar melhor a variação dos dados do que uma simples média. Valores intermediários mostram proporcionalmente qual fração da variação total é explicada pelo modelo. Sua fórmula é:

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

Onde $\bar{y}$ é a média dos valores reais das amostras, calculado como:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (8)$$

2.3 DADOS FALTANTES

Dados faltantes são recorrentes em séries temporais ambientais, sobretudo em medições obtidas por sensores. Segundo Pedersen *et al.* (2017), as ausências podem ser agrupadas em

três tipos principais: Dados Ausentes Completamente ao Acaso (MCAR, do inglês *Missing Completely at Random*), Dados Ausentes ao Acaso (MAR, do inglês *Missing at Random*) e Dados Ausentes Não ao Acaso (MNAR, do inglês *Missing Not at Random*) .

No caso MCAR, a ausência ocorre de forma totalmente aleatória, sem relação com dados observados ou não observados. Um exemplo seria a quebra acidental de um sensor, o que impede o registro de uma variável, independentemente do seu valor. Já no tipo MAR, os dados ausentes dependem de variáveis observadas. Por exemplo, se sensores apresentam mais falhas em determinados períodos do dia ou em condições climáticas específicas, mas a ausência não está relacionada diretamente ao valor da variável de interesse, trata-se de um caso MAR.

MNAR é o cenário em que a ausência está relacionada ao próprio valor não observado. Um exemplo seria quando sensores deixam de registrar dados justamente em situações de alta concentração de ozônio, tornando os dados faltantes sistematicamente ligados ao fenômeno que se pretende analisar. Identificar corretamente o tipo de ausência é importante, pois métodos como o MICE assumem o mecanismo MAR, enquanto dados MNAR exigem abordagens mais complexas, como o uso de modelos causais.

2.3.1 Técnicas de imputação

A imputação de dados é uma estratégia para lidar com valores ausentes em séries temporais de sensores, evitando perdas de informação nas análises. Faizin, Riasetiawan e Ashari (2019) agrupam os métodos de imputação em diferentes abordagens.

Os métodos mais simples incluem técnicas estatísticas como a média, mediana e o valor anterior (LOCF, do inglês *Last Observation Carried Forward*). Embora sejam computacionalmente eficientes, essas abordagens não capturam adequadamente a variabilidade temporal e podem distorcer a estrutura dos dados.

Técnicas baseadas em modelos temporais, como ARIMA (Autorregressivo Integrado de Médias Móveis (Vishwas; Patel, 2020), do inglês *Autoregressive Integrated Moving Average*), são eficazes em séries temporais univariadas, enquanto outras abordagens exploram métodos baseados em aprendizado profundo, como Redes Neurais Recorrentes (RNNs, do inglês *Recurrent Neural Networks*) e *autoencoders*, conseguem lidar com sequências temporais mais complexas.

Modelos de aprendizado de máquina, como *k-Nearest Neighbors* (kNN) e Random Forest (RF), também são utilizados para tratar dados faltantes. A imputação com Random Forest é eficaz mesmo em cenários com altos níveis de ausência e variáveis correlacionadas. O algoritmo constrói diversas árvores de decisão e utiliza médias ou votos majoritários para estimar valores ausentes de forma iterativa.

O kNN se destaca como um método robusto de imputação, ao utilizar a similaridade entre instâncias completas para estimar valores ausentes. Essa abordagem, apesar de não assumir a distribuição dos dados, pode preservar padrões locais.

A escolha do método depende do padrão de ausência, da estrutura dos dados e do objetivo da análise. Em situações em que os dados faltantes ocorrem de forma esparsa e aproximadamente

aleatória (MCAR), métodos simples, como imputação pela média, podem ter impacto limitado sobre as análises. No entanto, em cenários mais complexos, como falhas de sensores em séries ambientais, a ausência de dados tende a ocorrer de forma sistemática e em blocos (não aleatória), frequentemente associada a mecanismos do tipo MAR ou MNAR. Nesses casos, métodos simplificados podem introduzir viés significativo e distorcer a variabilidade e as dependências temporais dos dados.

Para dados ambientais, onde múltiplas variáveis meteorológicas frequentemente se relacionam de forma não-linear, técnicas multivariadas ou baseadas em aprendizado de máquina tendem a ser mais adequadas para capturar relações complexas (Stekhoven; Bühlmann, 2011), pois são capazes de explorar a correlação entre variáveis e capturar padrões complexos, mesmo na presença de grandes proporções de dados ausentes.

2.3.1.1 MICE

Os métodos multivariados iterativos, como a Imputação Múltipla por Equações Encadeadas (MICE, do inglês *Multiple Imputation by Chained Equations*) (Buuren; Groothuis-Oudshoorn, 2011), realizam imputações múltiplas em variáveis com dados faltantes por meio de um processo iterativo baseado em regressões condicionais.

A abordagem consiste em imputar cada variável incompleta usando as demais como preditoras, de forma encadeada. Valores iniciais são atribuídos aos dados ausentes. Em seguida, o algoritmo passa a atualizar esses valores em ciclos, utilizando modelos apropriados para cada tipo de variável (numérica, categórica, ordenada).

A repetição desse processo é feita algumas vezes (10 a 20 iterações) e é capaz de gerar valores plausíveis que preservam as relações estatísticas entre as variáveis. O MICE, embora seja robusto, assume que os dados ausentes são MAR e demandam maior custo computacional.

3 MAPEAMENTO SISTEMÁTICO

Esta revisão segue a metodologia de mapeamento sistemático de Petersen, Vakkalanka e Kuzniarz (2015) e busca identificar, analisar e comparar métodos de imputação de dados e abordagens preditivas para previsão de ozônio troposférico em séries temporais com dados faltantes.

Como complemento, o método PICO (Paciente ou Problema, Intervenção, Comparação e Resultado, do inglês *Patient or Problem, Intervention, Comparison, and Outcome*) (Petticrew; Roberts, 2006) foi adotado para formular as questões de pesquisa de maneira estruturada, considerando como população as séries temporais de dados atmosféricos provenientes de estações de monitoramento da qualidade do ar com dados faltantes; como intervenção, os modelos preditivos baseados em aprendizado de máquina e imputação de dados para previsão de ozônio troposférico; estabelecendo comparações com as métricas estatísticas entre diferentes algoritmos de aprendizado de máquina e técnicas de imputação de dados; e tendo como resultado (*outcome*) a análise do desempenho preditivo dos modelos e sua eficácia no tratamento de dados faltantes. Com base nessa estrutura, foram formuladas três questões de pesquisa (*Research Questions*, ou RQs), detalhadas a seguir:

RQ1: Quais são os métodos mais eficazes de predição da concentração de ozônio troposférico, considerando a aplicação de aprendizado de máquina?

RQ2: Quais as técnicas de imputação utilizadas para tratar dados faltantes?

RQ3: Quais métricas são mais frequentemente utilizadas para avaliar o desempenho de modelos preditivos de concentração de ozônio?

Entre os trabalhos relacionados, Yafouz *et al.* (2021) e Gryech *et al.* (2024) revisam o uso de métodos de aprendizado de máquina na predição de ozônio, mas não abordam técnicas de imputação de dados. Assim, esta revisão apresenta as abordagens para lidar com dados faltantes no contexto de predição de ozônio.

A estratégia de busca utilizou a combinação dos termos "*ozone*", "*predict*" (incluindo o sinônimo "*forecast*") e "*machine learning*". Todas as buscas empregaram o operador booleano "AND", resultando na string: "*ozone AND (predict OR forecast) AND machine learning*", aplicada como busca em texto completo quando suportado pelas bases de dados. As buscas foram realizadas em maio de 2025, sem restrições temporais, limitando-se ao histórico de cada base, permitindo uma cobertura mais ampla.

As bases IEEE Xplore Digital Library e ACM Digital Library foram selecionadas por concentrarem o maior número de estudos relevantes no levantamento de Gryech *et al.* (2024), e por estarem disponíveis para acesso institucional. A base MDPI foi incluída por ser uma plataforma de acesso aberto com ampla cobertura interdisciplinar, além de também constar entre as bases analisadas por Gryech *et al.* (2024). A busca inicial resultou em 669 estudos, com a maior contribuição vinda da ACM Digital Library (451), conforme detalhado na Tabela 3.1.

Tabela 3.1 – Quantidade de estudos coletados por base de dados.

Base de dados	Estudos coletados
ACM Digital Library	451
IEEE Xplore Digital Library	119
MDPI	99
Total	669

Fonte: Elaborada pela autora (2026).

Para garantir a qualidade e relevância dos estudos selecionados, foram estabelecidos critérios de inclusão e exclusão (Tabela 3.2). Os critérios de inclusão focaram em trabalhos que apresentassem contribuições práticas para a predição de ozônio troposférico utilizando técnicas de aprendizado de máquina, com atenção especial àqueles que descrevessem os métodos de imputação empregados no tratamento de dados faltantes durante a fase de pré-processamento dos dados.

Tabela 3.2 – Critérios de inclusão e exclusão dos estudos.

Critérios de inclusão	Critérios de exclusão
Artigos que apresentam modelos preditivos para concentração de ozônio troposférico	Documentos que não são artigos de periódicos
Estudos que empregam técnicas de imputação para tratamento de dados faltantes	Publicações anteriores a 2020
	Trabalhos puramente teóricos, sem aplicação em dados reais ou informações suficientes sobre dados e modelos
	Estudos focados exclusivamente em índices compostos de qualidade do ar (e.g., AQI) ou outros poluentes (e.g., PM _{2.5} , NO ₂)
	Estudos voltados à detecção/classificação de presença ou excedência de O ₃ (e.g., visão computacional, CNN, SVM), sem predição numérica da concentração

Fonte: Elaborada pela autora (2026).

Os critérios de exclusão foram definidos para filtrar estudos fora do escopo temporal e metodológico desta revisão. A decisão de excluir publicações anteriores a 2020 baseia-se na rápida evolução das técnicas de aprendizado de máquina observada nos últimos anos. Estudos recentes, como o de Gryech *et al.* (2024), demonstram no período de 2015 a 2022 houve uma adoção generalizada de modelos mais sofisticados, incluindo *deep learning* (LSTM), métodos de *ensemble* (XGBoost, Random Forest) e algoritmos híbridos, em substituição aos métodos estatísticos tradicionais, que apresentaram desempenho inferior aos modelos de aprendizado profundo na predição de poluentes atmosféricos, conforme demonstrado por Naz *et al.* (2023).

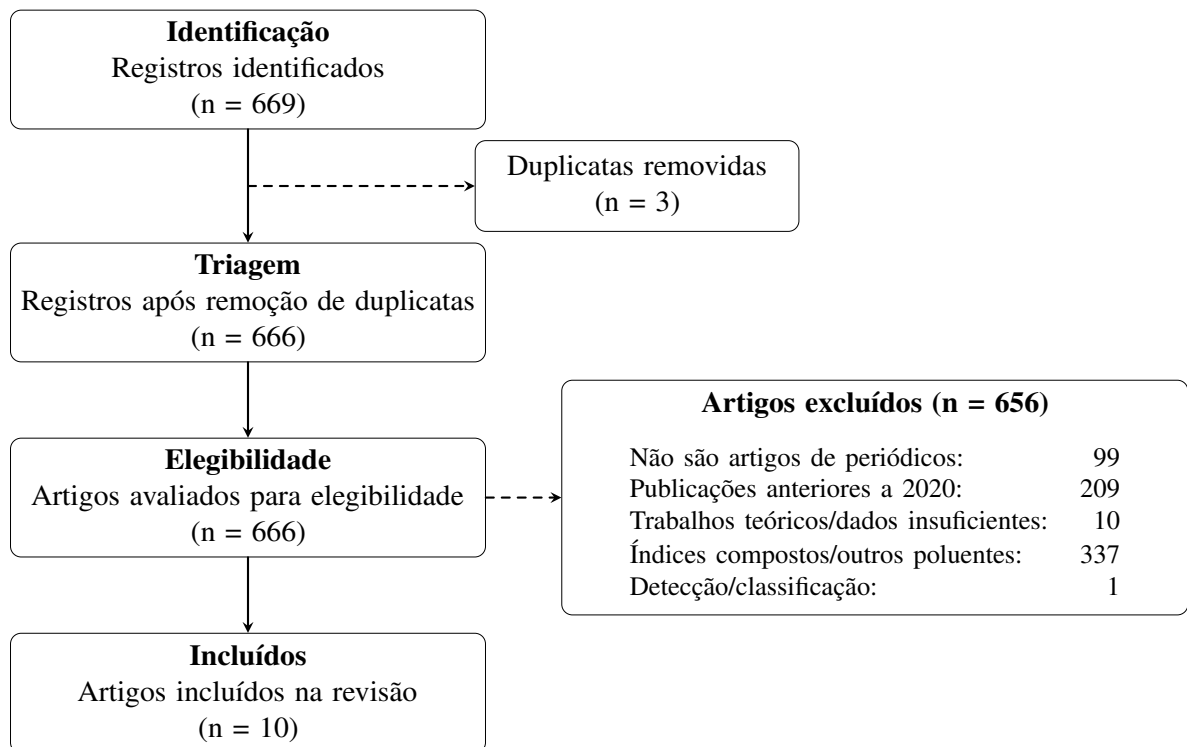
Da mesma forma, excluíram-se trabalhos cujo objetivo é meramente detectar ou classificar a presença de ozônio, seja com uso de visão computacional, CNNs ou classificadores como SVM e KNN, porque tais abordagens lidam com uma tarefa distinta. A exclusão de estudos

sobre outros poluentes justifica-se pela natureza única do ozônio troposférico como poluente secundário, formado por reações fotoquímicas específicas entre seus precursores sob influência meteorológica (Capítulo 2).

Além disso, estudos sobre o Air Quality Index (AQI), ou Índice de Qualidade do Ar, foram excluídos por tratarem de um índice composto que combina diversos poluentes em uma única métrica, o que mascara os processos particulares do ozônio.

A aplicação destes critérios resultou no processo de seleção ilustrado na Figura 3.1. Dos 669 registros inicialmente identificados nas bases de dados, 3 foram removidos por serem duplicatas, resultando em 666 artigos únicos para triagem. Durante a fase de elegibilidade, 656 artigos foram excluídos após análise do título e do resumo, aplicando-se os critérios de exclusão na ordem apresentada no fluxograma PRISMA (Page *et al.*, 2021): 99 por não serem artigos de periódicos revisados por pares; 209 por serem publicações anteriores a 2020; 10 por serem trabalhos puramente teóricos ou não apresentarem informações suficientes sobre os dados e modelos utilizados; 337 por focarem exclusivamente em índices compostos de qualidade do ar ou outros poluentes que não o ozônio troposférico; e 1 por se concentrar na detecção binária da presença de ozônio. Este processo de seleção resultou em 10 artigos que atenderam aos critérios de inclusão.

Figura 3.1 – Fluxograma PRISMA (Page *et al.*, 2021) do processo de seleção de artigos.



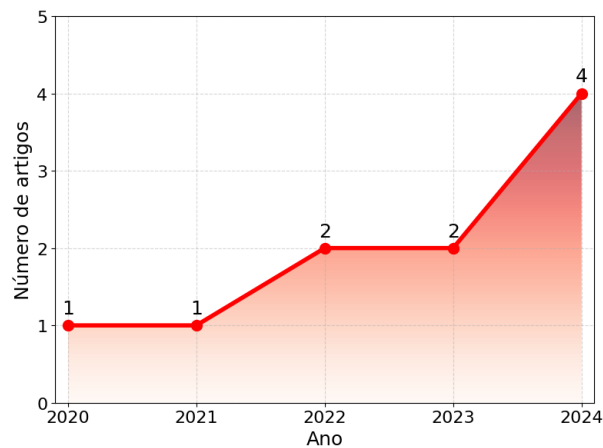
Fonte: Elaborada pela autora (2026).

Após a seleção final, os artigos foram submetidos a uma avaliação de qualidade considerando: (i) clareza na descrição dos métodos de aprendizado de máquina aplicados ao ozônio; (ii) descrição e justificativa do método utilizado tratar dados faltantes; (iii) adequação das métricas

utilizadas; e (iv) validação dos modelos. A avaliação de qualidade demonstrou que os dez estudos atendiam aos critérios de clareza, adequação das métricas e validação. No entanto, foi identificada uma lacuna no critério (ii), pois quatro estudos não descreveram o método utilizado para tratar dados faltantes. Dos artigos selecionados foram extraídas título, resumo, ano de publicação, algoritmos de aprendizado de máquina empregados, métodos de imputação utilizados para tratamento de dados faltantes, além das métricas de avaliação e resultados reportados.

A Figura 3.2 apresenta a distribuição temporal dos 10 artigos selecionados, com uma concentração de publicações em 2024 (4 artigos), seguida pelos anos de 2022 e 2023 com produção constante (2 artigos/ano), enquanto 2021 e 2020 registraram apenas uma publicação cada. Esta tendência ascendente indica a crescente aplicação de aprendizado de máquina para predição de ozônio troposférico.

Figura 3.2 – Distribuição anual dos estudos selecionados na revisão sistemática.

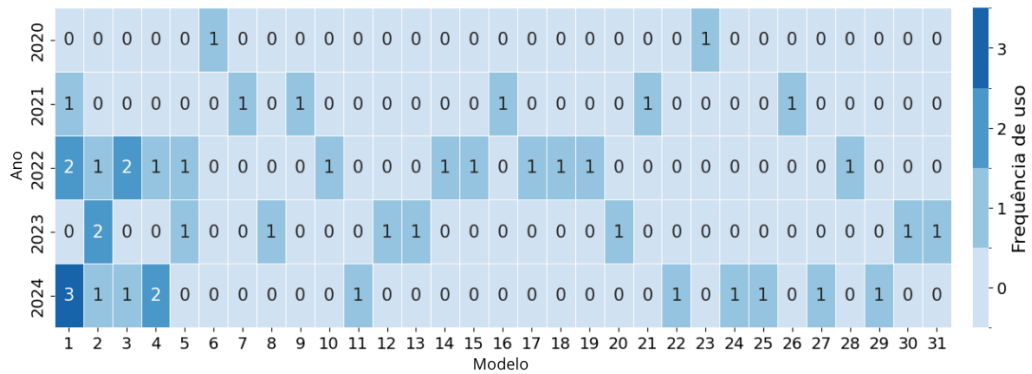


Fonte: Elaborada pela autora (2026).

A análise dos estudos selecionados revela que os métodos mais eficazes para predição de concentração de ozônio troposférico são predominantemente baseados em algoritmos de *ensemble* e arquiteturas de *deep learning*. A Figura 3.3 apresenta um mapa de calor que ilustra a frequência de uso dos diferentes algoritmos ao longo do período de 2020 a 2024. O Random Forest se destaca como o algoritmo mais utilizado. O LSTM e o XGBoost mostram padrões similares, com maior atividade entre 2022 e 2024. Outros algoritmos, como alguns modelos de redes neurais (MLP, MLR), aparecem com menor frequência distribuídos ao longo dos anos.

A Tabela 3.3 apresenta um panorama dos algoritmos e métricas de avaliação utilizados nos estudos analisados. Observa-se um padrão na escolha das métricas conforme a natureza dos algoritmos empregados. Os modelos baseados em árvore, como XGBoost, Random Forest e Decision Tree, são predominantemente avaliados através de métricas de erro (RMSE, MAE) e coeficiente de determinação (R^2). Por outro lado, os modelos de aprendizado profundo, particularmente LSTM, podem incorporar métricas adicionais como MAPE (Erro Percentual Absoluto Médio).

Figura 3.3 – Evolução do uso de modelos de aprendizado de máquina nos estudos selecionados.



1: Random Forest; 2: LSTM; 3: Decision Tree; 4: XGBoost; 5: CNN; 6: BRT; 7: Bagging; 8: CNN-LSTM; 9: CART; 10: ElasticNet; 11: DNN; 12: GRU; 13: InceptionTime; 14: AdaBoost; 15: ANN; 16: Lasso; 17: KNN; 18: LightGBM; 19: Regressão Linear; 20: MiniRocket; 21: Modelo Selecionado; 22: MLP; 23: MLR; 24: RFE-SVR-LSTM; 25: RFE-RF-LSTM; 26: SPLS; 27: RFE-XGBoost-LSTM; 28: SVM; 29: SVR; 30: Transformer-Encoder; e 31: Transformers.

Fonte: Elaborada pela autora (2026).

Tabela 3.3 – Síntese dos algoritmos de aprendizado de máquina e métricas de desempenho utilizados nos estudos selecionados.

Referência	Algoritmos	Métricas de Avaliação
Gagliardi e Andenna (2020)	BRT, MLR	R ² , MBE, MAE, RMSE, IoA
Oufdou <i>et al.</i> (2021)	SPLS, Lasso, CART, Bagging, RF, Modelo Selecionado	R ² , RMSE
Juarez e Petersen (2022)	XGBoost, Random Forest, KNN, SVM, Decision Tree, Linear Regression, AdaBoost, LSTM	Coefficiente de Correlação, R ² , R ² Ajustado, RMSE, MAE
Midyya e Roy (2022)	ANN, Decision Tree, ElasticNet, LightGBM, Random Forest, CNN	MAE, RMSE, Coeficiente de Correlação, R ²
AlShehhi e AlShehhi (2023)	LSTM, MiniRocket, InceptionTime, Transformer-decoder	RMSE, MAE, MAPE
Bekkar <i>et al.</i> (2023)	LSTM, GRU, CNN, CNN-LSTM, Transformers	MAE, RMSE, R ²
Jailani e Mara (2024)	MLP, DNN, XGBoost	R ² , MAE, MSE
Venuka <i>et al.</i> (2024)	Decision Tree, SVR, Random Forest, RF with Grid Search, XGBoost	R ² , RMSE
Chaurasia e Misra (2024)	Random Forest	R ² , RMSE
Ji <i>et al.</i> (2024)	LSTM, RFE-XGBoost-LSTM, RFE-RF-LSTM, RFE-SVR-LSTM	R ² , MAE, RMSE

Fonte: Elaborada pela autora (2026).

Com base na análise comparativa dos desempenhos apresentada na Tabela 3.4, o XGBoost aparece como o algoritmo de melhor desempenho, apresentando resultados superiores na maioria dos estudos, com destaque para Venuka *et al.* (2024) que alcançou R² de 0.957, e Jailani e Mara (2024) com R² de 0.93. As arquiteturas Transformer também demonstram excelente capacidade preditiva, conforme evidenciado por AlShehhi e AlShehhi (2023) e Bekkar *et al.* (2023), que

Tabela 3.4 – Resumo dos estudos organizados por desempenho (R^2).

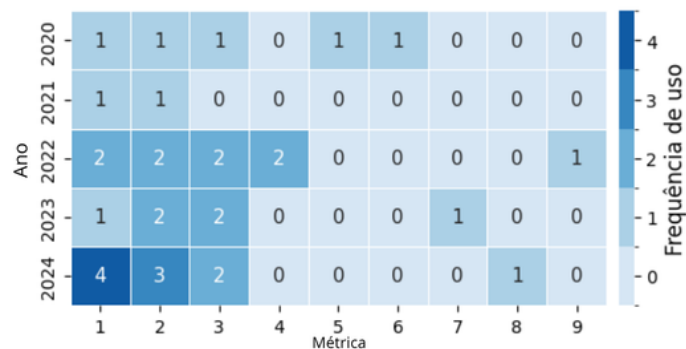
Referência	Melhor Algoritmo	Métricas e Resultados	R^2
Venuka <i>et al.</i> (2024)	XGBoost	$R^2=0.957$; RMSE=21.63	0.957
Jailani e Mara (2024)	XGBoost	$R^2=0.93$; MAE=0.93; MSE=2.86	0.930
Ji <i>et al.</i> (2024)	RFE-SVR-LSTM	$R^2=0.922$; MAE=11.552; RMSE=15.973	0.922
Bekkar <i>et al.</i> (2023)	Transformer	$R^2=0.92$; MAE=2.90; RMSE=3.99	0.920
Chaurasia e Misra (2024)	Random Forest	$R^2=0.87$; RMSE=7.79	0.870
Oufdou <i>et al.</i> (2021)	Modelo Selecionado	$R^2=0.86$; RMSE=9.60	0.860
Gagliardi e Andenna (2020)	BRT	$R^2=0.81$; MBE=3.58; MAE=9.84; RMSE=12.29; IoA=0.79	0.810
Middya e Roy (2022)	LightGBM	$R^2=0.79$; MAE=7.39; RMSE=10.17; Correlação=0.89	0.790
Juarez e Petersen (2022)	XGBoost	$R^2=0.6161$; RMSE=20.78; MAE=13.67; Correlação=0.784	0.616
AlShehhi e AlShehhi (2023)	Transformer	RMSE=0.25; MAE=0.22; MAPE=1.88	-

Fonte: Elaborada pela autora (2026).

reportaram valores de R^2 de 0.92 e métricas de erro baixas. Outros algoritmos relevantes incluem LightGBM, BRT (*Boosted Regression Trees*) e Random Forest, todos apresentando coeficientes de determinação superiores a 0.79, indicando a eficácia dos métodos baseados em árvores de decisão para este tipo de problema preditivo.

As métricas de avaliação predominantes nos estudos analisados incluem o coeficiente de determinação (R^2) e o RMSE, ambos presentes em 90% dos trabalhos, estabelecendo-se como indicadores primários de desempenho dos modelos. O MAE é utilizado em 70% dos casos. Outras métricas como erro de tendência médio (MBE), erro percentual absoluto médio (MAPE), coeficiente de correlação, erro quadrático médio (MSE) e índice de concordância (IoA) são aplicadas de forma mais seletiva, atendendo a objetivos específicos de cada pesquisa. A distribuição temporal dessas métricas ao longo dos anos pode ser observada na Figura 3.4.

Figura 3.4 – Evolução do uso das métricas de avaliação nos estudos selecionados.



1: R^2 ; 2: RMSE; 3: MAE; 4: Coeficiente de correlação; 5: IoA; 6: MBE; 7: MAPE; 8: MSE; e 9: R^2 Ajustado.

Fonte: Elaborada pela autora (2026).

Quanto às técnicas de imputação para tratamento de dados faltantes, conforme apresentado na Tabela 3.5, observa-se uma diversidade entre os estudos. A interpolação linear constitui a abordagem mais frequentemente empregada, utilizada por Ji *et al.* (2024) e Juarez e Petersen (2022), seguida por técnicas mais sofisticadas como MICE implementado por Bekkar *et al.* (2023) e *k-nearest neighbor* por Oufdou *et al.* (2021).

Destaca-se a metodologia de Middya e Roy (2022), que propõe seleção de dados de outros locais baseada em critérios de proximidade geográfica e agrupamentos, e a abordagem de AlShehhi e AlShehhi (2023) com médias móveis ponderadas para séries temporais univariadas. É relevante notar que 40% dos estudos não especificaram as técnicas de imputação. Essa omissão compromete a reprodutibilidade dos experimentos, pois a escolha do método de imputação pode influenciar diretamente o desempenho do modelo final.

Tabela 3.5 – Métodos de imputação utilizados nos estudos selecionados.

Referência	Método de Imputação	Tipo
Oufdou <i>et al.</i> (2021)	<i>k-nearest neighbor</i>	Baseado em vizinhança
Juarez e Petersen (2022)	Interpolação linear	Temporal
Middya e Roy (2022)	Propõe uma metodologia de seleção de dados de outros locais (por exemplo, baseadas em raio, divergência KL e baseadas em <i>clusters</i>)	Espacial
AlShehhi e AlShehhi (2023)	Médias móveis ponderadas para séries temporais univariadas	Temporal
Bekkar <i>et al.</i> (2023)	MICE	Multivariado
Ji <i>et al.</i> (2024)	Interpolação linear	Temporal

Fonte: Elaborada pela autora (2026).

Para avaliar o desempenho dos modelos de predição de ozônio troposférico, implementou-se uma classificação baseada no coeficiente de determinação utilizando o algoritmo de agrupamento *KMeans*. O algoritmo foi configurado para gerar três clusters distintos ($n_clusters=3$), agrupando automaticamente os modelos com base na similaridade de seus valores de R^2 . Os grupos resultantes foram rotulados como "Excelente", "Bom" e "Moderado", conforme apresentado na Tabela 3.6.

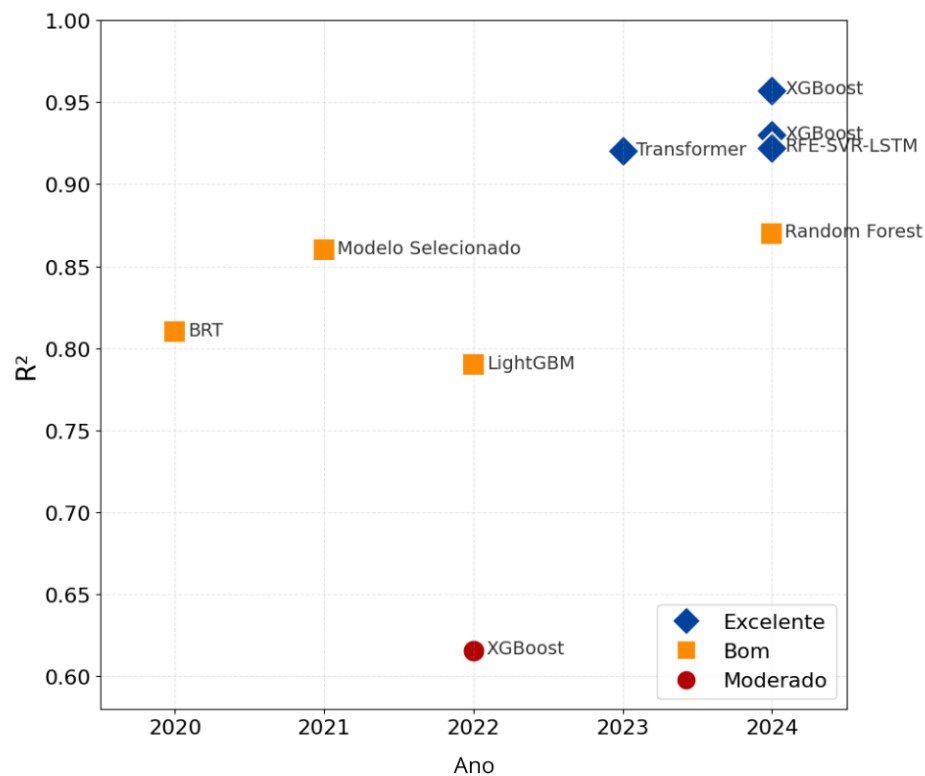
A Figura 3.5 apresenta a evolução do desempenho dos modelos, medido pelo R^2 , entre 2020 e 2024. A análise mostra uma transição de paradigmas de modelagem, que se reflete diretamente na melhoria dos resultados. Enquanto os estudos mais antigos se baseavam em modelos de *ensemble* tradicionais com desempenho moderado ou bom, as abordagens mais recentes (a partir de 2023) adotaram arquiteturas mais sofisticadas e de aprendizado profundo, como Transformer e XGBoost, sendo agrupadas na categoria de desempenho "Excelente", com valores de R^2 que formam o *cluster* superior identificado na análise.

Com base nesta análise sistemática da literatura, este trabalho adotará uma abordagem combinada para tratamento de dados faltantes. O método MICE será a estratégia principal para tratamento de dados faltantes, escolhido por preservar as relações multivariadas e sua eficácia

Tabela 3.6 – Desempenho (R^2) de diferentes estudos com classificação KMeans.

Referência	R^2	Desempenho KMeans
Venuka <i>et al.</i> (2024)	0.957	Excelente
Jailani e Mara (2024)	0.930	Excelente
Ji <i>et al.</i> (2024)	0.922	Excelente
Bekkar <i>et al.</i> (2023)	0.920	Excelente
Chaurasia e Misra (2024)	0.870	Bom
Oufdou <i>et al.</i> (2021)	0.860	Bom
Gagliardi e Andenna (2020)	0.810	Bom
Middya e Roy (2022)	0.790	Bom
Juarez e Petersen (2022)	0.616	Moderado

Fonte: Elaborada pela autora (2026).

Figura 3.5 – Evolução do desempenho (R^2) dos modelos por ano de publicação com agrupamento por desempenho.

Fonte: Elaborada pela autora (2026).

em estudos meteorológicos. Seguindo Middya e Roy (2022), também serão testados dados de estações próximas para imputar variáveis meteorológicas, mantendo o MICE para as demais variáveis.

Para modelagem preditiva, será utilizado como algoritmo principal o XGBoost, que demonstrou desempenho superior nos estudos analisados, com destaque para Venuka *et al.* (2024) ($R^2 = 0.957$) e Jailani e Mara (2024) ($R^2 = 0.930$). Como estratégia de comparação, também serão testados o Random Forest e o k-Nearest Neighbors. O Random Forest, algoritmo baseado em árvores de decisão, é o mais utilizado nos estudos mapeados e permite avaliar se

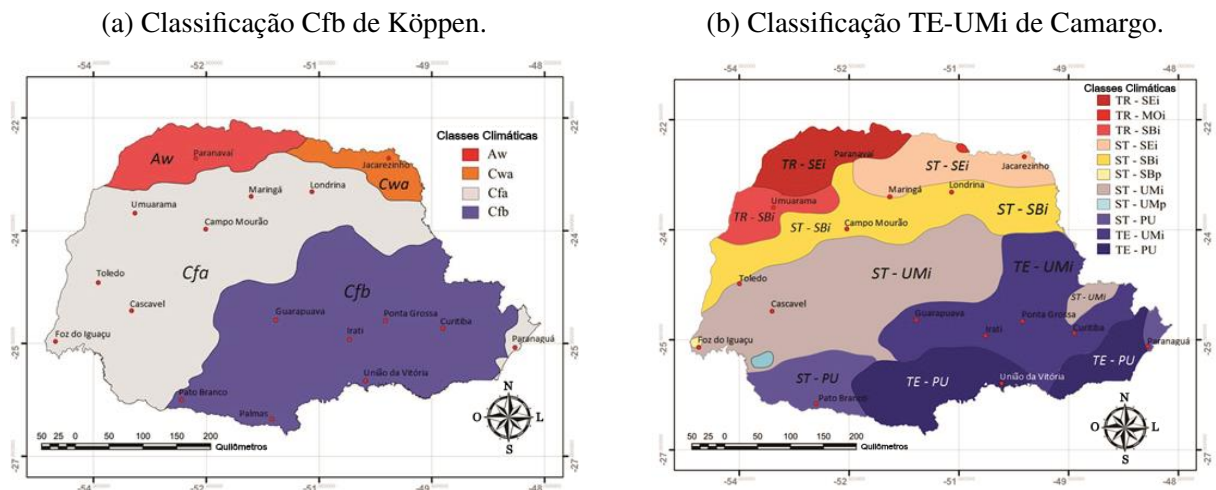
a abordagem de (*bagging*) oferece vantagens em relação ao *boosting* sequencial do XGBoost. Já o kNN, por ser um modelo baseado em instâncias e não em árvores, permite avaliar se relações locais de similaridade são mais adequadas que estruturas hierárquicas para capturar os processos de formação de ozônio troposférico. A avaliação dos modelos utilizará R^2 , RMSE e MAE, métricas padrão que avaliam variância explicada e erros preditivos. Adicionalmente, os resultados serão comparados para identificar os métodos e modelos mais eficazes na previsão de ozônio.

4 ANÁLISE EXPLORATÓRIA E MÉTODOS DE IMPUTAÇÃO

Este estudo foca em dados reais coletados em uma estação de qualidade do ar localizada na cidade de Ponta Grossa, no estado do Paraná, Brasil. A cidade de Ponta Grossa é localizada no Segundo Planalto Paranaense, a aproximadamente 115 km a noroeste de Curitiba, situada entre as coordenadas 25°05'58" de latitude sul e 50°09'30" de longitude oeste, com altitudes que variam entre 780 e 985 metros (Sobrinho, 2018). É importante destacar que estações de qualidade do ar monitoram a concentração de poluentes atmosféricos (incluindo O_3), enquanto estações meteorológicas medem variáveis climáticas como temperatura, umidade, vento e precipitação.

A cidade de Ponta Grossa possui clima Cfb (Figura 4.1a), classificado como temperado úmido, sem estação seca e com verões amenos (Aparecido *et al.*, 2016). Pelo sistema de Camargo (Figura 4.1b), enquadra-se em TE-UMi, definido como temperado úmido com leve deficiência hídrica no inverno (Aparecido *et al.*, 2016). Ambos os métodos possuem em comum o caráter frio e úmido dos planaltos onde Ponta Grossa se encontra.

Figura 4.1 – Classificações climáticas de Ponta Grossa segundo Köppen (a) e Camargo (b).



Aw: Clima tropical com estação seca no inverno; **Cwa:** Clima subtropical úmido com verões quentes e inverno seco; **Cfa:** Clima subtropical úmido com verões quentes; e **Cfb:** Clima temperado oceânico com verões suaves e precipitação bem distribuída.

Fonte: Adaptado de Aparecido *et al.* (2016).

TR - SEi: Tropical, Semiúmido; **TR - MOi:** Tropical, Moderadamente úmido; **TR - SBi:** Subtropical, Umidade baixa; **ST - SEi:** Subtropical, Semiúmido; **ST - SBp:** Subtropical, Umidade mediana; **ST - UMi:** Subtropical, Umidade moderada; **ST - UMp:** Subtropical, Umidade moderada a alta; **ST - PU:** Subtropical, Umidade alta; **TE - UMi:** Temperado, Umidade moderada; e **TE - PU:** Temperado, Umidade alta.

Fonte: Adaptado de Aparecido *et al.* (2016).

Essas classificações climáticas explicam o regime térmico e hídrico que modula a formação, dispersão e deposição de poluentes na região e fornecem maior coerência à interpretação das variações sazonais observadas nos dados da estação de monitoramento da qualidade do ar.

Para fins de análise sazonal ao longo do trabalho, adotou-se a convenção apresentada na Tabela 4.1, que define o início e o fim de cada estação do ano com base em terços proporcionais dos meses de transição.

Tabela 4.1 – Convenção para o período sazonal.

Estação	Período
Verão	$\frac{1}{3}$ DEZ + JAN + FEV + $\frac{2}{3}$ MAR
Outono	$\frac{1}{3}$ MAR + ABR + MAI + $\frac{1}{3}$ JUN
Inverno	$\frac{2}{3}$ JUN + JUL + AGO + $\frac{2}{3}$ SET
Primavera	$\frac{1}{3}$ SET + OUT + NOV + $\frac{2}{3}$ DEZ

Fonte: Adaptado de Aparecido *et al.* (2016).

4.1 DADOS DA ESTAÇÃO DE QUALIDADE DO AR

A Tabela 4.2 apresenta as estatísticas descritivas das variáveis ambientais coletadas de hora em hora ao longo de um período de quatro anos (25 de outubro de 2016 a 31 de dezembro de 2020), totalizando 36.686 observações. Variáveis meteorológicas como temperatura e umidade relativa apresentam distribuição relativamente centrada, com medianas próximas à média. Já variáveis como a radiação solar global e velocidade do vento exibem assimetrias mais acentuadas, indicadas por quartis distantes da média.

As variáveis relacionadas a poluentes atmosféricos, como O₃, NO_x, MP10 e SO₂, também mostram padrões distintos. O O₃, por exemplo, apresenta distribuição relativamente equilibrada, com média e mediana próximas (14,24 e 12,90 ppb, respectivamente), enquanto o CO apresenta um valor mínimo negativo (-289,00 ppm), indicando possível erro de medição. Esse tipo de anomalia, conhecida como *outlier*, reforça a importância de estratégias robustas de imputação, especialmente em estudos preditivos com séries temporais.

Tabela 4.2 – Estatísticas descritivas das variáveis medidas.

Variável	Qtd. de dados	Média	Desvio padrão	Mínimo	1º Quartil (25%)	Mediana (50%)	3º Quartil (75%)	Máximo
Chuva (mm)	20229	0.01	0.09	0.00	0.00	0.00	0.00	3.82
Direção do vento (°)	20678	162.74	122.41	0.00	35.42	143.18	296.85	360.00
Pressão atmosférica (hPa)	19900	911.01	3.46	900.08	908.68	910.83	913.26	922.32
RSG (W/m²)	20129	190.14	281.54	0.00	1.92	9.25	323.42	1205.20
Temperatura (°C)	19981	17.65	4.83	-1.38	14.44	17.48	20.79	33.96
Temperatura interna (°C)	7858	24.14	2.95	17.81	22.42	23.73	24.78	36.50
Umidade relativa (%)	19465	78.58	19.32	23.63	64.97	84.04	95.66	100.00
Velocidade do vento (m/s)	19925	1.88	0.87	0.00	1.20	1.77	2.45	7.05
CH ₄ (ppm)	9257	1.74	0.62	0.00	1.57	1.89	2.09	4.28
CO (ppm)	19339	0.55	2.15	-289.00	0.34	0.46	0.63	9.66
HCNM (ppm)	6413	0.67	0.69	0.00	0.02	0.33	1.37	2.00
HCT (ppm)	9255	2.42	1.09	0.00	1.85	2.20	3.26	5.72
MP10 (µg/m³)	16254	19.82	15.40	0.00	10.00	16.00	25.00	149.00
NO (ppb)	19788	4.81	7.39	0.00	0.13	1.75	7.45	128.02
NO ₂ (ppb)	19527	5.69	6.15	0.00	0.05	4.59	8.94	51.33
NO _x (ppb)	19814	10.36	12.71	0.00	0.13	6.75	16.53	145.43
O ₃ (ppb)	19775	14.24	8.05	0.00	8.42	12.90	18.96	60.06
PTS (µg/m³)	15676	35.09	26.89	1.00	17.00	28.00	45.00	239.00
SO ₂ (ppb)	5274	0.44	0.67	0.00	0.00	0.00	0.99	2.55

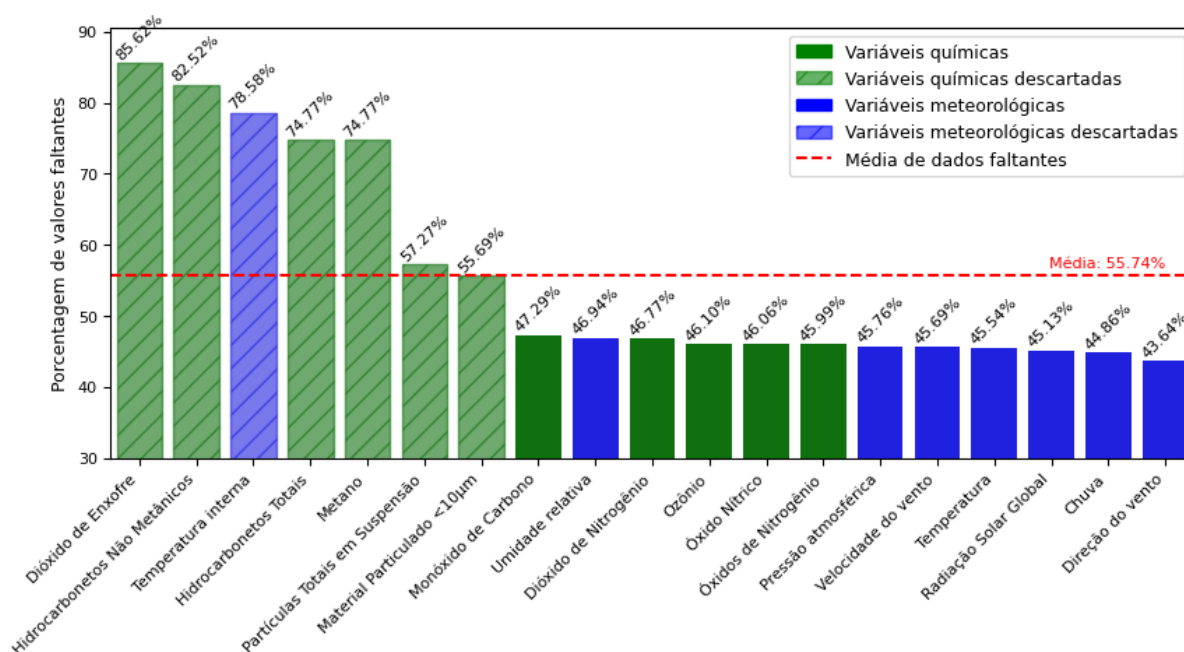
RSG: Radiação Solar Global; **CH₄:** Metano; **CO:** Monóxido de Carbono; **HCNM:** Hidrocarbonetos Não Metânicos; **HCT:** Hidrocarbonetos Totais; **MP10:** Material Particulado (diâmetro ≤ 10 µm); **NO:** Óxido Nítrico; **NO₂:** Dióxido de Nitrogênio; **NO_x:** Óxidos de Nitrogênio; **O₃:** Ozônio; **PTS:** Partículas Totais em Suspensão; e **SO₂:** Dióxido de Enxofre.

Fonte: Elaborada pela autora (2026).

A Figura 4.2 mostra que as variáveis SO_2 (85.62%), HCNM (82.52%), temperatura interna (78.58%), HCT (74.77%) e CH_4 (74.77%) apresentam um número muito alto de ausência de dados. Em virtude dessa limitação, optou-se por excluir essas variáveis dos modelos preditivos.

Adicionalmente, as variáveis PTS e MP10, embora apresentem percentuais moderados de dados faltantes (57.27% e 55.69%, respectivamente), também foram excluídas da etapa de imputação e modelagem. Embora possam influenciar o ozônio de forma indireta, por exemplo ao modificar a disponibilidade de radiação solar por espalhamento e atenuação, ou ao oferecer superfícies para reações heterogêneas que alteram a química de radicais e de NO_x , esses poluentes não atuam como precursores diretos do ozônio troposférico no sentido fotoquímico discutido no Capítulo 2.

Figura 4.2 – Proporção de dados faltantes por variável (%).



Fonte: Elaborada pela autora (2026).

Dentre as demais variáveis, observa-se que a maioria apresenta distribuições aproximadamente simétricas e concentradas, como pressão, RSG, umidade relativa, velocidade do vento, temperatura e ozônio (Apêndice B, Figura B.1). Além disso, os gráficos temporais (Apêndice B, Figura B.2) mostram a presença de lacunas em diversos períodos, a mais extensa estende-se de março de 2017 até junho de 2018. Apesar disso, muitas variáveis mantêm um comportamento relativamente estável ao longo do tempo, como temperatura, ozônio, umidade relativa e a radiação solar global.

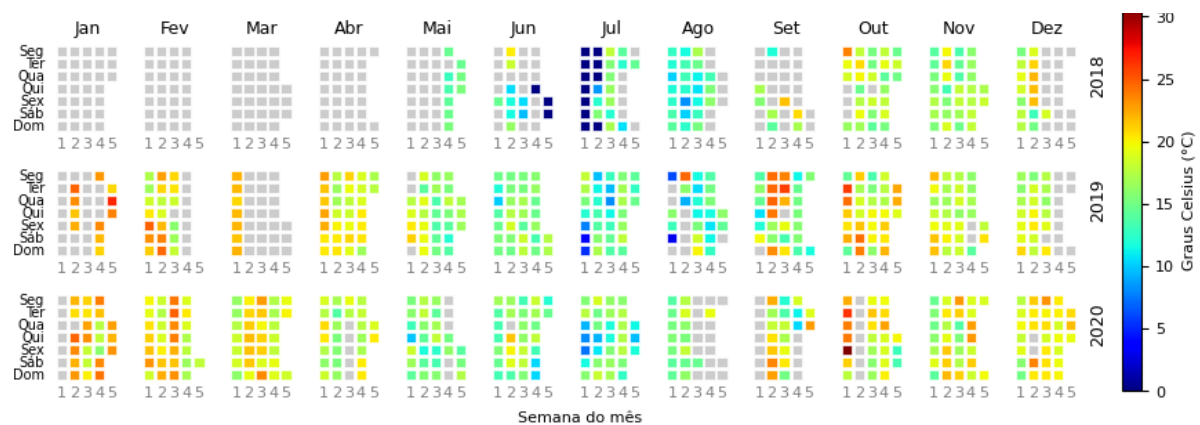
A Figura 4.3 mostra a variabilidade da temperatura no período 2018 a 2020. Entre dezembro e março ocorrem os valores mais altos, quase sempre acima de $20\text{ }^{\circ}\text{C}$. Durante a temporada de verão de 2019/2020, entre os meses de dezembro a março, quase todas as semanas ficaram na faixa de $20\text{--}30\text{ }^{\circ}\text{C}$. Em 2020 há um pico entre os meses de janeiro e fevereiro, que

registram médias próximas de 25 °C.

Julho de 2018 apresenta as temperaturas mínimas mais baixas de toda a série, com várias células entre 0 °C e 5 °C. Os invernos de 2019 e 2020 também trazem semanas frias, porém com valores ligeiramente mais altos, na faixa de 5–10 °C. Em 2019 a transição para a primavera ocorre em setembro. A temperatura nesse período sobe para cerca de 22 °C e em outubro volta a manter estável na faixa de 20-25 °C.

A temperatura média de todo o período é de aproximadamente 17 °C e os meses mais frios apresentam médias mensais abaixo de 13 °C. Estes valores são consistentes com a classificação climática de Camargo, que estabelece a classe TE para este intervalo de temperatura, conforme apresentado por Aparecido *et al.* (2016).

Figura 4.3 – Temperatura média diária (2018 – 2020). Os quadrados em cinza correspondem a dados faltantes ou não disponíveis.



Fonte: Elaborada pela autora (2026).

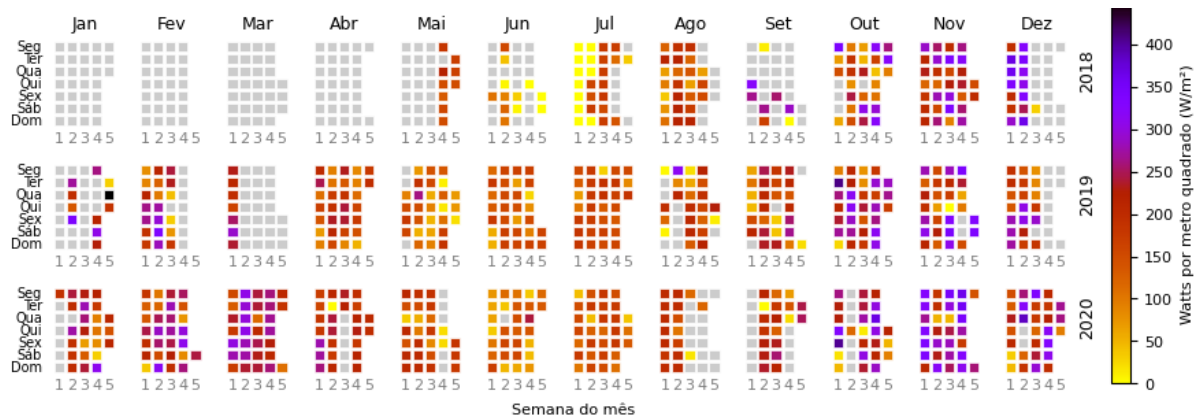
Entre setembro e março a radiação solar global atinge os maiores valores, com várias semanas acima de 400 W/m² (Figura 4.4). O pico mais claro acontece no verão de 2019/2020: de dezembro a fevereiro quase não há valores baixos, indicando dias longos e possivelmente céu pouco nublado. Nos meses de inverno, principalmente em junho e julho, a RSG cai de forma acentuada.

Julho de 2018 é o ponto mais baixo da série, com muitas células abaixo de 100 W/m²; este é o mesmo período que registrou as temperaturas mínimas entre 0 °C e 5 °C. Os invernos de 2019 e 2020 mostram radiação mais elevada em relação ao ano anterior, o que combina com as temperaturas um pouco mais altas observadas nesses anos.

As Figuras 4.3 e 4.4 deixam evidente a relação direta entre radiação e aquecimento. A RSG também ajuda a explicar por que os verões foram sucessivamente quentes e por que o inverno de 2018 se destacou como o mais rigoroso do período.

A Figura 4.5 mostra que a umidade relativa permanece elevada ao longo de todo o período 2018–2020, com a maioria das semanas acima de 70%, e são coerentes com a classificação climática de Köppen, que segundo Aparecido *et al.* (2016), indica ausência de estação seca para esta região.

Figura 4.4 – Radiação solar global média diária (2018 – 2020). Os quadrados em cinza correspondem a dados faltantes ou não disponíveis.

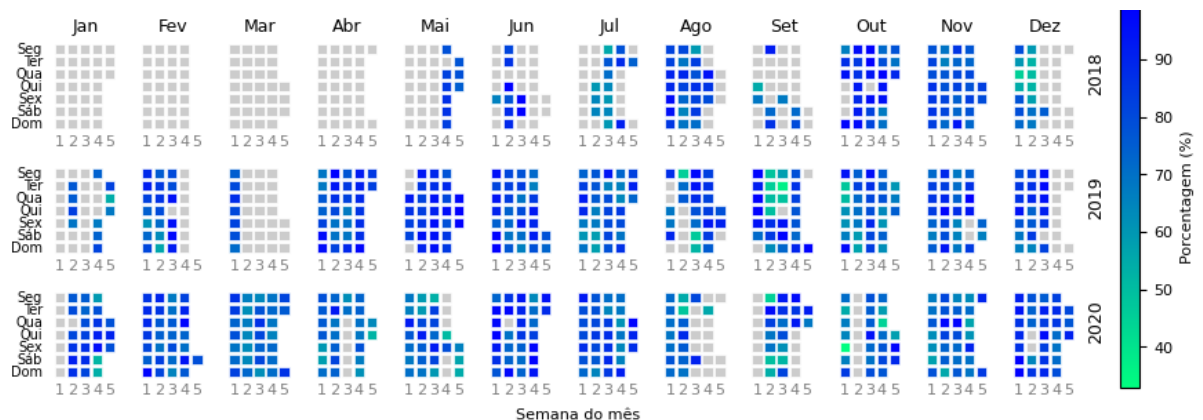


Fonte: Elaborada pela autora (2026).

Os valores máximos concentram-se no inverno, justamente quando a radiação solar global é menor e a temperatura atinge seus níveis mais baixos. Nas fases de radiação intensa, principalmente entre setembro e março, aparecem pequenos bolsões esverdeados (40–50%), indicando queda temporária da umidade durante dias mais quentes e ensolarados.

Assim, a relação da umidade relativa com a alta radiação e temperatura é indireta, o que favorece reduções pontuais da umidade, enquanto radiação baixa e frio mantêm o ar constantemente úmido.

Figura 4.5 – Umidade relativa média diária (2018 – 2020). Os quadrados em cinza correspondem a dados faltantes ou não disponíveis.



Fonte: Elaborada pela autora (2026).

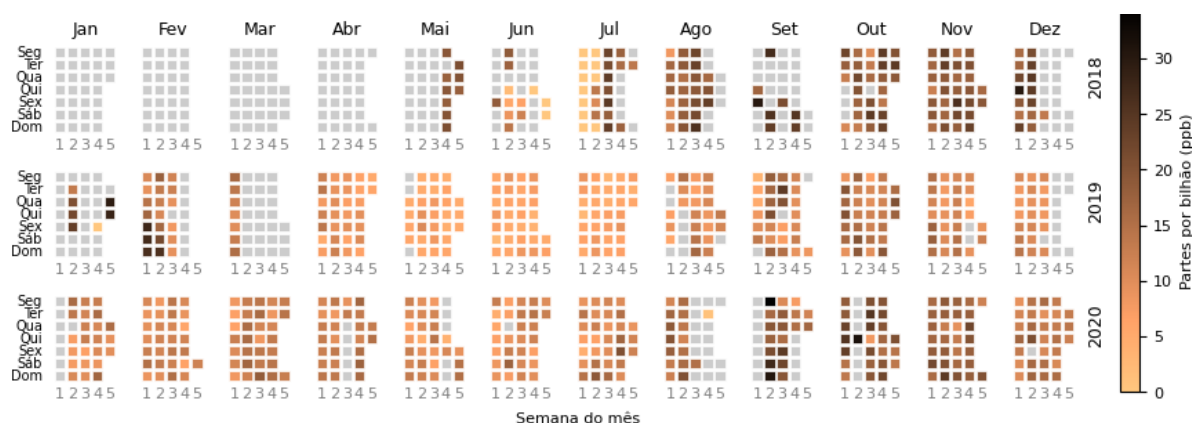
A concentração de ozônio também apresenta uma estrutura sazonal (Figura 4.6), com os níveis mais altos ocorrendo com frequência entre outubro e fevereiro, acompanhando os períodos mais quentes e secos do ano. Há destaque para os verões de 2018 e 2020, quando várias semanas ultrapassam 20 ppb, sugerindo episódios de fotodegradação mais intensa.

Já nos meses frios, principalmente em junho de 2019 a julho de 2019, os tons mais suaves indicam níveis reduzidos de ozônio, compatíveis com a menor temperatura nesse período. No

inverno, as concentrações também tendem a elevar-se pela menor dispersão dos poluentes pela turbulência atmosférica reduzida no frio (Santana; Moreira; Armani, 2020; Torres; Martins, 2005), acumulando ozônio em condições estáveis.

Nos meses em que a umidade relativa se mantém acima de 80%, principalmente no auge do inverno, as concentrações raramente passam de 10 ppb. Essa associação pode refletir um conjunto de fatores meteorológicos correlacionados, como menor disponibilidade de radiação solar, maior nebulosidade e condições de estabilidade da camada limite, que em conjunto reduzem a produção fotoquímica de ozônio. Vale destacar, ainda, que muitos sensores de O₃ apresentam sensibilidade à umidade relativa elevada, o que pode contribuir para leituras sistematicamente mais baixas em condições de inverno, sem que isso represente necessariamente uma redução real na concentração atmosférica do poluente.

Figura 4.6 – Concentração média diária de ozônio (2018 – 2020). Os quadrados em cinza correspondem a dados faltantes ou não disponíveis.



Fonte: Elaborada pela autora (2026).

Os diagramas de dispersão (Apêndice B, Figura B.3) apresenta as correlações entre o ozônio e as variáveis meteorológicas, com destaque para RSG, temperatura e umidade, e as variáveis precursoras. Verifica-se uma relação positiva entre ozônio e radiação solar global, confirmando que a radiação solar intensifica as reações fotoquímicas de formação de ozônio. A temperatura também mostra correlação positiva com ozônio, confirmando sua influência nos processos fotoquímicos.

Já a umidade apresenta uma relação mais complexa, com correlações inversas em alguns intervalos. Vale destacar que, embora altas taxas de umidade relativa possam sugerir condições mais favoráveis à deposição de poluentes, maiores níveis de umidade relativa podem estar associados a menor incidência de radiação solar, o que pode reduzir a formação fotoquímica de ozônio e explicar parte da correlação negativa observada.

Com base nas relações observadas entre o ozônio e as demais variáveis, a próxima etapa consiste na realização dos experimentos de imputação de dados e modelagem preditiva. Primeiramente, aplica-se um tratamento de valores atípicos por meio de limites físicos, removendo observações teoricamente impossíveis para cada variável, como concentrações nulas ou

negativas de poluentes que, em condições reais, tendem a apresentar níveis mínimos detectáveis na atmosfera, ainda que o limiar de detecção varie conforme a acurácia do sensor utilizado. Em seguida, os experimentos de imputação são conduzidos utilizando diferentes estratégias para preenchimento de dados faltantes, combinadas com modelos de aprendizado de máquina previamente otimizados. Os hiperparâmetros dos modelos foram determinados através de busca em grade (*grid search*), testando múltiplas combinações de valores para cada parâmetro configurável. A validação dos modelos é realizada por meio de validação cruzada adaptada para séries temporais (*TimeSeriesSplit*), que preserva a ordem cronológica dos dados e evita vazamento de informação entre conjuntos de treinamento e teste, garantindo avaliações mais realistas do desempenho preditivo.

A Tabela 4.3 apresenta as estatísticas descritivas das variáveis após a remoção de valores atípicos e antes da aplicação das estratégias de imputação. Comparando-a com a Tabela 4.2, observa-se que a limpeza eliminou valores fisicamente inconsistentes: o monóxido de carbono, que apresentava mínimo de -289,00 ppm, agora possui limite inferior de 0,10 ppm; os óxidos de nitrogênio (NO, NO₂ e NO_x) também tiveram seus valores mínimos ajustados para 0,30 ppb, refletindo concentrações residuais naturalmente presentes na atmosfera. Além disso, a remoção de valores extremos reduziu ligeiramente os desvios padrão de algumas variáveis, como temperatura e umidade relativa, tornando as distribuições mais próximas de suas características reais e reduzindo a influência de possíveis erros de medição nos modelos.

Tabela 4.3 – Estatísticas descritivas sem valores atípicos antes da imputação.

Variável	Média	Desvio Padrão	Mediana	Mínimo	Máximo
Chuva (mm)	0.01	0.07	0.00	0.00	2.70
Pressão Atmosférica (hPa)	911.26	3.39	911.12	900.08	922.32
Radiação Solar Global (W/m ²)	182.37	274.69	6.12	0.00	1189.83
Temperatura (°C)	17.32	4.92	17.12	-1.38	33.96
Umidade Relativa (%)	77.87	19.72	83.07	25.00	99.50
Velocidade do Vento (m/s)	1.86	0.87	1.75	0.00	6.88
Monóxido de Carbono (ppm)	0.59	0.59	0.45	0.10	9.66
Óxido Nítrico (ppb)	7.45	8.88	5.33	0.30	128.02
Dióxido de Nitrogênio (ppb)	9.02	6.10	7.80	0.30	51.33
Óxidos de Nitrogênio (ppb)	15.85	13.91	13.31	0.30	145.43
Ozônio (ppb)	13.15	7.47	11.82	0.10	60.06

Fonte: Elaborada pela autora (2026).

4.2 IMPUTAÇÃO COM MÉDIA

A imputação por média preenche cada valor faltante usando a média aritmética dos valores observados da própria variável (Equação 9), o que mantém a tendência central mas reduz a variabilidade e ignora a dinâmica temporal da série. Esse método é mais adequado para variáveis com distribuição aproximadamente simétrica e sem mudanças bruscas ao longo do tempo; em séries ambientais com sazonalidade marcada, ele tende a suavizar picos e vales de

forma excessiva, o que pode impactar a capacidade dos modelos de aprender eventos extremos de concentração de poluentes.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (9)$$

Onde n é o número de observações válidas e x_i representa cada valor observado da variável.

A Tabela 4.4 apresenta as estatísticas descritivas após a imputação por média. Em comparação com a Tabela 4.3, observa-se redução dos desvios padrão em todas as variáveis, o que é esperado, pois a substituição de valores faltantes pela média introduz observações exatamente na tendência central, reduzindo artificialmente a dispersão. No caso do ozônio, a média e a mediana permaneceram inalteradas, enquanto o desvio padrão diminuiu, refletindo a suavização da variabilidade decorrente do método de imputação.

Tabela 4.4 – Estatísticas descritivas depois da imputação por média.

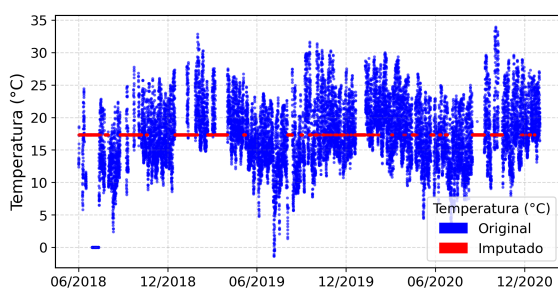
Variável	Média	Desvio Padrão	Mediana	Mínimo	Máximo
Chuva (mm)	0.01	0.06	0.00	0.00	2.70
Pressão Atmosférica (hPa)	911.26	2.89	911.26	900.08	922.32
Radiação Solar Global (W/m ²)	182.37	236.12	169.17	0.00	1189.83
Temperatura (°C)	17.32	4.20	17.32	-1.38	33.96
Umidade Relativa (%)	77.87	16.63	77.87	25.00	99.50
Velocidade do Vento (m/s)	1.86	0.74	1.86	0.00	6.88
Monóxido de Carbono (ppm)	0.59	0.50	0.59	0.10	9.66
Óxido Nítrico (ppb)	7.45	5.81	7.45	0.30	128.02
Dióxido de Nitrogênio (ppb)	9.02	3.94	9.02	0.30	51.33
Óxidos de Nitrogênio (ppb)	15.85	9.19	15.85	0.30	145.43
Ozônio (ppb)	13.15	6.34	13.15	0.10	60.06

Fonte: Elaborada pela autora (2026).

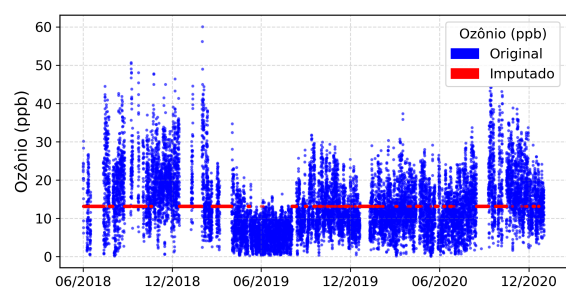
As Figuras 4.7a e 4.7b apresentam as séries temporais de temperatura e ozônio após a imputação por média. Na temperatura, os valores imputados (em vermelho) formam uma linha horizontal constante em aproximadamente 17,32 °C. O ozônio apresenta padrão similar, com valores imputados fixos em 13,15 ppb.

Figura 4.7 – Séries temporais após imputação por média para temperatura (a) e ozônio (b).

(a) Temperatura (°C).



(b) Ozônio (ppb).



Fonte: Elaborada pela autora (2026).

A Tabela 4.5 apresenta os resultados da predição da concentração de ozônio (ppb) para os três modelos avaliados, considerando as variáveis meteorológicas após imputação por média. O XGBoost obteve o melhor desempenho ($R^2 = 0.853$), seguido de pelo Random Forest ($R^2 = 0.851$), indicando que ambos os modelos baseados em árvores capturaram de forma similar as relações entre os preditores meteorológicos e o ozônio. O kNN apresentou desempenho inferior ($R^2 = 0.737$), o que é consistente com a sensibilidade da imputação pela média, que tende a reduzir a variabilidade dos dados.

Tabela 4.5 – Resultados da imputação por média ordenados por R^2 .

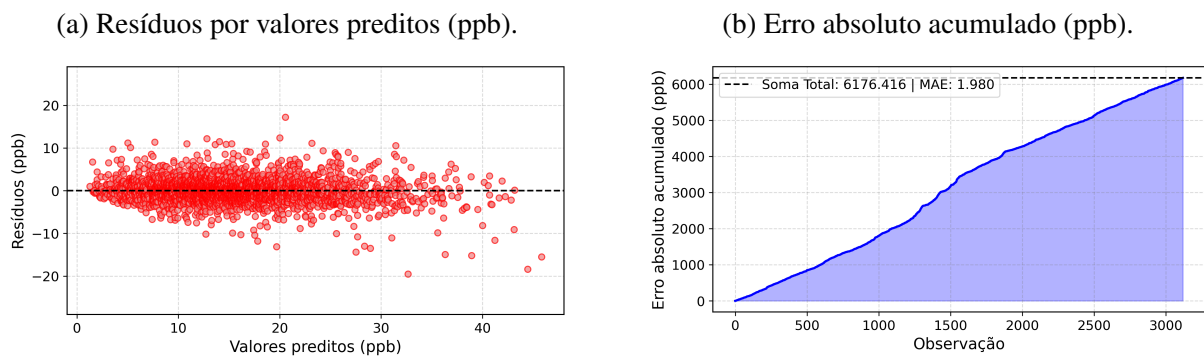
Modelo	R^2	MAE	RMSE
XGBoost	0.853	1.980	2.794
Random Forest	0.851	1.976	2.814
kNN	0.737	2.729	3.745

Fonte: Elaborada pela autora (2026).

A dispersão dos resíduos em função dos valores preditos (Figura 4.8a) apresenta observações com erros próximos a ± 20 ppb nas faixas de maior concentração de ozônio, que correspondem a resíduos atípicos ao longo da faixa de predição, o que mostra uma variação não constante do erro.

O gráfico de erro acumulado (Figura 4.8b) representa o acúmulo sequencial dos erros absolutos ao longo dos conjuntos de teste da validação cruzada em série temporal. Uma curva aproximadamente linear indica estabilidade relativa do erro médio ao longo das iterações. Se houvesse períodos específicos problemáticos, como estações do ano mal modeladas ou eventos meteorológicos atípicos, observaríamos mudanças abruptas na inclinação. Por fim, a soma total de 6176.416 ppb dividida pelo número de observações resulta no MAE reportado (1.980 ppb), uma vez que o MAE é definido como a média dos erros absolutos.

Figura 4.8 – Análise de resíduos (a) e erro acumulado (b) para imputação por média com XGBoost.



Fonte: Elaborada pela autora (2026).

4.3 IMPUTAÇÃO COM MEDIANA

A imputação por mediana substitui os valores ausentes pelo valor central da distribuição observada (Equação 10), sendo mais robusta que a média na presença de assimetrias e valores extremos. Em séries de poluentes atmosféricos, que costumam apresentar caudas longas à direita, este método preserva melhor o comportamento típico dos dados, mas ainda assim repete o mesmo valor em todas as lacunas e não incorpora informações sobre sazonalidade diária ou anual, o que limita o potencial dos modelos em capturar padrões.

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ é ímpar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{se } n \text{ é par} \end{cases} \quad (10)$$

Onde $x_{(i)}$ representa o i -ésimo valor na série ordenada.

A Tabela 4.6 apresenta as estatísticas descritivas após a imputação por mediana. Comparando com a Tabela 4.3, observa-se redução nos desvios padrão similar à imputação por média, porém com diferenças entre a média de ambas as séries.

Tabela 4.6 – Estatísticas descritivas depois da imputação por mediana.

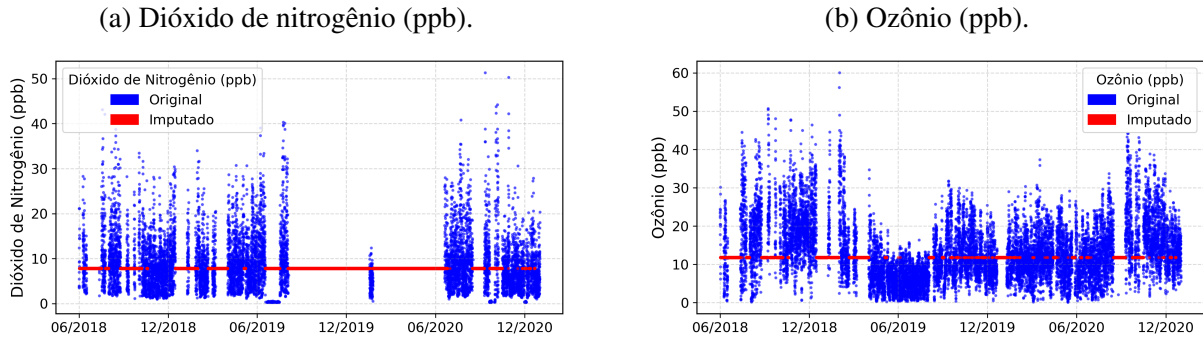
Variável	Média	Desvio Padrão	Mediana	Mínimo	Máximo
Chuva (mm)	0.01	0.06	0.00	0.00	2.70
Pressão Atmosférica (hPa)	911.22	2.89	911.12	900.08	922.32
Radiação Solar Global (W/m ²)	136.35	248.49	6.12	0.00	1189.83
Temperatura (°C)	17.26	4.20	17.12	-1.38	33.96
Umidade Relativa (%)	79.37	16.80	83.07	25.00	99.50
Velocidade do Vento (m/s)	1.83	0.74	1.75	0.00	6.88
Monóxido de Carbono (ppm)	0.55	0.51	0.45	0.10	9.66
Óxido Nítrico (ppb)	6.24	5.90	5.33	0.30	128.02
Dióxido de Nitrogênio (ppb)	8.31	3.98	7.80	0.30	51.33
Óxidos de Nitrogênio (ppb)	14.42	9.28	13.31	0.30	145.43
Ozônio (ppb)	12.78	6.37	11.82	0.10	60.06

Fonte: Elaborada pela autora (2026).

As Figuras 4.9a e 4.9b ilustram as séries de dióxido de nitrogênio e ozônio após imputação por mediana. A linha vermelha horizontal em valor fixo reproduz a mesma limitação observada na média, com ausência de variação temporal nos dados imputados e desconsideração da sazonalidade das séries ambientais.

A Tabela 4.7 apresenta os resultados da validação cruzada para os três modelos testados com dados imputados por mediana. O XGBoost alcançou o melhor desempenho com $R^2 = 0.857$. Comparado à imputação por média, o XGBoost com mediana apresentou melhoria no R^2 e redução no MAE, indicando pouco ganho ao usar a mediana em dados.

Figura 4.9 – Séries temporais após imputação por mediana para dióxido de nitrogênio (a) e ozônio (b).



Fonte: Elaborada pela autora (2026).

Tabela 4.7 – Resultados da imputação por mediana ordenados por R^2 .

Modelo	R^2	MAE	RMSE
XGBoost	0.857	1.936	2.761
Random Forest	0.850	1.978	2.823
kNN	0.736	2.742	3.750

Fonte: Elaborada pela autora (2026).

4.4 IMPUTAÇÃO COM INTERPOLAÇÃO

A interpolação linear preenche as lacunas calculando valores intermediários entre duas observações válidas consecutivas conforme a Equação 11, preservando a continuidade local da série e evitando saltos artificiais entre pontos observados. Esse método tende a funcionar bem para variáveis com variação gradual, porém pode subestimar picos e mudanças rápidas, especialmente quando a lacuna atravessa transições abruptas.

$$\hat{x}_t = x_{t-k} + \frac{k}{k+j} (x_{t+j} - x_{t-k}) \quad (11)$$

Onde x_{t-k} é a última observação válida antes da lacuna, x_{t+j} é a primeira observação válida após a lacuna, e t é o instante a ser imputado.

A implementação utilizou interpolação linear sem limite de janela, aplicada em ambas as direções (anterior e posterior), de modo que lacunas em bordas da série também fossem preenchidas a partir do vizinho válido mais próximo disponível.

A Tabela 4.8 apresenta as estatísticas descritivas após a imputação por interpolação linear. Comparando com a Tabela 4.3, observa-se que os desvios padrão mantiveram-se próximos aos valores originais.

As Figuras 4.10a e 4.10b mostram as séries temporais após a imputação por interpolação linear. Em ambas as variáveis, os trechos imputados (em vermelho) aparecem como segmentos lineares conectando pontos válidos, sem reproduzir a variabilidade observada na série original. Na temperatura, isso tende a manter coerência visual em lacunas curtas; já no ozônio, os segmentos imputados frequentemente cruzam regiões onde a série original apresenta alta dispersão e picos,

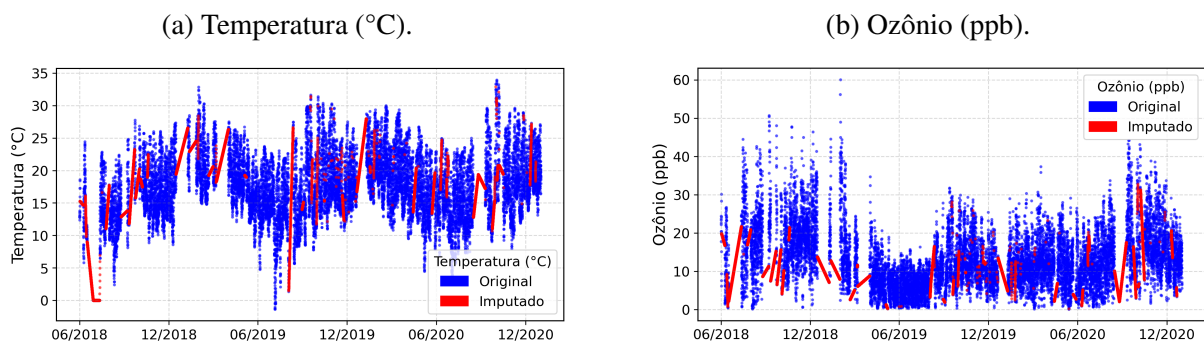
Tabela 4.8 – Estatísticas descritivas depois da imputação por interpolação linear.

Variável	Média	Desvio Padrão	Mediana	Mínimo	Máximo
Chuva (mm)	0.02	0.09	0.00	0.00	2.70
Pressão Atmosférica (hPa)	911.48	3.27	911.34	900.08	922.32
Radiação Solar Global (W/m ²)	203.48	265.76	68.50	0.00	1189.83
Temperatura (°C)	17.41	5.33	17.58	-1.38	33.96
Umidade Relativa (%)	77.50	18.31	81.00	25.00	99.50
Velocidade do Vento (m/s)	1.84	0.86	1.77	0.00	6.88
Monóxido de Carbono (ppm)	0.56	0.57	0.43	0.10	9.66
Óxido Nítrico (ppb)	6.41	6.36	5.30	0.30	128.02
Dióxido de Nitrogênio (ppb)	7.12	5.24	6.32	0.30	51.33
Óxidos de Nitrogênio (ppb)	12.17	11.33	10.79	0.30	145.43
Ozônio (ppb)	12.30	7.03	11.00	0.10	60.06

Fonte: Elaborada pela autora (2026).

o que indica potencial suavização de extremos e menos propensão a gerar valores atípicos artificiais.

Figura 4.10 – Séries temporais após interpolação linear para temperatura (a) e ozônio (b).



Fonte: Elaborada pela autora (2026).

A Tabela 4.9 apresenta os resultados da validação cruzada para os três modelos testados com dados imputados por interpolação linear. O XGBoost obteve $R^2 = 0.863$, representando o melhor desempenho entre os modelos avaliados. O Random Forest alcançou resultado próximo ($R^2 = 0.855$), enquanto o kNN apresentou desempenho inferior ($R^2 = 0.746$).

Tabela 4.9 – Resultados da imputação por interpolação linear ordenados por R^2 .

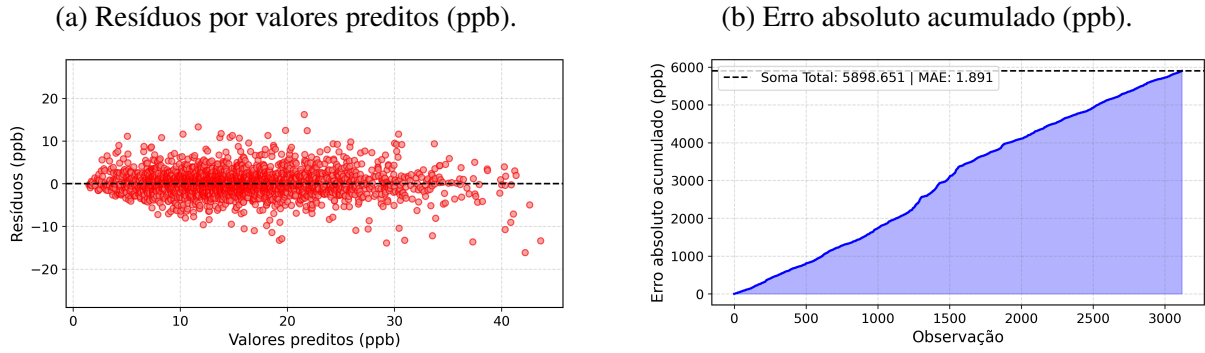
Modelo	R^2	MAE	RMSE
XGBoost	0.863	1.891	2.704
Random Forest	0.855	1.947	2.778
kNN	0.746	2.713	3.686

Fonte: Elaborada pela autora (2026).

A análise de resíduos (Figura 4.11a) apresenta padrão semelhante às imputações anteriores, com concentração dos pontos em torno de zero e presença de valores atípicos em altas concentrações de ozônio. O gráfico de erro acumulado (Figura 4.11b) mantém comportamento

linear com soma total de 5898,651 ppb, resultando no MAE reportado de 1,891 ppb e exibindo desempenho consistente ao longo dos períodos da validação cruzada.

Figura 4.11 – Análise de resíduos (a) e erro acumulado (b) para interpolação linear com XGBoost.



Fonte: Elaborada pela autora (2026).

4.5 IMPUTAÇÃO COM MÉDIA MÓVEL

A imputação por média móvel consiste em substituir valores ausentes pela média das observações contidas em uma janela temporal centrada no instante t (Equações 12 e 13). Esse método de imputação preserva a informação local da série e reduz discontinuidades artificiais causadas por lacunas nos dados. Por outro lado, a média móvel atua como um filtro de suavização e janelas maiores tendem a reduzir variações de curta duração e atenuar picos, o que pode resultar em perda de informação relevante em eventos extremos.

$$V_t = \left\{ i \in \left[t - \left\lfloor \frac{w}{2} \right\rfloor, t + \left\lfloor \frac{w}{2} \right\rfloor \right] : x_i \text{ observado} \right\} \quad (12)$$

$$\hat{x}_t = \frac{1}{|V_t|} \sum_{i \in V_t} x_i \quad (13)$$

Onde w é o tamanho da janela temporal e a soma considera apenas o conjunto de observações válidas (V_t) dentro da janela. Note que se o conjunto V_t é vazio, a média móvel é indefinida.

A implementação utilizou uma janela de tamanho 3, ou seja, para cada valor ausente foi calculada a média entre até três observações vizinhas centradas naquele instante, sendo suficiente a presença de ao menos uma observação válida dentro da janela para que a imputação fosse realizada. Para os valores que permanecessem ausentes após essa etapa, como os localizados nas extremidades da série ou em lacunas sem nenhum vizinho próximo disponível, utilizou-se o valor válido mais recente anterior ao dado faltante e, quando necessário, o valor válido imediatamente posterior.

A Tabela 4.10 apresenta as estatísticas descritivas após a imputação por média móvel, evidenciando a preservação da variabilidade local das séries através dos desvios padrão observados.

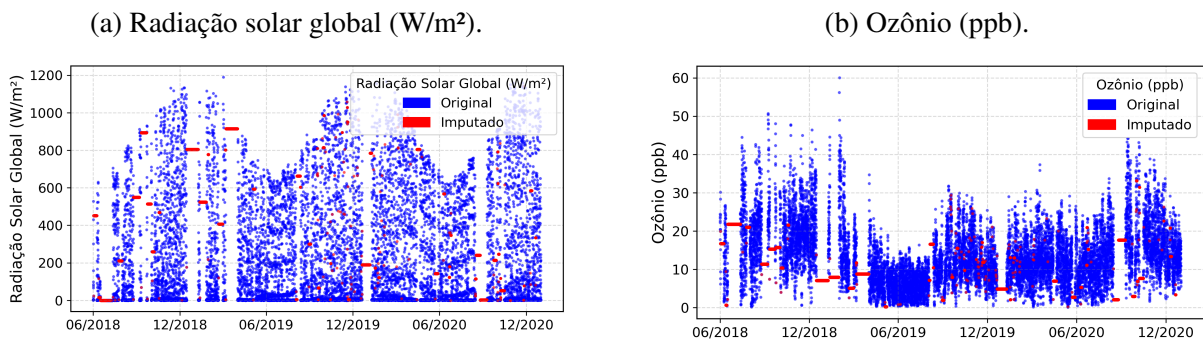
Tabela 4.10 – Estatísticas descritivas depois da imputação por média móvel.

Variável	Média	Desvio Padrão	Mediana	Mínimo	Máximo
Chuva (mm)	0.01	0.06	0.00	0.00	2.70
Pressão Atmosférica (hPa)	911.48	3.26	911.45	900.08	922.32
Radiação Solar Global (W/m ²)	253.76	311.52	82.88	0.00	1189.83
Temperatura (°C)	17.99	5.93	17.83	-1.38	33.96
Umidade Relativa (%)	74.85	19.57	76.64	25.00	99.50
Velocidade do Vento (m/s)	1.82	0.87	1.77	0.00	6.88
Monóxido de Carbono (ppm)	0.55	0.59	0.42	0.10	9.66
Óxido Nítrico (ppb)	8.05	6.54	8.80	0.30	128.02
Dióxido de Nitrogênio (ppb)	7.74	4.84	6.42	0.30	51.33
Óxidos de Nitrogênio (ppb)	13.19	11.39	14.06	0.30	145.43
Ozônio (ppb)	12.51	7.26	11.29	0.10	60.06

Fonte: Elaborada pela autora (2026).

As Figuras 4.12a e 4.12b apresentam as séries temporais após a aplicação da imputação por média móvel. Observa-se que os valores imputados mostra que o método preserva valores mais pontuais da série, mas não faz transições abruptas que ocorrem dentro da janela temporal da própria amostra, resultando em valores mais concentrados e próximos, e que visualmente formam pequenos segmentos em degraus (destacados em vermelho).

Figura 4.12 – Séries temporais após imputação por média móvel para radiação solar global (a) e ozônio (b).



Fonte: Elaborada pela autora (2026).

A Tabela 4.11 apresenta os resultados da validação cruzada para os três modelos testados com dados imputados por média móvel. O XGBoost obteve $R^2 = 0.860$, o Random Forest alcançou resultado de $R^2 = 0.857$, e o kNN apresentou desempenho inferior em relação à todos.

Tabela 4.11 – Resultados da imputação por média móvel ordenados por R^2 .

Modelo	R^2	MAE	RMSE
XGBoost	0.86	1.92	2.738
Random Forest	0.857	1.923	2.764
kNN	0.719	2.858	3.874

Fonte: Elaborada pela autora (2026).

4.6 IMPUTAÇÃO COM LOCF

A imputação por *Last Observation Carried Forward* (LOCF) preenche cada lacuna repetindo o último valor observado antes do dado faltante, conforme a Equação 14. Esse método preserva o nível mais recente da variável, mas impõe trechos constantes dentro das lacunas, o que pode reduzir a variação de curto prazo e suavizar mudanças rápidas, principalmente em variáveis com alta variabilidade, como o ozônio.

$$\hat{x}_t = x_{t-k} \quad (14)$$

Onde x_{t-k} é a última observação válida antes do instante t com valor ausente.

A Tabela 4.12 apresenta as estatísticas descritivas após a imputação por LOCF. Observa-se que o método manteve o desvio padrão do ozônio próximo ao valor original (7.46 ppb; Tabela 4.3); entretanto, houve redução da média, indicando um deslocamento da tendência central após a imputação, uma vez que o método replica valores anteriores e pode subestimar períodos em que o ozônio tenderia a assumir níveis mais elevados ou mais baixos.

Tabela 4.12 – Estatísticas descritivas depois da imputação por LOCF.

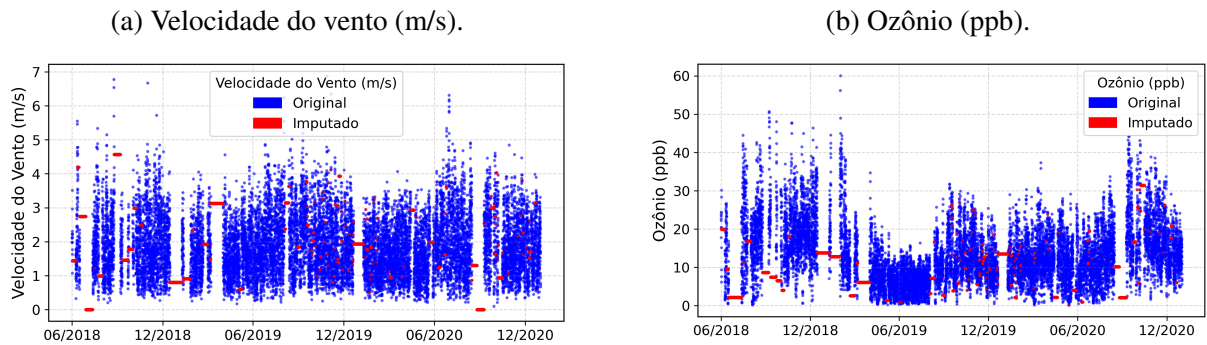
Variável	Média	Desvio Padrão	Mediana	Mínimo	Máximo
Chuva (mm)	0.03	0.13	0.00	0.00	2.70
Pressão Atmosférica (hPa)	911.49	3.43	911.20	900.08	922.32
Radiação Solar Global (W/m ²)	152.62	247.72	13.67	0.00	1189.83
Temperatura (°C)	16.84	5.08	16.90	-1.38	33.96
Umidade Relativa (%)	80.19	19.14	86.64	25.00	99.50
Velocidade do Vento (m/s)	1.87	0.97	1.78	0.00	6.88
Monóxido de Carbono (ppm)	0.57	0.56	0.45	0.10	9.66
Óxido Nítrico (ppb)	4.77	6.68	2.30	0.30	128.02
Dióxido de Nitrogênio (ppb)	6.49	5.96	6.28	0.30	51.33
Óxidos de Nitrogênio (ppb)	11.14	11.81	8.58	0.30	145.43
Ozônio (ppb)	12.09	7.45	11.12	0.10	60.06

Fonte: Elaborada pela autora (2026).

As Figuras 4.13a e 4.13b apresentam as séries temporais após a imputação por LOCF. Em ambas as variáveis, os valores imputados (em vermelho) aparecem como pequenos segmentos horizontais, mostrando que o método mantém o último valor e não cria transições bruscas dentro da lacuna.

A Tabela 4.13 apresenta os resultados da validação cruzada para os três modelos testados com dados imputados por LOCF. O XGBoost obteve $R^2 = 0.861$, representando o melhor

Figura 4.13 – Séries temporais após imputação por LOCF para velocidade do vento (a) e ozônio (b).



Fonte: Elaborada pela autora (2026).

desempenho entre os modelos avaliados. O Random Forest alcançou resultado próximo ($R^2 = 0.858$), enquanto o kNN apresentou desempenho inferior ($R^2 = 0.755$).

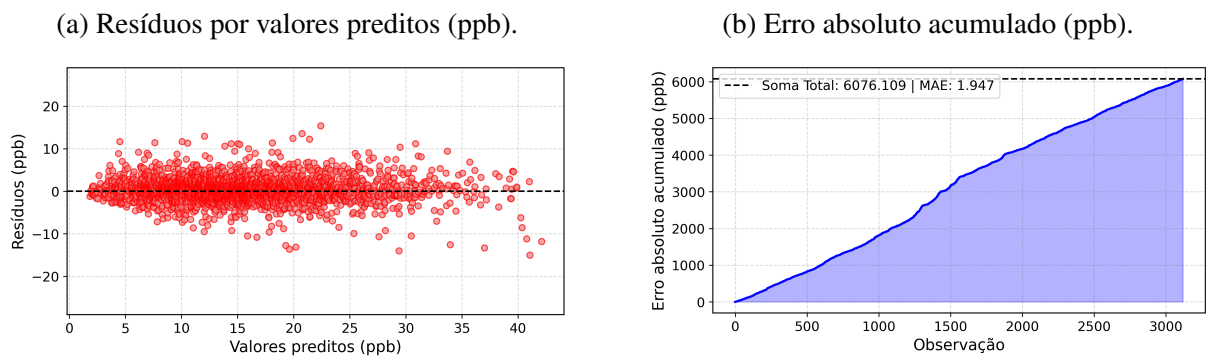
Tabela 4.13 – Resultados da imputação por LOCF ordenados por R^2 .

Modelo	R^2	MAE	RMSE
XGBoost	0.861	1.947	2.724
Random Forest	0.858	1.939	2.756
kNN	0.755	2.648	3.618

Fonte: Elaborada pela autora (2026).

A Figura 4.14a apresenta resíduos mais centrados em zero, com menos dispersão na faixa de previsões intermediárias, mas aumento da dispersão para valores preditos mais elevados (acima de 30 ppb). Esse padrão sugere que o modelo tem menos probabilidade de erro na faixa aproximada de 5 a 20 ppb e maior incerteza e erro de predição em concentrações mais altas, onde também se observam resíduos mais extremos, próximos a -20 ppb.

Figura 4.14 – Análise de resíduos (a) e erro acumulado (b) para imputação por LOCF com XGBoost.



Fonte: Elaborada pela autora (2026).

4.7 IMPUTAÇÃO COM MICE

A implementação do método MICE foi feita ajustando os parâmetros para otimizar a qualidade da imputação. Um dos ajustes feitos foi nos parâmetros de limite, onde os valores mínimos e máximos foram configurados especificamente para cada variável utilizando os próprios limites observados nos dados originais. Dessa forma foram preservadas as características únicas de cada variável, evitando imputações irrealistas que poderiam ocorrer se fossem aplicados limites genéricos a variáveis de escalas e comportamentos diferentes.

A Tabela 4.14 apresenta as estatísticas descritivas após a imputação por MICE. Comparando com a Tabela 4.3, observa-se que o MICE preservou as estatísticas de tendência central, mantendo a média do ozônio próxima ao valor original (13,10 ppb, 13,15 ppb original). Os desvios do ozônio reduziu de 7,47 para 6,36 ppb (14,9% de redução), a pressão atmosférica de 3,39 para 2,92 hPa (13,9% de redução), e a temperatura de 4,92 para 4,20 °C (14,6% de redução), indicando que o método iterativo do MICE compreendeu as relações multivariadas. Diferentemente dos métodos univariados (média, mediana), o MICE utiliza as correlações entre variáveis para estimar valores faltantes.

Tabela 4.14 – Estatísticas descritivas depois da imputação por Mice.

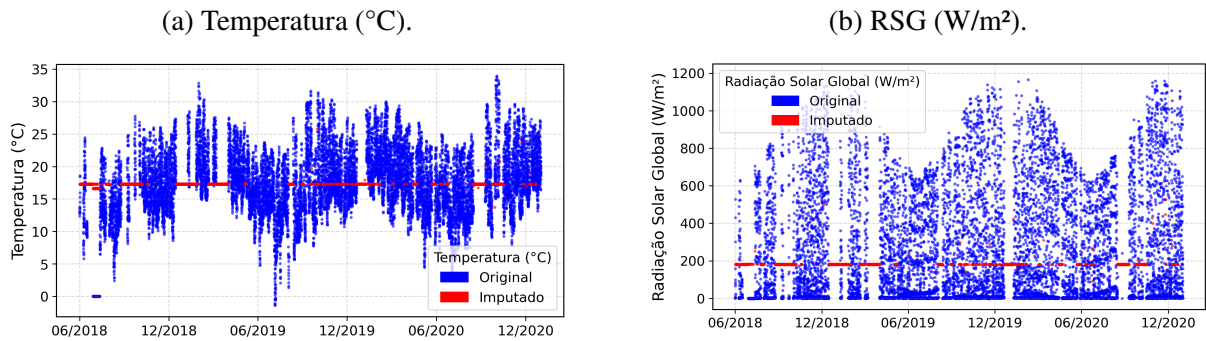
Variável	Média	Desvio Padrão	Mediana	Mínimo	Máximo
Chuva (mm)	0.01	0.06	0.00	0.00	2.70
Pressão Atmosférica (hPa)	911.29	2.92	911.29	900.08	922.32
Radiação Solar Global (W/m ²)	182.23	236.17	168.92	0.00	1189.83
Temperatura (°C)	17.30	4.20	17.28	-1.38	33.96
Umidade Relativa (%)	78.19	16.83	78.32	25.00	99.50
Velocidade do Vento (m/s)	1.86	0.74	1.86	0.00	6.88
Monóxido de Carbono (ppm)	0.59	0.50	0.58	0.10	9.66
Óxido Nítrico (ppb)	7.49	5.99	7.51	0.30	128.02
Dióxido de Nitrogênio (ppb)	8.78	4.53	8.79	0.30	51.33
Óxidos de Nitrogênio (ppb)	15.86	9.57	15.92	0.30	145.43
Ozônio (ppb)	13.10	6.36	13.08	0.10	60.06

Fonte: Elaborada pela autora (2026).

As séries temporais da Figura 4.15 mostram a sobreposição entre os valores observados e os valores imputados pelo MICE para temperatura e radiação solar global. Observa-se que nos valores imputados não aparecem com valores extremos artificiais, se aproximando às linhas de imputação por média, devido ao ajuste de parâmetros que otimizados para este método de imputação.

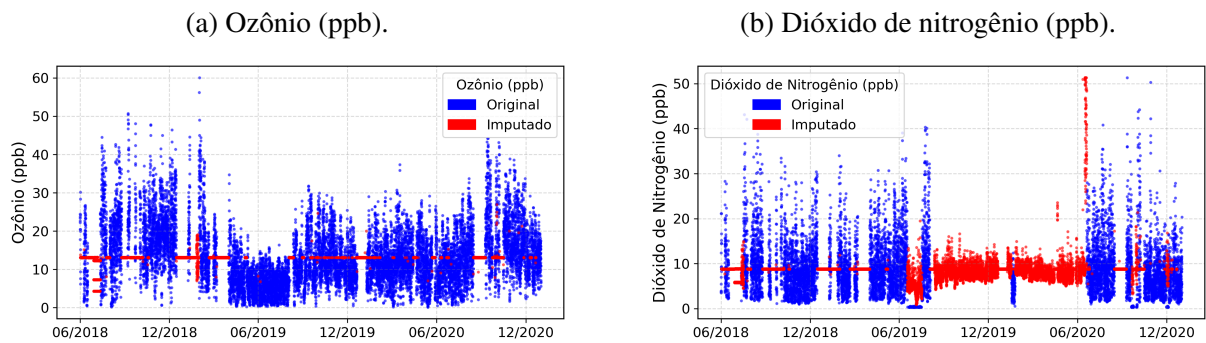
A Figura 4.16 apresenta as séries temporais do ozônio e do dióxido de nitrogênio após a imputação por MICE. Nota-se que os valores imputados acompanham a escala dos dados observados, mantendo a coerência física entre os poluentes, e com valores próximos à média para o ozônio. Para o dióxido de nitrogênio, os valores imputados concentram-se em níveis intermediários, mas apresentam maior variabilidade em lacunas sequenciais com grande falta de dados.

Figura 4.15 – Séries temporais após imputação por MICE para temperatura (a) e RSG (b).



Fonte: Elaborada pela autora (2026).

Figura 4.16 – Séries temporais após imputação por MICE para ozônio (a) e dióxido de nitrogênio (b).



Fonte: Elaborada pela autora (2026).

A Tabela 4.15 mostra diferenças no desempenho dos modelos aplicados aos dados imputados por MICE. O XGBoost apresentou o melhor ajuste, com maior valor de R^2 e menores erros (MAE e RMSE), indicando maior capacidade de explicar a variabilidade dos dados e maior precisão preditiva. O Random Forest apresentou desempenho muito próximo, o que indica que o modelo se beneficiou da estrutura feita pela imputação por MICE. O kNN obteve desempenho inferior, com menor R^2 e erros mais elevados, indicando menor capacidade de generalização.

Tabela 4.15 – Resultados da imputação por MICE ordenados por R^2 .

Modelo	R^2	MAE	RMSE
XGBoost	0.855	1.951	2.779
Random Forest	0.849	1.983	2.833
kNN	0.732	2.736	3.779

Fonte: Elaborada pela autora (2026).

5 IMPUTAÇÃO COM DADOS EXTERNOS

Neste capítulo, investigou-se a utilização de dados provenientes de estações meteorológicas (dados externos) como estratégia para a imputação das variáveis meteorológicas. Os dados externos foram obtidos por meio da *Weatherbit Historical Weather Hourly API* (Weatherbit, 2025), que fornece séries horárias a partir de uma rede de mais de 120.000 estações meteorológicas físicas, integrando fontes em um raio de aproximadamente 15 km do ponto de interesse, representando portanto medições reais e não saídas de modelos numéricos. Para as variáveis de qualidade do ar, optou-se pelo método MICE como etapa complementar, com o objetivo de avaliar se uma base meteorológica imputada com medições reais, em vez de valores sintéticos, potencializaria a capacidade do MICE de capturar relações multivariadas entre as variáveis químicas e atmosféricas. A Tabela 5.1 apresenta as variáveis utilizadas neste trabalho, com suas respectivas unidades, conforme disponibilizadas pela API. Ao comparar essas variáveis com aquelas presentes no conjunto de dados original, observa-se que não há divergência significativa nas unidades utilizadas.

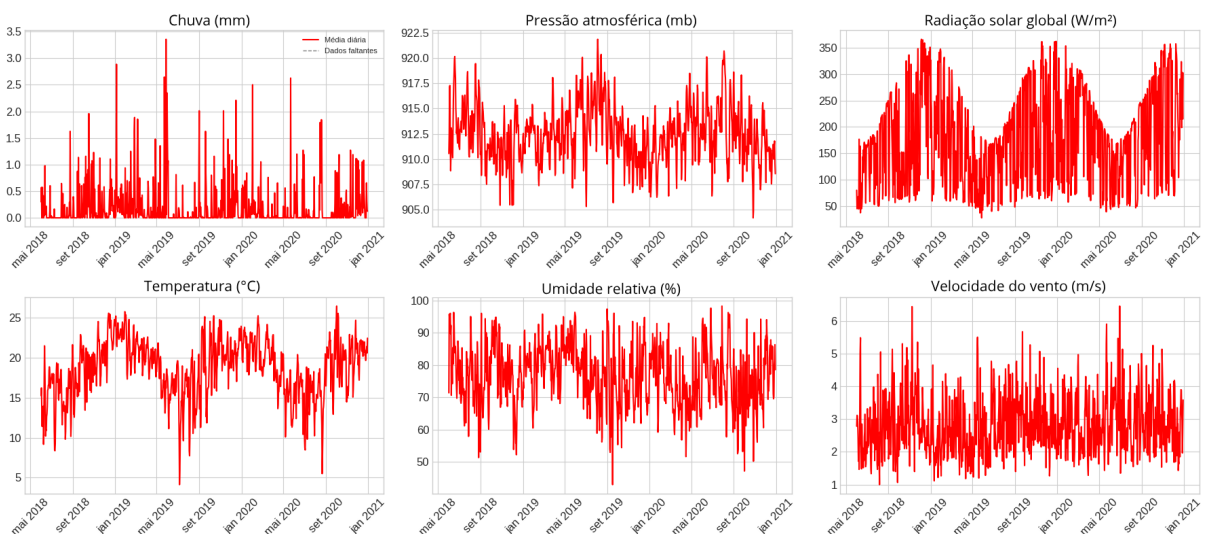
Tabela 5.1 – Comparação de unidades entre os conjuntos de dados.

Variável	Dados dos sensores	Dados externos	Observação
Chuva	mm	mm	
Pressão atmosférica	hPa	mb	1 hPa = 1 mb
Radiação Solar Global	W m^{-2}	W m^{-2}	
Umidade relativa	%	%	
Velocidade do vento	m s^{-1}	m s^{-1}	

Fonte: Elaborada pela autora (2026).

A Figura 5.1 mostra as séries temporais completas das variáveis meteorológicas externas sem dados faltantes para o período de março/2018 a janeiro/2021.

Figura 5.1 – Dados meteorológicos externos (2018-2021).



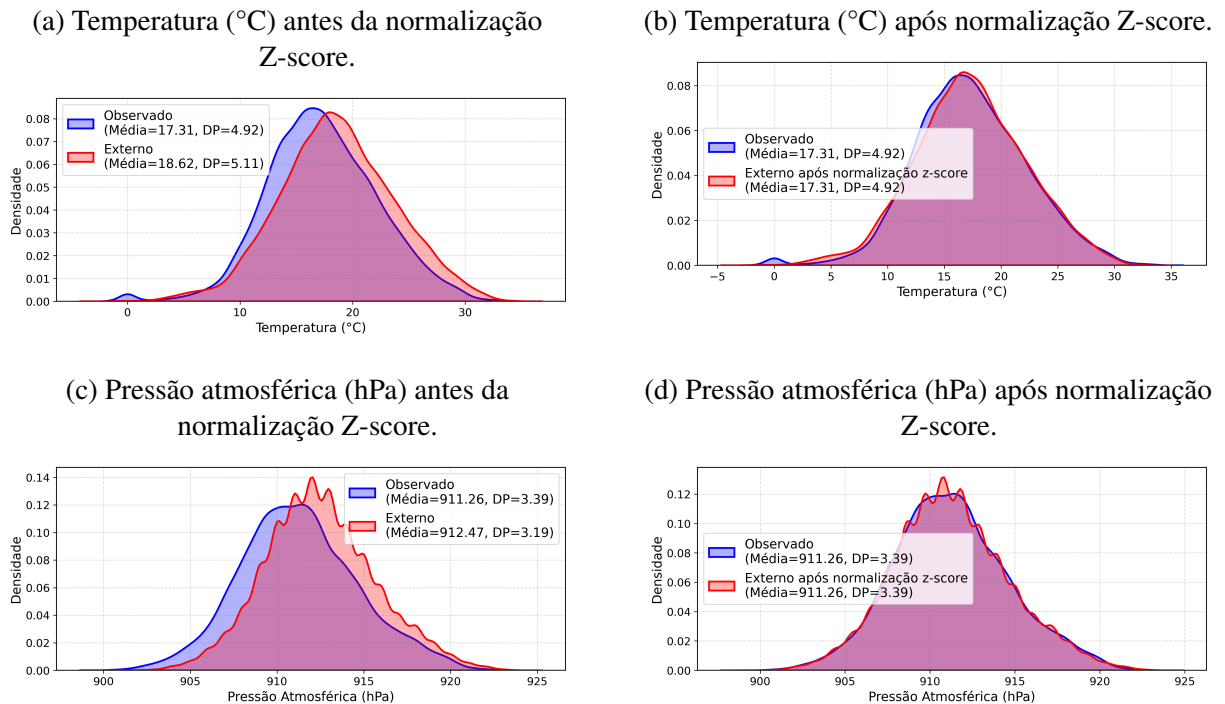
Fonte: Elaborada pela autora (2026).

Para garantir a imputação de forma correta, foi feita uma compatibilidade estatística entre os dados externos e originais existentes da banco de dados oriundo da estação de qualidade do ar. Esta compatibilização estatística foi então utilizada para imputar, a partir dos dados externos compatibilizados, os dados faltantes no banco de dados original. Implementou-se então um processo de compatibilização em três etapas: (1) normalização Z-score para centralização (Equação 15), (2) transformações específicas por variável (logarítmica para chuva e raiz quadrada para RSG) e, por fim, (3) normalização quantílica para alinhar os quantis dos dados externos aos originais.

$$Z = \frac{X - \mu}{\sigma} \quad (15)$$

A Figura 5.2 ilustra o processo de compatibilização estatística para as variáveis temperatura e pressão atmosférica, considerando os dados existentes do banco de dados original. As Figuras 5.2a e 5.2c apresentam as distribuições antes da normalização Z-score. Observa-se um deslocamento entre os dados observados e os dados externos para ambas as variáveis. Após a aplicação da normalização Z-score (Figuras 5.2b e 5.2d), as distribuições apresentam sobreposição completa. Diferentemente das técnicas de imputação apresentadas no Capítulo 4, onde em alguns casos as médias eram próximas mas os desvios padrão permaneciam distintos, ou em outros casos o desvio padrão era satisfatório mas a média não, a normalização Z-score garantiu que ambas as séries compartilhem média e desvio padrão idênticos, preservando as características estatísticas dos dados originais.

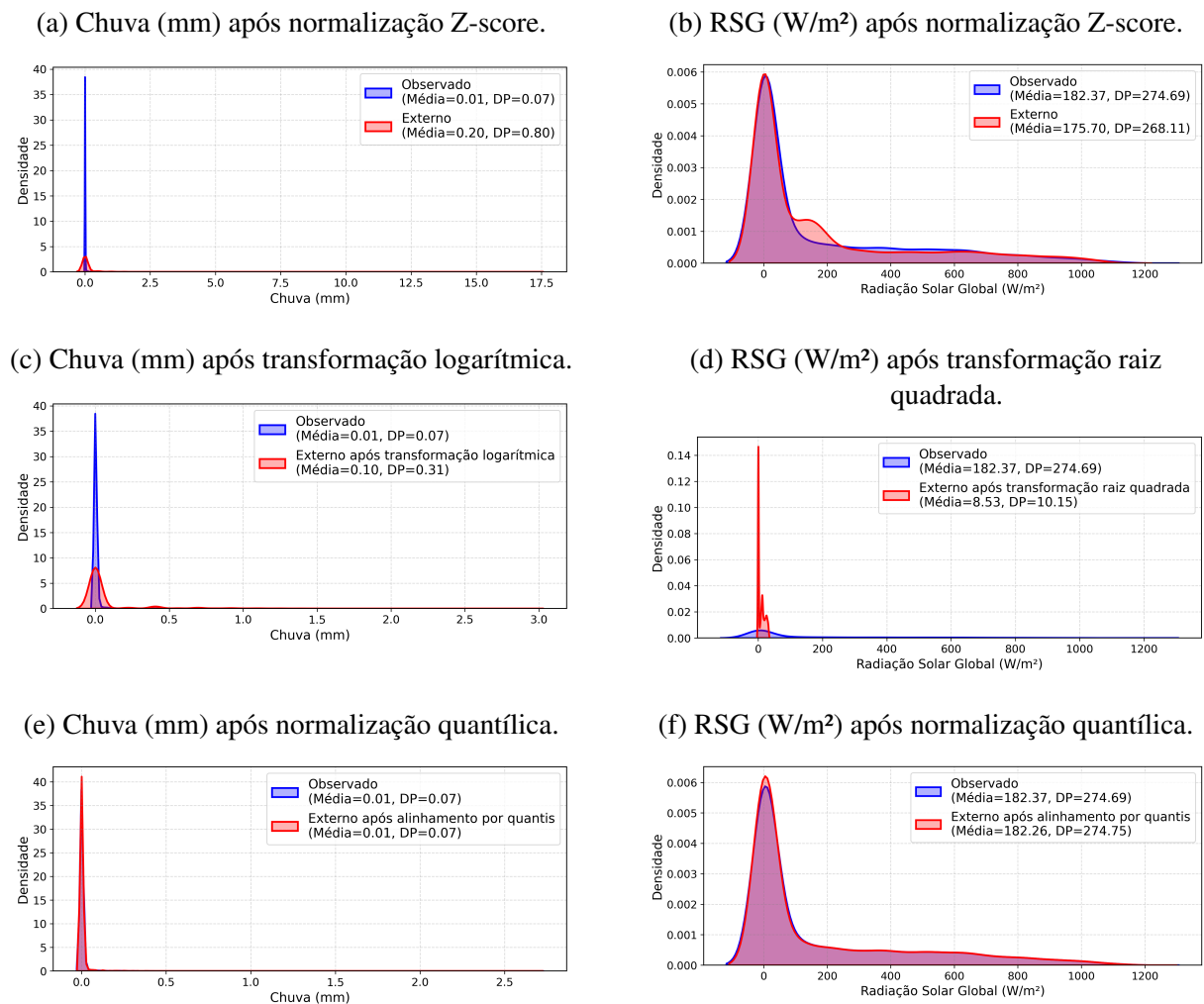
Figura 5.2 – Compatibilização estatística de variáveis meteorológicas através de normalização Z-score.



Fonte: Elaborada pela autora (2026).

A Figura 5.3 apresenta o processo de compatibilização para as variáveis Chuva e RSG, que exigiram etapas adicionais além da normalização Z-score. As Figuras 5.3a e 5.3b mostram as distribuições após a aplicação inicial do Z-score, onde ainda persistem diferenças na forma das distribuições entre os dados observados e externos. Foram então aplicadas transformações específicas: logarítmica para Chuva (Figura 5.3c) e raiz quadrada para RSG (Figura 5.3d). Por fim, as Figuras 5.3e e 5.3f apresentam os resultados após a normalização quantílica, onde observa-se sobreposição completa das distribuições. Essa sequência de transformações aproximou as características estatísticas das distribuições, embora com resultados distintos entre algumas variáveis.

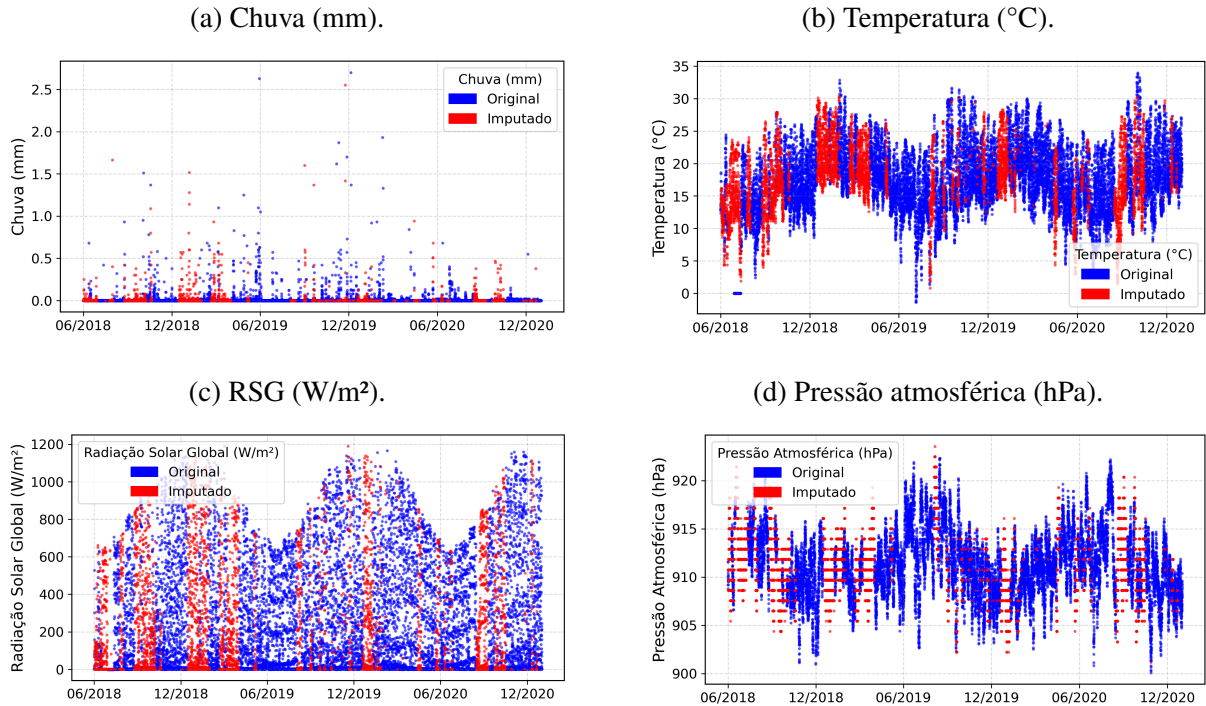
Figura 5.3 – Transformações específicas aplicadas às variáveis Chuva e RSG após normalização Z-score.



Fonte: Elaborada pela autora (2026).

A Figura 5.4 compara as distribuições dos dados originais (azul) com os dados externos após a compatibilização estatística (vermelho). Observa-se que as distribuições de chuva e umidade exibem padrões muito próximos; temperatura, pressão e RSG também preservam o comportamento sazonal.

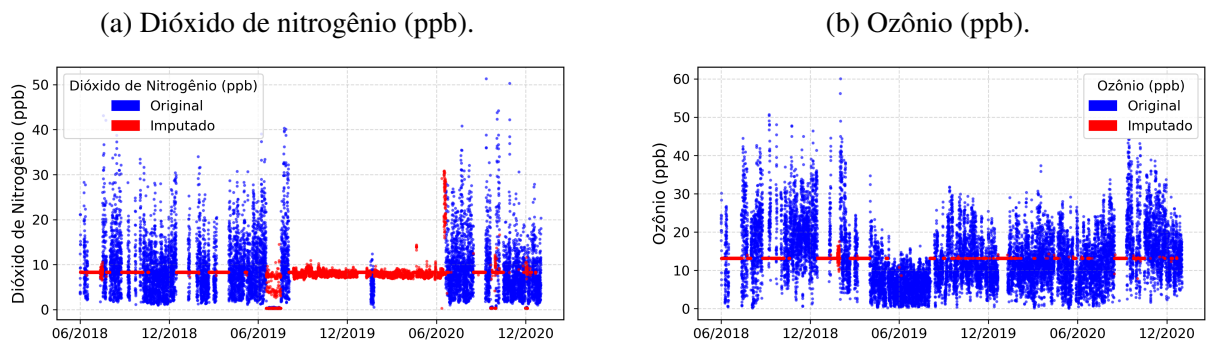
Figura 5.4 – Séries temporais após imputação com dados externos.



Fonte: Elaborada pela autora (2026).

No Capítulo 4 utilizou-se o MICE para todas as variáveis do conjunto. Anteriormente, o dióxido de nitrogênio apresentou um pico acentuado próximo ao mês de julho de 2020 (Figura 4.16b). Ao aplicar dados externos para as variáveis meteorológicas e, posteriormente, o MICE para as variáveis químicas remanescentes, esse pico foi significativamente suavizado (Figura 5.5a), reduzindo a presença de valores atípicos.

Figura 5.5 – Séries temporais das variáveis precursoras do ozônio após imputação das meteorológicas com dados externos e aplicação do MICE: dióxido de nitrogênio (a) e ozônio (b).



Fonte: Elaborada pela autora (2026).

Os resultados de predição de ozônio (ppb) utilizando dados meteorológicos imputados externos são apresentados na Tabela 5.2. Entre os três modelos avaliados, XGBoost alcançou desempenho superior com R^2 de 0.858, valor próximo ao Random Forest ($R^2 = 0.850$). Esses resultados reforçam a boa capacidade dos algoritmos baseados em árvores de decisão em modelar

a complexidade da relação entre variáveis meteorológicas e concentração de ozônio, assim como observado anteriormente no Capítulo 4.

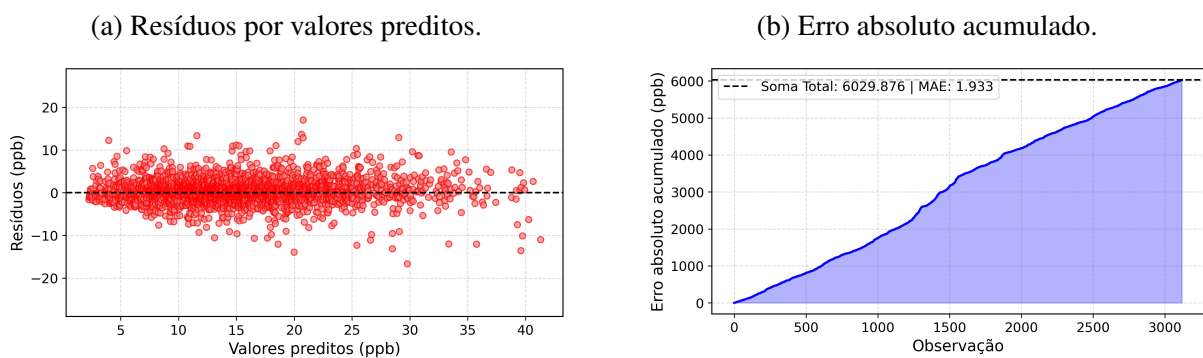
Tabela 5.2 – Resultados da imputação por dados externos ordenados por R^2 .

Modelo	R^2	MAE	RMSE
XGBoost	0.858	1.933	2.754
Random Forest	0.850	1.992	2.828
kNN	0.731	2.776	3.784

Fonte: Elaborada pela autora (2026).

Na Figura 5.6a, o gráfico de resíduos por valores preditos mostra a presença de alguns valores atípicos com erros de aproximadamente ± 20 ppb, concentrados predominantemente em regiões de alta concentração de ozônio e também em valores acima da faixa de 25 ppb.

Figura 5.6 – Análise de resíduos (a) e erro acumulado (b) para imputação por dados externos com XGBoost.



Fonte: Elaborada pela autora (2026).

5.1 VIABILIDADE DO USO DE DADOS EXTERNOS PARA IMPUTAÇÃO

Inicialmente, analisou-se o desempenho médio de cada algoritmo considerando todos os métodos de imputação. O XGBoost demonstrou superioridade consistente, apresentando R^2 médio de 0.858, MAE de 1.937 e RMSE de 2.751 (Tabela 5.3). O desvio padrão entre os métodos de imputação é extremamente baixo: 0.003 para R^2 , 0.028 para MAE e 0.031 para RMSE. Isso indica que o XGBoost mantém performance estável independentemente do método de imputação escolhido. O Random Forest apresentou comportamento similar, com desvio padrão de 0.004 em R^2 . O kNN exibiu maior sensibilidade ao método de imputação, com desvio padrão de 0.011 em R^2 , aproximadamente 3.5 vezes maior que o XGBoost. Essa estabilidade dos algoritmos baseados em árvores indica que a escolha do método de imputação tem impacto limitado no desempenho final.

Dado o desempenho superior e a estabilidade do XGBoost, as análises subsequentes concentram-se neste algoritmo. A Interpolação Linear obteve o melhor desempenho ($R^2 = 0.863$), seguida por LOCF ($R^2 = 0.861$) e Média Móvel ($R^2 = 0.860$). Os dados externos ocuparam a

Tabela 5.3 – Desempenho médio por algoritmo de aprendizado considerando sete métodos de imputação.

Modelo	R^2		MAE		RMSE	
	Média	DP	Média	DP	Média	DP
XGBoost	0.858	0.003	1.937	0.028	2.751	0.031
Random Forest	0.853	0.004	1.963	0.026	2.799	0.032
kNN	0.737	0.011	2.743	0.064	3.748	0.081

Fonte: Elaborada pela autora (2026).

quarta posição com $R^2 = 0.858$, apenas 0.005 pontos abaixo do melhor método. A diferença entre o melhor e o pior método é de 0.010 pontos em R^2 , reforçando a estabilidade previamente observada. A Tabela 5.4 apresenta a ordenação completa.

Tabela 5.4 – Ordenação dos métodos de imputação no XGBoost por desempenho.

Posição	Método	R^2	MAE	RMSE
1	Interpolação Linear	0.863	1.891	2.704
2	LOCF	0.861	1.947	2.724
3	Média Móvel	0.860	1.920	2.738
4	Dados Externos	0.858	1.933	2.754
5	Mediana	0.857	1.936	2.761
6	MICE	0.855	1.951	2.779
7	Média	0.853	1.980	2.794

Fonte: Elaborada pela autora (2026).

A Tabela 5.5 quantifica as diferenças entre Interpolação Linear e Dados Externos. A diferença em R^2 é de 0.005 pontos absolutos (0.58%), em MAE é de 0.042 unidades (2.22%) e em RMSE é de 0.050 unidades (1.85%). Essas diferenças indicam que ambos os métodos explicam praticamente a mesma proporção da variância dos dados, com erro de predição quase idêntico. Em termos práticos, o modelo com dados externos erra em média apenas 0.042 unidades a mais por predição, significando uma magnitude baixa considerando a escala dos erros observados, como por exemplo o MAE na ordem de 1.9 unidades dos dados externos.

Tabela 5.5 – Diferença de desempenho entre Interpolação Linear e Dados Externos no XGBoost.

Método	R^2	MAE	RMSE
Interpolação Linear	0.863	1.891	2.704
Dados Externos	0.858	1.933	2.754
Diferença Absoluta	-0.005	+0.042	+0.050
Diferença Relativa (%)	-0.58%	+2.22%	+1.85%

Fonte: Elaborada pela autora (2026).

Aplicou-se um teste de significância estatística para verificar se essas diferenças são consistentes ou resultado de variação aleatória. O teste comparou os erros de predição ponto a ponto entre todos os métodos no XGBoost, resultando em $\chi^2 = 39.97$ com $p < 0.0001$. Esse resultado indica que existem diferenças estatisticamente detectáveis entre os métodos. Entretanto,

a magnitude dessas diferenças permanece inferior a 1% em R^2 e cerca de 2% nas métricas de erro, como demonstrado na tabela anterior.

Para complementar a análise, avaliou-se o desempenho médio dos métodos de imputação considerando os três algoritmos conjuntamente. O LOCF apresentou o melhor desempenho médio ($R^2 = 0.825$), enquanto dados externos obtiveram $R^2 = 0.813$ (Tabela 5.6).

Tabela 5.6 – Diferença de desempenho entre LOCF e Dados Externos considerando três algoritmos.

Método	R^2	MAE	RMSE
LOCF	0.825	2.178	3.033
Dados Externos	0.813	2.234	3.122
Diferença Absoluta	-0.012	+0.056	+0.089
Diferença Relativa (%)	-1.41%	+2.56%	+2.95%

Fonte: Elaborada pela autora (2026).

Aplicou-se também um teste de significância estatística sobre os resultados médios dos três algoritmos. Os resultados indicaram $p = 0.077$ para R^2 , $p = 0.334$ para MAE e $p = 0.076$ para RMSE. Todos os valores indicam que as diferenças entre os métodos não são estatisticamente significativas quando considerados múltiplos algoritmos. Isso demonstra que a variabilidade entre algoritmos supera a variabilidade entre métodos de imputação, indicando que a escolha do método de imputação tem impacto reduzido no desempenho final quando comparada à escolha do algoritmo de aprendizado.

Dados externos apresentam desempenho comparável aos métodos clássicos de imputação, com diferenças inferiores a 3% em todas as métricas. Diferente de métodos que geram valores sintéticos, dados externos utilizam medições reais de estações independentes, oferecendo maior interpretabilidade dos valores imputados.

Um valor imputado por Interpolação Linear ou MICE é uma estimativa matemática sem correspondência física direta, enquanto um valor de dados externos representa uma medição real coletada por instrumentação calibrada em condições atmosféricas similares.

Essa característica facilita a validação dos dados imputados e aumenta a confiança em análises que exigem rastreabilidade das medições. Além disso, dados externos operam independentemente do histórico local e fornecem valores em tempo real sem janelas de observação. Métodos como MICE dependem de múltiplas observações adjacentes para estimar valores por regressão múltipla, enquanto dados externos não apresentam essa limitação.

Considerando que a diferença de desempenho no melhor algoritmo é inferior a 1% em R^2 e que as diferenças não são estatisticamente significativas na análise conjunta, o uso de dados externos é viável em aplicações que demandam robustez operacional, independência de histórico e aplicabilidade em tempo real.

5.2 DESEMPENHO DOS MODELOS PREDITIVOS

A Tabela 5.3 apresenta o desempenho médio dos três algoritmos avaliados, considerando todos os métodos de imputação testados. O XGBoost apresentou o melhor desempenho médio, com $R^2 = 0.858$, seguido do Random Forest, que obteve resultados próximos ($R^2 = 0.853$), e por fim o kNN, que apresentou desempenho inferior, com $R^2 = 0.737$. A diferença entre XGBoost e Random Forest é de 0.005 pontos em R^2 , enquanto a diferença entre XGBoost e kNN é de 0.121 pontos.

As métricas de erro absoluto indicam que o XGBoost erra em média 1.937 unidades por predição, enquanto o kNN erra 2.743 unidades. O RMSE do XGBoost (2.751) é aproximadamente 27% menor que o do kNN (3.748), indicando que o XGBoost também produz menos erros extremos. O Random Forest mantém desempenho intermediário em todas as métricas, com valores próximos ao XGBoost.

O R^2 de 0.858 do XGBoost indica que o modelo explica 85.8% da variância dos dados de ozônio. O MAE de 1.937 mostra que, em média, as predições desviam 1.9 unidades do valor real, o que representa aproximadamente 14.7% da concentração média observada (13.15 unidades, Tabela 4.3). Adicionalmente, o RMSE de 2.751 é inferior ao desvio padrão dos dados de ozônio (7.47), indicando que as predições do modelo são consideravelmente mais precisas do que simplesmente utilizar a média histórica como estimativa. O RMSE, por ser apenas 42% maior que o MAE, sugere que os erros são relativamente uniformes. Se houvesse erros muito grandes frequentes, o RMSE seria significativamente maior que o MAE. Isso mostra que o XGBoost é capaz de gerar predições consistentes para aplicações de monitoramento contínuo de ozônio troposférico.

6 CONSIDERAÇÕES FINAIS

A presença de lacunas nos dados de qualidade do ar representa um obstáculo crítico para o monitoramento contínuo e confiável de poluentes atmosféricos, afetando diretamente estudos como o de Santana, Moreira e Armani (2020), que precisou descartar períodos significativos de medição devido a falhas operacionais e valores atípicos. Este trabalho avaliou sete métodos de imputação para lidar com dados faltantes em séries temporais de ozônio troposférico, testando-os com três algoritmos de aprendizado de máquina. Os resultados mostraram que o XGBoost mantém desempenho estável independentemente do método de imputação, com variação inferior a 1% em R^2 .

Dados externos ocuparam posição mediana entre os métodos de imputação, mas com diferenças inferiores a 3% em todas as métricas, sem significância estatística quando considerados múltiplos algoritmos. Diferente dos métodos clássicos, dados externos operam com medições reais de estações independentes, eliminando dependência de histórico local. Métodos como Interpolação Linear e LOCF propagam padrões existentes, enquanto MICE exige janelas de observação para regressão múltipla. Dados externos podem ser aplicados imediatamente após falhas de sensor, em novos pontos de monitoramento sem histórico prévio, ou quando toda a série local apresenta lacunas prolongadas de dados faltantes.

A formação de ozônio troposférico envolve reações fotoquímicas não-lineares dependentes de múltiplos fatores (NO_x, VOCs, temperatura, radiação solar, umidade), cujas interações não são adequadamente capturadas por métodos estatísticos simples como média e mediana. Embora esses métodos apresentem bom desempenho em bases pequenas com poucos valores faltantes, em bases maiores ou com lacunas extensas, a imputação repetida de valores constantes reduz artificialmente a variabilidade natural dos dados, subestima a variância real da série temporal e pode induzir viés ao criar padrões de constância que não existem nos fenômenos atmosféricos.

Dados externos, por utilizarem medições reais que preservam a complexidade e variabilidade natural das condições meteorológicas, apresentam potencial promissor mesmo sem otimização específica. A expansão da base de dados com séries temporais mais longas, incluindo múltiplos anos de observação, permitiria capturar maior diversidade de regimes climáticos, eventos extremos e padrões sazonais, potencialmente ampliando a capacidade preditiva e a robustez dos modelos baseados em dados externos.

Dessa forma, a imputação de dados meteorológicos a partir de outras bases de dados permite reconstruir séries temporais completas utilizáveis em múltiplas análises além da previsão de ozônio. Variáveis como temperatura, umidade e radiação solar imputadas podem preencher lacunas em medições de outros poluentes correlacionados (NO, NO₂, CO), viabilizando estudos de interação entre múltiplos contaminantes.

Séries contínuas facilitam identificação de padrões sazonais, ciclos fotoquímicos diários e relações não-lineares entre variáveis químicas. Estudos que mapeiam fontes de emissão e dispersão atmosférica em regiões industriais e portuárias demandam séries longas sem interrup-

ções. Modelos de transporte químico atmosférico, que simulam formação e dispersão de ozônio secundário, exigem dados de variáveis meteorológicas contínuas como entrada. Em regiões com infraestrutura limitada, onde sensores de ozônio são escassos por custo e complexidade operacional, reconstruir variáveis meteorológicas via dados externos amplia a capacidade de gerar estimativas espacialmente distribuídas e confiáveis. Isso possibilita análises de exposição populacional e suporte a decisões de controle de qualidade do ar mesmo em locais sem monitoramento direto de O_3 .

Este trabalho demonstra que pequenas perdas de desempenho preditivo são compensadas por ganhos operacionais ao utilizar dados externos. A combinação de XGBoost com imputação baseada em medições independentes permite prever ozônio com alta acurácia ($R^2 = 0.858$), indicando que o modelo explica 85.8% da variância dos dados, e construir bases de dados completas para múltiplas aplicações. Trabalhos futuros podem explorar engenharia de variáveis (*feature engineering*) sobre os dados imputados, considerando que a capacidade do modelo de capturar relações complexas entre ozônio, poluentes precursores e variáveis meteorológicas sugere que transformações adicionais (interações não-lineares, defasagens temporais, agregações sazonais) podem extrair informações ainda não exploradas. A segmentação de modelos por estações do ano, análise de ciclos completos de variáveis meteorológicas e investigação de transferência de aprendizado entre estações próximas podem ampliar a compreensão sobre dinâmica de formação de ozônio. Além disso, a expansão da série temporal, incorporando dados de múltiplos anos, pode revelar tendências de longo prazo, melhorar a robustez dos modelos em diferentes regimes climáticos e permitir análises de sazonalidade e eventos extremos com maior confiabilidade estatística.

REFERÊNCIAS

- AHMAD, Aimi Nursyahirah *et al.* Modeling ground level ozone (o₃) of air pollution using regression technique. **Malaysian Journal of Medicine and Health Sciences**, v. 18, n. SUPP 8, p. 97–103, 2022. Citado na página 21.
- AHMAD, Aimi Nursyahirah *et al.* Assessment of daytime and nighttime ground level ozone pollution in malaysian urban areas. **Malaysian Journal of Medicine and Health Sciences**, v. 19, n. 6, p. 242 – 248, 2023. Citado 2 vezes nas páginas 20 e 21.
- ALSHEHHI, F. A.; ALSHEHHI, A. M. Artificial intelligence for ground-level ozone concentration forecasting using data from the ground stations of the abu dhabi environment agency. In: **2023 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)**. [S.l.: s.n.], 2023. p. 1017–1021. Citado 3 vezes nas páginas 33, 34 e 35.
- ALVES, Luciano da Silva; ALVES, Lara Camila da Silva. Formação e concentração do ozônio troposférico no município de lamarão do passé – ba: Estudo das transformações de nox e os possíveis impactos a saúde humana. **Revista Geama**, v. 5, n. 1, p. 30–35, abr. 2019. Disponível em: <https://www.journals.ufrpe.br/index.php/geama/article/view/2465>. Citado na página 14.
- APARECIDO, Lucas Eduardo de Oliveira *et al.* Köppen, thornthwaite and camargo climate classifications for climatic zoning in the state of paraná, brazil. **Ciência e Agrotecnologia**, Editora da Universidade Federal de Lavras, v. 40, n. 4, p. 405–417, Jul 2016. ISSN 1413-7054. Disponível em: <https://doi.org/10.1590/1413-70542016404003916>. Citado 3 vezes nas páginas 38, 39 e 41.
- ATKINSON, Roger. Atmospheric chemistry of vocs and nox. **Atmospheric Environment**, v. 34, n. 12, p. 2063–2101, 2000. ISSN 1352-2310. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1352231099004604>. Citado 3 vezes nas páginas 20, 21 e 22.
- AVINO, Pasquale; MANIGRASSO, Maurizio. Ozone formation in relation with combustion processes in highly populated urban areas. **AIMS Environmental Science**, v. 2, n. 3, p. 764–781, 2015. ISSN 2372-0352. Disponível em: <https://www.aimspress.com/article/doi/10.3934/environsci.2015.3.764>. Citado na página 20.
- AW, Jeremy; KLEEMAN, Michael J. Evaluating the first-order effect of intraannual temperature variability on urban air pollution. **Journal of Geophysical Research: Atmospheres**, v. 108, n. D12, 2003. Disponível em: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD002688>. Citado na página 22.
- BARRY, Roger G.; CHORLEY, Richard J. **Atmosfera, tempo e clima**. 9. ed. Porto Alegre: Bookman, 2009. 49 p. E-book. Acesso em: 13 mai. 2025. ISBN 9788565837392. Disponível em: <https://app.minhabiblioteca.com.br/reader/books/9788565837392/>. Citado na página 20.
- BEKKAR, Abdellatif *et al.* Transformer-based model for multi-horizon forecasting ozone in marrakech city, morocco. In: **2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)**. [S.l.: s.n.], 2023. p. 1–8. Citado 4 vezes nas páginas 33, 34, 35 e 36.
- BENCHRIFA, Mohammed *et al.* Study of the evolution of ozone concentration as a function of nitrogen monoxide and dioxide in ambient air. In: **Integrated Solutions for Smart and Sustainable Environmental Conservation**. [S.l.]: Springer Nature, 2024. p. 65–78. Citado na página 21.

BRASIL. **Lei nº 14.850, de 2 de maio de 2024. Institui a Política Nacional de Qualidade do Ar.** 2024. Diário Oficial da União: seção 1, Brasília, DF, 3 maio 2024. Acesso em: 14 abr. 2025. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/L14850.htm. Citado na página 15.

BUUREN, Stef van; GROOTHUIS-OUDSHOORN, Karin. mice: Multivariate imputation by chained equations in r. **Journal of Statistical Software**, v. 45, n. 3, p. 1–67, 2011. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v045i03>. Citado na página 28.

CHANDRAPPA, Ramesha; KULSHRESTHA, Umesh Chandra. Major issues of air pollution. In: _____. **Sustainable Air Pollution Management: Theory and Practice**. Cham: Springer International Publishing, 2016. p. 1–48. ISBN 978-3-319-21596-9. Disponível em: https://doi.org/10.1007/978-3-319-21596-9_1. Citado na página 20.

CHAURASIA, Pallavi; MISRA, Prakhar. Tropospheric ozone estimation using machine learning of remote sensing data: a case study of delhi ncr. In: . [S.l.: s.n.], 2024. p. 1–4. Citado 3 vezes nas páginas 33, 34 e 36.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939785>. Citado na página 23.

CHU, Wanghui *et al.* Research on ozone formation sensitivity based on observational methods: Development history, methodology, and application and prospects in china. **Journal of Environmental Sciences**, v. 138, p. 543–560, 2024. ISSN 1001-0742. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1001074223000980>. Citado na página 15.

CORREA, Marcele de Jesus *et al.* Filling of faults in climatological air temperature series in brazilian state capitals from 1980 to 2017. **Revista Brasileira de Climatologia**, v. 29, p. 251–272, set. 2021. Disponível em: <https://ojs.ufgd.edu.br/rbclima/article/view/15176>. Citado na página 14.

DOMÍNGUEZ-GARCÍA, Rodrigo; ARELLANO-VÁZQUEZ, Magali. Evaluation of machine learning models for ozone concentration forecasting in the metropolitan valley of mexico. **Applied Sciences**, v. 14, n. 4, 2024. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/14/4/1408>. Citado na página 15.

DUQUE, Laura; POELMAN, Erik H.; STEFFAN-DEWENTER, Ingolf. Effects of ozone stress on flowering phenology, plant-pollinator interactions and plant reproductive success. **Environmental Pollution**, v. 272, p. 115953, 2021. ISSN 0269-7491. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0269749120366422>. Citado na página 14.

EDEN, Amnon H. Three paradigms of computer science. **Minds and Machines**, Springer, v. 17, n. 2, p. 135–167, 2007. Citado na página 17.

FAIZIN, Rahmat Nur; RIASETIAWAN, Mardhani; ASHARI, Ahmad. A review of missing sensor data imputation methods. In: **2019 5th International Conference on Science and Technology (ICST)**. [S.l.: s.n.], 2019. v. 1, p. 1–6. Citado na página 27.

GAGLIARDI, Roberta Valentina; ANDENNA, Claudio. A machine learning approach to investigate the surface ozone behavior. **Atmosphere**, v. 11, n. 11, 2020. ISSN 2073-4433. Disponível em: <https://www.mdpi.com/2073-4433/11/11/1173>. Citado 3 vezes nas páginas 33, 34 e 36.

GAO, Hui *et al.* A systematic review and meta-analysis of short-term ambient ozone exposure and copd hospitalizations. **International Journal of Environmental Research and Public Health**, v. 17, n. 6, 2020. ISSN 1660-4601. Disponível em: <https://www.mdpi.com/1660-4601/17/6/2130>. Citado na página 14.

GIL, Antonio Carlos. Métodos e técnicas de pesquisa social. In: _____. 7. ed. Rio de Janeiro: Atlas, 2019. p. 55–57. ISBN 9788597020991. E-book. Disponível em: <https://app.minhabiblioteca.com.br/reader/books/9788597020991/>. Citado na página 18.

GIL, Antonio Carlos. Como elaborar projetos de pesquisa. In: _____. 7. ed. Rio de Janeiro: Atlas, 2022. p. 41. ISBN 9786559771653. E-book. Disponível em: <https://app.minhabiblioteca.com.br/reader/books/9786559771653/>. Citado na página 17.

GRYECH, Ihsane *et al.* Applications of machine learning and internet of things for outdoor air pollution monitoring and prediction: A systematic literature review. **Engineering Applications of Artificial Intelligence**, v. 137, p. 109182, 2024. ISSN 0952-1976. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095219762401340X>. Citado 2 vezes nas páginas 29 e 30.

JACOB, Daniel J.; WINNER, Darrell A. Effect of climate change on air quality. **Atmospheric Environment**, v. 43, n. 1, p. 51–63, 2009. ISSN 1352-2310. Atmospheric Environment - Fifty Years of Endeavour. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1352231008008571>. Citado na página 22.

JAFARZADEH, Hamid *et al.* Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and polsar data: A comparative evaluation. **Remote Sensing**, v. 13, n. 21, 2021. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/13/21/4405>. Citado 2 vezes nas páginas 23 e 24.

JAILANI, N Mohamed Abdul Kader; MARA, Geeta C. Ozone concentration forecasting: Assessing the efficacy of mlp, dnn, and xgboost in environmental bench-aq dataset. In: **2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)**. [S.l.: s.n.], 2024. v. 1, p. 1–5. Citado 3 vezes nas páginas 33, 34 e 36.

Jl, Wenquan *et al.* Based on the rfe-lstm ozone prediction research. In: **2024 9th International Conference on Electronic Technology and Information Science (ICETIS)**. [S.l.: s.n.], 2024. p. 773–776. Citado 4 vezes nas páginas 33, 34, 35 e 36.

JUAREZ, Eliana Kai; PETERSEN, Mark R. A comparison of machine learning methods to forecast tropospheric ozone levels in delhi. **Atmosphere**, v. 13, n. 1, 2022. ISSN 2073-4433. Disponível em: <https://www.mdpi.com/2073-4433/13/1/46>. Citado 4 vezes nas páginas 33, 34, 35 e 36.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. Fundamentos de metodologia científica. In: _____. 9. ed. Rio de Janeiro: Atlas, 2021. ISBN 9788597026580. E-book. Disponível em: <https://app.minhabiblioteca.com.br/reader/books/9788597026580/>. Citado na página 18.

LATIF, Sarmad Dashti *et al.* Ozone concentration forecasting utilizing leveraging of regression machine learnings: A case study at klang valley, malaysia. **Results in Engineering**, v. 21, p. 101872, 2024. ISSN 2590-1230. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2590123024001257>. Citado na página 15.

LI, Jin *et al.* Quantifying the contributions of meteorology, emissions, and transport to ground-level ozone in the pearl river delta, china. **Science of The Total Environment**, v. 932, p. 173011, 2024. ISSN 0048-9697. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0048969724031589>. Citado na página 22.

LORENA, Ana *et al.* **Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2a edição)**. [S.l.: s.n.], 2021. ISBN 9788521637493. Citado 2 vezes nas páginas 24 e 25.

LU, Xiao; ZHANG, Lin; SHEN, Lu. Meteorology and climate influences on tropospheric ozone: a review of natural sources, chemistry, and transport patterns. **Current Pollution Reports**, v. 5, n. 4, p. 238–260, 2019. ISSN 2198-6592. Disponível em: <https://doi.org/10.1007/s40726-019-00118-3>. Citado 2 vezes nas páginas 22 e 23.

MANAHAN, Stanley E. **Química ambiental**. 9. ed. Porto Alegre: Bookman, 2013. 278–317, 401–426 p. ISBN 9788565837064. Citado na página 14.

MANOJ, P.; MANJUNATH, K. C.; KOTAGI, Punith B. Machine learning models for proactive road safety: Evaluating regression and ensemble techniques based on road geometry. **Engineering, Technology and Applied Science Research**, v. 15, n. 4, p. 25951–25958, Aug. 2025. Disponível em: <https://etasr.com/index.php/ETASR/article/view/11569>. Citado na página 24.

MARSHALL, Clare P.; FAIRBRIDGE, Rhodes W. **Ozone**. 1. ed. Dordrecht: Springer, 1999. 2, 467 p. (Encyclopedia of Earth Sciences Series). ISBN 978-0-412-75500-2. Disponível em: <https://doi.org/10.1007/1-4020-4496-8>. Citado 2 vezes nas páginas 20 e 21.

MIDDYA, Asif Iqbal; ROY, Sarbani. Predicting missing value for ozone data based on optimal site-selection strategy. In: **2022 IEEE Calcutta Conference (CALCON)**. [S.l.: s.n.], 2022. p. 288–292. Citado 4 vezes nas páginas 33, 34, 35 e 36.

NAZ, Fareena *et al.* Comparative analysis of deep learning and statistical models for air pollutants prediction in urban areas. **IEEE Access**, v. 11, p. 64016–64025, 2023. Citado na página 30.

NETTO, Fabiola Dayane *et al.* Formação de ozônio troposférico: Uma revisão da literatura. **Blucher Engineering Proceedings**, v. 4, n. 1, p. 411 – 428, 2017. ISSN 2357-7592. Disponível em: www.proceedings.blucher.com.br/article-details/formao-de-oznio-troposfrico-uma-revisao-da-literatura-26558. Citado na página 14.

OUFDOU, Halima *et al.* Forecasting daily of surface ozone concentration in the grand casablanca region using parametric and nonparametric statistical models. **Atmosphere**, v. 12, n. 6, 2021. ISSN 2073-4433. Disponível em: <https://www.mdpi.com/2073-4433/12/6/666>. Citado 4 vezes nas páginas 33, 34, 35 e 36.

PAGE, Matthew J *et al.* The prisma 2020 statement: an updated guideline for reporting systematic reviews. **BMJ**, BMJ Publishing Group Ltd, v. 372, 2021. Disponível em: <https://www.bmj.com/content/372/bmj.n71>. Citado 2 vezes nas páginas 8 e 31.

PAN, Qilong; HARROU, Fouzi; SUN, Ying. A comparison of machine learning methods for ozone pollution prediction. **Journal of Big Data**, v. 10, n. 1, p. 63, 2023. ISSN 2196-1115. Disponível em: <https://doi.org/10.1186/s40537-023-00748-x>. Citado na página 23.

PEDERSEN, Alma B. *et al.* Missing data and multiple imputation in clinical epidemiological research. **Clinical Epidemiology**, v. 9, p. 157 – 166, 2017. Cited by: 676; All Open Access, Gold Open Access, Green Open Access. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85015735233&doi=10.2147%2fCLEP.S129785&partnerID=40&md5=d9a527caeb318e269850a5efcda2eb6d>. Citado na página 26.

PEDREGOSA, Fabian *et al.* Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, n. 85, p. 2825–2830, 2011. Disponível em: <http://jmlr.org/papers/v12/pedregosa11a.html>. Citado na página 18.

PEREIRA, Ana Rute Batista *et al.* Preenchimento de falhas em séries de dados meteorológicos de estações automáticas. **Revista Brasileira de Climatologia**, v. 35, n. 20, p. 22–44, jul. 2024. Disponível em: <https://ojs.ufgd.edu.br/rbclima/article/view/17599>. Citado na página 14.

PETERSEN, Kai; VAKKALANKA, Sairam; KUZNIARZ, Ludwik. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and Software Technology**, v. 64, p. 1–18, 2015. ISSN 0950-5849. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950584915000646>. Citado na página 29.

PETTICREW, Mark; ROBERTS, Helen. **Systematic Reviews in the Social Sciences: A Practical Guide**. [S.l.]: Blackwell Publishing, 2006. Online ISBN: 9780470754887. ISBN 9781405121101. Citado na página 29.

QU, K. *et al.* A comparative study to reveal the influence of typhoons on the transport, production and accumulation of O₃ in the pearl river delta, china. **Atmospheric Chemistry and Physics**, v. 21, n. 15, p. 11593–11612, 2021. Disponível em: <https://acp.copernicus.org/articles/21/11593/2021/>. Citado na página 21.

RÄSS, Stephan; LEUENBERGER, Markus C. Analysis and prediction of atmospheric ozone concentrations using machine learning. **Frontiers in Big Data**, v. 7, p. 1469809, jan. 2025. Citado 2 vezes nas páginas 15 e 23.

SANTANA, Gabriela Pereira; MOREIRA, Virnei Silva; ARMANI, Fernando Augusto Silveira. Poluição atmosférica e doenças respiratórias: um estudo de caso em paranaguá, paraná. **Revista Ibero-Americana de Ciências Ambientais**, v. 11, n. 5, p. 352–361, jun. 2020. Disponível em: <https://sustenere.inf.br/index.php/rica/article/view/CBPC2179-6858.2020.005.0033>. Citado 4 vezes nas páginas 15, 16, 43 e 64.

SEINFELD, John H.; PANDIS, Spyros N. **Atmospheric Chemistry and Physics: From Air Pollution to Climate Change**. 3rd. ed. Hoboken, NJ: John Wiley & Sons, 2016. 175–259 p. ISBN 978-1-118-94740-1. Citado na página 21.

SHER, Eran. Environmental aspects of air pollution. In: SHER, Eran (Ed.). **Handbook of Air Pollution from Internal Combustion Engines**. San Diego: Academic Press, 1998. p. 27–41. ISBN 9780126398557. Acesso em: 18 abr. 2025. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780126398557500417>. Citado na página 14.

SOBRINHO, Hamilton Dias Ribeiro. Análise geográfica dos destelhamentos na área urbana de ponta grossa (pr) entre 2005 e 2015 com base nos dados do corpo de bombeiros. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 9, n. 24, p. 138–168, 2018. ISSN 2175-1846. Disponível em: <https://revistas.utfpr.edu.br/recit/article/view/e-8387>. Citado na página 38.

SOKOLOVA, Olga; YURGENSON, Anastasia; SHAKHOV, Vladimir. Development of air quality monitoring systems: Balancing infrastructure investment and user satisfaction policies. **Sensors**, v. 25, n. 3, 2025. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/25/3/875>. Citado na página 15.

STEKHOFEN, Daniel J.; BÜHLMANN, Peter. Missforest—non-parametric missing value imputation for mixed-type data. **Bioinformatics**, v. 28, n. 1, p. 112–118, 10 2011. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btr597>. Citado na página 28.

TORRES, Fillipe Tamiozzo Pereira; MARTINS, Luiz Alberto. Fatores que influenciam na concentração do material particulado inalável na cidade de juiz de fora (mg). **Caminhos de Geografia**, v. 6, n. 16, p. 23–39, 2005. Citado na página 43.

TRASHKO, Mariia D. *et al.* Development of an environmental monitoring system for predicting and preventing photochemical smog. In: **2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)**. [S.l.: s.n.], 2022. p. 1624–1627. Citado na página 14.

VENUKA, T *et al.* Prediction of ground-level ozone in bengaluru using machine learning techniques. In: **2024 IEEE 9th International Conference for Convergence in Technology (I2CT)**. [S.l.: s.n.], 2024. p. 1–8. Citado 4 vezes nas páginas 15, 33, 34 e 36.

VISHWAS, B. V.; PATEL, Ashish. Time-series characteristics. In: _____. **Hands-on Time Series Analysis with Python: From Basics to Bleeding Edge Techniques**. Berkeley, CA: Apress, 2020. p. 1–21. ISBN 978-1-4842-5992-4. Disponível em: https://doi.org/10.1007/978-1-4842-5992-4_1. Citado 2 vezes nas páginas 23 e 27.

VORMITTAG, Evangelina da Motta P. A. de Araújo *et al.* Análise do monitoramento da qualidade do ar no brasil. **Estudos Avançados**, Instituto de Estudos Avançados da Universidade de São Paulo, v. 35, n. 102, p. 7–30, May 2021. ISSN 0103-4014. Disponível em: <https://doi.org/10.1590/s0103-4014.2021.35102.002>. Citado 2 vezes nas páginas 14 e 15.

WAZLAWICK, Raul Sidnei. Metodologia de pesquisa para ciência da computação. In: _____. 3. ed. Rio de Janeiro: GEN LTC, 2020. ISBN 9788595157712. E-book. Disponível em: <https://app.minhabiblioteca.com.br/reader/books/9788595157712/>. Citado 2 vezes nas páginas 17 e 19.

WEATHERBIT. **Hourly Historical Weather API**. 2025. <https://www.weatherbit.io/api/historical-weather-hourly>. Citado na página 56.

WEIß, Christian H. Time series modelling. **Entropy**, v. 23, n. 9, 2021. ISSN 1099-4300. Disponível em: <https://www.mdpi.com/1099-4300/23/9/1163>. Citado na página 23.

YAFUZZ, Ayman *et al.* Ozone concentration forecasting based on artificial intelligence techniques: A systematic review. **Water, Air, & Soil Pollution**, v. 232, n. 2, p. 79, 2021. Disponível em: <https://doi.org/10.1007/s11270-021-04989-5>. Citado na página 29.

ZHANG, Xiangxue *et al.* Short-term health impacts related to ozone in china before and after implementation of policy measures: A systematic review and meta-analysis. **Science of The Total Environment**, v. 847, p. 157588, 2022. ISSN 0048-9697. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0048969722046861>. Citado na página 14.

APÊNDICE A – ILUSTRAÇÕES E CARACTERÍSTICAS TÉCNICAS DE SENSORES PORTÁTEIS PARA DETECÇÃO DE OZÔNIO

Figura A.1 – Sensor portátil Aeroqual S500 Starter Kit com capacidade de medição de ozônio.



Fonte: Imagem obtida do site do fabricante Aeroqual, disponível em <https://www.cleanair.com/product/aeroqual-s500-starter-kit/>, com acesso em 30 abr. 2025.

Figura A.2 – Sensor fixo Serinus[®] 10, voltado à análise contínua de ozônio em ambientes.



Fonte: Imagem obtida do site do fabricante Acoem, disponível em <https://www.acoem.com/brasil/pt-br/products/ar-ambiente/analisadores-de-gases-de-qualidade-do-ar/serinus-10/>, com acesso em 30 abr. 2025.

Figura A.3 – Sensor pessoal POM[™], desenvolvido para monitoramento individual de ozônio.



Fonte: Imagem obtida do site da 2B Technologies, disponível em <https://2btech.io/items/ambient-ozone-monitors/personal-ozone-monitor-pom/>, com acesso em 30 abr. 2025.

Tabela A.1 – Especificações gerais de sensores portáteis para monitoramento de ozônio.

Modelo	Parâmetros Monitorados	Faixa de O ₃ (ppm)	Limite Inferior	Tempo de Amostragem	Alimentação	Transferência de Dados	Preço (USD)	Certificações
Aeroqual S500 Starter Kit	PM ₁₀ , PM _{2.5} , O ₃ , NO ₂ , temperatura e umidade relativa	0–0,15	-	Não informado	Bateria de lítio recarregável	Cabo USB	US\$ 3.690,85	Não informado
Aeroqual S500 Handheld Pro Kit	PM _{2.5} , PM ₁₀ , NO ₂ , O ₃ , CO, COV, temperatura e umidade relativa	0–0,15	-	Não informado	Bateria de lítio recarregável	Cabo USB	US\$ 6.275,90	Não informado
Serinus® 10	Ozônio (O ₃)	0–20	0,5 ppb	Não informado	Bateria de lítio recarregável	Pen drive USB removível	Não informado	EPA (EUA), EN TÜV, AS 3580.6.1
Personal Ozone Monitor™ (POM™)	Ozônio (O ₃)	0–10	0,1 ppb	1 min, 5 min ou 1 h (ajustável)	Bateria de lítio recarregável	Cabo USB	~US\$ 4.430,00	FEM (EPA/EUA)

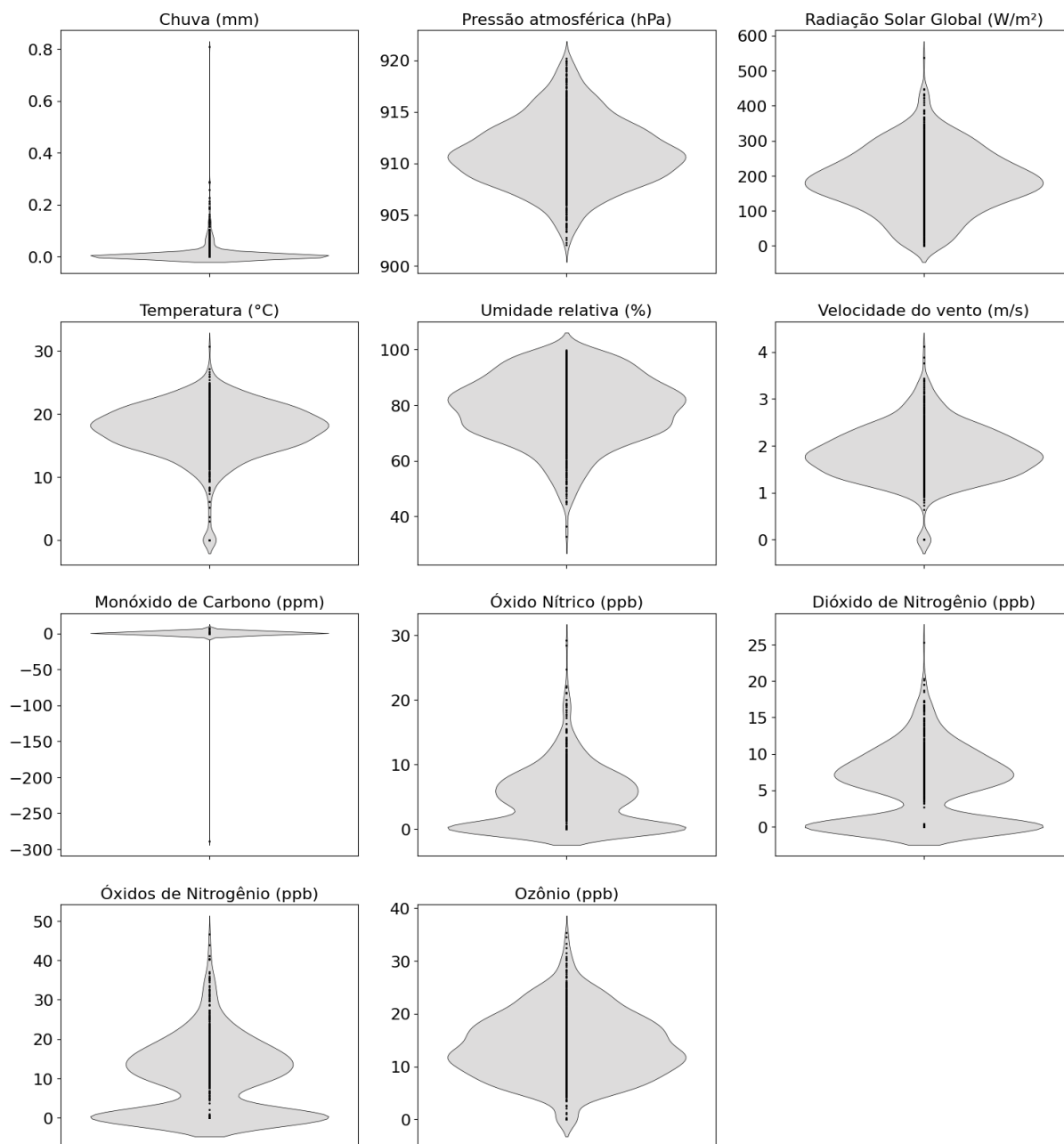
As informações foram compiladas com base em especificações técnicas dos fabricantes Aeroqual, Acoem e 2B Technologies. O valor do equipamento POM foi estimado em Dólar (US\$) a partir de oferta pública em site internacional de comércio eletrônico, em virtude da ausência de preço oficial no site da fabricante

Fonte: Elaborada pela autora (2026).

Esta página foi intencionalmente deixada em branco.

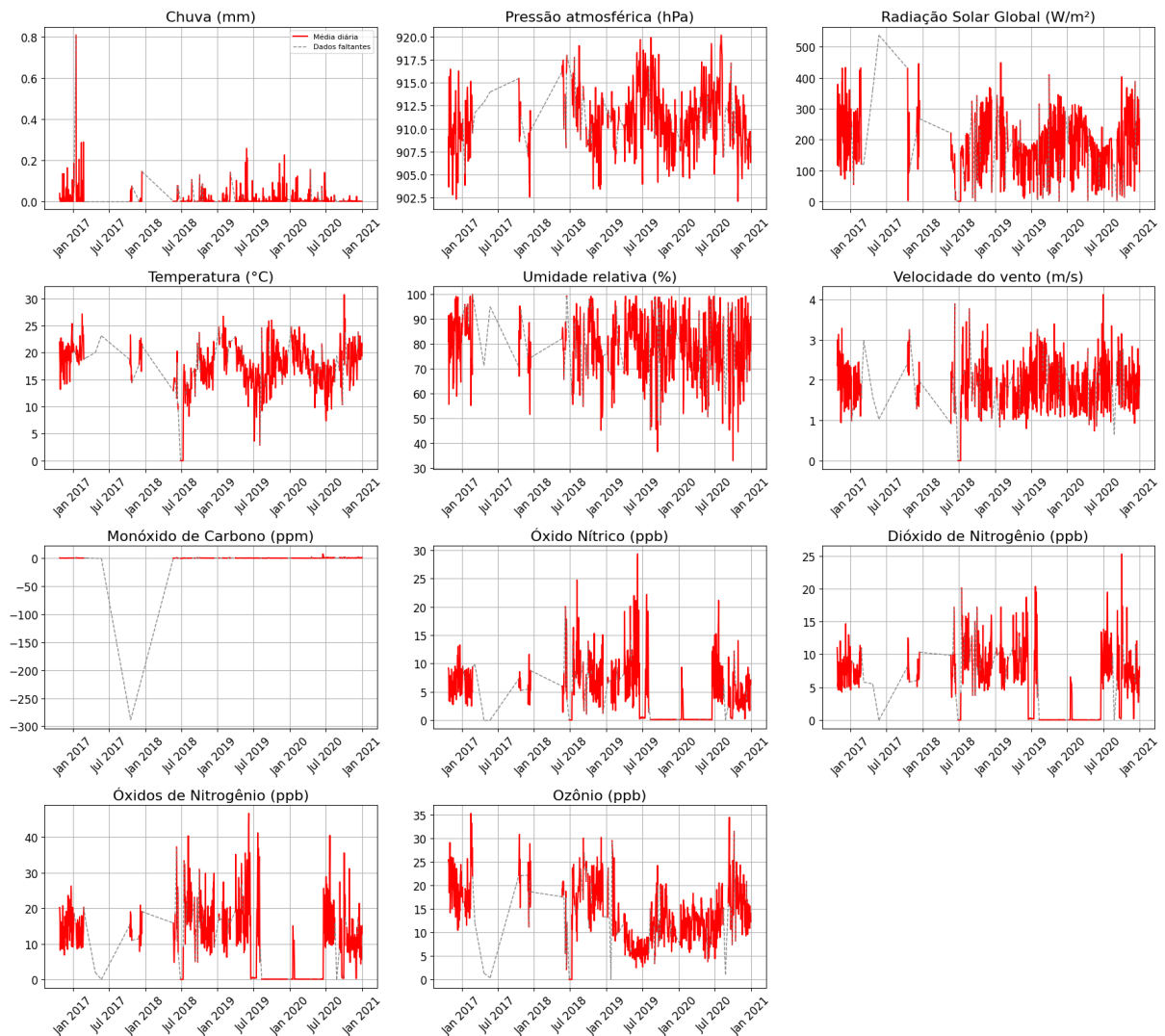
APÊNDICE B – DISTRIBUIÇÕES E RELAÇÕES ENTRE VARIÁVEIS

Figura B.1 – Distribuições de variáveis meteorológicas e químicas atmosféricas. Destaque para os padrões de ozônio, que revelam que o ozônio tem maiores concentrações entre 10-20 ppb; temperatura, com ampla variabilidade (0-25 °C) e mediana próxima a 17 °C; umidade relativa com concentrações entre 70-90%.



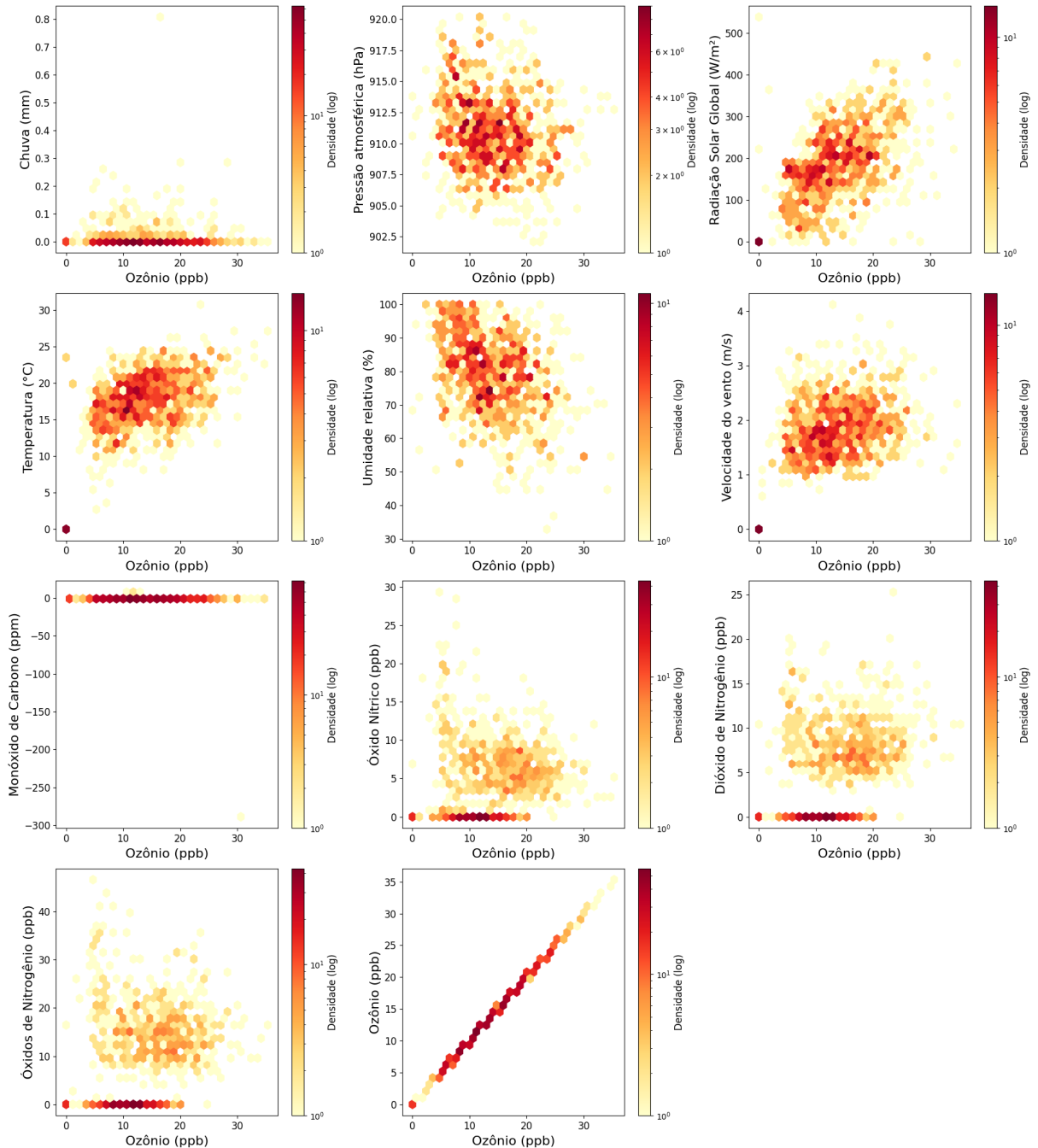
Fonte: Elaborada pela autora (2026).

Figura B.2 – Séries temporais de variáveis meteorológicas e químicas atmosféricas. O ozônio varia sazonalmente (5-35 ppb) acompanhando temperatura e radiação solar, enquanto os óxidos de nitrogênio (NO, NO₂, NO_x) mostram picos irregulares típicos de emissões veiculares urbanas; a umidade se comporta inversamente ao ozônio (40-100%), e eventos de chuva esporádicos reduzem as concentrações de poluentes. Períodos sem chuva coincidem com acúmulo de poluentes, especialmente no inverno quando baixas temperaturas e alta umidade favorecem a concentração de NO_x próximo ao solo.



Fonte: Elaborada pela autora (2026).

Figura B.3 – Diagramas de dispersão entre ozônio e variáveis meteorológicas e químicas atmosféricas. O ozônio mostra correlação positiva com temperatura (concentrações máximas de O_3 ocorrem entre 15-25 °C) e radiação solar (maior densidade de pontos em 200-300 W/m^2), enquanto apresenta relação inversa com umidade relativa (maiores concentrações de O_3 em umidades de 60-80%). A relação com óxidos de nitrogênio revela o comportamento não-linear da química atmosférica.



Fonte: Elaborada pela autora (2026).

UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
BIBLIOTECA UNIVERSITÁRIA
REPOSITÓRIO INSTITUCIONAL

CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT

ATESTADO DE VERSÃO FINAL

Eu, Yuri Kaszubowski Lopes, professor do curso de Programa de Pós-Graduação em Computação Aplicada, declaro que esta é a versão final aprovada pela comissão julgadora da dissertação/tese intitulada: **“Previsão Inteligente de Ozônio Troposférico em Estações de Qualidade do Ar com Modelos Preditivos Resilientes a Dados Faltantes”** de autoria do(a) acadêmico Miraceli Bonjardim Waldemar.

Joinville, 16 de Abril de 2026.

Assinatura digital do orientador:

Documento assinado digitalmente
 **YURI KASZUBOWSKI LOPES**
Data: 18/04/2026 14:00:44-0300
Verifique em <https://validar.iti.gov.br>

Yuri Kaszubowski Lopes