MODELAGEM DA CONCENTRAÇÃO DE OZÔNIO TROPOSFÉRICO PARA A REGIÃO NORDESTE COM USO DE APRENDIZADO DE MÁQUINA

Antonio Campos Neto[1]
Humberto Cordeiro da Silva Nelo[2]

RESUMO
Este estudo teve como objetivo desenvolver um modelo de aprendizado de máquina para estimar a concentração de ozônio troposférico (O_3) na Região Nordeste do Brasil, a partir de dados do Copernicus Atmosphere Monitoring Service (CAMS-Reanalysis) para o período de 2003 a 2018. Foram utilizadas variáveis preditoras relacionadas a poluentes atmosféricos e parâmetros meteorológicos, e o algoritmo Random Forest foi empregado para a modelagem, após pré-processamento e tratamento dos dados. O desempenho do modelo foi avaliado por métricas estatísticas, apresentando Erro Percentual Médio Absoluto (MAPE) de 9% e correlação de Spearman de 0,897, indicando boa precisão e forte relação entre valores observados e estimados. Os resultados demonstram a capacidade do modelo em prever concentrações de O_3 e apontam seu potencial como ferramenta de apoio à gestão da qualidade do ar e a estudos de saúde, com possibilidade de aplicação em políticas ambientais e climáticas.

PALAVRAS-CHAVE: Machine Learning; Poluentes; Saúde; Predição.

MODELING TROPOSPHERIC OZONE CONCENTRATION FOR BRAZIL'S NORTHEAST REGION USING MACHINE LEARNING

ABSTRACT
This study aimed to develop a machine learning model to estimate tropospheric ozone (O_3) concentration in the Northeast Region of Brazil, using data from the Copernicus Atmosphere Monitoring Service (CAMS-Reanalysis) for the period 2003 to 2018. Predictor variables related to atmospheric conditions and meteorological configurations were used, and the Random Forest algorithm was employed for modeling after data preprocessing and treatment. The model's performance was evaluated using statistical analysis, showing a Mean Absolute Percent Error (MAPE) of 9% and a Spearman brightness of 0.897, indicating good scores and a strong correlation between collected and estimated values. The results demonstrate the model's capability in O_3 prevention and highlight its potential as a tool to support air quality management and health studies, with possible applications in environmental and climate policies.

KEYWORDS: Pollutants; Health; Prediction.

[1] Licenciando em Geografia, Universidade Federal do Rio Grande do Norte - UFRN, Natal, RN, Email: antoniocamposneto9@gmail.com
[2] Mestrando em Geografia, Universidade Federal do Rio Grande do Norte - UFRN, Natal, RN, Email: humberto.cordeiro.701@ufrn.edu.br

MODELADO DE LA CONCENTRACIÓN DE OZONO TROPOSFÉRICO PARA LA REGIÓN NORESTE DE BRASIL MEDIANTE APRENDIZAJE AUTOMÁTICO

RESUMEN
Este estudio tuvo como objetivo desarrollar un modelo de aprendizaje automático para estimar la concentración de ozono troposférico (O_3) en la Región Noreste de Brasil, utilizando datos del Servicio de Monitoreo Atmosférico de Copernicus (CAMS-Reanálisis) para el período 2003-2018. Se emplearon variables predictoras relacionadas con contaminantes atmosféricos y parámetros meteorológicos, y se utilizó el algoritmo Random Forest para el modelado tras el preprocesamiento y tratamiento de los datos. El desempeño del modelo se evaluó mediante métricas estadísticas, presentando un Error Porcentual Absoluto Medio (MAPE) del 9% y una correlación de Spearman de 0,897, lo que indica una buena precisión y una fuerte relación entre los valores observados y estimados. Los resultados demuestran la capacidad del modelo para predecir concentraciones de O_3 y resaltan su potencial como herramienta para apoyar la gestión de la calidad del aire y los estudios de salud, con posibles aplicaciones en políticas ambientales y climáticas.

PALABRAS CLAVE: Aprendizaje automático; Contaminantes; Salud; Predicción.

1.INTRODUÇÃO

O ozônio é um poluente atmosférico secundário, que quando na baixa atmosfera, possui impactos significativos no clima, na saúde humana e nos ecossistemas, também sendo um gás de efeito estufa. Lu, Zhang e Shen (2019) apontam como as variações nas condições meteorológicas e o hábito humano afetam diretamente a formação e a concentração deste gás, por meio das mudanças climáticas, alterações nas emissões de precursores naturais, na química, na deposição e nos padrões de transporte. Na saúde humana, Huang *et. al.* (2019) e Zhang (2019) discorrem como a exposição prolongada ao ozônio está associada a um aumento do risco de mortalidade cardiovascular e respiratória.

Trabalhos como o de Alonso, Gouveia e Santos (2025) identificaram, especificamente para Portugal, que as variáveis mais significantes na concentração do ozônio troposférico foram a amplitude térmica, temperatura máxima, altura da camada limite (*boundary layer height*), dióxido de nitrogênio (NO_2), radiação e recorte temporal. Para este trabalho, o conjunto de dados utilizado no modelo foi obtido do *Copernicus Atmosphere Monitoring Service* (CAMS)-*Reanalysis* (do Centro Europeu de Previsões Meteorológicas de Médio Prazo-ECMWF), com 16 anos de dados diários (2003-2018) para todos os municípios da Região Nordeste do Brasil.

Se pretende modelar a variação do Ozônio (O_3) troposférico para o Nordeste, gerando um modelo de predição com aprendizado de máquina. Estudar o comportamento do O_3 pode ser útil para entender se há episódios de alta concentração deste poluente, suas principais causas e onde ocorrem, ademais, sua utilidade é relevante para complementar banco de dados que disponham de variáveis que possuam relação com o Ozônio, mas não ele próprio, pois sua aferição *in locu* é rara, mesmo em grandes cidades, onde os efeitos danosos à saúde humana são mais proeminentes (Castelhano, Requia, 2024). Seus níveis e tendências são influenciados tanto por processos naturais quanto por atividades humanas, e compreender sua evolução é crucial para a gestão da qualidade do ar e para as políticas climáticas e de saúde.

2. METODOLOGIA

A abordagem metodológica deste estudo combinou uma revisão bibliográfica com a aplicação de técnicas de modelagem preditiva para estimar a concentração de ozônio troposférico na Região Nordeste do Brasil, com a etapa de modelagem sendo integralmente desenvolvida no software RStudio (RStudio Team, 2025), utilizando o pacote Ranger (Wright, 2023) que possibilita o emprego do método *Random Forest* para aprendizado de máquina. Tal método consiste na construção de múltiplas "árvores" de decisão utilizando amostras aleatórias de observações para cada árvore e, em cada ponto de divisão, amostras aleatórias de preditores. A "floresta" resultante dessas árvores fornece valores ajustados que são mais precisos do que os de qualquer árvore individual (Fernández-de-las-Peñas *et al*.,2015), permitindo desta forma que se possa indiciar o valor de um dado específico, com base em outras informações, com certa acurácia.

O conjunto de dados utilizado no modelo foi composto pela variável concentração de ozônio em partes por bilhão (PPB), e por um grupo de variáveis preditoras: Monóxido de Carbono e dióxido de Nitrogênio (PPB), Material Particulado 2.5 e Dióxido de Enxofre em microgramas por metro cúbico (μg/m³), precipitação em milímetros por dia (mm/dia), temperatura média em graus Celsius (°C), Velocidade do Vento em metros por segundo (m/s), mês e dia da semana, pois se observou por trabalhos como o de (Tiwari *et. al*., 2008; Geng *et. al.*, 2015) que a variação sazonal/mensal do ozônio é mais expressiva que a diária ou anual. O "mês" e "dia da semana" são variáveis extraídas de outra, que continha a data formatada em

dia/mês/ano, e se optou por usá-las por se constatar um aumento na significância do modelo ao fazê-lo, em especial a "dia da semana", pois ela carrega em si a diferença entre domingo, segunda, terça e assim por diante, e com isso, as variações nos padrões do transporte e atividade urbana, significantes para a formação do Ozônio troposférico, informação que se obtida por dados diretos/primários, seria de grande dificuldade para abranger toda a área de estudo.

A variável a ser modelada foi o Ozônio, pois, dentre as presentes no banco de dados, foi a que apresentou sua variação mais bem explicada pelas demais, observado pelo emprego de uma regressão linear múltipla, apontando 60% do comportamento do O_3 contemplado (r quadrado ajustado = 0,60). A regressão linear múltipla é uma técnica estatística que usa várias variáveis explicativas para prever o resultado de uma variável resposta (neste caso, o Ozônio), ajudando a quantificar a força da relação entre cada variável independente e a variável dependente (Montgomery, Peck, Vining, 2012).

Para avaliação de valores atípicos (*outliers*), tratamento e análise da consistência dos dados, foi realizada uma etapa de pré-processamento. A saber, a regressão linear múltipla teve um incremento no r^2 ajustado de 0,54 para 0,60 ao se omitir os valores *outliers* e os dados faltantes da análise.

A escolha do algoritmo de aprendizado de máquina *Random Forest* se deu pela capacidade de modelar relações não-lineares complexas entre as variáveis e por sua robustez à multicolinearidade (SILVA *et al*., 2023), com o modelo treinado para identificar os padrões existentes entre as variáveis preditoras e a concentração de ozônio. O banco de dados inicialmente possuia mais de 13 milhões de linhas, e após o tratamento para exclusão de *outliers* e dados faltantes, o mesmo continha 11 milhões. Além disso, por algumas variáveis precisarem ser tratadas como fatores, não meramente valores numéricos, modelar se prova uma tarefa de grande demanda computacional: o uso do *Random Forest* requer a separação do banco de dados em duas partes, uma para treinar o modelo e uma para testá-lo. Tipicamente, esta divisão é feita na proporção 80/20 (80% dos dados geram o modelo, o treinam, e os demais 20% são usados para verificar a acurácia), como retratam Nguyen *et.al*.; Vrigazova *et. al*. (2021), porém, os computadores disponíveis para a execução desta abordagem não suportam uma divisão para treino maior que 5/95.

Um conjunto menor para testes pode se apresentar enquanto um problema, pois a modelagem pode ser comprometida por falta de dados, mas, neste caso, o banco original é grande, e apenas 5% dele já representam mais de 500.000 dados, uma quantidade considerável, para que o recorte metodológico aplicado não desqualificasse o rigor da análise. Também se verificou que, o incremento na acurácia do modelo era pouca quando se aumentava gradualmente o conjunto teste (de 3% para 4%, e de 4% para os 5%, por exemplo), logo, o volume alto de dados que os cinco por cento representam não ocasionam comprometimento neste sentido, diante os resultados obtidos e devidamente avaliados.

Para avaliar o desempenho e a capacidade de generalização do modelo, se valeu da validação por métricas complementares: utilizou-se a Correlação de Spearman (ρ) para mensurar a força e a direção da relação monotônica entre predições e observações, escolhida por sua robustez a dados com distribuição não normal. Para quantificar a magnitude do erro na mesma unidade variável alvo, foram empregados o Erro Médio Absoluto (MAE) e a Raiz do Erro Quadrático Médio (RMSE), que se distingue por penalizar com maior rigor os erros de grande magnitude. Adicionalmente, o Erro Percentual Médio Absoluto (MAPE) foi calculado para fornecer uma avaliação relativa da acurácia do modelo, independente da escala dos dados (Viscondi, 2022).

3. RESULTADOS

O modelo de predição para o Ozônio troposférico se mostrou robusto ao exibir as seguintes métricas (Quadro 1):

Quadro 01 – Resultados dos Métodos de Avaliação do Modelo

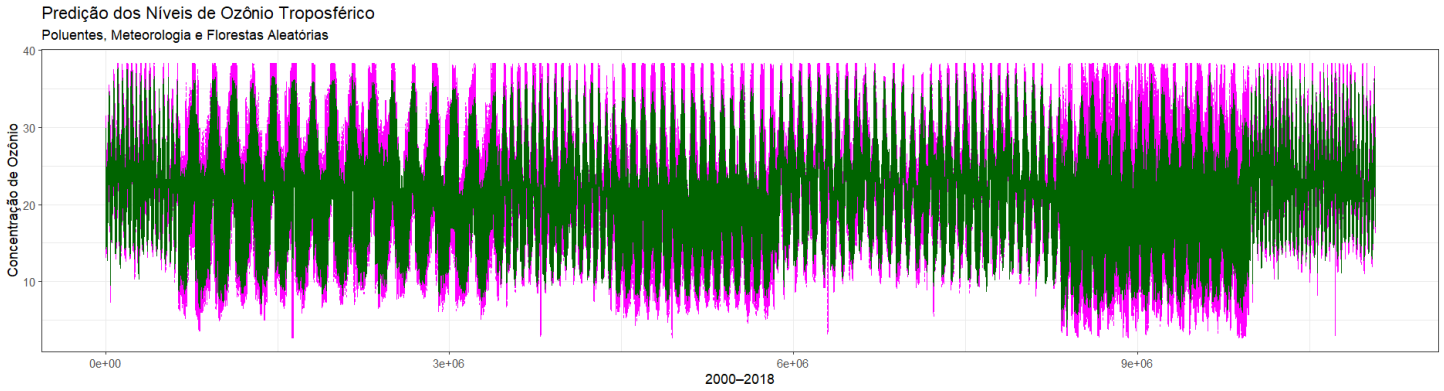
Método	Resultado
RMSE	2,4 ppb
MEA	1,8
MAPE	9%
SPEARMAN	0.8973599

Fonte: Autoria Própria (2025).

A interpretação da tabela 1 precisa ser contextualizada, principalmente por se tratar da avaliação de uma modelagem de uma variável caótica: a diferença entre o RMSE (2,4 ppb) e o MAE (1,8 ppb) sugere a presença de alguns erros de maior magnitude, indicando que nosso modelo apresenta boa precisão geral, mas com alguns desvios pontuais relevantes. O RMSE aponta um erro médio em cada amostra de 2,4 ppb na concentração do Ozônio, que, quando se leva em conta que seus valores reais variam de 0-110.7 ppb, é um erro médio baixo. O MAPE aponta que a diferença entre a concentração de O_3 modelada e a medida é de apenas 9%, significando uma boa precisão no contexto do Ozônio, e a correlação de *Spearman* aponta forte correlação entre a concentração de Ozônio predita e a observada (valores próximos a 1 indicam correlação forte e positiva). Retomando a regressão linear múltipla e as características da variável estudada, os resultados obtidos são satisfatórios.

Tais resultados podem ser mais bem visualizados pelo gráfico na figura 1, que por si só é reducionista, mas com o contexto dado anteriormente, torna-se um bom recurso visual para mostrar as conclusões da modelagem.

Figura 01 – Gráfico Ozônio (Dito e Predito) x Tempo



Fonte: Autoria Própria (2025).

É nítida a semelhança entre as duas linhas no gráfico, onde a rosa representa os valores do O_3 do banco de dados, e a linha verde os resultados do modelo. Idealmente, as linhas teriam uma sobreposição completa, mas por se tratar de uma predição com base num recorte da realidade, a semelhança é o que se busca aqui, e há uma evidente aproximação nas curvas das

duas linhas, sendo menor nos valores extremos, comum na modelagem, pois há maior dificuldade em se prever valores que fogem da média.

4. CONCLUSÕES

Apesar da dificuldade inicial em lidar com o volume do banco de dados, o refino das formas de se modelar ao longo desta pesquisa contribuíram para tornar este trabalho possível, e geraram dúvidas que devem contribuir no avanço do mesmo, como testar se o uso da Análise dos Componentes Principais possibilitaria o emprego de um volume maior dos dados sem a perca de significância dos resultados, ou ainda, se pensar meios de aumentar as variáveis do banco (como informações sobre uso e ocupação da terra), almejando uma base mais sólida para as predições. Também se poderia incluir dados primários sobre a emissão dos gases formadores do ozônio e sobre a radiação solar.

Para aprimorar o modelo, recomenda-se a inclusão de variáveis que representam a vegetação, pois a bioquímica das plantas influencia diretamente a formação e a remoção do ozônio troposférico (O_3). Isso ocorre tanto pela emissão de Compostos Orgânicos Voláteis, que são precursores do O_3, quanto pela deposição seca do gás nas superfícies foliares. A inclusão desses processos vegetativos é considerada relevante para uma predição ainda mais acurada das concentrações de Ozônio (Wedow; Ainsworth; Li, 2021).

É discutível também o uso de outras escalas para o emprego desta metodologia adotada, mais ou menos abrangentes, pois a municipalidade em série temporal que compõe o banco de dados empregado possibilita trabalhar com suas variáveis de nível local à nacional, e ao tratar a escala como a janela de maior visibilidade de um fenômeno, o ozônio troposférico pode ser estudado em diversos recortes.

Embora tais melhoras possam enriquecer o modelo, é bom ter em mente a dificuldade inata de projetar cenários totalmente certeiros quando se trata de dinâmicas climáticas, ainda mais em macroescala como para todo o Nordeste. Além disso, o poder computacional para uma análise mais próxima do ideal já se mostrou uma questão em virtude do volume de dados analisado, quiçá num banco de dados mais complexo.

De modo geral, a metodologia se mostrou eficiente em utilizar os dados disponíveis, mesmo com espaço para aperfeiçoamento tanto nos métodos de modelagem, máquinas

utilizadas para execução e principalmente, maior disponibilidade de dados primários que aportem todo o Nordeste. O modelo gerado é robusto com base nas avaliações feitas, e pode ser replicado para a gestão da qualidade do ar e para as políticas climáticas e de saúde.

REFERÊNCIAS

ALONSO, Catarina; GOUVEIA, Célia M.; SANTOS, João A. **Analysis of tropospheric ozone concentration and their predictors in mainland Portugal**. *Atmospheric Research*, v. 314, 107766, 2025. ISSN 0169-8095.

CASTELHANO, Francisco Jablinski; RÉQUIA, Weeberb J. **Weather impact on ambient air pollution and its association with land use types/activities over 5,572 municipalities in Brazil**. *Heliyon*, [S. l.], v. 10, n. 24, p. e31857, 2024.

FERNÁNDEZ-DE-LAS-PEÑAS, César et al. **Random Forests for medical data analysis: a practical overview and some recent advances**. *Computational and Structural Biotechnology Journal*, [S.l.], v. 13, p. 1–19, 2015. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4710485/. Acesso em: 6 ago. 2025.

GENG, F. et al. **Multi-year ozone concentration and its spectra in Shanghai, China**. *The Science of the Total Environment*, v. 521-522, p. 135-143, 2015. Disponível em: https://doi.org/10.1016/j.scitotenv.2015.03.082. Acesso em: 6 ago. 2025.

HUANG, C.; GAO, H.; FENG, R.; ZHENG, H.; Y.; ZHANG, A. **Unveiling tropospheric ozone by the traditional atmospheric model and machine learning, and their comparison: a case study in Hangzhou, China**. *Environmental Pollution*, v. 252, pt. A, p. 366–378, 2019.

LU, X.; ZHANG, L.; SHEN, L. **Meteorology and climate influences on tropospheric ozone: a review of natural sources, chemistry, and transport patterns**. *Current Pollution Reports*, v. 5, p. 238-260, 2019.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to Linear Regression Analysis**. 5. ed. Hoboken: Wiley, 2012.

NGUYEN, Q. H.; LY, H. B.; HO, L. S.; AL-ANSARI, N.; LE, H. V.; TRAN, V. Q.; PRAKASH, I.; PHAM, B. T. **Influence of data splitting on performance of machine learning models in prediction of shear strength of soil**. *Mathematical Problems in Engineering*, [S.l.], v. 2021, art. ID 4832864, p. 1–15, 2021. DOI: 10.1155/2021/4832864. Disponível em: https://doi.org/10.1155/2021/4832864. Acesso em: 6 ago. 2025.

RSTUDIO TEAM. **RStudio: integrated development environment for R**. Boston, MA: RStudio, PBC, 2023. Disponível em: https://posit.co/products/open-source/rstudio/.

Silva, C. B., Gomes, F. F. B., Santos, J. V. C., & Santos, C. S. e. **UMA ANÁLISE COMPARATIVA DAS TÉCNICAS DE MACHINE LEARNING: Regressão Logística, Árvores de Decisão, Random Forest e SVM**. *Apoena*, 7, 501–511, 2023. Recuperado de https://publicacoes.unijorge.com.br/apoena/article/view/183.

TIWARI, S.; RAI, R.; AGRAWAL, M. **Annual and seasonal variations in tropospheric ozone concentrations around Varanasi**. *International Journal of Remote Sensing*, v. 29, p. 4499-4514, 2008. Disponível em: https://doi.org/10.1080/01431160801961391. Acesso em: 6 ago. 2025.

VISCONDI, Gabriel de Freitas. **Predição de radiação solar usando algoritmos de aprendizagem de máquina e parâmetros meteorológicos**. 2022. Tese de Doutorado. Universidade de São Paulo. DOI: 10.11606/D.3.2022.tde-22052023-111138

VRIGAZOVA, B. **Understanding the reasons for the train-test split ratio convention in machine learning**. *PeerJ Computer Science*, [S.l.], v. 7, e618, 2021. DOI: 10.7717/peerj-cs.618. Disponível em: https://peerj.com/articles/cs-618/. Acesso em: 6 ago. 2025.

WEDOW, Jessica M.; AINSWORTH, Elizabeth A.; LI, Shuai. **Plant biochemistry influences tropospheric ozone formation, destruction, deposition, and response**. *Trends in Biochemical Sciences*, v. 46, n. 12, p. 992-1002, dez. 2021. DOI: 10.1016/j.tibs.2021.06.007.

WRIGHT, Marvin N. **ranger: A fast implementation of random forests**. *R package version 0.16.0*, 2023. Disponível em: https://CRAN.R-project.org/package=ranger. Acesso em: 6 ago. 2025.

ZHANG, J.; WEI, Y.; FANG, Z. **Ozone Pollution: A Major Health Hazard Worldwide**. *Frontiers in Immunology*, v. 10, 2019. Disponível em: https://doi.org/10.3389/fimmu.2019.02518. Acesso em: 6 ago. 2025.