# UTILIZAÇÃO DE RANDOM FORESTS PARA PREDIÇÃO DA CONCENTRAÇÃO DE MATERIAL PARTICULADO

**Palavras-Chave: aprendizado de máquina, $MP_{2.5}$, poluição atmosférica, MERRA-2**

**Autores:**
**SEAN TORRES DOS SANTOS, FT, UNICAMP**
**Prof. Dr. GUILHERME PALERMO COELHO (orientador) FT, UNICAMP**
**Prof.ª Dr.ª SIMONE ANDRÉA POZZA (coorientadora) FT, UNICAMP**

## INTRODUÇÃO:

O Material Particulado (MP) é um poluente atmosférico, caracterizado como um conjunto de partículas, sólidas ou líquidas, suspensas na atmosfera. O diâmetro aerodinâmico dessas partículas, que é inferior a 100 µm, lhes confere a capacidade de serem transportadas pela atmosfera. Logo, não é incomum encontrar essas partículas em algum ponto distante de seu local de emissão, que pode ser tanto por fontes naturais como antropogênicas (DA SILVA BRITO; FABRIZ SODRÉ; DE ALMEIDA, 2018).

O escopo deste trabalho está limitado às partículas com diâmetro aerodinâmico médio de 2,5 µm ($MP_{2,5}$). Essas partículas estão associadas a doenças respiratórias, uma vez que são pequenas o bastante para serem inaladas e instalarem-se no sistema respiratório (GUARIEIRO; VASCONCELLOS; SOLCI, 2011).

Embora o Conselho Nacional de Meio Ambiente (CONAMA), por meio da Resolução nº 506 de 2024 (BRASIL, 2024) e de normativas anteriores, reconheça os impactos da poluição atmosférica na saúde pública e tenha delegado ao Ministério do Meio Ambiente e Mudança do Clima (MMA) a responsabilidade de monitorar a qualidade do ar, o Relatório Anual de Acompanhamento da Qualidade do Ar (Ministério do Meio Ambiente e Mudança do Clima, 2024, p. 19) mostra que as estações de monitoramento estão concentradas em poucas regiões do país. Essa limitação compromete uma avaliação abrangente da qualidade do ar em todo o território nacional. No estado de São Paulo, por exemplo, de acordo com o Relatório de Qualidade do Ar de 2023 da Companhia Ambiental do Estado de São Paulo (CETESB), apenas 42 dos 645 municípios possuem algum tipo de monitoramento (CETESB, 2024).

Diante disso, este projeto busca propor uma metodologia alternativa para complementar as atividades de monitoramento em localidades onde não existem estações. A ideia é processar séries históricas de dados obtidos por satélites para estimar as concentrações de poluentes na superfície, utilizando algoritmos de aprendizado de máquina como apoio.

Foram escolhidas aqui as *Random Forests*, que são modelos que já vêm sendo utilizados com sucesso na literatura em modelagens similares de predição de MP (Alkhodaidi et al.,

2024). Para validação dos resultados, foram utilizados dados coletados pelas estações de monitoramento do Sistema de Informações da Qualidade do Ar (QUALAR) da CETESB (CETESB, 2025). E, como alternativa às estações de monitoramento, dados meteorológicos e de concentração de $MP_{2,5}$ foram obtidos a partir da reanálise *Modern-Era Retrospective Analysis for Research and Applications, versão 2* (MERRA-2 – GMAO, 2024).

## METODOLOGIA:

A literatura propõe diversas abordagens para modelar a predição de $MP_{2,5}$, como por exemplo, usar exclusivamente os dados sobre o poluente, obtidos através de satélites, ou combinar dados de satélites com dados medidos por estações em terra, como dados meteorológicos (Kumar et al., 2023). Além disso, a Densidade Óptica dos Aerossóis, uma propriedade observada pelos satélites, é vista como potencial variável para apoiar tarefas de predição (MAO; SHEN; FENG, 2017). Sendo assim, foram realizados cinco experimentos, em uma tentativa de abranger as modelagens mais frequentes, combinando dados obtidos a partir do sistema QUALAR e da reanálise MERRA-2.

Dentre as cidades que possuem estações de monitoramento da CETESB, foram escolhidas três: Guarulhos (GRU), Rio Claro (RC) e São José do Rio Preto (SJRP). A coordenada geográfica de cada estação foi usada para obtenção dos dados do MERRA-2.

Os critérios para escolha das cidades foram a disponibilidade de séries temporais de $MP_{2,5}$ com poucos dados faltantes e configurações similares das estações da CETESB. Todas as estações possuem uma área de abrangência de 50 x 50 km, e possuem cerca de 90 a 95% dos dados preservados.

Após análise da distribuição dos dados, foram aplicados *thresholds* (limiares) inferiores e superiores, removendo-se valores abaixo de 3 μg/m³ e acima de 150 μg/m³, e períodos com dados faltantes foram excluídos.

Os experimentos foram realizados inicialmente para a cidade de Guarulhos, sendo as medições da estação local utilizadas como “valor alvo” dos modelos. Nessa etapa, 90% dos dados disponíveis foram empregados para o treinamento e a validação, enquanto os 10% restantes, assim como os dados das demais cidades, foram reservados para a etapa de testes.

Os experimentos foram organizados para que cada modelo recebesse um determinado número de observações passadas de suas respectivas variáveis (*features*), e usasse essa “janela temporal” para prever a concentração de $MP_{2,5}$ sempre uma hora à frente do fim da janela ($t$+1). As janelas variaram em tamanho entre 12, 24, 48 e 72 medições (horárias) anteriores.

Cada experimento foi avaliado com uma validação do tipo *Time Series Split*, que envolve dividir a série, preservando a ordem temporal, e analisar o desempenho dos modelos em cada subdivisão. Ao todo, foram realizadas dez divisões, e a cada uma foi realizada uma medição de erro quadrático médio (*mean squared error*, MSE) e de erro médio absoluto (*mean absolute error*, MAE). Essas métricas foram avaliadas ao longo de 30 execuções para cada variação de janela de cada experimento.

A Tabela 1 oferece uma visão geral dos experimentos. A coluna *features* lista quais informações os modelos tiveram acesso durante a etapa de treinamento.

Tabela 1 – Descrição dos experimentos realizados em Guarulhos

| Experimento | Descrição | *Features* |
|---|---|---|
| E1 | Apenas variáveis do MERRA-2 relacionadas ao $MP_{2.5}$ | BCSMASS, DUSMASS25, OCSMASS, SO4SMASS, SSSMASS25 |
| E2 | E1 com acréscimo de AOD | E1 + AODANA |
| E3 | E1 com acréscimo de variáveis meteorológicas do MERRA-2 | E1 + QV10M, Q500, Q850, T10M, T500, T850, PBLH, TOTEXTTAU |
| E4 | E1 com acréscimo de variáveis meteorológicas do QUALAR | Direção do vento, Direção do vento global, Pressão atmosférica, Radiacao Solar Global, Temperatura do Ar, Umidade Relativa Do Ar, Velocidade Do Vento |
| E5 | Medições de $MP_{2,5}$ de outra estação próxima | $MP_{2,5}$ de Itaim Paulista |
| Baseline | Fórmula sugerida pela documentação do MERRA-2 | PM2.5 = DUSMASS25+ OCSMASS+ BCSMASS + SSSMASS25+ SO4SMASS* (132.14/96.06) |

Para os experimentos E1 a E3, os dados de entrada foram obtidos exclusivamente a partir do MERRA-2, e as siglas da Tabela 1 são as mesmas dos conjuntos “tavg1_2d_aer_Nx”, “tavg1_2d_slv_Nx” e “tavg1_2d_flx_Nx” das bases de dados da reanálise.

Como os dados do MERRA-2 são disponibilizados seguindo o fuso horário UTC, com medições horárias que se iniciam às 0:30, foi realizada então uma conversão para GMT-3 (Horário de Brasília), para alinhamento da série com os dados do QUALAR. E como o QUALAR disponibiliza medições em horas completas, uma interpolação linear foi aplicada aos dados do MERRA-2.

O experimento E4 tem como objetivo avaliar se a combinação de dados de satélite com dados meteorológicos medidos localmente é mais eficaz do que utilizar exclusivamente informações provenientes de satélites (como no experimento E3).

O melhor modelo de cada experimento, selecionado na etapa de validação, foi novamente avaliado na etapa de testes, utilizando dados não vistos durante o treinamento. Os melhores modelos treinados com os dados de Guarulhos foram aplicados às demais cidades, recebendo as mesmas combinações de *features* utilizadas no treinamento, mas agora com informações referentes a Rio Claro e São José do Rio Preto, com o objetivo de prever as concentrações de $MP_{2,5}$ dessas cidades. Essa etapa teve o intuito de avaliar se um modelo treinado em uma localidade com dados conhecidos de MP consegue generalizar seu aprendizado, permitindo sua aplicação em regiões onde apenas dados de satélite estão disponíveis.

Considerando as propriedades do $MP_{2.5}$, que permitem seu transporte a longas distâncias, o experimento E5 busca avaliar se as medições em uma estação podem explicar as concentrações registradas em outra, mais distante. Na etapa de validação, as medições da estação do bairro de Itaim Paulista, da cidade de São Paulo, foram utilizadas para estimar os valores em Guarulhos, como descrito na Tabela 1. Contudo, para a etapa de teste, novos modelos foram treinados, mantendo apenas as configurações utilizadas em Guarulhos, descritas na Tabela 2. Foram realizados dois experimentos: no primeiro, um modelo treinado com dados de Rio

Claro foi usado para estimar as medições em São José do Rio Preto; no segundo, um modelo treinado com dados de São José do Rio Preto foi usado para estimar as medições em Rio Claro. Esse pareamento oferece dois contextos distintos: Guarulhos e Itaim possuem estações na mesma área de abrangência, enquanto Rio Claro e São José do Rio Preto estão situadas a uma distância maior.

Por fim, o *baseline* é um modelo simples de ajuste dos dados a partir de uma fórmula sugerida na documentação do MERRA-2, para o cálculo da concentração de $MP_{2,5}$ na superfície. Para fins de comparação, os valores do *baseline* também foram divididos nos mesmos períodos de validação e teste, e foram avaliados ao longo do tempo, com o *Time Series Split*.

Os hiperparâmetros de cada modelo foram definidos através de *Random Search*, implementado em Python com o auxílio da biblioteca Scikit-learn (Pedregosa et al., 2011). A Tabela 2 apresenta os melhores hiperparâmetros retornados pelo algoritmo para cada experimento.

## RESULTADOS E DISCUSSÃO:

A Tabela 3 apresenta as médias e desvios-padrão de MAE e MSE obtidos na validação, associadas à janela temporal ("*t*") que gerou os melhores resultados em cada experimento.

Os modelos apresentaram resultados positivos em comparação com o *baseline* durante a etapa de validação e teste. No teste (Tabela 4), destacam-se os modelos E3 e E4, que incluem dados meteorológicos e foram ligeiramente melhores que os modelos E1 e E2, que utilizam apenas dados dos poluentes.

Tabela 2 – Melhores hiperparâmetros para cada experimento

| Experimento | Hiperparâmetros |
|---|---|
| E1 | max_depth =80, max_features = log2, min_samples_leaf=12, min_samples_split=2, n_estimators = 273 |
| E2 | max_depth = 100, max_features = log2, min_samples_leaf=12, min_samples_split=20, n_estimators=300 |
| E3 | max_depth = 37, max_features = log2, min_samples_leaf=12, min_samples_split=20, n_estimators = 224 |
| E4 | max_depth = 84, max_features = None, min_samples_leaf=10, min_samples_split=17, n_estimators=298 |
| E5 | max_depth = 99, max_features= sqrt, min_samples_leaf=14, min_samples_split=6, n_estimators=300 |

Tabela 3 – Resultados da etapa de validação em Guarulhos

| Experimentos | MSE | MAE | *t* |
|---|---|---|---|
| **E1** | 169,39±0,56 | 10,16±0,01 | -48 |
| **E2** | 159,18±0,62 | 9,79±0,01 | -48 |
| **E3** | 153,14±0,86 | 9,5±0,02 | -48 |
| **E4** | 169,77±1,18 | 9,63±0,03 | -48 |
| **E5 (Itaim)** | 164,09±0,29 | 9,69±0,01 | -12 |
| **Baseline** | 250,57±91,21 | 10,60±2,13 | NA |

Tabela 4 – Resultados da etapa de testes para Guarulhos

| Experimentos | MSE | RMSE |
|---|---|---|
| **E1** | 392,09 | 14,65 |
| **E2** | 387,02 | 14,60 |
| **E3** | 310,29 | 13,17 |
| **E4** | 301,36 | 12,45 |
| **E5 (Itaim)** | 276,41 | 12,16 |
| **Baseline** | 440,55 | 15,16 |

Já a avaliação dos modelos para a cidade de Rio Claro é dada na Tabela 5. É possível notar que os modelos não generalizam tão bem o seu aprendizado para explicar de forma satisfatória as medições na cidade de Rio Claro. Já para a cidade de São José do Rio Preto (Tabela 6), os resultados foram positivos.

Tabela 5 – Resultados da etapa de teste para Rio Claro

| Experimentos | MSE | RMSE |
|---|---|---|
| **E1 (GRU)** | 349,24 | 13,52 |
| **E2 (GRU)** | 290,79 | 12,64 |
| **E3 (GRU)** | 294,44 | 12,79 |
| **E4 (GRU)** | 328,16 | 13,75 |
| **E5 (SJRP)** | 443,17 | 14,49 |
| **Baseline** | 285,71 | 11,48 |

Tabela 6 – Resultados da etapa de teste para S. J. do Rio Preto

| Experimentos | MSE | MAE |
|---|---|---|
| **E1 (GRU)** | 147,61 | 9,01 |
| **E2 (GRU)** | 145,23 | 9,04 |
| **E3 (GRU)** | 204,28 | 9,40 |
| **E4 (GRU)** | 153,20 | 9,29 |
| **E5 (RC)** | 168,01 | 8,98 |
| **Baseline** | 217,12 | 10,01 |

Uma possível explicação para os resultados de E1 a E4 pode estar nas características específicas de cada cidade. É possível que haja maior similaridade entre Guarulhos e S. J. do Rio Preto que entre Guarulhos e Rio Claro. Nesse caso, essas diferenças regionais impactam a capacidade de generalização do modelo, que será capaz de realizar predições de forma satisfatória apenas para localidades com características semelhantes àquelas usadas em seu treinamento.

Por fim, em Guarulhos e em S. J. do Rio Preto, os experimentos E5, que utilizaram apenas séries de MP coletadas em localidades próximas, demonstraram bons resultados. Isso indica que estações já existentes podem prover *features* úteis para os modelos.

## CONCLUSÕES:

Esse trabalho reforça o uso de modelos de Random Forest para predição de Material Particulado Atmosférico e indica que dados coletados de satélites têm potencial para serem usados como *features*.

Como trabalhos futuros, observou-se que os modelos podem vir a se beneficiar tanto de uma melhor compreensão de como as variáveis meteorológicas se relacionam com as medições de MP quanto da inclusão de aspectos relacionados à dinâmica de massas de ar na atmosfera.

## BIBLIOGRAFIA


Alkhodaidi, A., Attiah, A., Mhawish, A., & Hakeem, A. (2024). **The Role of Machine Learning in Enhancing Particulate Matter Estimation: A Systematic Literature Review. Technologies**, 12(10). https://doi.org/10.3390/technologies12100198

BRASIL. CONSELHO NACIONAL DO MEIO AMBIENTE (CONAMA). **Resolução nº 506 5 de julho de 2024. Estabelece padrões de qualidade do ar**. Publicação no Diário Oficial da União de 09/07/2024.

CETESB - Companhia Ambiental do Estado de São Paulo. **QUALAR – Sistema de Informações da Qualidade do Ar**. Disponível em: https://cetesb.sp.gov.br/ar/qualar/. Acesso em: 2 fev. 2025.

CETESB - Companhia Ambiental do Estado de São Paulo. **QUALIDADE DO AR NO ESTADO DE SÃO PAULO 2023**. São Paulo: CETESB, 2024. Disponível em: https://cetesb.sp.gov.br/ar/publicacoes-relatorios/. ISBN 978-65-5577-083-4.

DA SILVA BRITO, G. F.; FABRIZ SODRÉ, F.; DE ALMEIDA, F. V. **Impact of particulate matter on air quality**. Revista Virtual de Quimica, v. 10, n. 5, p. 1335–1354, 2018.

GELARO, R. et al. **The modern-era retrospective analysis for research and applications, version 2 (MERRA-2)**. Journal of Climate, v. 30, n. 14, p. 5419–5454, 1 jul. 2017.

Global Modeling and Assimilation Office**. MERRA-2: Modern-Era Retrospective analysis for Research and Applications, Version 2**. Disponível em: https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/. Acesso em: 2 fev. 2025.

Global Modeling and Assimilation Office (GMAO) (2015), **MERRA-2 tavg1_2d_aer_Nx: 2d,1-Hourly,Time-averaged,Single-Level,Assimilation,Aerosol Diagnostics V5.12.4**, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). DOI: 10.5067/KLICLTZ8EM9D

Global Modeling and Assimilation Office (GMAO) (2015), **MERRA-2 tavg1_2d_slv_Nx: 2d,1-Hourly,Time-averaged,Single-Level,Assimilation,Single-Level Diagnostics V5.12.4**, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). DOI: 10.5067/VJAFPLI1CSIV

Global Modeling and Assimilation Office (GMAO) (2015), **MERRA-2 tavg1_2d_flx_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Surface Flux Diagnostics V5.12.4**, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). DOI: 10.5067/7MCPBJ41Y0K6

GUARIEIRO, L. L. N.; VASCONCELLOS, P. C.; SOLCI, M. C. **Air pollutants from the burning of fossil fuels and biofuels: A brief review**. Revista Virtual de Quimica Sociedade Brasileira de Quimica, 2011.

KUMAR, V. et al. **Spatiotemporal analysis of fine particulate matter for India (1980–2021) from MERRA-2 using ensemble machine learning.** Atmospheric Pollution Research, v. 14, n. 8, p. 101834, 2023. DOI: 10.1016/j.apr.2023.101834. ISBN 1309-1042.

MAO, Xi; SHEN, Tao; FENG, Xiao. **Prediction of hourly ground-level PM2.5 concentrations 3 days in advance using neural networks with satellite data in eastern China**. Atmospheric Pollution Research, v. 8, n. 6, p. 1005–1015, 2017. DOI: https://doi.org/10.1016/j.apr.2017.04.002.

MMA. Ministério do Meio Ambiente e Qualidade do Ar**. RELATÓRIO ANUAL DE ACOMPANHAMENTO DA QUALIDADE DO AR - 2023.** Brasília: Ministério do Meio Ambiente, 2024.

PEDREGOSA, F. et al. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011.