Calibration of Low-Cost Air Quality Sensors Using Machine Learning Models

Andrey S. Castro[1], Thiago I. A. Souza[1], Jairo I. C. Brito[1], Danielo G. Gomes[2], Bruno V. Bertoncini[1]

[1]Universidade Federal do Ceará
Núcleo de Pesquisa em Transportes e Meio Ambiente (TRAMA)

[2]Universidade Federal do Ceará
Grupo de Redes, Engenharia de Software e Sistemas (GREat)
Fortaleza – CE – Brasil

andrey.soares@det.ufc.br, thiagoiachiley@ufc.br, jairoivo.brito@det.ufc.br,

danielo@ufc.br, bruviber@det.ufc.br

Abstract. *Low-cost air quality sensors have emerged as an affordable alternative to reference monitoring stations; however, their measurements are affected by environmental conditions. This paper proposes a machine learning-based approach for calibrating particulate matter sensors ($PM_{2.5}$ and PM_{10}) using real-world data collected in a city in northeastern Brazil. Ridge Regression, Random Forest, XGBoost, and Neural Networks were evaluated using five-fold cross-validation. The results show that ensemble models, particularly XGBoost, achieve superior performance, with R^2 values of 0.4480 for $PM_{2.5}$ and 0.6320 for PM_{10}. The proposed approach enhances data reliability and supports scalable low-cost air quality monitoring in urban environments.*

Resumo. *Sensores de baixo custo para monitoramento da qualidade do ar têm se destacado como alternativa acessível às estações de referência, porém apresentam limitações de precisão devido à influência de variáveis ambientais. Este trabalho propõe uma abordagem baseada em aprendizado de máquina para a calibração de sensores de material particulado ($PM_{2.5}$ e PM_{10}), utilizando dados reais coletados em uma cidade do nordeste do Brasil. Foram avaliados os modelos Ridge Regression, Random Forest, XGBoost e Redes Neurais por meio de validação cruzada com cinco dobras. Os resultados indicam que modelos ensemble, especialmente o XGBoost, apresentam melhor desempenho, com R^2 de 0,4480 para $PM_{2.5}$ e 0,6320 para PM_{10}. A abordagem proposta melhora a confiabilidade dos dados e contribui para o monitoramento ambiental de baixo custo em áreas urbanas.*

1. Introdução

Nos últimos anos, a rápida urbanização e o crescimento populacional têm intensificado a poluição do ar em grandes centros urbanos. Estima-se que cerca de 90% da população mundial vive em áreas que não atendem aos padrões de qualidade do ar da Organização Mundial da Saúde (OMS) [WHO 2019]. Nesse contexto, poluentes atmosféricos, especialmente o material particulado (MP), destacam-se como importantes

ameaças à saúde pública, estando associados a doenças respiratórias e cardiovasculares e ao aumento da mortalidade global [Xiong et al. 2022].

Diante desse cenário, o monitoramento da qualidade do ar torna-se fundamental. Embora estações de referência ofereçam alta precisão, seu elevado custo e baixa cobertura espacial limitam sua utilização em larga escala [Snyder et al. 2013, Mead et al. 2013]. Como alternativa, sensores de baixo custo com recursos de aquisição remota de dados estão sendo desenvolvidos como uma alternativa às redes convencionais. Sensores de captura de parâmetros relacionados a qualidade do ar, aliados aos avanços em computação e comunicação, têm proporcionado a capacidade de monitorar poluentes atmosféricos com resoluções espaciais e temporais mais altas a baixos custos.

Portanto, sensores de qualidade do ar de baixo custo estão ganhando popularidade. Esses dispositivos são compactos, acessíveis, baratos, alimentados por bateria, leves, consomem menos energia e são fáceis de implantar, permitindo ampla adoção e o estabelecimento de densas redes de sensores. Sua capacidade de fornecer dados em tempo real pode aumentar significativamente a conscientização pública e os processos de formulação de políticas relacionados à poluição do ar [Ionascu et al. 2024].

No entanto, a principal limitação desses sensores está na variabilidade de suas medições, influenciada por fatores ambientais, como temperatura e umidade relativa, o que compromete sua acurácia [Badura et al. 2018, Agrawal et al. 2024]. Nesse contexto, técnicas de aprendizado de máquina têm se mostrado eficazes para a calibração desses dispositivos, devido à sua capacidade de modelar relações não lineares e capturar a complexidade dos dados ambientais.

Apesar dos avanços recentes, ainda persistem limitações relacionadas à robustez dos modelos em cenários reais e à incorporação adequada de variáveis ambientais no processo de calibração. Nesse contexto, este artigo investiga a aplicação de modelos de aprendizado de máquina para a calibração de sensores de baixo custo utilizados no monitoramento de material particulado fino ($PM_{2.5}$) e grosso (PM_{10}), com base em dados coletados em uma cidade do nordeste do Brasil. Para isso, são analisados aproximadamente três meses de medições, comparando as leituras dos sensores com dados provenientes de uma estação oficial de referência.

2. Trabalhos Relacionados

Diversos estudos têm investigado a calibração de sensores de qualidade do ar de baixo custo. Conforme discutido em [Agrawal et al. 2024], pesquisas anteriores baseavam-se predominantemente em modelos de regressão linear simples e múltipla para calibrar esses sensores. No entanto, nos últimos anos, abordagens baseadas em aprendizado de máquina (ML), aprendizado profundo (DL) e redes neurais artificiais (ANNs) têm sido amplamente exploradas devido à sua capacidade de modelar relações não lineares e capturar a complexidade dos fenômenos ambientais.

Nesse contexto, os estudos apresentados em [Lovric et al. 2021, Peng et al. 2025, Vovk et al. 2024] destacam o potencial do modelo Random Forest na calibração de dados provenientes de sensores de baixo custo em diferentes cenários. Na Áustria, essa abordagem permitiu identificar os efeitos do lockdown da COVID-19, evidenciando uma redução significativa nos níveis de NO_2 e PM_{10} associada à diminuição do tráfego veicular [Lovric et al. 2021]. Em Xi'an, China, a aplicação combinada do modelo Random

Forest e do método SHAP (SHapley Additive exPlanations) possibilitou a calibração das concentrações internas de $PM_{2.5}$ e a identificação dos principais fatores ambientais que influenciam essas medições [Peng et al. 2025]. Adicionalmente, [Vovk et al. 2024] utilizou o modelo Random Forest para aprimorar a calibração das concentrações de $PM_{2.5}$ derivadas do modelo EMEP4PL na Polônia, com foco na mitigação dos impactos da poluição atmosférica na saúde pública.

Em [Albala et al. 2025], o módulo sensor de baixo custo ZPHS01B, capaz de monitorar temperatura (T), umidade relativa (UR), material particulado (MP), CO, CO_2, NO_2, O_3, CH_2O e compostos orgânicos voláteis (VOCs), foi avaliado e validado experimentalmente, sendo proposto um processo de calibração contínua baseado em aprendizado de máquina. Esse processo foi executado automaticamente a cada 24 horas para os sensores de $PM_{2.5}$ e O_3, alcançando um coeficiente de determinação (R^2) de aproximadamente 0,8.

De forma complementar, [Smith et al. 2025] investigou como diferentes modelos de aprendizado de máquina e pacotes de software open-source influenciam a acurácia dos dados de $PM_{2.5}$ coletados por sensores PurpleAir. Foram avaliados onze modelos de ML e cinco pacotes distintos. Os resultados indicaram que tanto a escolha do modelo quanto do ambiente computacional impactam significativamente o desempenho preditivo, enquanto a variação na divisão entre conjuntos de treinamento e teste apresentou influência marginal. Modelos *Long Short-Term Memory* (LSTM), treinados com RStudio e TensorFlow, alcançaram os maiores valores de R^2 (0,856 e 0,857), juntamente com baixos valores de RMSE.

Apesar dos avanços observados, a literatura ainda apresenta limitações importantes. Grande parte dos estudos concentra-se na avaliação de um único modelo, em plataformas específicas de sensores ou em cenários experimentais restritos. Além disso, poucos trabalhos exploram de forma sistemática a incorporação de variáveis ambientais, como temperatura e umidade relativa, no processo de calibração, apesar de sua influência direta na resposta dos sensores. Adicionalmente, há uma lacuna significativa de estudos conduzidos em regiões do Sul Global, especialmente em cidades de médio porte.

Nesse contexto, este trabalho propõe uma abordagem comparativa e orientada a dados, avaliando sistematicamente múltiplos modelos de aprendizado de máquina baseados em regressão por meio de um protocolo unificado de validação cruzada com cinco dobras. O estudo também incorpora variáveis ambientais no processo de calibração, buscando melhorar a robustez e a capacidade de generalização dos modelos. Por fim, apresenta evidências empíricas obtidas a partir de dados coletados em uma cidade do nordeste do Brasil, contribuindo para o avanço de soluções confiáveis, escaláveis e de baixo custo para o monitoramento da qualidade do ar em ambientes urbanos inteligentes e sustentáveis.

3. Metodologia

A Figura 1 ilustra o arcabouço metodológico proposto, abrangendo todo o processo desde a aquisição dos dados ambientais até a avaliação dos modelos preditivos de aprendizado de máquina. O processo é estruturado em quatro etapas principais: (i) construção da base de dados, (ii) pré-processamento e sincronização temporal, (iii) seleção de atributos e treinamento dos modelos, e (iv) avaliação de desempenho.

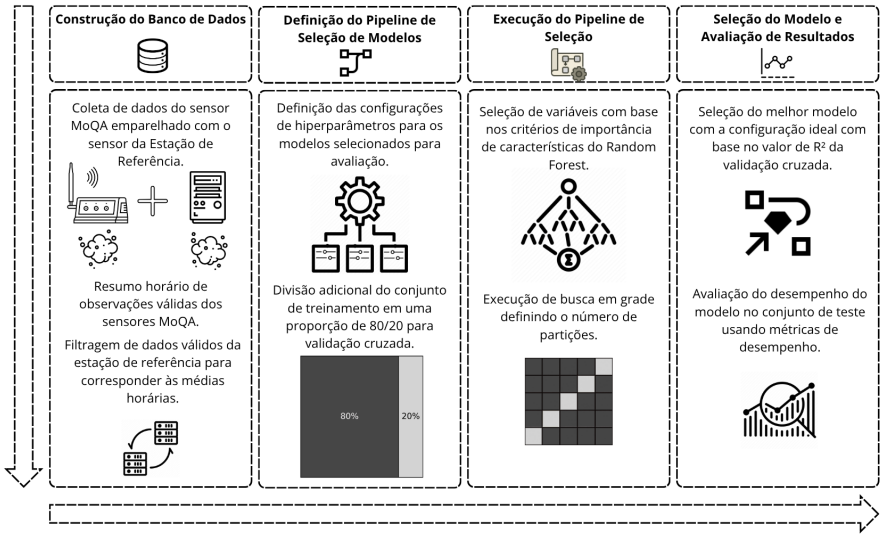


Figura 1. Fluxograma do desenvolvimento e avaliação dos métodos de calibração.

3.1. Descrição dos Dados

Neste estudo, foi desenvolvido um método para gerar funções de calibração para dados reais coletados por uma rede espacialmente distribuída de sensores de baixo custo na cidade de Fortaleza-CE , localizada no nordeste do Brasil (ver Figura 2). Na Figura 2, marcadores circulares indicam a localização dos dispositivos de monitoramento denominados de MoQA (Monitor da Qualidade do Ar) [Souza et al. 2025b, Souza et al. 2025a], dispositivo de baixo custo desenvolvido pelo Núcleo de Pesquisa em Transportes e Meio Ambiente (TRAMA) da Universidade Federal do Ceará (UFC) para monitorar a qualidade do ar de ambientes externos[1].

Distribuídos próximos a áreas escolares, corredores de tráfego, centros comerciais, equipamentos de saúde, espaços públicos e áreas industriais da cidade de Fortaleza, o dispositivo MoQA realiza medições contínuas de variáveis ambientais e poluentes atmosféricos em intervalos de um segundo, incluindo material particulado fino ($PM_{2.5}$) e material particulado grosso (PM_{10}). Os dados coletados foram agregados em médias horárias, reduzindo ruídos e tornando os dados compatíveis com a resolução temporal da estação de referência.

Os dados da estação de referência foram filtrados e sincronizados temporalmente com os dados dos sensores MoQA, garantindo correspondência exata entre as observações utilizadas no processo de calibração.

3.2. Pré-processamento dos Dados

O pré-processamento incluiu uma decisão metodológica fundamental, a construção de conjuntos de dados específicos por poluente, em vez da utilização de um único conjunto de dados unificado. Essa estratégia foi adotada para maximizar a quantidade de dados válidos disponíveis para o treinamento dos modelos.

[1]https://www.fortaleza.ce.gov.br/noticias/prefeitura-lanca-politica-e-rede-de-monitoramento-da-qualidade-do-ar-de-fortaleza

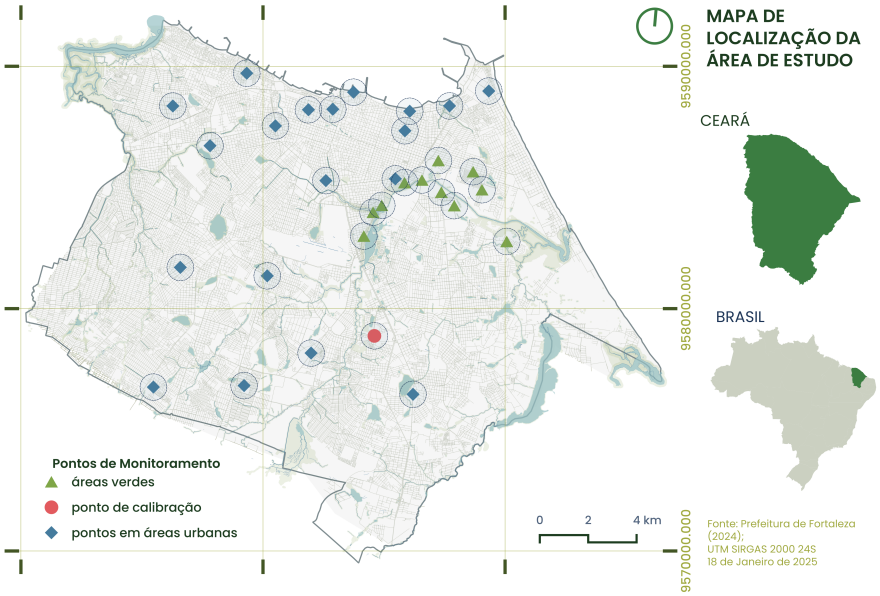


Figura 2. Área de calibração.

Como a estação de referência pode apresentar valores ausentes para determinados poluentes, o uso de um conjunto único implicaria na remoção desnecessária de observações válidas. Dessa forma, para cada variável $x \in \{PM_{2.5}, PM_{10}\}$, foi construído um conjunto de dados dedicado contendo apenas os instantes em que a medição de referência estava disponível.

3.3. Seleção de Atributos

A análise de importância de atributos é um componente essencial deste estudo, pois permite identificar as variáveis mais relevantes que influenciam os modelos, além de fornecer insights sobre o impacto de cada atributo no conjunto de dados [Xiang et al. 2025].

Para esse fim, foi empregada a técnica *Recursive Feature Elimination with Cross-Validation* (RFECV). O pipeline de seleção de atributos foi executado com base nos critérios de importância fornecidos pelo algoritmo Random Forest. Em seguida, foi aplicada uma estratégia de busca em grade (*Grid Search*) para a otimização dos hiperparâmetros dos modelos de aprendizado de máquina [Bai et al. 2025]. Ressalta-se que o número de dobras na validação cruzada foi mantido fixo em $k = 5$ ao longo de todo o processo. Dessa forma, o *Grid Search* foi utilizado exclusivamente para explorar diferentes combinações de hiperparâmetros, sendo cada configuração avaliada por meio do esquema de validação cruzada previamente definido.

A contribuição de um atributo f utilizado na divisão de um nó j (seja à esquerda ou à direita) é definida como:

$$I(m) = \frac{1}{N_m} \sum_{i \in m} (y_i - \bar{y}_m)^2 \tag{1}$$

onde y_i representa o valor alvo da i-ésima observação, e $\bar{y}_m$ corresponde à média dos

valores do alvo no nó considerado. Dessa forma, a importância de um atributo f em uma árvore de decisão é obtida a partir da redução total da variância proporcionada quando esse atributo é utilizado como critério de divisão.

3.4. Modelos de Calibração

Após a consolidação e sincronização temporal dos dados, o conjunto de observações foi dividido em treinamento (80%) e teste (20%), sendo este último mantido isolado para avaliação final.

No conjunto de treinamento, foi empregada validação cruzada com cinco dobras ($k = 5$), na qual, a cada iteração, quatro subconjuntos são utilizados para treinamento e um para validação, garantindo o uso completo dos dados durante o processo.

Neste estudo, adotou-se uma abordagem de regressão supervisionada estática, em que cada observação horária é tratada de forma independente. Essa estratégia permite avaliar a capacidade dos modelos em capturar relações instantâneas entre as variáveis ambientais e as leituras dos sensores de baixo custo, sendo adequada para cenários de aplicação em tempo real.

3.4.1. Regressão Ridge

A regressão *Ridge* é uma extensão da regressão linear que incorpora um termo de regularização para controlar a magnitude dos coeficientes e reduzir o risco de sobreajuste [Rawlings et al. 1998]. Esse modelo minimiza uma função de custo composta pelo erro quadrático e por um termo de penalização associado aos coeficientes:

$$S_r = \sum_{i=1}^{n} \left(y_i - (a_0 + a_1 x_{1_i} + \cdots + a_m x_{m_i})\right)^2 + \lambda \sum_{i=0}^{m} a_i^2 \tag{2}$$

onde $\lambda \geq 0$ é o parâmetro de regularização que controla o compromisso entre o ajuste aos dados e a complexidade do modelo. Quando $\lambda = 0$, o modelo equivale à regressão linear clássica, podendo apresentar alta variância.

3.4.2. XGBoost

O XGBoost é um algoritmo de *boosting* baseado em um conjunto de árvores de decisão construídas de forma iterativa, no qual cada nova árvore busca corrigir os erros das predições anteriores [Chen and Guestrin 2016]. A predição final é obtida pela soma ponderada das saídas dessas árvores:

$$\hat{y} = \phi(x) = \sum_{i=1}^{N} \eta f_i(x), \quad f_i \in \mathcal{F} \tag{3}$$

onde $\mathcal{F}$ representa o espaço das árvores de regressão e η é a taxa de aprendizado, responsável por controlar a contribuição de cada árvore no modelo final.

3.4.3. Redes Neurais

As Redes Neurais Artificiais (ANNs) são modelos de aprendizado profundo compostos por neurônios organizados em camadas de entrada, ocultas e saída [Koziel et al. 2025]. Cada neurônio aplica uma função de ativação para transformar os dados e propagar a informação ao longo da rede.

Neste estudo, foi utilizada a função de ativação ReLU, amplamente empregada em tarefas de regressão devido à sua simplicidade e eficiência, definida como:

$$\text{ReLU}(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases} \tag{4}$$

o que essencialmente define como zero quaisquer valores negativos calculados pelo neurônio.

3.5. Avaliação dos Modelos de Calibração

A avaliação dos modelos de calibração foi realizada por meio de validação cruzada, uma técnica amplamente utilizada para estimar o desempenho em dados não observados e mitigar o sobreajuste [Refaeilzadeh et al. 2009]. O conjunto de dados foi particionado em $k = 5$ subconjuntos, nos quais os modelos foram treinados e avaliados de forma iterativa, sendo o desempenho final obtido pela média dos resultados.

Para esse fim, foi utilizada a métrica *Root Mean Squared Error* (RMSE), que consiste na raiz quadrada da média das diferenças ao quadrado entre os valores preditos e os valores reais:

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

Adicionalmente, foi utilizado o coeficiente de determinação (R^2), que quantifica a proporção da variância da variável alvo explicada pelo modelo:

$$\mathbf{R^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

4. Resultados e Discussões

Este artigo adotou uma abordagem baseada em aprendizado de máquina para a predição da qualidade do ar utilizando um conjunto de dados reais. Para isso, todo o estudo foi conduzido em *Python* (v.3.10), em um computador equipado com processador i3, GPU e 12 GB de memória RAM.

Os resultados apresentados nesta seção evidenciam a capacidade de calibração de diferentes algoritmos de regressão aplicados à estimativa das concentrações de material particulado ($PM_{2.5}$) e (PM_{10}), ambos amplamente reconhecidos como prejudiciais à saúde pública segundo estudos ambientais [CETESB 2024]. Os modelos foram avaliados com base no coeficiente de determinação (R^2).

Os valores de R^2 obtidos em cada *fold* são apresentados com o objetivo de demonstrar a estabilidade dos modelos sob diferentes particionamentos dos dados, enquanto a média dos cinco *folds* foi adotada como principal critério para comparação e seleção dos modelos. Os resultados completos são apresentados na Tabela 1.

Tabela 1. Desempenho dos Modelos Preditivos

Modelo	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Média
MP2.5 (Ridge)	0.47	0.42	0.40	0.48	0.39	0.4320
MP2.5 (RF)	0.48	0.44	0.41	0.53	0.37	0.4460
MP2.5 (**XGB**)	0.48	0.42	0.44	0.54	0.36	**0.4480**
MP2.5 (NN)	0.52	0.34	0.38	0.50	0.34	0.4160
MP10 (Ridge)	0.61	0.65	0.67	0.59	0.48	0.6000
MP10 (RF)	0.62	0.62	0.63	0.60	0.61	0.6160
MP10 (**XGB**)	0.62	0.64	0.65	0.62	0.63	**0.6320**
MP10 (NN)	0.60	0.64	0.66	0.57	0.52	0.5980

Para a predição de $PM_{2.5}$, observou-se que o modelo XGBoost apresentou o melhor desempenho médio entre os modelos comparados, com R^2 igual a 0,4480, superando ligeiramente o modelo Random Forest, que obteve R^2 de 0,4460. O modelo de regressão linear Ridge apresentou desempenho inferior (R^2 = 0,4320), indicando que métodos baseados em árvores possuem maior capacidade de capturar padrões não lineares nos dados atmosféricos analisados. Por sua vez, o modelo de Rede Neural (NN) apresentou um R^2 médio de 0,4160, sugerindo um desempenho mais instável na predição de $PM_{2.5}$, possivelmente devido à sensibilidade aos hiperparâmetros e à necessidade de um conjunto de dados maior ou de um ajuste mais refinado para melhor generalização.

No caso do PM_{10}, os modelos apresentaram, de modo geral, desempenho superior ao observado para $PM_{2.5}$, o que pode estar associado à menor variabilidade e a relações mais lineares presentes nos dados desse poluente. Novamente, o modelo XGBoost obteve o melhor desempenho médio, com R^2 de 0,6320, seguido pelas duas configurações do *Random Forest*, com valores médios de R^2 de 0,6160 e 0,6120, respectivamente. Esses resultados reforçam a robustez de métodos *ensemble* baseados em árvores para dados ambientais caracterizados por ruído e possíveis correlações entre variáveis. A regressão *Ridge* apresentou R^2 de 0,6000, enquanto o modelo de Rede Neural obteve R^2 de 0,5980, indicando desempenho competitivo, embora ligeiramente inferior aos modelos baseados em técnicas de *bagging* e *boosting*.

A Figura 3 apresenta a curva de aprendizado do modelo de regressão para a concentração de PM_{10}, ilustrando o comportamento do RMSE em função do número de árvores (iterações) durante as fases de treinamento e validação. Observa-se que o RMSE de validação (curva laranja), após uma redução inicial acentuada, apresenta estabilização a partir de aproximadamente 40 iterações, convergindo para valores em torno de 15–16. Por outro lado, após a 50ª iteração, a continuidade do treinamento, embora reduza o erro no conjunto de treinamento, não resulta em melhorias significativas na capacidade de generalização do modelo para dados não observados.

A Figura 4 apresenta a comparação entre as medições de PM_{10} obtidas pelo sensor de baixo custo, os valores de referência e as estimativas produzidas pelo modelo de

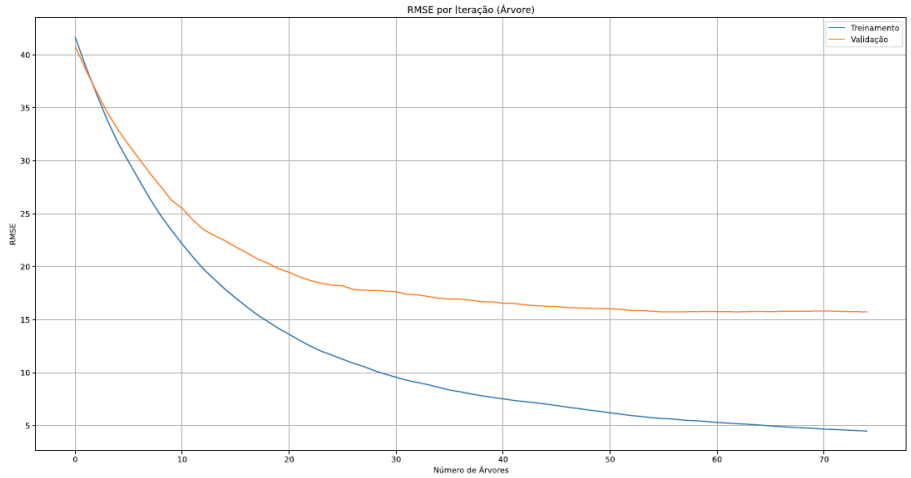


Figura 3. RMSE do PM_{10}.

calibração. Observa-se que o modelo é capaz de reproduzir adequadamente a dinâmica temporal das concentrações, acompanhando de forma consistente as variações observadas nos dados de referência. No gráfico comparativo, nota-se uma boa aderência entre as curvas, indicando que o modelo captura as principais tendências do sinal. O gráfico de dispersão confirma esse comportamento, com os pontos distribuídos próximos à linha $y = x$, evidenciando boa concordância entre valores estimados e observados. Entretanto, percebe-se uma leve dispersão crescente em concentrações mais elevadas, sugerindo a presença de heterocedasticidade e possíveis limitações do modelo ou dos sensores em cenários de maior concentração de material particulado.

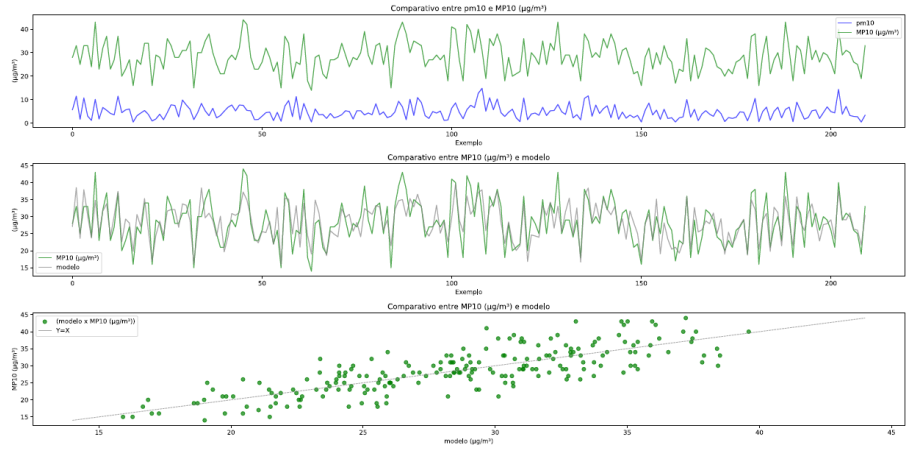


Figura 4. Curvas de calibração aplicadas ao conjunto de validação (PM_{10}).

A Figura 5 apresenta a evolução do erro quadrático médio (RMSE) em função do número de árvores durante o treinamento do modelo XGBoost para a calibração de $PM_{2.5}$. Observa-se que o RMSE de treinamento decresce de forma contínua e gradual à medida que o número de árvores aumenta, atingindo valores próximos de 1,55 após 100 iterações, indicando que o modelo ajusta progressivamente melhor os dados de treinamento. Em relação à validação (curva laranja), após uma redução inicial mais acentuada, observa-se uma estabilização por volta de 60 iterações, indicando que a capacidade de

generalização do modelo para dados não observados atinge um platô, com valores entre aproximadamente 2,40 e 2,45.

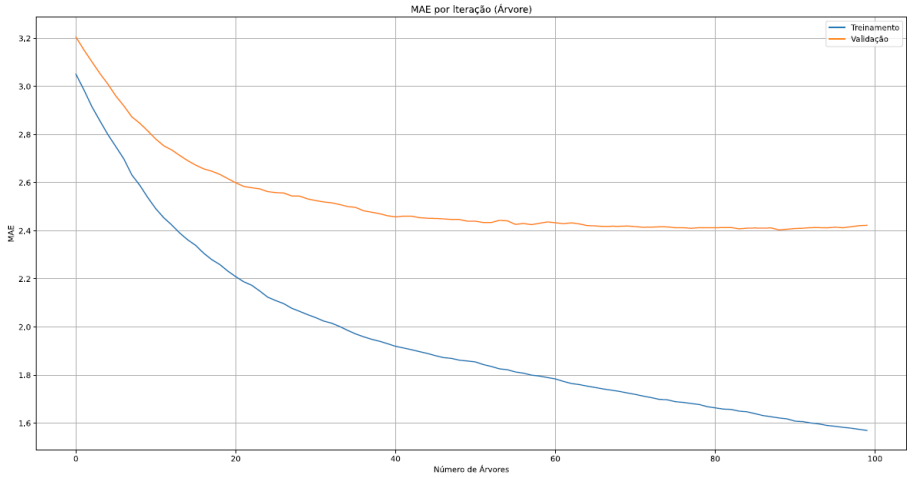


Figura 5. RMSE do $PM_{2.5}$.

Finalmente, a Figura 6 apresenta a comparação entre as medições de $PM_{2.5}$ e as estimativas do modelo de calibração. De forma similar ao PM_{10}, observa-se que o modelo acompanha adequadamente a tendência geral dos dados, sobretudo em níveis baixos e intermediários de concentração, indicando maior variabilidade dos erros em concentrações elevadas, caracterizando comportamento heterocedástico e limitações do modelo em cenários mais extremos.

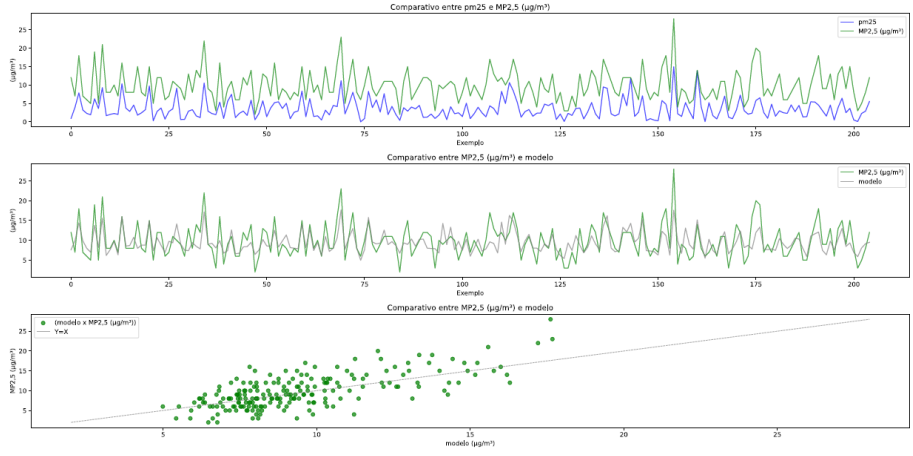


Figura 6. Curvas de calibração aplicadas ao conjunto de validação ($PM_{2.5}$).

5. Conclusão

Este trabalho avaliou o desempenho de quatro algoritmos de regressão para a calibração de sensores de baixo custo utilizados na medição de material particulado fino ($PM_{2.5}$) e grosso (PM_{10}), com o objetivo de alinhar suas medições às de estações de monitoramento de referência. Por meio de validação cruzada com cinco dobras, foram comparados os modelos *Ridge Regression*, *Random Forest*, XGBoost e Redes Neurais Artificiais, adotando-se o coeficiente de determinação (R^2) como métrica de avaliação.

Modelos baseados em *ensemble*, especialmente o XGBoost, apresentaram desempenho superior, alcançando valores médios de R^2 de 0,4480 para $PM_{2.5}$ e 0,6320 para PM_{10}, devido à sua capacidade de lidar com ruído e relações não lineares típicas de dados provenientes de sensores de baixo custo.

O método de calibração proposto demonstrou a confiabilidade dos dados coletados pelo sistema MoQA , mesmo sob variações de temperatura e umidade relativa. Esse avanço contribui para a democratização do acesso a informações ambientais de alta qualidade, ao oferecer uma solução acessível e escalável para o monitoramento da poluição do ar em áreas urbanas.

Como trabalhos futuros, recomenda-se a ampliação do conjunto de dados, a avaliação de sensores de diferentes fabricantes e a incorporação de abordagens baseadas em aprendizado profundo, como redes neurais convolucionais, que podem capturar padrões temporais e contextuais de forma mais eficaz. De modo geral, a abordagem apresentada contribui para o avanço de soluções acessíveis, confiáveis e escaláveis para o monitoramento da qualidade do ar em ambientes urbanos.

Referências

Agrawal, D., Saini, K., A., Rai, A. C., and Kala, P. (2024). In chamber calibration and performance evaluation of air quality low-cost sensors. *Atmospheric Pollution Research*, 15:102299.

Albala, E. M., Faet, G. M., Castell, S. F., Solano, J. J. P., and Jordan, R. F. (2025). Assessment of a multisensor zphs01b-based low-cost air quality monitoring system: Case study. *Electronics*, 14:1531.

Badura, M., Batog, P., Osiadacz, A. D., and Modzel, P. (2018). Evaluation of low-cost sensors for ambient pm2.5 monitoring. *Journal of Sensors*, pages 1–16.

Bai, Y., Dong, Z., and Liu, L. (2025). Hybrid feature selection-based machine learning methods for thermal preference prediction in diverse seasons and building environments. *Building and Environment*, 269:112450.

CETESB (2024). Poluentes atmosféricos. Technical report, Companhia Ambiental do Estado de São Paulo. Disponível em: https://cetesb.sp.gov.br/ar/poluentes/. Acesso em: 16 jun 2025.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA*.

Ionascu, M.-E., Marcu, M., Bogdan, R., and Darie, M. (2024). Calibration of no, so2, and pm using airify: A low-cost sensor cluster for air quality monitoring. *Atmospheric Environment*, 339:120841.

Koziel, S., Dabrowska, A., Wojcikowski, M., and Pankiewicz, B. (2025). Efficient calibration of cost-efficient particulate matter sensors using machine learning and time-series alignment. *Knowledge-Based Systems*, 295:111879.

Lovric, M., Pavlovic, K., Vukovic, M., Grange, S. K., Haberl, M., and Kern, R. (2021). Understanding the true effects of the covid-19 lockdown on air pollution by means of machine learning. *Environmental Pollution*, 274:115900.

Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J. J., McLeod, M. W., Hodgson, T. F., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J. R., and Jones, R. L. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, 70:186–203.

Peng, Z., Yang, J., Sun, J., Duan, J., Chen, Z., Niu, X., Hu, T., Huang, Y., Xu, H., Cao, J., and Shen, Z. (2025). Exploring indoor pm2.5 pollution characteristics in xi'an city and its health implications using interpretable machine learning. *Sustainable Horizons*, 13:100131.

Rawlings, J. O., Dickey, D. A., and Pantula, S. G. (1998). *Applied regression analysis: A research tool*. New York: Ny Springer.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). *Cross-Validation*. Encyclopedia of Database Systems. Springer, US, Boston, MA.

Smith, S., Trefonides, T., Malarvizhi, A. S., LaGarde, S., Liu, J., Jia, X., Wang, Z., Cain, J., Huang, T., Pourhomayoun, M., Llewellyn, G., Phyo, W., Hasheminassab, S., Roberts, J., Marlis, K., Duffy, D. Q., and Yang, C. (2025). A systematic study of popular software packages and ai/ml models for calibrating in situ air quality data: An example with purple air sensors. *Sensors*, 25:1028.

Snyder, E. G., Watkins, T. H., S., A., P., Thoma, E. D., Williams, R. W., Hagler, G. S. W., Shelow, D., Hindin, D. A., Kilaru, V. J., and Preuss, P. W. (2013). The changing paradigm of air pollution monitoring. *Environmental Science and Technology*, 47:11369–11377.

Souza, G. R. P., Souza, T. I. A., Batista, N. S., Bertoncini, B. V., and Gomes, D. G. (2025a). Predição de poluentes atmosféricos via métodos de aprendizado de máquina. In *Proceedings of the XIX Brazilian e-Science Workshop (BreSci)*.

Souza, T. I. A., Brito, J. I. C., Bertoncini, B. V., and Gomes, D. G. (2025b). Repoair: Repositório de dados para monitoramento em tempo real da qualidade do ar em ambientes externos. In *Proceedings of the XIX Brazilian e-Science Workshop (BreSci)*.

Vovk, T., Kryza, M., and Werner, M. (2024). Using random forest to improve emep4pl model estimates of daily pm2.5 in poland. *Atmospheric Environment*, 332:120615.

WHO (2019). Vision 2050: A strategy to decarbonize the global transport sector by mid-century. Technical report, Disponível em: https://theicct.org/publication/vision-2050-a-strategy-to-decarbonize-the-global-transport-sector-by-mid-century/. Acesso em: 17 jun 2025.

Xiang, F., Zhao, Y., Zhang, Z., Zuo, Y., Zou, X., and Tao, F. (2025). Ensemble learning-based stability improvement method for feature selection towards performance prediction. *Journal of Manufacturing Systems*, 74:55–67.

Xiong, J., Li, A., Liu, Y., Wang, L., Qin, X., and Yu, J. (2022). Multi-scale nanoarchitectured fibrous networks for high-performance, self-sterilization, and recyclable face masks. *Small*, 18.