



UNIVERSIDADE FEDERAL DO PARANÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA AMBIENTAL

Charles Juan Fouto de Lima

**Correção dos Dados do Sensor PurpleAir PA-II-FLEX para Monitoramento
de Material Particulado Ultrafino ($PM_{1.0}$)**

Curitiba/PR

2025



CORREÇÃO DOS DADOS DO SENSOR PURPLEAIR PA-II-FLEX PARA MONITORAMENTO DE MATERIAL PARTICULADO ULTRAFINO ($PM_{1.0}$)

Charles Juan Fouto de Lima

Trabalho de Conclusão de Curso II submetido ao
Curso de Graduação em Engenharia Ambiental,
da Universidade Federal do Paraná.

Orientador: Ricardo Henrique Moreton Godoi

Co-orientador: Felipe Baglioli

Curitiba/PR

2025



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM ENGENHARIA AMBIENTAL

TERMO DE APROVAÇÃO DE PROJETO FINAL

CHARLES JUAN FOUTO DE LIMA

CORREÇÃO DOS DADOS DO SENSOR PURPLEAIR PA-II-FLEX PARA MONITORAMENTO DO MATERIAL PARTICULADO ULTRAFINO (PM_{1.0})

Projeto Final de Curso, aprovado como requisito parcial para a obtenção do Diploma de Bacharel em Engenharia Ambiental no Curso de Graduação em Engenharia Ambiental do Setor de Tecnologia da Universidade Federal do Paraná, com nota 95, pela seguinte banca examinadora:



Prof. Dr. Ricardo H. M. Godoi
Engenharia Ambiental
Universidade Federal do Paraná - UFPR
e-mail: rhmgodoi@ufpr.br

Orientador(a): _____

Prof. Dr. Ricardo Henrique Moreton Godoi
Departamento de Engenharia Ambiental / UFPR

Documento assinado digitalmente
 MICHAEL MANNICH
Data: 15/07/2025 15:52:41-0300
Verifique em <https://validar.iti.gov.br>

Membro(a) 1: _____

Prof. Dr. Michael Mannich
Departamento de Engenharia Ambiental / UFPR

Documento assinado digitalmente
 YARA DE SOUZA TADANO
Data: 17/07/2025 10:30:22-0300
Verifique em <https://validar.iti.gov.br>

Membro(a) 2: _____

Profa. Dra. Yara de Souza Tadano
Departamento de Matemática / UTFPR

Curitiba, 14 de julho de 2025

Dedico este trabalho à minha família, pelo amor incondicional, pelos conselhos sinceros, pelo apoio nos momentos difíceis e pelo suporte financeiro fundamental, sem o qual esta jornada acadêmica não teria sido possível. O incentivo, a compreensão e a confiança de vocês foram essenciais para que eu pudesse chegar até aqui e conquistar mais este objetivo.

Agradecimentos

A realização deste trabalho marca a concretização de uma etapa fundamental da minha trajetória acadêmica e pessoal. Expresso minha profunda gratidão, em primeiro lugar, à minha família, pelo amor incondicional, apoio constante e incentivo em todos os momentos, especialmente nos desafios ao longo da graduação.

Agradeço de forma especial ao Professor Ricardo Henrique Moreton Godoi, meu orientador, pela orientação, disponibilidade e confiança durante todas as etapas deste trabalho. Ao coorientador Felipe Baglioli, sou grato pelo suporte, pelas contribuições valiosas e pelo olhar aos detalhes que enriqueceram a pesquisa.

Registro minha gratidão à Isabelle Oliveira Silva pelo auxílio fundamental durante a amostragem e análise gravimétrica, cuja colaboração foi indispensável para o desenvolvimento experimental do estudo.

Agradeço ao Laboratório de Análise e Qualidade do Ar (LABAIR) da UFPR pela infraestrutura, disponibilidade de equipamentos e ambiente de pesquisa, essenciais para a realização desta pesquisa.

Por fim, agradeço aos amigos do curso de Engenharia Ambiental da UFPR pelo companheirismo, pelas trocas de conhecimento, pelo apoio mútuo e pelas boas memórias que tornaram a caminhada universitária mais alegre e divertida. A cada um que fez parte desta jornada, deixo o meu sincero reconhecimento.

"Olhar para trás após uma longa caminhada pode fazer perder a noção da distância que percorremos, mas se nos detivermos em nossa imagem, quando a iniciamos e ao término, certamente nos lembraremos o quanto nos custou chegar até o ponto final, e hoje temos a impressão de que tudo começou ontem."

Guimarães Rosa

Resumo

A crescente demanda por sensores acessíveis e precisos para o monitoramento ambiental impulsionou a avaliação de instrumentos de baixo custo como alternativas viáveis para a quantificação do material particulado ultrafino (PM_{1.0}) em ambientes urbanos. Este trabalho avaliou o desempenho do sensor óptico PurpleAir PA-II-FLEX (PMS6003) em comparação ao método gravimétrico de referência (Impactador Harvard MS&T), durante uma campanha de amostragem conduzida na cidade de Curitiba-PR. Os dados brutos do sensor evidenciaram tendência à superestimação das concentrações de PM_{1.0}, especialmente em cenários de alta umidade relativa, reforçando a necessidade de calibração local e ajustes metodológicos para garantir a confiabilidade das medições.

A metodologia empregou a aplicação de modelos de regressão linear múltipla (MLR), algoritmos do tipo Random Forest (RF) e um método alternativo baseado em conversão de contagem de partículas. Para garantir robustez estatística, foram adotados dois procedimentos de validação: separação *holdout* (80% treino e 20% teste) e validação cruzada *k-Fold* ($k = 5$). As principais métricas de desempenho analisadas incluíram coeficiente de determinação (R^2), raiz do erro quadrático médio (RMSE), erro absoluto médio (MAE) e viés normalizado médio (MNB).

Os resultados mostraram que ambos os modelos calibrados reduziram significativamente o erro e o viés em relação aos dados brutos do sensor, elevando o desempenho preditivo para níveis compatíveis com padrões internacionais. O modelo de Regressão Linear Simples (MLR-A) apresentou maior consistência ($R^2=0.58 \pm 0.09$, RMSE= $1.74 \pm 0.06 \mu\text{g}/\text{m}^3$ e MAE= $1.31 \pm 0.10 \mu\text{g}/\text{m}^3$), especialmente na validação cruzada, enquanto o *Random Forest* ofereceu ganhos pontuais em conjuntos específicos de teste, mas apresentou maior variabilidade para faixas de concentração elevadas ($R^2=0.54 \pm 0.11$, RMSE= $1.81 \pm 0.07 \mu\text{g}/\text{m}^3$ e MAE= $1.37 \pm 0.10 \mu\text{g}/\text{m}^3$). A análise segmentada revelou que os modelos calibrados acompanham adequadamente o método de referência em intermediárias de concentrações, ainda que apresentem limitações para concentrações mais altas.

O estudo destaca a importância de incorporar variáveis ambientais no processo de calibração e validação de sensores ópticos. Entre as limitações, ressalta-se o tamanho amostral reduzido e a necessidade de avaliações multicêntricas para ampliar a aplicabilidade dos resultados. Em síntese, a calibração criteriosa dos sensores de baixo custo, associada a modelos estatísticos e validação, representa um avanço para o monitoramento ambiental de PM_{1.0}, contribuindo para a geração de dados mais confiáveis e subsidiando políticas de gestão da qualidade do ar e saúde pública.

Palavras-chave: Sensores de Baixo custo, Material Particulado Ultrafino, Saúde Respiratória.

Abstract

The growing demand for accessible and accurate sensors for environmental monitoring has driven the evaluation of low-cost instruments as viable alternatives for the quantification of ultrafine particulate matter (PM_{1.0}) in urban environments. This study assessed the performance of the PurpleAir PA-II-FLEX optical sensor (PMS6003) in comparison to the gravimetric reference method (Harvard MS&T Impactor) during a sampling campaign conducted in Curitiba, Brazil. The raw sensor data revealed a tendency to overestimate PM_{1.0} concentrations, particularly under high relative humidity scenarios, reinforcing the need for local calibration and methodological adjustments to ensure measurement reliability.

The methodology employed multiple linear regression (MLR) models, Random Forest (RF) algorithms, and an alternative method based on particle count conversion. To ensure statistical robustness, two validation approaches were adopted: holdout separation (80% training and 20% testing) and *k*-Fold cross-validation (*k* = 5). The main performance metrics analyzed included coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean normalized bias (MNB).

Results showed that both calibrated models significantly reduced error and bias relative to the raw sensor data, raising predictive performance to levels consistent with international standards. The simple linear regression model (MLR-A) demonstrated greater consistency ($R^2=0.58 \pm 0.09$, RMSE= $1.74 \pm 0.06 \mu\text{g}/\text{m}^3$, MAE= $1.31 \pm 0.10 \mu\text{g}/\text{m}^3$), especially in cross-validation, while the Random Forest model offered specific gains in selected test sets but showed greater variability for higher concentration ranges ($R^2=0.54 \pm 0.11$, RMSE= $1.81 \pm 0.07 \mu\text{g}/\text{m}^3$, MAE= $1.37 \pm 0.10 \mu\text{g}/\text{m}^3$). Segmented analysis revealed that both calibrated models closely tracked the reference method at intermediate concentration ranges, although limitations were observed at higher concentrations.

This study highlights the importance of incorporating environmental variables in the calibration and validation of optical sensors, particularly in regions with high climatic variability. Among the limitations, the relatively small sample size and the need for multicenter evaluations to expand the applicability of the results stand out. In summary, the careful calibration of low-cost sensors, combined with robust statistical modeling and rigorous validation, represents an advance for PM_{1.0} environmental monitoring, contributing to the generation of more reliable data and supporting air quality management and public health policies.

Keywords: Low-cost Sensors, Ultrafine Particulate Matter, Respiratory Health.

Lista de ilustrações

Figura 1 – Classificação por tamanho e composição do material particulado atmosférico.	19
Figura 2 – Comparação de tamanho das partículas inaláveis.	20
Figura 3 – Tempo de residência em função do tamanho do material particulado.	21
Figura 4 – Regiões de deposição do material particulado no sistema respiratório.	23
Figura 5 – Mapa da Localização de Amostragem.	30
Figura 7 – Sensor PurpleAir PA-II-FLEX.	31
Figura 6 – Funcionamento de um contador óptico de partículas.	31
Figura 8 – Esquema do sistema de amostragem utilizando Impactador Harvard, bomba de vácuo, medidor volumétrico e fluxímetro.	36
Figura 9 – Harvard Impactor MS&T.	36
Figura 10 – Esquema conceitual do funcionamento do Random Forest, ilustrando a combinação de múltiplas árvores de decisão independentes e agregação dos resultados.	39
Figura 11 – Esquema ilustrativo da validação cruzada k -fold ($k = 5$), mostrando a alternância entre conjuntos de treino (verde) e teste (vermelho) em cada <i>fold</i> e o cálculo dos resultados médios e variabilidade das métricas.	41
Figura 12 – Série temporal das concentrações de $PM_{1.0}$ registradas pelos canais A e B do sensor PurpleAir PA-II-FLEX no período de amostragem.	46
Figura 13 – Histogramas sobrepostos das concentrações de $PM_{1.0}$ medidas pelo sensor óptico PurpleAir (PA-II-FLEX) e pelo método gravimétrico de referência (HI MS/T).	46
Figura 14 – Boxplot das concentrações de $PM_{1.0}$ obtidas pelo sensor óptico (PA-II-FLEX) e método gravimétrico de referência (HI MS/T). As linhas pontilhadas representam as médias das respectivas distribuições.	47
Figura 15 – Série temporal das concentrações médias diárias de $PM_{1.0}$ obtidas pelo sensor PA-II-FLEX e pelo método gravimétrico HI MS&T, no período entre abril de 2024 e março de 2025, em Curitiba (PR).	48
Figura 16 – Gráfico de dispersão das concentrações de $PM_{1.0}$ do sensor PA-II-FLEX e do método gravimétrico HI MS&T.	49
Figura 17 – Matriz de correlação de Pearson entre as concentrações de $PM_{1.0}$ (sensor e referência) e variáveis ambientais.	51
Figura 18 – Matriz de correlação entre HI MS&T ($PM_{1.0}$ manual) e as estatísticas diárias das variáveis ambientais e do sensor PA-II-FLEX.	53

Figura 19 – Matriz de dispersão entre as concentrações de $PM_{1.0}$ (sensor e referência) e variáveis ambientais: umidade relativa, temperatura, precipitação e pressão atmosférica média.	54
Figura 20 – Dispersão e ajuste linear entre $PM_{1.0}$ PA-II-FLEX e HI MS&T (linha 1:1 em vermelho tracejado).	56
Figura 21 – Dispersão e ajuste linear entre $PM_{1.0}$ PA-II-FLEX e umidade relativa média (%).	56
Figura 22 – Dispersão e ajuste linear entre $PM_{1.0}$ PA-II-FLEX e temperatura média ($^{\circ}C$).	57
Figura 23 – Dispersão e ajuste linear entre $PM_{1.0}$ PA-II-FLEX e pressão atmosférica média (hPa).	57
Figura 24 – Dispersão e ajuste linear entre $PM_{1.0}$ PA-II-FLEX e precipitação acumulada (mm).	58
Figura 25 – Dispersão dos dados de teste para diferentes modelos MLR para as médias de amostragem.	62
Figura 26 – Dispersão dos dados de teste para diferentes modelos MLR com inclusão de variáveis estatísticas.	65
Figura 27 – Random Forest – Dispersão dos dados de teste para o modelo Q (40 árvores).	66
Figura 28 – Desempenho do Random Forest em função do número de árvores: evolução dos indicadores R^2 e RMSE para o modelo Q.	67
Figura 29 – Importância relativa das variáveis no modelo <i>Random Forest</i> de melhor desempenho.	67
Figura 31 – Dispersão dos dados de teste para diferentes modelos Random Forest (continuação).	70
Figura 32 – R^2 e RMSE da validação cruzada ($k = 5$) para cada modelo de regressão múltipla linear (MLR).	72
Figura 33 – R^2 e RMSE da validação cruzada ($k = 5$) para cada modelo <i>Random Forest</i> (RF).	73
Figura 34 – Curva de aprendizado para o modelo MLR-A: evolução de R^2 e RMSE em função do tamanho do conjunto de treinamento.	76
Figura 35 – Curva de aprendizado para o modelo RF-R: evolução de R^2 e RMSE em função do tamanho do conjunto de treinamento.	76
Figura 36 – Boxplot dos resíduos dos modelos MLR-A e RF-R.	79
Figura 37 – Distribuição de densidade dos resíduos dos modelos MLR-A e RF-R	79
Figura 38 – Autocorrelação dos resíduos para MLR-A e RF-R.	80
Figura 39 – Gráficos de resíduos versus valores previstos para os modelos MLR-A e RF-R.	80

Figura 40 – Gráficos de dispersão das previsões calibradas dos modelos MLR-A e RF-R, coloridos conforme a temperatura média durante o período amostral. A linha tracejada indica a correspondência 1:1.	81
Figura 41 – Gráficos de dispersão das previsões calibradas dos modelos MLR-A e RF-R, coloridos conforme a umidade relativa média durante o período amostral. A linha tracejada indica a correspondência 1:1.	82
Figura 42 – Gráficos de dispersão das previsões calibradas dos modelos MLR-A e RF-R, coloridos conforme a pressão atmosférica média durante o período amostral. A linha tracejada indica a correspondência 1:1.	82
Figura 43 – Dispersão das concentrações $PM_{1.0}$ calculadas pelo método alternativo (média geométrica, densidade $2,50\text{ g/cm}^3$) em relação ao método de referência HI MS&T nos dados de teste (<i>holdout</i>).	84
Figura 44 – Boxplot das concentrações de $PM_{1.0}$ da referência (HI MS&T), modelo MLR-A e modelo RF-R. As linhas tracejadas indicam as médias. . . .	85
Figura 45 – Série temporal das concentrações de $PM_{1.0}$: referência, modelos calibrados (MLR-A e RF-R) e sensor bruto (PA-II-FLEX).	85
Figura 46 – Série temporal dos resíduos para os modelos MLR-A e RF-R (diferença entre previsão e referência em cada campanha).	86
Figura 47 – Gráficos de dispersão das previsões calibradas dos modelos MLR-N e RF-Q nos dados de teste (<i>holdout</i>), em relação à referência HI MS&T.	88
Figura 48 – Gráficos de dispersão das previsões calibradas dos modelos MLR-A (esquerda) e RF-R (direita) por validação cruzada, em relação à referência gravimétrica HI MS&T. A linha tracejada representa a correspondência 1:1.	88
Figura 49 – Média das concentrações previstas pelos modelos MLR-A e RF-R em comparação com a referência gravimétrica HI MS&T, segmentadas por faixas de concentração de referência. As barras de erro representam o desvio padrão em cada faixa.	89
Figura 50 – Gráfico de dispersão das previsões dos modelos MLR-A (azul) e RF-R (verde) em relação à referência. A linha tracejada representa a correspondência 1:1.	89

Lista de tabelas

Tabela 1 – Tempo para deposição de partículas por deposição seca em aproximadamente 1 km da superfície terrestre.	21
Tabela 2 – Especificações técnicas do sensor PMS6003.	32
Tabela 3 – Resumo das métricas de desempenho dos modelos de regressão para diferentes combinações de variáveis independentes (conjunto de teste).	61
Tabela 4 – Resumo das métricas de desempenho dos modelos de regressão com variáveis estatísticas (conjunto de teste).	64
Tabela 5 – Resumo das métricas de desempenho dos modelos Random Forest para diferentes combinações de variáveis independentes (conjunto de teste).	71
Tabela 6 – Comparação das métricas de desempenho (validação cruzada k -fold) para os modelos de regressão linear múltipla (MLR) e <i>Random Forest</i> (RF).	75
Tabela 7 – Métricas de erro médias e desvios-padrão dos modelos ajustados ($n = 79$, $k = 5$).	77
Tabela 8 – Estatísticas descritivas dos resíduos para os modelos MLR-A e RF-R da validação cruzada.	78
Tabela 9 – Desempenho das combinações de médias e densidades para o cálculo alternativo de $PM_{1.0}$ (método Hanout), comparando o ajuste por holdout (20% teste) e validação cruzada ($K=5$).	83
Tabela 10 – Resumo dos modelos finais avaliados para MLR e RF, considerando validação cruzada (k -Fold, $k = 5$) e <i>holdout</i> (treino/teste 80/20).	90

Lista de abreviaturas e siglas

PM: Particulate Matter (Material Particulado).

*PM*_{1,0}: Material Particulado Ultrafino.

*PM*_{2,5}: Material Particulado Fino.

*PM*₁₀: Material Particulado Grosso.

WHO: World Health Organization (Organização Mundial da Saúde).

EPA: U.S Environmental Protection Agency (Agência de Proteção Ambiental dos Estados Unidos).

HEI: Health Effects Institute (instituto de efeitos à saúde humana).

NO_x: óxidos de nitrogênio.

*SO*₂: Dióxido de Enxofre.

*O*₃: Ozônio Troposférico.

VOC's: Compostos Orgânicos Voláteis.

UR: Umidade Relativa.

DPOC: Doença Pulmonar Obstrutiva Crônica.

AVC: Acidente Vascular Cerebral.

LCS: Low-Cost Sensors (Sensores de Baixo Custo).

INMET: Instituto Nacional de Meteorologia.

UFPR: Universidade Federal do Paraná.

IBGE: Instituto Brasileiro de Geografia e Estatística.

HI: Havard Impactor (Impactador Havard).

MQO: Mínimos Quadrados Ordinários.

MLR: Regressao Linear Multivariada.

ULR: Regressao Linear Univariada.

Sumário

1	INTRODUÇÃO	15
1.1	Objetivo	16
1.1.1	Objetivo Geral	16
1.1.2	Objetivos Específicos	17
2	REVISÃO DA LITERATURA	18
2.1	Poluição Atmosférica	18
2.1.1	Material Particulado	19
2.1.1.1	Impactos do PM _{1.0} na Saúde Humana	21
2.1.1.2	Sensores de Baixo Custo (LCS)	24
2.1.1.3	Modelos de Correção de Sensores	27
3	METODOLOGIA	29
3.1	Local de Amostragem	29
3.2	Amostragem	29
3.2.1	Amostragem Online PurpleAir PA-II-FLEX	29
3.2.1.1	Coleta e Tratamento dos Dados	32
3.2.2	Estimativa da Concentração Mássica por Método Alternativo (ALT)	33
3.2.3	Amostrador de Referência: <i>Harvard Impactor</i> (MS&T)	35
3.2.3.1	Análise Gravimétrica	37
3.3	Desenvolvimento do Modelo de Correção	37
3.3.1	Regressão Linear Simples (ULR)	38
3.3.2	Regressão Linear Múltipla (MLR)	38
3.3.3	<i>Random Forest Regressor</i>	39
3.4	Validação e Avaliação dos Modelos	41
3.4.1	Divisão dos dados: <i>Holdout</i> e <i>K-Fold Cross Validation</i>	41
3.4.2	Métricas de Desempenho dos Modelos	42
3.4.3	Análise dos Resíduos e Diagnóstico dos Modelos	43
4	RESULTADOS	45
4.1	Análise dos Dados Brutos	45
4.2	Correlação Inicial entre Variáveis	49
4.2.1	Matriz de Correlação entre Variáveis Ambientais e Concentrações de PM _{1.0}	49
4.2.2	Correlação entre HI MS&T e Estatísticas Diárias das Variáveis	51

4.2.3	Matriz de Dispersão das Variáveis	52
4.2.4	Análise das Relações Lineares entre as Variáveis	55
4.3	Aplicação e Avaliação dos Modelos	58
4.3.1	Modelos Lineares Simples e Múltiplos	58
4.3.1.1	Regressão Linear Simples	58
4.3.1.2	Modelos de Regressão Linear Múltipla	59
4.3.1.3	Regressão Linear Múltipla com Inclusão de Variáveis Estatísticas	61
4.3.2	<i>Machine Learning</i>	66
4.3.2.1	<i>Random Forest Regressor</i>	66
4.3.3	Avaliação dos Modelos	71
4.3.3.1	Validação Cruzada (<i>k-fold</i>)	71
4.3.3.2	Curvas de Aprendizado dos Modelos Seleccionados	75
4.3.4	Análise dos Erros dos Modelos	77
4.3.4.1	Análise dos Resíduos	78
4.3.5	Influência das Condições Meteorológicas nos Modelos	81
4.4	Concentração pelo método ALT	82
4.5	Análise Temporal dos Resultados	84
4.6	Comparação dos Modelos	87
5	APLICAÇÃO PRÁTICA DOS MODELOS	91
6	CONCLUSÃO	93
	REFERÊNCIAS	95

1 INTRODUÇÃO

A poluição atmosférica é uma das principais causas de doenças e mortes prematuras no mundo, afetando especialmente regiões urbanas e industrializadas. A Organização Mundial da Saúde (WHO) estima que a poluição do ar é responsável por aproximadamente 7 milhões de mortes anuais, sendo que a exposição ao material particulado (PM) desempenha um papel central nesse quadro (WHO, 2021). As partículas finas e ultrafinas, como o $PM_{2.5}$ e o $PM_{1.0}$, são consideradas extremamente perigosas devido à sua capacidade de penetração nos pulmões e na corrente sanguínea, causando uma série de problemas respiratórios, cardiovasculares e neurológicos (BROOK et al., 2010; POPE; EZZATI; DOCKERY, 2015).

Entre os principais poluentes atmosféricos, o material particulado (PM) é composto por uma complexa mistura de substâncias sólidas e líquidas, incluindo poeira, metais pesados, compostos orgânicos e sulfatos. As partículas podem ser classificadas como naturais ou antropogênicas, sendo a última mais relacionada a fontes industriais. Entre elas, o processo de geração de energia por combustão de combustível fóssil e de carvão é um dos maiores contribuidores para a poluição do ar (WHO, 2021). Essas partículas variam em tamanho e origem, podendo ser emitidas diretamente (primárias) por fontes, como veículos e indústrias e processos naturais, ou formadas por reações químicas (secundárias) na atmosfera ((EEA), 2019).

Estudos apontam que a exposição crônica ao PM aumenta os riscos de infartos, acidentes vasculares cerebrais (AVC) e doenças pulmonares obstrutivas crônicas (DPOC), reforçando a necessidade de políticas eficazes de controle de emissão (WHO, 2018; (EEA), 2019). As desigualdades sociais e econômicas influenciam diretamente a exposição à poluição do ar, uma vez que populações de baixa renda muitas vezes vivem em regiões de maior densidade de fontes, devido ao maior agrupamento de instalações industriais ((EEA), 2019; WHO, 2018). A poluição do ar impacta de maneira desproporcional crianças, idosos e indivíduos com condições crônicas de saúde, tornando-se um fator crítico para a saúde pública global (HEI, 2024).

O material particulado ultrafino, conhecido como $PM_{1.0}$, é composto por partículas com diâmetro aerodinâmico menor ou igual a 1 μm , sendo potencialmente mais prejudicial à saúde humana do que as frações mais grossas. Partículas ultrafinas podem alcançar as áreas mais profundas dos pulmões e, em certos casos, entrar na circulação sanguínea (OBERDÖRSTER; OBERDÖRSTER; OBERDÖRSTER, 2005; SCHRAUFNAGEL et al., 2019). Pesquisas indicam que o PM está fortemente associado a agravos em indivíduos com comorbidades pré-existentes, como diabetes (KIM et al., 2015a).

Devido a relação supracitada entre poluentes e saúde, os avanços no monitoramento

da qualidade do ar são essenciais para detectar e controlar níveis elevados de poluição, especialmente em áreas urbanas. No entanto, as tecnologias tradicionais de monitoramento, são caras e de resolução temporal inviável para certas aplicações, limitando sua aplicabilidade em escala ampla. Em resposta a essa limitação, sensores de baixo custo, como o PurpleAir, surgem como alternativas viáveis para aumentar a cobertura de monitoramento em regiões que carecem de infraestrutura. Esses sensores, levando em conta suas limitações, permitem a construção de redes comunitárias de monitoramento, facilitando o acesso a dados de qualidade do ar em tempo real (CASTELL et al., 2017a).

Apesar das vantagens, os sensores de baixo custo ainda enfrentam desafios significativos de precisão, especialmente em condições ambientais variáveis. A umidade relativa, por exemplo, pode afetar as leituras de PM, devido levando a uma superestimação das concentrações de partículas em ambientes úmidos, pois partículas higroscópicas absorvem água, aumentando de tamanho e intensificando a dispersão de luz captada pelos sensores ópticos. (JAYARATNE, 2018; BULOT et al., 2019). Dessa forma, é essencial o desenvolvimento de modelos de correção que considerem as variáveis ambientais para garantir a confiabilidade dos dados fornecidos por esses sensores.

A correção de dados de sensores de baixo custo é uma abordagem promissora para aprimorar a qualidade das medições, permitindo uma integração mais efetiva com redes de monitoramento. Estudos recentes têm focado na criação de modelos que ajustem os dados de sensores em função das condições ambientais, como temperatura e umidade, melhorando significativamente a precisão das medições em comparação com as medições convencionais (MISKELL; SALMOND; WILLIAMS, 2018).

Dada a relevância dos dados sobre poluição do ar para a implementação de redes de monitoramento acessíveis e precisas, a correção dos dados é fundamental. Essas redes de baixo custo, ajustadas por algoritmos de correção, podem desempenhar um papel vital no monitoramento da poluição do ar. Especialmente em áreas onde o monitoramento tradicional é financeiramente inviável, tais redes podem atuar de forma a ampliar o acesso à informações essenciais para a saúde pública (MALINGS et al., 2020b; EPA, 2013).

1.1 Objetivo

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver uma correção para os dados do amostrador Purple Air, equipado com sensores PMS6003, visando aprimorar a precisão no monitoramento do material particulado ultrafino ($PM_{1.0}$), e mensurar os efeitos de variáveis ambientais nos resultados originalmente gerados por estes sensores. Essas variáveis climáticas podem influenciar diretamente os dados do sensor, alterando as propriedades

físicas do material particulado.

Para abordar esses desafios, este trabalho considera a inclusão de variáveis como temperatura, pressão e umidade relativa em modelos de regressão linear e *random forest*, que buscam corrigir os dados do sensor. Assim, espera-se garantir maior acurácia e reprodutibilidade nos dados gerados, facilitando seu uso em estudos de qualidade do ar e estudos epidemiológicos.

1.1.2 Objetivos Específicos

Considerando o desenvolvimento do trabalho e o objetivo geral apresentado, coloca-se os seguintes objetivos específicos:

- Comparar estatisticamente as leituras do sensor PurpleAir e o impactador Harvard, utilizado como amostrador padrão de referência para $PM_{1.0}$;
- Avaliar a influência de variáveis ambientais nas medições do sensor de baixo custo, comparando estatisticamente a precisão do sensor com a inclusão dessas variáveis no modelo;
- Aplicar diferentes modelos para corrigir os dados do sensor em relação aos valores de referência;
- Testar a eficiência do modelo de correção proposto, utilizando ferramentas estatísticas de avaliação;

2 Revisão da Literatura

Este capítulo aborda os conceitos para embasamento do tema, relacionados ao material particulado. Primeiramente, é apresentado o tema de poluição atmosférica, seguido por material particulado. Posteriormente, são apresentados os impactos na saúde humana. Também é apresentada uma revisão sobre o monitoramento do material particulado através de sensores de baixo custo e, por fim, uma breve revisão sobre os modelos de correção.

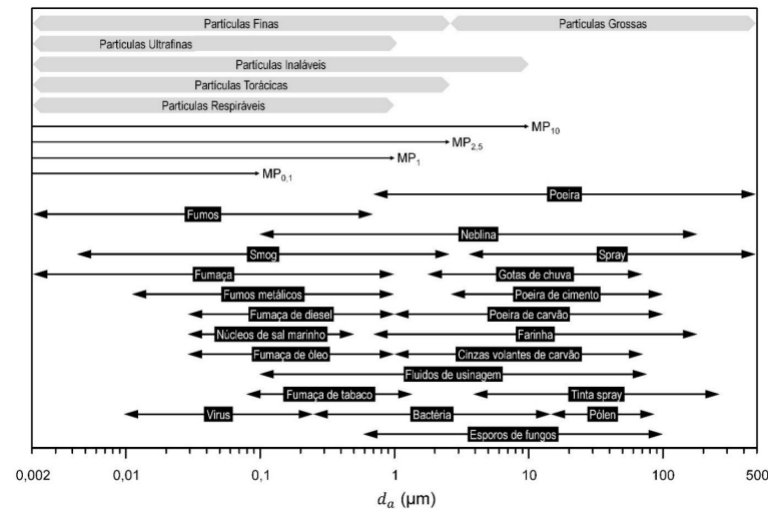
2.1 Poluição Atmosférica

A poluição atmosférica é resultado da emissão de poluentes provenientes de fontes naturais e antropogênicas, como atividades industriais, tráfego veicular, queima de combustíveis fósseis e processos agrícolas (SEINFELD; PANDIS, 2016). Os poluentes atmosféricos podem ser classificados em dois estados físicos principais: gasoso e aerossol. Além disso, os poluentes podem ser classificados como primários, quando emitidos diretamente para a atmosfera, e secundários, quando formados na atmosfera por reações químicas (WHO, 2021).

A atmosfera é considerada poluída quando a concentração de substâncias (ou energia) emitidas por atividades antropogênicas está acima dos níveis naturais do ambiente, ocasionando impactos na saúde humana e no ambiente (SEINFELD; PANDIS, 2016). Esses poluentes incluem óxidos de nitrogênio (NO_x), dióxido de enxofre (SO_2), ozônio troposférico (O_3), compostos orgânicos voláteis (VOCs) e o material particulado. O processo de geração de energia por combustão de combustível fóssil e carvão é um dos maiores contribuintes para emissão de poluição atmosférica (WHO, 2021)(EPA, 2013).

Os poluentes atmosféricos apresentam grande diversidade nas escalas de atuação, desde a microescala de uma indústria até a escala global. A determinação da região de influência de um poluente atmosférico está intrinsecamente relacionada ao seu tempo de residência na atmosfera, que é a duração média que uma substância permanece no sistema atmosférico antes de ser removida por processos como deposição ou transformação química. O tempo de residência é essencial para estimar a dispersão de poluentes e prever seus impactos ambientais, pois define se o poluente terá uma influência local, regional ou até global (ALVES, 2005; MOREIRA, 2008). Além disso, alguns fatores que podem influenciar o risco de danos à saúde são: localização geográfica, as fontes e a composição do poluente (BRITO, 2018).

Figura 1 – Classificação por tamanho e composição do material particulado atmosférico.



Fonte: Adaptado de Muhlfield et al., 2008.

2.1.1 Material Particulado

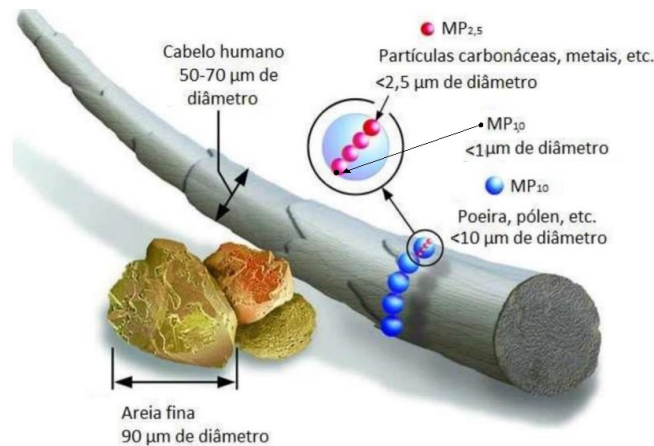
O material particulado (PM) é um aerossol que pode conter partículas sólidas e líquidas, formado por uma mistura de ácidos, compostos orgânicos, partículas do solo, metais e partículas biogênicas, e pode ter diferentes tamanhos e formas (EPA, 2013).

O PM pode ser primário, quando emitido diretamente, ou secundário, quando sofre transformação na atmosfera. O conhecimento do tipo de formação do PM pode ajudar na identificação e controle das emissões. Existem uma grande variedade de fontes de PM, tanto de origem natural quanto antrópica, sendo que os aerossóis naturais possuem semelhança ao material de origem (JACOBSON et al., 2012).

A Figura 1 apresenta a classificação das partículas presentes na atmosfera em função do seu diâmetro aerodinâmico (d_a), destacando as principais categorias adotadas na literatura, como partículas respiráveis, torácicas, inaláveis, ultrafinas, finas e grossas. Além disso, a figura ilustra a faixa de tamanhos de partículas emitidas por diferentes fontes, como fumaça, poeira, névoa, vírus, bactérias, pólen, entre outros agentes biológicos e industriais. Esta representação evidencia a diversidade de origens e tamanhos das partículas em suspensão, ressaltando a complexidade envolvida no monitoramento e avaliação de impactos à saúde (JACOBSON et al., 2012).

A classificação do PM pode ser feita pelo tipo de formação, origem e pelo tamanho. O tamanho comumente associado à classificação do PM é o diâmetro aerodinâmico médio das partículas. O diâmetro aerodinâmico (d_a) é o diâmetro da esfera de densidade unitária que tem a mesma velocidade de sedimentação que a partícula. As partículas podem ser chamadas de PM quando tiverem um diâmetro aerodinâmico entre $0,001 \mu\text{m}$ e $100 \mu\text{m}$, pois as partículas com diâmetros inferiores a $0,001 \mu\text{m}$ têm comportamento semelhante ao

Figura 2 – Comparação de tamanho das partículas inaláveis.



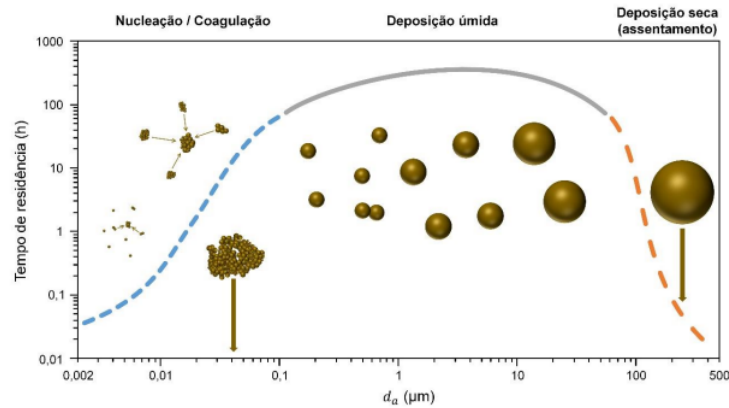
Fonte: Adaptado de (EPA, 2013).

de gases, e partículas maiores que 100 µm são depositadas rapidamente (EPA, 2013). As partículas inaláveis são classificadas como: material particulado grosso (PM₁₀) (<10 µm), material particulado fino (PM_{2,5}) (<2,5 µm), material particulado ultrafino (PM₁) (<1 µm) e nanopartículas (<0,01 µm) (WHO, 2006). Essa hierarquia reflete que as partículas menores estão contidas dentro das maiores, contribuindo para a fração total do material particulado (EPA, 2024).

O transporte do material particulado está diretamente associado ao tamanho das partículas, visto que quanto maior o tamanho da partícula, maior o assentamento devido à ação da força da gravidade, conhecido como deposição seca (VANLOON; DUFFY, 2000). Além disso, partículas menores que 0,01 µm estão fortemente associadas a um processo denominado coagulação, que envolve a remoção de partículas pequenas por meio da formação de agregados de maior diâmetro aerodinâmico (d_a). A coagulação é fortemente influenciada pelos movimentos brownianos, promovendo condições de colisão e, conseqüentemente, formando agregados maiores (VANLOON; DUFFY, 2000).

Outro evento importante de remoção de partículas suspensas na atmosfera é a deposição úmida, em que, por efeito de precipitação, as partículas são capturadas por gotas, denominado *washout*. Quando as partículas suspensas servem de núcleo de condensação, chama-se *rainout* (BRITO, 2018). A Tabela 1 mostra o tempo de residência para diferentes tamanhos de partículas. Já na Figura 3, podemos observar os principais fatores para a remoção das partículas suspensas.

Figura 3 – Tempo de residência em função do tamanho do material particulado.



Fonte: Adaptado de (VANLOON; DUFFY, 2000).

Tabela 1 – Tempo para deposição de partículas por deposição seca em aproximadamente 1 km da superfície terrestre.

d_a (μm)	Tempo para deposição
0,02	228 anos
0,1	36 anos
1	328 dias
10	3,6 dias
100	1,1 hora
1000	4 minutos
5000	1,8 minutos

Fonte: Adaptado de (JACOBSON et al., 2012).

2.1.1.1 Impactos do $\text{PM}_{1,0}$ na Saúde Humana

O material particulado com diâmetro aerodinâmico inferior a $1 \mu\text{m}$ ($\text{PM}_{1,0}$) constitui uma fração ultrafina dos aerossóis atmosféricos, reconhecida por sua elevada infiltração no trato respiratório inferior e associação a impactos adversos à saúde, especialmente em populações sensíveis como crianças, idosos e pessoas com doenças respiratórias preexistentes (YEE et al., 2021; YAN et al., 2021; KIM et al., 2015b). Ao contrário das frações mais amplamente monitoradas, como $\text{PM}_{2,5}$ e PM_{10} , o $\text{PM}_{1,0}$ permanece menos estudado, apesar de evidências crescentes apontarem riscos substanciais para a saúde pública.

Diversos estudos demonstram que a exposição ao $\text{PM}_{1,0}$ está correlacionada ao aumento da morbidade e mortalidade por doenças respiratórias e cardiovasculares. Uma revisão sistemática e metanálise recente identificou que cada aumento de $10 \mu\text{g}/\text{m}^3$ em $\text{PM}_{1,0}$ eleva em 1,25 vezes o risco de asma e em 1,07 vezes o risco de pneumonia, destacando efeitos mais severos para exposições prolongadas (YEE et al., 2021). Em crianças e adolescentes, os impactos são ainda mais marcantes: a exposição ao $\text{PM}_{1,0}$ resulta em declínios significativos nos parâmetros de função pulmonar, como Capacidade

Vital Forçada (CVF) e Volume Expiratório Forçado no primeiro segundo (VEF_1) (YAN et al., 2021). Idosos e pessoas com doenças respiratórias preexistentes também apresentam maior suscetibilidade, com aumento do risco de hospitalizações e mortalidade associadas a exacerbações agudas da asma, para partículas (KIM et al., 2015b).

Apesar dos avanços, a literatura destaca limitações importantes: (i) a maioria dos estudos se concentra em $PM_{2.5}$, enquanto pesquisas específicas para $PM_{1.0}$ ainda são incipientes; (ii) há heterogeneidade metodológica, dificultando comparações diretas entre resultados; (iii) os efeitos do $PM_{1.0}$ em grupos vulneráveis, como crianças, idosos e portadores de doenças crônicas, carecem de maior detalhamento; e (iv) faltam estudos longitudinais e multicêntricos em países de baixa e média renda (LIU et al., 2019; YAN et al., 2021).

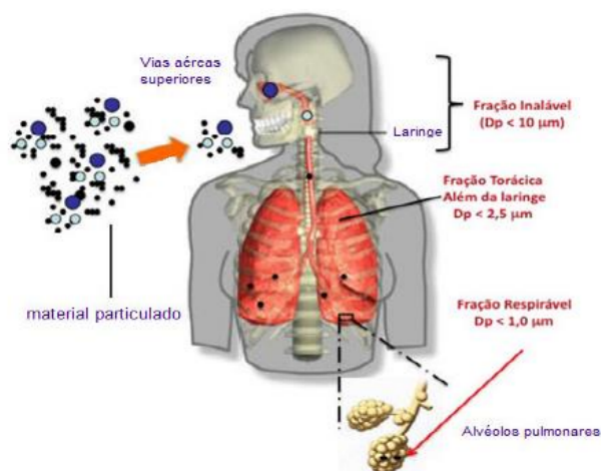
O PM pode ocasionar muitos efeitos adversos na saúde, podendo variar com o tamanho e composição. A capacidade de infiltração no corpo humano depende do tamanho da partícula. O $PM_{1.0}$ compõe uma subfração do material particulado e consiste nas partículas respiráveis, que podem atingir regiões profundas do sistema respiratório, chegando até os alvéolos pulmonares e, em certos casos, entrar na circulação sanguínea (FERNANDES et al., 2010).

Além dos impactos respiratórios e cardiovasculares já mencionados, pesquisas recentes têm apontado a presença de componentes químicos específicos no $PM_{1.0}$ que aumentam substancialmente seus riscos toxicológicos. Estudos realizados na região do Delta do Rio das Pérolas, na China, identificaram concentrações significativas de substâncias perfluoroalquiladas e polifluoroalquiladas (PFAS) em partículas $PM_{1.0}$ coletadas em ambientes escolares. Estes compostos sintéticos, amplamente utilizados na indústria química, têm sido associados a distúrbios endócrinos, hepatotoxicidade, comprometimento imunológico e aumento do risco de câncer (YANG et al., 2025). Em um estudo conduzido na Grécia, análises químicas detalhadas identificaram que hidrocarbonetos aromáticos policíclicos (PAHs) presentes no $PM_{1.0}$ apresentaram padrões carcinogênicos e mutagênicos distintos daqueles encontrados em frações maiores, como o $PM_{2.5}$, destacando a importância específica desta fração fina na avaliação de riscos à saúde pública (PATERAKI et al., 2020).

Estudos adicionais sugerem que o $PM_{1.0}$ possui uma composição química particularmente rica em carbono negro, carbono orgânico, metais pesados e poluentes orgânicos persistentes (POPs), conferindo-lhe uma toxicidade intrinsecamente maior em comparação com frações maiores do material particulado (KULSHRESTHA, 2018). Essas características indicam que estratégias regulatórias e de monitoramento que visem especificamente o $PM_{1.0}$ podem ser mais eficazes na prevenção de danos à saúde humana e redução dos impactos socioeconômicos associados à exposição a partículas ultrafinas.

Evidências relatam fortemente relações entre a exposição ao material particulado fino $PM_{2.5}$ e a mortalidade por diversas causas, dentre elas infecções respiratórias agudas,

Figura 4 – Regiões de deposição do material particulado no sistema respiratório.



Fonte: (GUARIEIRO; VASCONCELLOS; SOLCI, 2011)

doença pulmonar obstrutiva crônica (DPOC), doença cardíaca isquêmica (DCI), câncer no pulmão e acidente vascular cerebral (AVC) (WHO, 2018). Além disso, também se sugerem relações com o diabetes tipo 2, mortalidade neonatal e por nascimentos prematuros (IHME, 2020).

Além dos efeitos respiratórios, a exposição de longo prazo ao $PM_{1.0}$ está associada ao aumento da mortalidade por todas as causas e ao risco de doenças cardiovasculares, como hipertensão (YIN et al., 2017; ZHANG et al., 2024). Um estudo prospectivo realizado na China com mais de 33 mil participantes relatou uma relação em forma de J entre exposição ao $PM_{1.0}$ e mortalidade, com razão de risco de 1,114 para cada $10 \mu g/m^3$ de aumento (YIN et al., 2017). Outro estudo mostrou aumento significativo no risco de hospitalização por hipertensão em regiões com maiores níveis de $PM_{1.0}$, sendo mulheres e indivíduos fisicamente ativos os mais vulneráveis (ZHANG et al., 2024).

Os impactos econômicos e sociais também são relevantes: pesquisas realizadas em Pequim evidenciaram que aumentos nas concentrações de $PM_{1.0}$ resultam em elevações substanciais no número de internações hospitalares e nos custos de saúde pública (XU et al., 2024).

A Figura 4 retrata a infiltração do PM no sistema respiratório humano. A fração inalável é constituída por partículas $<10 \mu m$, que tendem a ficar retidas principalmente no trato respiratório superior. A fração torácica inclui partículas $<2,5 \mu m$, que penetram além da laringe. A fração respirável inclui partículas $<1 \mu m$, que são capazes de alcançar os alvéolos pulmonares (FERNANDES et al., 2010).

A exposição a partículas ultrafinas, como o $PM_{1.0}$, está associada a diversos efeitos adversos à saúde, que podem variar de acordo com a duração da exposição. No curto prazo, destacam-se reações inflamatórias nos pulmões, surgimento ou agravamento de

sintomas respiratórios, efeitos negativos sobre o sistema cardiovascular, aumento do uso de medicamentos, elevação das internações hospitalares e crescimento das taxas de mortalidade. Por outro lado, a exposição prolongada pode levar ao aumento dos sintomas respiratórios, redução da capacidade pulmonar em crianças e adultos, crescimento dos casos de doença pulmonar obstrutiva crônica, diminuição da expectativa de vida (principalmente em razão da mortalidade cardiopulmonar) e elevação do risco de câncer de pulmão (JOHARI, 2017).

Do ponto de vista regulatório, não existem padrões de qualidade do ar específicos para $PM_{1.0}$ na maior parte dos países, devido justamente à limitação de dados sistematizados e metodologias de medição padronizadas. Esta lacuna ressalta a necessidade de avanços em métodos de monitoramento, incluindo o uso de sensores de baixo custo e protocolos de calibração robustos frente a métodos de referência (BADURA et al., 2018).

Diante desse panorama, este estudo se propõe a contribuir com dados originais e avaliação sistemática da calibração de sensores ópticos para $PM_{1.0}$, aplicando modelos de regressão e análise comparativa com métodos de referência, visando suprir lacunas metodológicas e reforçar a base científica para monitoramento e gestão da qualidade do ar.

2.1.1.2 Sensores de Baixo Custo (LCS)

Os sensores de baixo custo (*Low-Cost Sensor* – LCS) para material particulado (PM) têm revolucionado o monitoramento da qualidade do ar, permitindo alta densidade espacial e temporal de medições em ambientes urbanos e rurais, a um custo significativamente inferior ao das estações de referência (GIORDANO et al., 2021; SAYAHI et al., 2019; WALLACE, 2022). Esses dispositivos utilizam majoritariamente o princípio da dispersão de luz, empregando uma fonte emissora (LED ou laser), um fotodiodo (ou fototransistor) e, muitas vezes, lentes de focalização (WANG et al., 2015; LIANG et al., 2021). O objetivo dos LCS é fornecer leituras próximas às obtidas por amostradores tradicionais, porém com maior acessibilidade e frequência de dados (LIANG et al., 2021; HOLDER et al., 2020).

Os sensores ópticos de PM, como o Plantower PMS6003 presente no PurpleAir PA-II-FLEX, funcionam a partir da detecção da luz espalhada por partículas em suspensão, convertendo o sinal óptico em concentração de massa por meio de algoritmos proprietários ou abertos (WALLACE, 2022; FAROOQUI; BISWAS; SAHA, 2023). A intensidade do sinal depende do número, tamanho, forma, composição e índice de refração das partículas, tornando a calibração um desafio em função da variabilidade morfológica dos aerossóis (HAGAN, 2020; SAYAHI et al., 2019). Adicionalmente, o desempenho desses sensores pode variar entre diferentes lotes, com diferenças de até 20% entre unidades do mesmo modelo (GIORDANO et al., 2021; LIANG, 2021).

Além disso, avaliações de campo do sensor PurpleAir PA-II-FLEX mostram que o desempenho varia conforme a fração de PM analisada e o tempo de integração dos dados. Os sensores apresentam forte correlação para $PM_{1.0}$ (R^2 de 0,91 a 0,96), mas

desempenho significativamente inferior para PM_{10} (R^2 geralmente abaixo de 0,5), mesmo após calibração. Esse padrão persiste tanto para médias de 5 minutos quanto de 24 horas, com inclinação de regressão inferior a 1 e intercepto positivo, indicando subestimação sistemática dos valores reais de PM pelo sensor (AQMD, 2022).

Embora o desempenho para $PM_{1.0}$ seja considerado elevado em algumas avaliações, outros estudos apontam maior variabilidade entre sensores e tendência à subestimação sistemática das concentrações reais, especialmente em faixas de baixas concentrações (TRYNER et al., 2020a). Além disso, fatores de calibração desenvolvidos para $PM_{2.5}$ podem não se aplicar diretamente à fração $PM_{1.0}$, sendo necessária a derivação de equações de correção específicas para cada faixa granulométrica e contexto ambiental avaliado.

Diversos fatores ambientais influenciam o desempenho desses sensores, como temperatura, umidade relativa (UR), pressão, velocidade do vento e até mesmo a presença de outros poluentes atmosféricos, devido a sensibilidades cruzadas (LIANG, 2021; SAYAHI et al., 2019; JOHNSON et al., 2018; HOLDER et al., 2020). O aumento da temperatura interna do próprio sensor, devido à dissipação de calor do diodo emissor, pode alterar a resposta do sensor, motivando a incorporação de sensores auxiliares de temperatura e umidade em modelos comerciais atuais (GIORDANO et al., 2021; WALLACE, 2022). Além disso, altas taxas de UR podem provocar a higroscopicidade das partículas, afetando tanto a dispersão de luz quanto a massa medida (GIORDANO et al., 2021; SAYAHI et al., 2019). Estudos mostram que, para $PM_{2.5}$, a umidade relativa acima de 80% pode causar desvios superiores a 15% nos resultados não corrigidos (FAROOQUI; BISWAS; SAHA, 2023; HOLDER et al., 2020).

O desempenho dos LCS geralmente é inferior em ambientes de baixa concentração de PM. Estudos demonstram baixa concordância entre sensores de baixo custo e equipamentos de referência em concentrações menores que $40 \mu g m^{-3}$, com melhora da correlação em níveis mais altos de poluição atmosférica (JOHNSON et al., 2018; WALLACE, 2022). Esse comportamento é atribuído ao limite de detecção¹ relativamente elevado dos sensores ópticos, que pode variar dependendo do algoritmo e modelo utilizado (WALLACE, 2022; FAROOQUI; BISWAS; SAHA, 2023). Resultados de avaliações independentes apontam erro médio absoluto (MAE) entre 5 e $12 \mu g m^{-3}$ para $PM_{2.5}$, podendo ultrapassar $15 \mu g m^{-3}$ sob condições de alta umidade ou baixas concentrações (GIORDANO et al., 2021; LIANG, 2021).

Isso é particularmente relevante para $PM_{1.0}$, onde o limite de detecção² e o ruído relativo podem comprometer ainda mais a confiabilidade dos dados em locais com baixa poluição, como evidenciado por avaliações de campo (TRYNER et al., 2020a).

¹ Menor valor que pode ser identificado como “diferente de zero” com confiança estatística.

² Nas especificações do sensor PMS6003 da Tabela 2, observa-se a eficiência de apenas 50% na contagem das partículas de até $0,3 \mu m$.

Além da utilização em ambientes externos, os sensores de baixo custo vêm sendo aplicados em medições *indoor* (escolas, residências etc.) e em estudos de exposição pessoal, permitindo identificar fontes locais e padrões de variabilidade intra-ambiente que seriam imperceptíveis em estações convencionais (GIORDANO et al., 2021; FAROOQUI; BISWAS; SAHA, 2023). A aplicação em projetos de ciência cidadã tem permitido o mapeamento de poluição do ar em alta resolução, engajando comunidades e apoiando estudos epidemiológicos (WALLACE, 2022; FAROOQUI; BISWAS; SAHA, 2023; LIANG, 2021).

A maior vantagem dos LCS é a possibilidade de implantação em larga escala devido ao baixo custo dos sensores, que variam de US\$200 a US\$2000 pelo sistema completo, valores consideravelmente inferiores aos dos equipamentos de referência (LIANG et al., 2021; SAYAHI et al., 2019). No caso do PurpleAir PA-II-FLEX, o custo é aproximadamente 120 vezes inferior ao de um analisador de referência FEM (*Federal Equivalent Method*) como o GRIMM, que pode variar de US\$20.000 a US\$50.000, e mais de 6 vezes inferior ao do amostrador Impactador Havard (HI MS&T) de referência utilizado neste estudo. Isso viabiliza o monitoramento em comunidades de baixa renda, zonas periurbanas, áreas industriais e locais sem infraestrutura de monitoramento (GIORDANO et al., 2021; LIANG, 2021; HOLDER et al., 2020). No entanto, o custo reduzido implica desafios: a qualidade dos dados brutos pode ser afetada por vieses sistemáticos, baixa seletividade de tamanho de partículas, perda de precisão em condições ambientais adversas e deriva do sensor ao longo do tempo (SAYAHI et al., 2019; FAROOQUI; BISWAS; SAHA, 2023; HOLDER et al., 2020). Observa-se que muitos sensores requerem recalibração ou substituição após 1–2 anos de operação contínua devido à deriva óptica (*drift*)³ ou envelhecimento dos componentes (LIANG, 2021; SAYAHI et al., 2019).

Fatores de calibração determinados em laboratório frequentemente apresentam desempenho inferior quando aplicados a ambientes reais, devido à variação nas propriedades físico-químicas dos aerossóis atmosféricos e condições ambientais locais (TRYNER et al., 2020a).

Apesar de suas vantagens, é importante ressaltar limitações específicas dos sensores de baixo custo, como o PurpleAir. Primeiramente, há tendência conhecida de superestimação das concentrações em condições de alta umidade relativa, devido ao crescimento higroscópico das partículas, que afeta o sinal óptico. Além disso, o desempenho para a fração PM_{1,0} tende a ser inferior em comparação à PM_{2,5}, dado que muitos algoritmos e calibrações originais priorizam partículas maiores. Por fim, a dependência de calibração local é uma limitação central, visto que os sensores requerem ajustes customizados para diferentes ambientes, tipos de aerossol e condições meteorológicas, limitando a aplicação

³ O *drift* pode ocorrer devido à deposição de partículas e degradação do LED ou laser interno, como observado por Tryner et al. (2020) (TRYNER et al., 2020a), afetando a linearidade da resposta óptica.

direta de fatores ou equações desenvolvidos em outras regiões.

Essas limitações operacionais e ambientais reforçam a necessidade de desenvolvimento e aplicação criteriosa de modelos de correção local específicos, tema discutido a seguir.

2.1.1.3 Modelos de Correção de Sensores

A popularização dos sensores ópticos de baixo custo para monitoramento de material particulado (PM) proporcionou avanços significativos em estudos ambientais, exposições pessoais e gestão da qualidade do ar. Entretanto, a acurácia dos dados obtidos por esses dispositivos é impactada por limitações intrínsecas, como a influência das condições meteorológicas (principalmente umidade relativa e temperatura), características do aerossol, envelhecimento óptico dos componentes e variações entre lotes de fabricação (CASTELL et al., 2017b; CRILLEY, 2020; TRYNER et al., 2020b; GIORDANO et al., 2021). Essas fontes de erro tornam indispensável a aplicação de modelos de correção que ajustem as leituras do sensor para valores mais próximos aos métodos de referência, como gravimétricos ou equivalentes federais (FEM).

A literatura enfatiza que a calibração dos sensores deve ser rigorosa, contínua e adaptada tanto às condições de laboratório quanto de campo, de modo a capturar o comportamento real dos dispositivos sob diferentes cenários ambientais (SAYAH et al., 2019; LIANG, 2021). Além disso, ressalta-se a necessidade do desenvolvimento de equações de correção customizadas para cada localidade, tipo de aerossol e até mesmo para cada unidade sensora, dada a variabilidade interinstrumental observada (diferença de fabricação).

O modelo de correção mais amplamente adotado é a regressão linear simples, na qual a concentração fornecida pelo sensor é ajustada a partir de uma equação do tipo $y = a + bx$, onde y representa o valor corrigido e x a leitura bruta do sensor. A simplicidade e a fácil interpretação tornam essa abordagem popular, principalmente para ambientes controlados ou faixas limitadas de concentração (BADURA et al., 2018; MALINGS et al., 2020a). No entanto, diversos estudos mostram que a relação entre o sinal óptico e a concentração real pode apresentar comportamento não linear sob diferentes faixas de umidade, temperatura e composição do aerossol (JAYARATNE, 2018; CRILLEY, 2020; WALLACE, 2022), limitando o desempenho da calibração linear em situações ambientais complexas.

No contexto da regressão linear simples, avaliações conduzidas pela South Coast Air Quality Management District (AQMD) nos Estados Unidos mostraram que sensores ópticos de baixo custo, como o PurpleAir PA-II-FLEX equipado com PMS6003, podem apresentar bons desempenhos para médias diárias de $PM_{1.0}$ após calibração, com R^2 entre 0,90 e 0,93 e RMSE variando de 1,2 a 1,7 $\mu g/m^3$ (AQMD, 2022). No entanto, ressalta-se

a influência de fatores ambientais, diferenças nos tipos de aerossol e limitações intrínsecas do método óptico, indicando a necessidade de calibração local e validação rigorosa.

Para superar essas limitações, a literatura destaca a utilização de modelos de regressão linear múltipla (MLR), nos quais são incluídas variáveis independentes adicionais, como temperatura, umidade relativa, pressão atmosférica e o próprio sinal do sensor. Essa abordagem possibilita um ajuste mais refinado, pois incorpora fontes adicionais de variabilidade, melhorando a correção especialmente sob condições ambientais variáveis (MALINGS et al., 2020a; CRILLEY, 2020; GIORDANO et al., 2021). Vários trabalhos relatam ganhos de desempenho na calibração ao incluir a umidade relativa como covariável, visto que o crescimento higroscópico das partículas pode aumentar artificialmente a leitura dos sensores ópticos em ambientes úmidos⁴ (JAYARATNE, 2018; PSOTKA; TRACEY; SICA, 2024; CRILLEY, 2020).

Novas abordagens de calibração têm incorporado técnicas de aprendizado de máquina, como *Random Forest* e *Gaussian Mixture Regression*, capazes de modelar relações não lineares e incluir múltiplas variáveis ambientais (temperatura, umidade, localização, influência do tráfego, entre outras). Tais modelos mostram ganhos expressivos em R^2 e redução do erro médio absoluto (MAE), especialmente em ambientes urbanos complexos, superando a regressão linear tradicional e múltipla (CONSIDINE et al., 2021; TRYNER et al., 2020a; WANG; LUNG; LIU, 2020; CONSIDINE et al., 2024). Esses modelos não lineares são capazes de capturar relações complexas entre múltiplas variáveis, características do sensor e valores de referência, superando as limitações dos modelos lineares tradicionais em ambientes com elevada variabilidade (MALINGS et al., 2020a; GIORDANO et al., 2021; WANG; LUNG; LIU, 2020; LIANG, 2021). O uso dessas abordagens mais avançadas permite incorporar parâmetros ambientais e operacionais ao ajuste dos dados, tornando a calibração mais robusta e adaptável às particularidades de cada localidade.

Em síntese, embora os sensores ópticos de baixo custo sejam ferramentas valiosas para o monitoramento ambiental, sua confiabilidade depende fortemente da adoção de modelos de correção robustos, adaptados às condições locais e validados frente a métodos de referência, aliados a estratégias rigorosas de controle de qualidade e calibração contínua, garantindo assim a precisão e a utilidade dos dados produzidos.

⁴ Fatores de correção empíricos para o efeito da umidade em sensores ópticos já foram propostos e validados especificamente para PM_{2.5}, como discutido em (PSOTKA; TRACEY; SICA, 2024; CRILLEY, 2020).

3 Metodologia

3.1 Local de Amostragem

O estudo foi conduzido na estação meteorológica A807 do Instituto Nacional de Meteorologia (INMET), localizada no Centro Politécnico da Universidade Federal do Paraná (UFPR), em Curitiba, Paraná, Brasil ($25^{\circ}26'53''$ S, $49^{\circ}13'12''$ W), retratada no mapa da Figura 5. Essa estação está situada próximo à rodovias federais (BR-476 e BR-277) de alta densidade de tráfego veicular. A proximidade dos amostradores com os equipamentos meteorológicos de medição, permite melhor precisão na associação das variáveis meteorológicas nos modelos de correção.

O Município de Curitiba, apresenta uma área territorial de 434,892 km² (IBGE, 2022) e uma população residente de 1.773.718 pessoas (IBGE, 2022), com uma altitude média de 934 metros, clima subtropical úmido com alta ocorrência de precipitação ao longo do ano. A precipitação e as mudanças de temperatura e umidade relativa (UR) são benéficas para a calibração de sensores de partículas, pois fornecem uma ampla variabilidade das condições ambientais. Essas variações meteorológicas podem ajudar na avaliação da precisão nas medições do sensor em cenários variados.

A estação meteorológica do INMET está equipada com instrumentos padronizados para medições precisas e contínuas de variáveis climáticas. A proximidade da estação meteorológica com a área densamente populada de Curitiba, faz dela um ponto chave para o monitoramento do material particulado na cidade. Além disso, a infraestrutura disponível no campus da UFPR facilita a instalação de equipamentos adicionais, garantindo condições para experimentos comparativos.

3.2 Amostragem

3.2.1 Amostragem Online PurpleAir PA-II-FLEX

O PurpleAir PA-II-FLEX contém dois sensores PMS6003 (*Plantower*, Pequim Co. Ltd.), rotulados como canais A e B. O PMS6003 é um sensor óptico de baixo custo que utiliza o princípio de espalhamento de luz: um feixe de laser incide sobre partículas suspensas no ar e a luz dispersada é detectada por um fotodiodo, gerando um sinal elétrico proporcional à concentração de partículas, conforme ilustrado na Figura 6. A similaridade entre os resultados dos dois canais (A e B) é utilizada como estratégia de controle de qualidade e autodiagnóstico: diferenças significativas podem indicar falhas em um dos sensores ou obstruções.

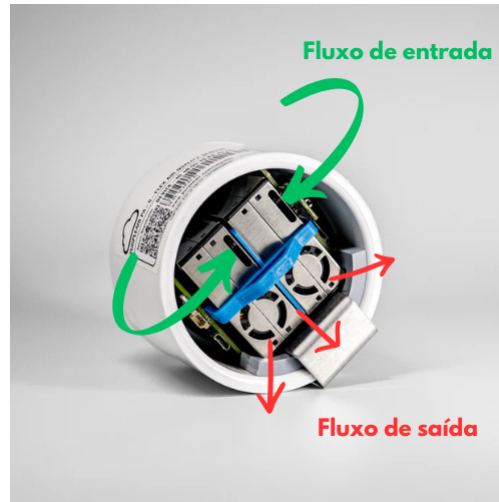
Figura 5 – Mapa da Localização de Amostragem.



O sensor oferece alta resolução temporal, operando em ciclos de 10 segundos e fornecendo dados médios a cada 2 minutos. Essa característica o torna amplamente utilizado em redes comunitárias de monitoramento de qualidade do ar, embora a precisão dos dados seja impactada por variáveis ambientais como umidade relativa, temperatura e variações na concentração de partículas atmosféricas (LIANG et al., 2021; ANTONIO et al., 2018).

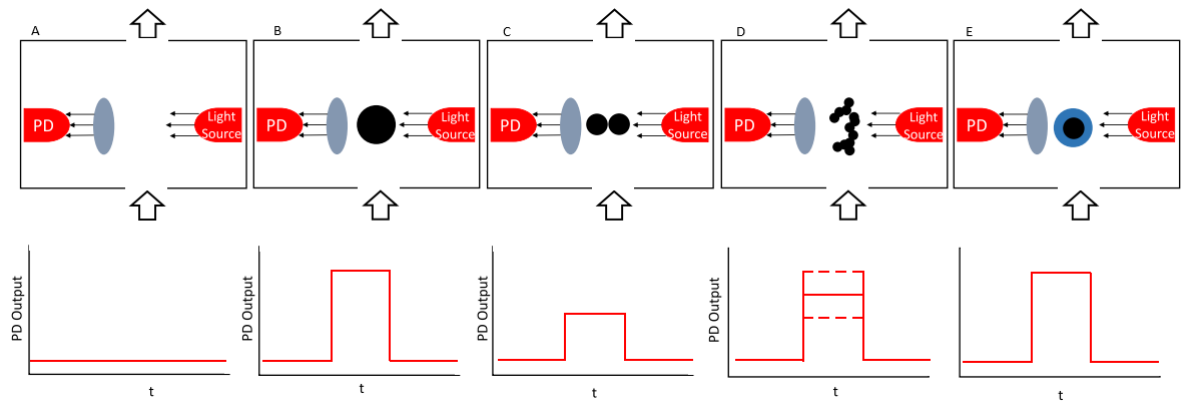
A Figura 6 representa o funcionamento de um sensor óptico por espalhamento de luz, destacando como sinais de partículas de diferentes tamanhos e estados (incluindo partículas úmidas) são interpretados pelo fotodiodo (PD). O aumento da umidade relativa, por exemplo, pode alterar substancialmente a resposta do sensor devido ao crescimento higroscópico das partículas, o que evidencia a necessidade de correções apropriadas (GIORDANO et al., 2021).

Figura 7 – Sensor PurpleAir PA-II-FLEX.



Fonte: Adaptado de PurpleAir Site, 2024

Figura 6 – Funcionamento de um contador óptico de partículas.



Fonte: Adaptado de M.R. Giordano et al., 2021.

Legenda: A) Mostra o sinal do fotodiodo na ausência de partículas; B) indica o sinal produzido por uma partícula de grande tamanho; C) corresponde ao sinal para duas partículas de mesmo tamanho; D) representa o sinal quando há diversas partículas presentes; E) sinal característico para uma partícula úmida.

Os sensores PMS6003 e demais componentes eletrônicos são protegidos por uma cúpula branca (Figura 7), que resguarda o sistema contra precipitação. A mesma figura também mostra o fluxo de entrada e saída de ar na parte inferior do amostrador. As especificações técnicas fornecidas pelo fabricante podem ser consultadas na Tabela 2.

O PA-II-FLEX também incorpora o sensor BME688 (Bosch Sensortech), responsável pela medição de pressão atmosférica, temperatura e umidade relativa. Os dados coletados são enviados via conexão Wi-Fi para a plataforma ThingSpeak ou podem ser armazenados localmente em um cartão MicroSD, garantindo flexibilidade de operação em ambientes sem acesso à internet.

Os sensores Plantower são calibrados de fábrica utilizando material particulado de ambientes urbanos na China (MALINGS et al., 2020b) e reportam concentrações estimadas de partículas nos intervalos de diâmetro aerodinâmico $PM_{10.0}$, $PM_{2.5}$ e $PM_{1.0}$. Os dados de saída são disponibilizados em dois modos: sem correção (cf_1) e corrigidos por um fator empírico (cf_{atm} ¹), ambos definidos pelo fabricante. Em ambientes com concentrações inferiores a $25 \mu\text{g}/\text{m}^3$, ambas as saídas são próximas, com razão $cf_{atm}/cf_1 = 1$; para concentrações mais elevadas, a proporção se aproxima de $2/3$.

Tabela 2 – Especificações técnicas do sensor PMS6003.

Parâmetro	Índice
Intervalo de contagem de partículas	$>0,3\mu\text{m}$; $>0,5\mu\text{m}$; $>1,0\mu\text{m}$; $>2,5\mu\text{m}$; $>10\mu\text{m}$
Eficiência de contagem de partículas	50% até $0,3\mu\text{m}$; 98% $\geq 0,5\mu\text{m}$
Faixa efetiva de partículas (padrão $PM_{2.5}$)	$0-500\mu\text{g}/\text{m}^3$
Faixa máxima de partículas (padrão $PM_{2.5}$)	$\geq 1000\mu\text{g}/\text{m}^3$
Resolução de partículas	$1 \mu\text{g}/\text{m}^3$
Erro máximo de consistência de partículas (dados padrão $PM_{2.5}$)	$\pm 10\mu\text{g}/\text{m}^3 @ 0-100\mu\text{g}/\text{m}^3$
Vazão	$0,1 \text{ L}/\text{min}$
Tempo de resposta único	$<1 \text{ s}$
Tempo de resposta total	$<10 \text{ s}$
Faixa de temperatura de trabalho	$-10 \sim +60^\circ\text{C}$
Faixa de umidade de trabalho	$0-95\%$ (sem condensação)
Faixa de temperatura de armazenamento	$-40 \sim +85^\circ\text{C}$
MTTF (Tempo Médio Antes de Falha)	$\geq 10 \text{ anos}$
Dimensões físicas	$50\text{mm} \times 38\text{mm} \times 21\text{mm}$

Fonte: Plantower, 2022.

3.2.1.1 Coleta e Tratamento dos Dados

A coleta dos dados do PurpleAir PA-II-FLEX foi realizada com resolução temporal de 2 minutos, registrando as leituras do canal cf_1 , além das variáveis meteorológicas

¹ cf_{atm} corresponde aos valores de concentração já corrigidos por um fator empírico implementado pelo próprio fabricante (Plantower), considerando condições típicas de ambientes urbanos.

internas (temperatura, pressão e umidade relativa). Paralelamente, foram coletados dados meteorológicos da estação A807 do INMET, principalmente para análise da influencia da precipitação.

Como o método de referência gravimétrico HI (*Harvard Impactor*) MS&T opera em regime de amostragem diária (ou seja, integrando as partículas depositadas ao longo de 24 horas ou mais), tornou-se necessário compatibilizar as diferentes escalas temporais dos dados para fins de calibração e modelagem. A sincronização temporal foi realizada por meio da extração de estatísticas descritivas (média, mediana, quartis, máximo, mínimo e desvio padrão) das séries temporais do PA-II-FLEX, considerando exclusivamente o período de coleta de cada amostra gravimétrica (intervalo entre data/hora de início e término de cada filtro do HI MS&T).

Esse procedimento permitiu obter, para cada amostra do método de referência, um conjunto de indicadores estatísticos representativos da variabilidade do sensor no mesmo intervalo temporal. Assim, mesmo que o sensor registrasse valores a cada 2 minutos, todos os dados correspondentes ao período de amostragem gravimétrica foram agregados, garantindo a comparabilidade entre as metodologias e possibilitando investigar como diferentes parâmetros estatísticos do sensor influenciam a calibração frente ao valor gravimétrico.

A sincronização temporal dos dados e o cálculo das estatísticas descritivas foram realizados em [Python](#), utilizando as bibliotecas `pandas` e `numpy`, em notebook do *Google Colab*. Dessa forma, cada registro gravimétrico de $PM_{1.0}$ (HI MS&T) possui uma série de variáveis estatísticas extraídas dos dados de alta frequência do sensor, alinhando as bases para aplicação dos modelos de calibração e análise comparativa.

3.2.2 Estimativa da Concentração Mássica por Método Alternativo (ALT)

Além das concentrações fornecidas cf_1 e cf_{atm} , o sensor PMS6003 também disponibiliza como saída a contagem de partículas presentes em 100 mL de ar, discriminadas por intervalos de diâmetro aerodinâmico. Considerando essa possibilidade, a concentração mássica de $PM_{1.0}$ também foi calculada a partir das contagens utilizando o método alternativo (ALT), originalmente proposto por (WALLACE, 2022). Este método foi desenvolvido para contornar as limitações dos algoritmos proprietários dos sensores *Plantower*, que frequentemente subestimam concentrações de material particulado, especialmente sob diferentes condições ambientais. Embora sua aplicação mais detalhada na literatura seja para $PM_{2.5}$, o procedimento foi adaptado neste trabalho para estimar $PM_{1.0}$, tendo em vista as características do sensor e o objetivo do estudo.

O PMS6003 reporta a quantidade de partículas acumuladas por volume nas interva-

los de diâmetro acima de $0,3 \mu\text{m}$, $0,5 \mu\text{m}$ e $1,0 \mu\text{m}$. Para obter o número de partículas em cada intervalo específico, foram calculadas as diferenças entre as contagens acumuladas, de modo que a quantidade de partículas na faixa entre $0,3$ e $0,5 \mu\text{m}$ correspondeu à diferença entre as contagens acima de $0,3 \mu\text{m}$ e $0,5 \mu\text{m}$, e a faixa de $0,5$ a $1,0 \mu\text{m}$ resultou da diferença entre as contagens acima de $0,5 \mu\text{m}$ e $1,0 \mu\text{m}$, denominada contagens diferenciais $N_{0.3-0.5}$ e $N_{0.5-1.0}$.

Supôs-se que todas as partículas presentes em cada intervalo eram esféricas e possuíam diâmetro médio representativo da faixa. Este diâmetro médio foi calculado utilizando diferentes tipos de média (geométrica, aritmética ou harmônica), conforme as Equações 3.2, 3.3 e 3.4, sendo a média geométrica a mais recomendada pela literatura para representar faixas logarítmicas de tamanho (WALLACE, 2022). A escolha da média influencia diretamente a estimativa final de massa, sendo as médias utilizadas:

$$d_{\text{m,arit}} = \frac{d_1 + d_2}{2} \quad (3.1)$$

$$d_{\text{m,geom}} = \sqrt{d_1 \cdot d_2} \quad (3.2)$$

$$d_{\text{m,harm}} = \frac{2d_1d_2}{d_1 + d_2} \quad (3.3)$$

$$(3.4)$$

Com o número de partículas em cada subfaixa e o diâmetro médio estimado, calculou-se o volume total das partículas em cada faixa a partir da expressão para o volume de uma esfera, multiplicado pelo número de partículas: $V = N \cdot (\pi/6)d_{\text{m}}^3$. Para converter o volume em massa, considerou-se a densidade típica dos aerossóis atmosféricos, variando entre $1,0$ e $2,5 \text{ g/cm}^3$, sendo $1,65 \text{ g/cm}^3$ um valor frequentemente adotado em estudos similares (GAUTAM et al., 2023; MARKOWICZ; CHILINSKI, 2020; HINDS, 1999). Vale ressaltar que, como a densidade média do material particulado local não foi determinada experimentalmente, a escolha desses valores pode introduzir um erro significativo na estimativa da concentração mássica, especialmente devido à grande variação entre as densidades adotadas.

A massa resultante foi então multiplicada por um fator de calibração (CF), cujo valor padrão sugerido por Wallace et al. (WALLACE, 2022) é 3 para $\text{PM}_{2.5}$, valor também adotado neste estudo para $\text{PM}_{1.0}$ devido à ausência de recomendações específicas na literatura. Por fim, para expressar o resultado em $\mu\text{g/m}^3$, realizou-se a conversão de unidades, levando em conta a relação entre volume de partículas (em μm^3) e massa (em μg), na forma de:

$$\text{PM}_{1.0}^{\text{ALT}} = [m_{0.3-0.5} + m_{0.5-1.0}] \cdot CF. \quad (3.5)$$

Todas essas operações foram implementadas em *Python*, possibilitando testar diferentes combinações de densidade, tipo de média do diâmetro e fator de calibração.

Para identificar a melhor combinação de densidade e tipo de média do diâmetro a ser utilizada na conversão, foi realizada uma avaliação estatística baseada em regressão linear simples entre as estimativas do método alternativo (ALT) e as concentrações obtidas pelo amostrador de referência. Assim, os parâmetros que apresentaram melhor desempenho em termos de R^2 , RMSE e MAE foram selecionados para as análises subsequentes, garantindo que a adaptação metodológica estivesse ajustada à realidade dos dados locais e ao contexto do experimento.

3.2.3 Amostrador de Referência: *Harvard Impactor* (MS&T)

A quantificação gravimétrica de $PM_{1.0}$ foi realizada com o impactador Harvard, apresentado na Figura 9, originalmente desenvolvido na *Harvard School of Public Health* e amplamente empregado em pesquisas ambientais e epidemiológicas (TURNER; OLSON; ALLEN, 2000; Air Diagnostics and Engineering, Inc.,). O princípio de funcionamento do impactador baseia-se na separação inercial: o fluxo de ar é acelerado pelo bico, direcionando as partículas maiores ($d_a > 1 \mu m$) contra uma placa de impação revestida com óleo mineral, enquanto as partículas menores seguem o fluxo e são coletadas em um filtro de membrana de policarbonato (37 mm).

O bico do HI utilizado neste estudo foi projetado e calibrado para operar a 23 L/min, garantindo um ponto de corte aerodinâmico (d_{50}) de $1,0 \mu m$. Essa configuração específica foi detalhada por Turner et al. (TURNER; OLSON; ALLEN, 2000), que apresentam curvas de eficiência de coleta mostrando transição acentuada ("*sharp cut*") e baixa dispersão ($GSD^2 \approx 1,22$ para $PM_{1.0}$), assegurando seletividade adequada para amostragem de partículas ultrafinas. As perdas entre estágios (*interstage losses*) foram reportadas como inferiores a 4% nessa configuração, indicando alta eficiência.

O sistema de amostragem consistiu em: HI acoplado a uma bomba de vácuo calibrada, um fluxímetro digital para ajuste e conferência do fluxo de amostragem, e um medidor volumétrico (*gas meter*), conforme ilustrado na Figura 8. O ajuste preciso da vazão foi realizado depois da coleta da amostra. O volume total de ar amostrado, horários, datas e eventuais observações foram registrados em planilha específica.

O óleo mineral aplicado na placa de impação minimiza o risco de reentrada ou ressuspensão de partículas, aspecto especialmente relevante para a coleta de frações ultrafinas. O HI, apesar de não possuir certificação EPA como método FEM, tem

² O GSD (*Geometric Standard Deviation*) representa a largura da distribuição dos tamanhos das partículas retidas ou permitidas pelo amostrador; valores próximos de 1 indicam uma separação mais clara.

desempenho amplamente validado e é largamente aceito em pesquisas científicas para monitoramento *outdoor* e *indoor* (TURNER; OLSON; ALLEN, 2000).

Diversas configurações de bico estão disponíveis para diferentes faixas granulométricas ($PM_{2.5}$, PM_{10}), mas neste trabalho optou-se pelo bico calibrado para $PM_{1.0}$ visando a seletividade à fração de interesse. Essa abordagem permite comparar diretamente os dados gravimétricos de referência com os dados do sensor óptico na mesma faixa de diâmetro aerodinâmico. O procedimento de coleta seguiu as recomendações do fabricante e da literatura especializada, garantindo rastreabilidade e robustez metodológica dos dados gravimétricos obtidos.

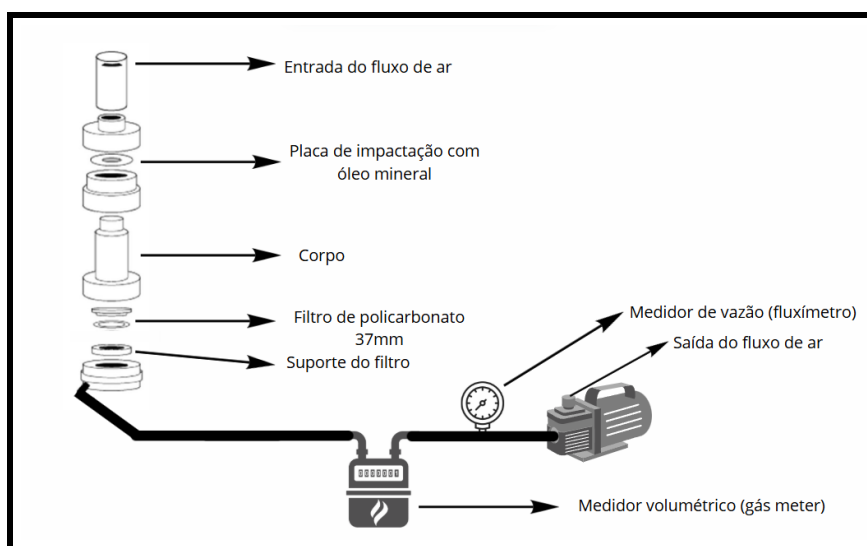


Figura 8 – Esquema do sistema de amostragem utilizando Impactador Harvard, bomba de vácuo, medidor volumétrico e fluxímetro.



Figura 9 – Harvard Impactor MS&T.

Fonte: (Air Diagnostics and Engineering, Inc.,).

3.2.3.1 Análise Gravimétrica

A determinação da concentração mássica de referência de $PM_{1.0}^{HI}$ ($PM_{1.0}^{HI}$) foi realizada a partir da massa depositada no filtro, juntamente com o controle da vazão e do tempo de amostragem. Para isso, a massa depositada foi obtida pela diferença entre os pesos do filtro antes e após a amostragem, utilizando uma microbalança analítica Sartorius Cubis MSA 2.7000-DF, com resolução de $0,1 \mu g$, equipada com eliminador de carga eletrostática (Sartorius Stat-Pen). Antes de cada pesagem, os filtros foram mantidos em dessecador sob temperatura controlada ($20 \pm 1^\circ C$) e umidade relativa ($50 \pm 5\%$) por 24 horas, seguindo o protocolo NIOSH 0500 (1994), visando garantir a remoção de umidade residual. Durante todo o procedimento, o ambiente de pesagem permaneceu nas mesmas condições controladas, de modo a assegurar a estabilidade higroscópica do material.

Cada filtro, incluindo os filtros amostrados e os filtros brancos, foi pesado em triplicata³ antes e depois da amostragem. O desvio padrão das medições foi mantido inferior a $8 \mu g$, assegurando a precisão dos resultados e a confiabilidade dos dados. Os filtros brancos com suspeita de contaminações decorrentes do manuseio, transporte ou do próprio ambiente de pesagem, foram eliminados, garantindo, assim, o controle de qualidade.

A concentração mássica do material particulado ultrafino foi calculada pela seguinte expressão:

$$PM_{1.0}^{HI} (\mu g m^{-3}) = \frac{m (\mu g) \times 1000}{Q (L min^{-1}) \times t (min)} \quad (3.6)$$

onde $PM_{1.0}^{HI}$ é a concentração mássica de $PM_{1.0}$ em $\mu g m^{-3}$, m é a massa depositada no filtro (em μg), Q é a vazão de amostragem (em $L min^{-1}$) e t é o tempo de amostragem (em minutos).

3.3 Desenvolvimento do Modelo de Correção

A calibração de sensores ópticos de baixo custo (LCS) para material particulado é baseada, predominantemente, em modelos de regressão linear. Diversos trabalhos demonstram que, devido à alta linearidade das leituras dos LCS em relação aos métodos de referência, a regressão linear simples ou múltipla é a abordagem mais empregada para a correção dos dados desses sensores (MALINGS et al., 2020b; GIORDANO et al., 2021). Modelos não-lineares podem ser testados para investigar comportamentos não-lineares, mas geralmente não promovem ganhos substanciais em relação ao modelo linear na maioria dos contextos de campo.

³ A pesagem em triplicata consiste em realizar três medições independentes de massa para cada filtro, antes e depois da amostragem, adotando-se como valor final a média aritmética dos resultados.

No presente trabalho, as análises das regressões lineares foram conduzidas utilizando o pacote `scikit-learn` no *Google Colab*, empregando o algoritmo de Mínimos Quadrados Ordinários (MQO) implementado por `LinearRegression` (PEDREGOSA et al., 2011). O processo de ajuste e validação dos modelos foi realizado conforme descrito no notebook em *Python*.

3.3.1 Regressão Linear Simples (ULR)

A regressão linear simples estabelece uma relação direta entre a concentração de $PM_{1.0}$ determinada pelo instrumento de referência ($PM_{1.0}^{HI}$) e a reportada pelo sensor ($PM_{1.0}^{PA}$), conforme a Equação 3.7:

$$PM_{1.0}^{HI} = \beta + \alpha \cdot PM_{1.0}^{PA} + \epsilon \quad (3.7)$$

Em que β é o coeficiente linear, α é o coeficiente angular (inclinação da reta) e ϵ representa o termo de erro aleatório.

Os parâmetros são estimados por MQO, minimizando a soma dos quadrados dos resíduos entre os valores observados e os estimados pelo modelo (STOCK; WATSON, 2010). Apesar de sua simplicidade, este modelo pode apresentar limitações quando variáveis ambientais têm influência significativa, tornando seu uso mais apropriado para condições controladas ou faixas restritas de concentração.

3.3.2 Regressão Linear Múltipla (MLR)

A regressão linear múltipla permite a inclusão de variáveis meteorológicas e ambientais como preditoras adicionais, o que frequentemente resulta em ganhos de desempenho (OKAFOR, 2023; SPINELLE et al., 2014; GIORDANO et al., 2021). A forma geral pode ser expressa pela Equação 3.8:

$$PM_{1.0}^{HI} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon \quad (3.8)$$

onde X_i representa as variáveis independentes utilizadas (como $PM_{1.0}^{PA}$, umidade relativa, temperatura, pressão atmosférica, entre outras). A inclusão dessas variáveis auxilia na correção de vieses introduzidos por condições ambientais variáveis, comum em aplicações urbanas e de campo.

Porém, esses métodos consideram uma relação de linearidade entre as variáveis, o que pode comprometer a precisão dos resultados em cenários onde há relações não lineares entre as variáveis (SPINELLE et al., 2014).

3.3.3 *Random Forest Regressor*

O método *Random Forest Regressor* (RF) foi empregado para o desenvolvimento dos modelos de calibração do sensor PurpleAir PMS6003, visando aprimorar a acurácia na estimativa da concentração de material particulado $PM_{1.0}$. Os *Random Forests*, introduzidos por Breiman (BREIMAN, 2001), consistem em um conjunto (*ensemble*) de árvores de decisão independentes, cada uma treinada sobre diferentes subconjuntos dos dados e variáveis, combinando seus resultados para aumentar a robustez e reduzir a variância preditiva. Cada árvore é treinada a partir de uma amostra dos dados originais (*bagging*), e a escolha dos pontos de divisão nos nós é restrita a subconjuntos aleatórios das variáveis predictoras, diminuindo a correlação entre as árvores e fortalecendo a generalização do modelo.

A Figura 10 ilustra esquematicamente o funcionamento do método *Random Forest*, mostrando como diferentes árvores de decisão processam subconjuntos dos dados e variáveis e, em seguida, produzem uma predição agregada por média ou moda, finalizando em um resultado final mais robusto e preciso.

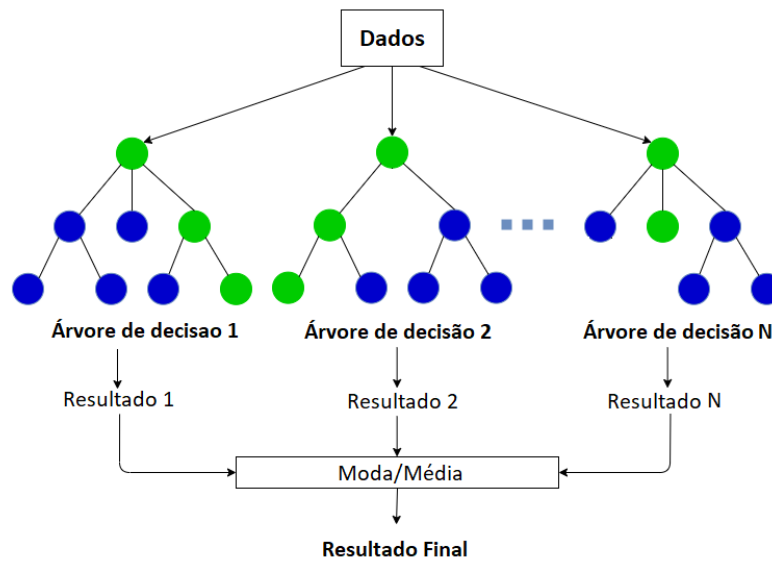


Figura 10 – Esquema conceitual do funcionamento do Random Forest, ilustrando a combinação de múltiplas árvores de decisão independentes e agregação dos resultados.

Fonte: statplace blog, 2021.

A implementação também foi realizada no ambiente [Python](#), utilizando o pacote `scikit-learn` (`sklearn.ensemble.RandomForestRegressor`) (PEDREGOSA et al., 2011). O algoritmo constrói um conjunto de árvores de decisão para problemas de regressão e retorna, como predição final, a média dos valores preditos por todas as árvores da floresta. Essa abordagem é especialmente recomendada para lidar com relações não

lineares, dados ruidosos e interações complexas entre variáveis, cenários típicos de séries temporais ambientais.

A escolha das variáveis preditoras no modelo incluiu não apenas as concentrações de $PM_{1.0}$ fornecidas pelo sensor óptico, mas também variáveis meteorológicas (temperatura, umidade relativa e pressão) registradas pelo próprio sensor, com o intuito de contemplar fatores influentes sobre o desempenho dos sensores ópticos em ambientes reais (GIORDANO et al., 2021; MALINGS et al., 2020b).

Buscando prevenir sobreajuste (*overfitting*) e otimizar o desempenho do modelo, foi conduzida uma análise sistemática do número de árvores ($n_estimators$) no *ensemble* (BREIMAN, 2001; PEDREGOSA et al., 2011). O número ótimo de árvores foi definido com base na análise da convergência das métricas de desempenho (R^2 , RMSE) para diferentes valores de $n_estimators$.

Outro aspecto fundamental foi a avaliação da importância das variáveis (*feature importance*), estimada automaticamente pelo algoritmo, com base na redução acumulada do erro quadrático médio (*mean squared error reduction*) atribuída a cada variável ao longo das divisões dos nós das árvores. Esta métrica fornece *insights* relevantes sobre os fatores de influência e apoia a seleção e interpretação dos resultados. Ressalta-se, no entanto, que a importância relativa das variáveis pode apresentar variações entre execuções, devido à natureza estocástica⁴ do algoritmo e à possível presença de multicolinearidade entre preditores.

Os resultados do *Random Forest Regressor* foram comparados aos modelos lineares tradicionais (regressão linear simples e múltipla), tanto em termos de ajuste quanto de capacidade preditiva, utilizando métricas descritas no próximo capítulo (HOLDER et al., 2020; GIORDANO et al., 2021).

Embora modelos de *Random Forest* apresentem vantagens em robustez, flexibilidade e capacidade de capturar não linearidades, sua aplicação requer cuidados em conjuntos de dados muito pequenos ou altamente correlacionados, contextos nos quais a interpretação das importâncias das variáveis e a própria generalização do modelo podem ser limitadas (BREIMAN, 2001; HOLDER et al., 2020). Por fim, destaca-se que todas as variáveis foram utilizadas em sua escala original, aproveitando a insensibilidade do algoritmo à normalização ou padronização de variáveis contínuas.

⁴ No *Random Forest*, isso ocorre na seleção de amostras e das variáveis preditoras a cada divisão de nó, de modo que execuções repetidas com os mesmos dados podem resultar em saídas diferentes.

3.4 Validação e Avaliação dos Modelos

3.4.1 Divisão dos dados: *Holdout* e *K-Fold Cross Validation*

A avaliação rigorosa do desempenho preditivo dos modelos de calibração de sensores requer a aplicação de técnicas de divisão de dados para simular cenários de dados desconhecidos e evitar o sobreajuste (*overfitting*). Neste estudo, foram empregados dois métodos amplamente reconhecidos na literatura: a divisão simples (*holdout*) e a validação cruzada *k*-fold (*k-Fold Cross Validation*), ambos implementados via pacote `scikit-learn` (PEDREGOSA et al., 2011), por meio das funções `train_test_split` e `KFold`.

No método *holdout*, o conjunto de dados original foi particionado aleatoriamente em dois subconjuntos mutuamente exclusivos, com 80% dos dados reservados para treino e 20% para teste. O modelo foi ajustado utilizando apenas os dados de treino e sua performance posteriormente avaliada no conjunto de teste, fornecendo uma estimativa direta da capacidade preditiva do modelo para novos dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Já na validação cruzada *k*-fold, adotou-se $k = 5$, pois valores mais elevados poderiam resultar em subconjuntos muito pequenos, o que aumenta a variabilidade dos resultados e reduz a representatividade de cada partição. O conjunto de dados foi dividido em cinco subconjuntos de tamanho aproximadamente igual, sendo que, em cada rodada, quatro deles foram utilizados para treino e um para teste. Esse processo foi repetido até que cada subconjunto tivesse sido utilizado uma vez como teste, e os resultados finais foram obtidos pela média das métricas de desempenho em cada rodada. A Figura 11 ilustra esquematicamente o procedimento adotado para $k = 5$, destacando os conjuntos de treino e teste em cada *fold*, bem como o cálculo da média e desvio padrão do coeficiente de determinação (R^2) entre as rodadas.

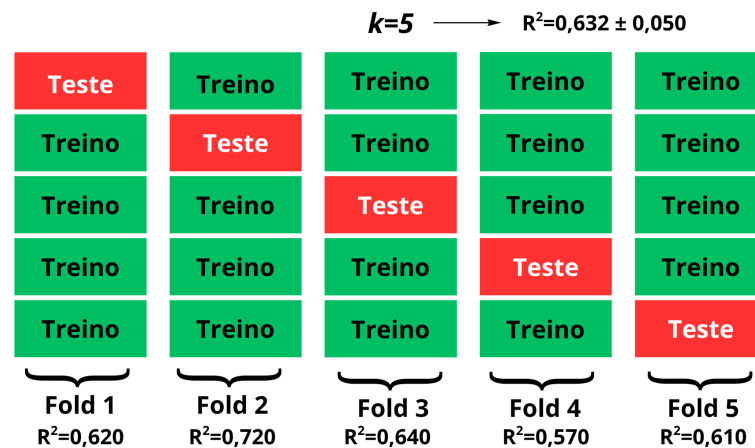


Figura 11 – Esquema ilustrativo da validação cruzada *k*-fold ($k = 5$), mostrando a alternância entre conjuntos de treino (verde) e teste (vermelho) em cada *fold* e o cálculo dos resultados médios e variabilidade das métricas.

Ambas as estratégias foram aplicadas para cada modelo e para cada combinação de variáveis testada (incluindo modelos lineares e não lineares), assegurando uma comparação justa e padronizada entre diferentes abordagens de calibração e seleção de variáveis. Essa abordagem reflete as recomendações da literatura internacional para evitar vieses na avaliação, especialmente em conjuntos de dados de dimensão limitada, frequentes em estudos experimentais com amostragem gravimétrica (GIORDANO et al., 2021; MALINGS et al., 2020b).

3.4.2 Métricas de Desempenho dos Modelos

A avaliação do desempenho dos modelos de calibração foi conduzida a partir de um conjunto de métricas estatísticas amplamente utilizadas para estudos de calibração de sensores ambientais (GIORDANO et al., 2021; HOLDER et al., 2020). As principais métricas consideradas foram o Coeficiente de Determinação (R^2), a Raiz do Erro Quadrático Médio (RMSE), o Erro Absoluto Médio (MAE) e o Viés Normalizado Médio (MNB). Tais métricas permitiram quantificar não apenas a força da relação entre as variáveis, mas também a precisão e o viés dos modelos ajustados.

O Coeficiente de Determinação (R^2) foi empregado para quantificar a proporção da variabilidade dos dados de referência explicada pelas estimativas dos modelos de correção. Ressalta-se, entretanto, que o R^2 não necessariamente expressa a concordância ponto a ponto entre os valores previstos e observados, podendo ser elevado mesmo na presença de desvios sistemáticos (ALEXANDER; TROPSHA; WINKLER, 2015). Por esse motivo, a avaliação foi complementada com métricas de erro absoluto e quadrático, que fornecem uma visão mais abrangente da acurácia dos modelos.

O Viés Normalizado Médio (MNB) foi calculado para detectar tendências sistemáticas de superestimação ou subestimação dos modelos ao longo do período avaliado. Valores de MNB próximos de zero indicam baixo viés, enquanto valores positivos ou negativos revelam, respectivamente, tendência à superestimação ou subestimação.

O Erro Absoluto Médio (MAE) quantifica o erro médio absoluto das previsões, sendo robusto frente a grandes outliers. O cálculo do MAE foi realizado considerando o valor absoluto da diferença entre os valores estimados e observados, conforme a Equação :

$$MAE = \frac{1}{n} \sum_{i=1}^n |PM_{1.0}^{PA} - PM_{1.0}^{HI}| \quad (3.9)$$

A Raiz do Erro Quadrático Médio (RMSE) foi utilizada para penalizar erros de maior magnitude, sendo particularmente útil para destacar dispersões elevadas entre previsto e observado, dado pela Equação :

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(PM_{1.0}^{PA} - PM_{1.0}^{HI} \right)^2} \quad (3.10)$$

O Viés Normalizado Médio (MNB) foi calculado segundo a Equação :

$$MNB = \frac{\sum_{i=1}^n \left(PM_{1.0}^{PA} - PM_{1.0}^{HI} \right)}{\sum_{i=1}^n PM_{1.0}^{HI}} \quad (3.11)$$

Adicionalmente, para possibilitar a comparação dos resultados com trabalhos realizados em outras localidades e faixas de concentração distintas, foram calculadas as versões normalizadas das métricas de erro, a saber: o MAE normalizado pelo coeficiente de variação (CvMAE) e o RMSE normalizado (nRMSE). Essas métricas são recomendadas para análise interlaboratorial e comparações internacionais, já que eliminam o efeito das diferentes magnitudes de concentração média em cada local, dada pelas Equações e :

$$CvMAE = \frac{MAE}{\overline{PM_{1.0}^{HI}}} \quad (3.12)$$

$$nRMSE = \frac{RMSE}{\overline{PM_{1.0}^{HI}}} \quad (3.13)$$

Onde $\overline{PM_{1.0}^{HI}}$ representa a média das concentrações de referência obtidas pelo amostrador gravimétrico.

Dessa forma, as métricas convencionais (R^2 , RMSE, MAE e MNB) foram utilizadas como principais indicadores de desempenho, enquanto as métricas normalizadas (CvMAE e nRMSE) foram apresentadas principalmente para fins de comparação futura com outros estudos desenvolvidos sob diferentes faixas de concentração.

3.4.3 Análise dos Resíduos e Diagnóstico dos Modelos

A avaliação criteriosa dos resíduos dos modelos de calibração é essencial para validar os pressupostos da regressão e garantir a robustez das conclusões estatísticas (MONTGOMERY; PECK; VINING, 2012). Os resíduos (e_i) são definidos como as diferenças entre os valores observados (y_i) e os valores previstos pelo modelo ($\hat{y}_i$), ou seja,

$$e_i = y_i - \hat{y}_i. \quad (3.14)$$

O diagnóstico dos resíduos permite identificar a presença de padrões sistemáticos, tendências ou desvios que possam indicar inadequações do modelo, como não linearidade, heterocedasticidade ou autocorrelação.

Entre as métricas utilizadas, o erro absoluto dos resíduos é uma medida fundamental da dispersão dos erros, na forma:

$$|e_i| = |y_i - \hat{y}_i| \quad (3.15)$$

Esse valor, quando analisado em conjunto com outras métricas (MAE, RMSE), contribui para uma visão abrangente da qualidade do ajuste.

Para examinar a normalidade dos resíduos, requisito para a validade dos intervalos de confiança e testes paramétricos, aplicou-se o teste de Shapiro-Wilk (SHAPIRO; WILK, 1965). O teste avalia a hipótese nula de que a amostra provém de uma distribuição normal; valores de p -valor inferiores a 0,05 indicam desvio significativo da normalidade. Os resíduos também foram avaliados visualmente por histogramas e gráficos.

A homocedasticidade⁵ foi verificada com o teste de Breusch-Pagan (BREUSCH; PAGAN, 1979), que examina a dependência da variância dos resíduos em relação aos valores ajustados do modelo. A rejeição da hipótese nula (constância da variância) sugere a necessidade de ajustes no modelo, como transformações de variáveis ou adoção de métodos robustos.

Além dos testes formais, indicadores estatísticos de assimetria (*skewness*) e curtose (*kurtosis*) foram calculados, fornecendo informações adicionais sobre a simetria e a propensão a valores extremos nos resíduos. Valores de assimetria próximos de zero e curtose próxima de três são indicativos de distribuição aproximadamente normal.

O diagnóstico detalhado dos resíduos possibilitou identificar eventuais limitações do ajuste e serviu de base para as interpretações dos resultados e recomendações de aprimoramento do modelo.

⁵ Pressuposto de que a variância dos resíduos permanece constante para todos os níveis dos valores previstos do modelo, e quando é violado, a precisão dos intervalos de confiança e testes de hipóteses é comprometida, podendo tornar os estimadores ineficientes.

4 Resultados

4.1 Análise dos Dados Brutos

Antes da implementação dos modelos de calibração, realizou-se uma análise exploratória dos dados brutos obtidos pelo sensor óptico do PurpleAir PA-II-FLEX e pelo método gravimétrico de referência (Impactador Harvard MS&T). Essa etapa é fundamental para caracterizar o comportamento inicial das medições, identificar padrões, detectar eventuais inconsistências e embasar a necessidade de procedimentos de correção.

A Figura 12 apresenta a série temporal das concentrações de $PM_{1.0}$ registradas pelos sensores A e B do PurpleAir PA-II-FLEX ao longo do período de amostragem. É possível observar que, enquanto o sensor A apresenta um comportamento estável e valores compatíveis com o esperado para o ambiente monitorado, o sensor B registrou um aumento abrupto e persistente nas concentrações, com valores anômalos que ultrapassam $3.000 \mu g/m^3$ durante um extenso intervalo de tempo, além de outros picos isolados ao longo da campanha.

Além das discrepâncias observadas na concentração de $PM_{1.0}$, constatou-se que os dados meteorológicos associados ao sensor B (temperatura, pressão e umidade relativa) permaneceram zerados durante todo o período de amostragem, o que indica um possível mau funcionamento do sensor ou falha na aquisição de dados deste canal. Esse padrão anômalo inviabiliza o uso dos dados do Sensor B tanto para análise de séries temporais quanto para aplicação de modelos de calibração que dependem de variáveis meteorológicas.

Diante desse contexto, optou-se por utilizar exclusivamente os dados do Sensor A para o desenvolvimento e calibração dos modelos de correção. Essa escolha assegura a confiabilidade das análises estatísticas e a consistência na incorporação das variáveis ambientais relevantes.

A Figura 13 apresenta os histogramas sobrepostos das concentrações de material particulado ultrafino ($PM_{1.0}$), obtidas tanto pelo sensor óptico de baixo custo PurpleAir PA-II-FLEX quanto pelo método gravimétrico de referência (Impactador Harvard MS&T). Nesta análise exploratória, os valores são expressos em frequência relativa (%), o que permite uma comparação direta entre as distribuições.

Observa-se que as concentrações medidas pelo sensor PurpleAir apresentam uma distribuição significativamente mais ampla e dispersa, alcançando valores superiores a $50 \mu g/m^3$. Em contraste, os resultados obtidos pela referência gravimétrica concentram-se majoritariamente em faixas mais baixas, tipicamente inferiores a $12 \mu g/m^3$. Essa diferença sugere que o sensor óptico tende a registrar uma proporção maior de valores elevados em

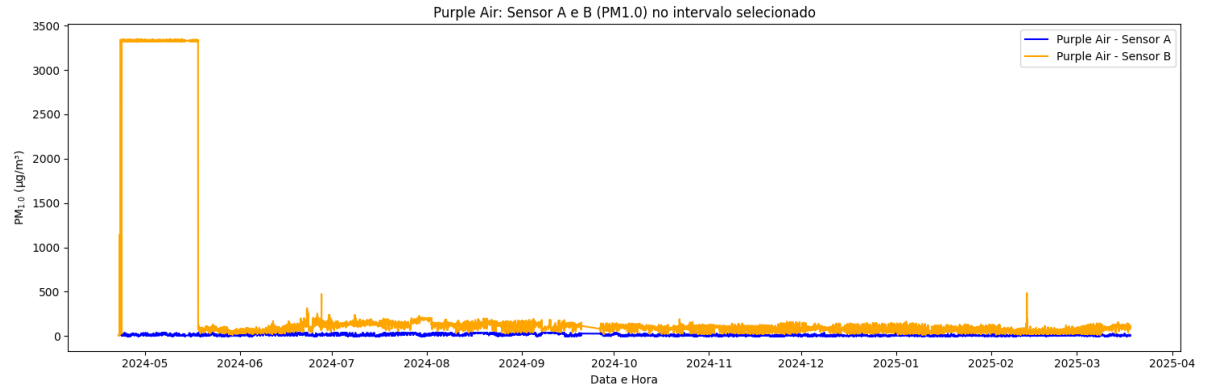


Figura 12 – Série temporal das concentrações de $PM_{1.0}$ registradas pelos canais A e B do sensor PurpleAir PA-II-FLEX no período de amostragem.

Fonte: O autor, 2025.

comparação ao método gravimétrico padrão.

As diferenças identificadas nessa etapa exploratória demonstram que, apesar da utilidade do sensor para medições em tempo real e em larga escala, ainda são necessárias correções e procedimentos de calibração para que seus resultados se aproximem daqueles obtidos por métodos de referência. A etapa seguinte deste estudo será dedicada à avaliação e aplicação de técnicas de calibração adequadas, visando aprimorar a precisão e a comparabilidade dos dados fornecidos pelo PurpleAir em relação ao método gravimétrico.

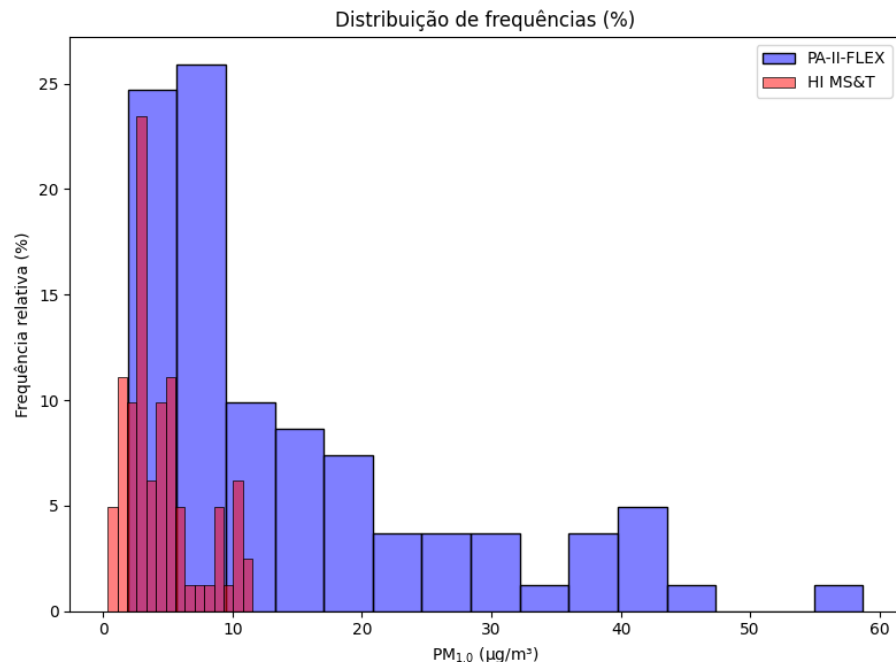


Figura 13 – Histogramas sobrepostos das concentrações de $PM_{1.0}$ medidas pelo sensor óptico PurpleAir (PA-II-FLEX) e pelo método gravimétrico de referência (HI MS/T).

A Figura 14 apresenta o boxplot comparativo das concentrações de $PM_{1.0}$ obtidas

pelo PurpleAir PA-II-FLEX e pelo método de referência gravimétrico (Impactador Harvard MST). O gráfico fornece informações importantes sobre a dispersão, tendência central, presença de valores discrepantes (*outliers*) e diferenças gerais entre os métodos de medição.

Observa-se que o PA-II-FLEX apresenta uma mediana e média mais elevadas (média = $14,94 \mu\text{g}/\text{m}^3$), além de uma ampla faixa de variação e presença de vários valores considerados *outliers*. Esses valores discrepantes acima de $40 \mu\text{g}/\text{m}^3$ indicam leituras significativamente superiores às concentrações típicas observadas pelo método de referência.

Por outro lado, o método gravimétrico (HI MS/T) exibe valores consideravelmente menores e menos dispersos, com uma média de $4,46 \mu\text{g}/\text{m}^3$. A caixa menor e menos dispersa do método de referência sugere maior estabilidade nas medições e reforça a diferença sistemática observada entre os dois métodos.

Essas diferenças destacam claramente a necessidade e relevância da aplicação de técnicas de calibração para corrigir e ajustar as medições feitas pelo sensor óptico. Tal ajuste visa reduzir a variabilidade, melhorar a exatidão das medições e aumentar a comparabilidade com os dados obtidos pelo método de referência padrão.

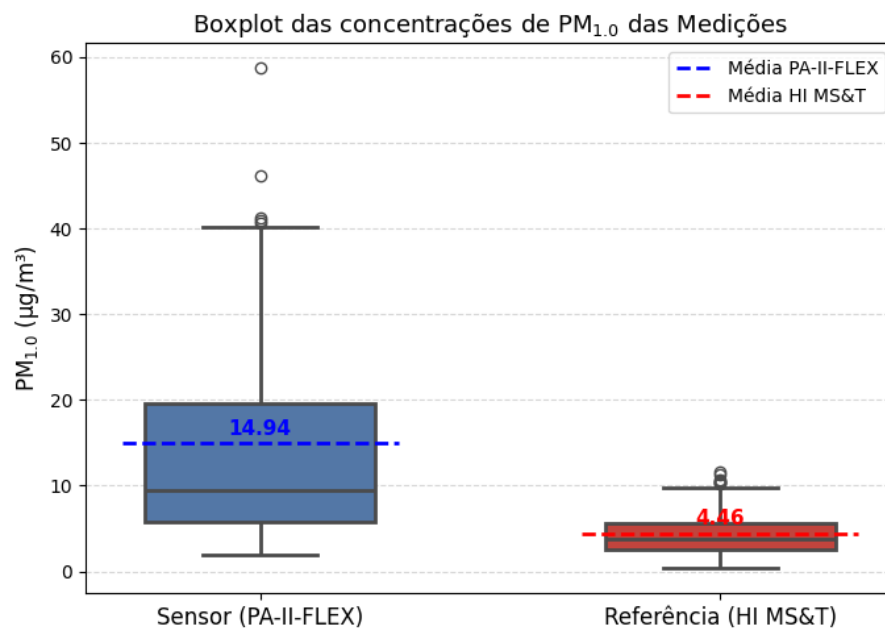


Figura 14 – Boxplot das concentrações de $\text{PM}_{1.0}$ obtidas pelo sensor óptico (PA-II-FLEX) e método gravimétrico de referência (HI MS/T). As linhas pontilhadas representam as médias das respectivas distribuições.

A Figura 15 apresenta a série temporal das concentrações de $\text{PM}_{1.0}$ obtidas pelo sensor óptico PurpleAir PA-II-FLEX e pelo método gravimétrico de referência (Impactador Harvard MST), durante o período de amostragem realizado entre abril de 2024 e março de 2025, em Curitiba, Paraná.

Observa-se uma clara discrepância entre as concentrações medidas pelo sensor

e as medições de referência ao longo do tempo. O sensor PA-II-FLEX frequentemente registrou valores superiores, especialmente nos meses de junho a setembro de 2024. Esse comportamento pode ser atribuído à maior estabilidade atmosférica durante o inverno, associada a inversões térmicas, redução da altura da camada de mistura e menor ventilação, o que dificulta a dispersão de poluentes na atmosfera urbana (Ministério do Meio Ambiente, 2014; CETESB, 2023). Além disso, sensores ópticos de baixo custo, como o PA-II-FLEX, tendem a superestimar os valores de $PM_{1.0}$ em ambientes com umidade elevada ou condições estáveis sem calibração adequada (MALINGS et al., 2020a; CASTELL et al., 2017b).

A partir de outubro de 2024, com o início da primavera e posteriormente o verão, observa-se uma redução nas concentrações medidas tanto pelo sensor quanto pelo método de referência. Esse comportamento é consistente com o aumento da precipitação e da convecção atmosférica nessas estações, favorecendo maior dispersão e deposição úmida dos poluentes.

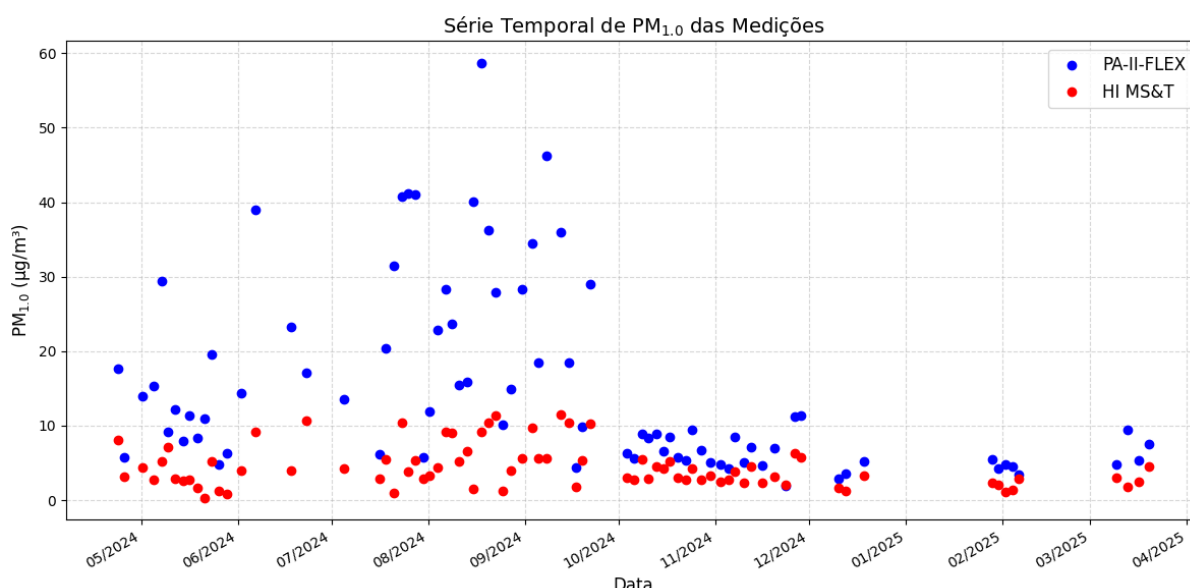


Figura 15 – Série temporal das concentrações médias diárias de $PM_{1.0}$ obtidas pelo sensor PA-II-FLEX e pelo método gravimétrico HI MS&T, no período entre abril de 2024 e março de 2025, em Curitiba (PR).

A Figura 16 apresenta o gráfico de dispersão entre as concentrações de $PM_{1.0}$ obtidas pelo sensor óptico PA-II-FLEX (eixo y) e pelo método gravimétrico de referência HI MS&T (eixo x) para todo o período de amostragem. Cada ponto do gráfico representa uma amostra coletada em um determinado intervalo temporal. Observa-se que a maioria dos pontos está concentrada acima da linha de identidade (linha 1:1), indicando que o sensor PA-II-FLEX sistematicamente reporta concentrações de $PM_{1.0}$ superiores às observadas pelo método de referência.

Além disso, a dispersão dos pontos evidencia que a diferença entre as medições dos dois métodos se acentua especialmente em concentrações mais elevadas, sugerindo uma

resposta não linear ou sensibilidade diferenciada do sensor PA-II-FLEX em determinados intervalos de concentração. Esse padrão já foi relatado em estudos prévios envolvendo sensores ópticos de baixo custo, que frequentemente apresentam superestimação em condições ambientais específicas ou na presença de partículas higroscópicas (MALINGS et al., 2020a; CASTELL et al., 2017b).

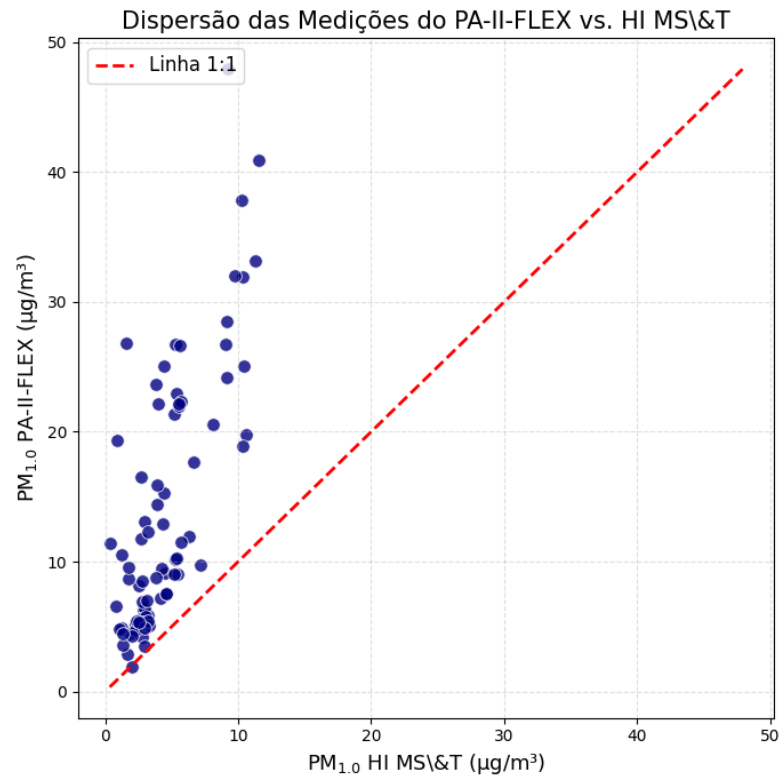


Figura 16 – Gráfico de dispersão das concentrações de $PM_{1.0}$ do sensor PA-II-FLEX e do método gravimétrico HI MS&T.

4.2 Correlação Inicial entre Variáveis

4.2.1 Matriz de Correlação entre Variáveis Ambientais e Concentrações de $PM_{1.0}$

A Figura 17 apresenta a matriz de correlação de Pearson entre as concentrações de $PM_{1.0}$ medidas pelo sensor óptico PA-II-FLEX e pelo método de referência HI MS&T, juntamente com as principais variáveis ambientais: umidade relativa, temperatura média, precipitação acumulada e pressão atmosférica média durante cada período amostral.

Observa-se que existe uma correlação positiva moderada entre o sensor PA-II-FLEX e o método de referência HI MS&T ($r = 0,62$), indicando que ambos captam tendências similares de variação temporal do material particulado ultrafino, ainda que com diferenças quantitativas já discutidas anteriormente.

Um dos achados mais relevantes da matriz é a forte correlação negativa entre as concentrações do sensor PA-II-FLEX e a umidade relativa ($r = -0,72$), enquanto o método de referência apresenta uma correlação negativa um pouco menor com essa variável ($r = -0,61$). Esse resultado corrobora com a literatura, que aponta que sensores ópticos de baixo custo tendem a superestimar ou apresentar maior incerteza nas medições em ambientes com alta umidade, devido ao crescimento higroscópico das partículas e à interferência na leitura óptica (MALINGS et al., 2020a; CASTELL et al., 2017b).

A pressão atmosférica média apresentou uma correlação positiva moderada com o sensor ($r = 0,31$), e praticamente nula com o método de referência, enquanto a temperatura média exibiu uma correlação baixa ou negativa com as demais variáveis, sendo negativa com a pressão ($r = -0,64$) e com a umidade ($r = -0,19$). Isso indica que variações de temperatura e pressão podem influenciar, de forma secundária, o comportamento dos sensores, mas não são tão relevantes quanto a umidade relativa.

A precipitação acumulada teve correlação negativa com as concentrações de $PM_{1.0}$ em ambos os métodos ($r = -0,31$ para o sensor e $r = -0,27$ para o método de referência), o que está de acordo com o mecanismo de deposição úmida, no qual eventos de chuva promovem a remoção das partículas suspensas do ar.

Esses resultados apontam que, para aprimorar a calibração dos sensores ópticos de baixo custo, é essencial considerar especialmente a umidade relativa e, em menor grau, a precipitação acumulada e a pressão atmosférica como variáveis preditoras nos modelos de correção.

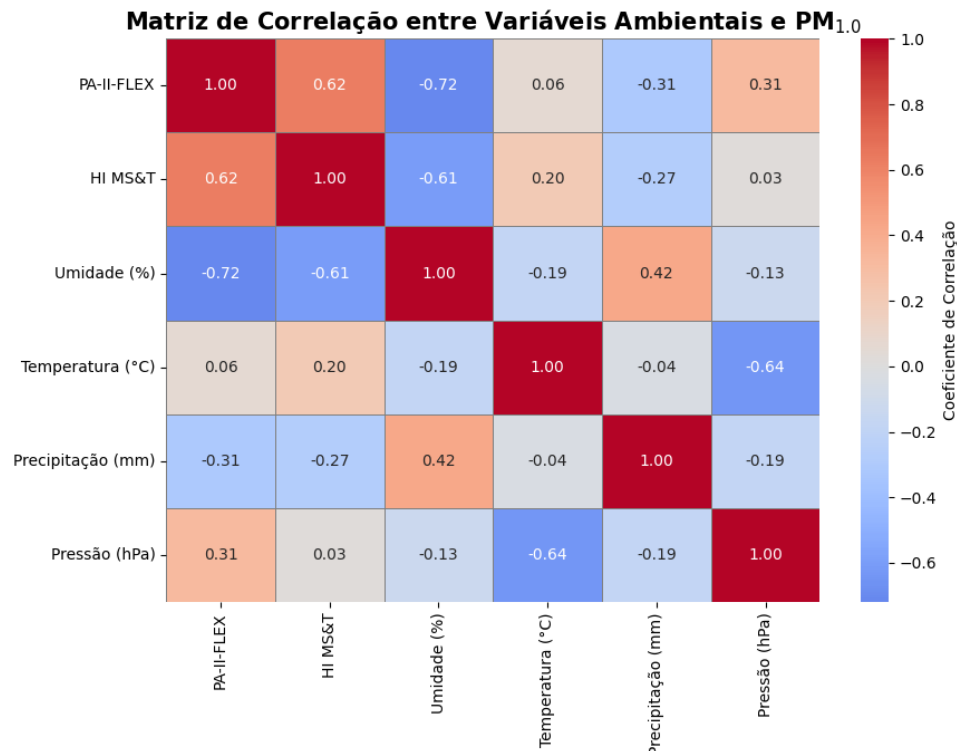


Figura 17 – Matriz de correlação de Pearson entre as concentrações de PM_{1.0} (sensor e referência) e variáveis ambientais.

4.2.2 Correlação entre HI MS&T e Estatísticas Diárias das Variáveis

A Figura 18 apresenta a matriz de correlação entre os valores diários de PM_{1.0} obtidos pelo método de referência HI MS&T e as principais estatísticas descritivas das variáveis ambientais (umidade relativa, temperatura e pressão atmosférica) e do sensor PA-II-FLEX, considerando média, mediana, máximos, mínimos, desvio padrão e quartis para cada período de amostragem.

Os resultados evidenciam que as estatísticas derivadas do PA-II-FLEX, especialmente o primeiro quartil das concentrações de PM_{1.0}, a mediana, a média e o valor mínimo, apresentam as correlações positivas mais altas com o valor de referência, destacando-se o primeiro quartil das concentrações do sensor ($r = 0,69$), a mediana das concentrações de PM_{1.0} ($r = 0,64$), a média das concentrações de PM_{1.0} ($r = 0,62$) e o valor mínimo registrado de PM_{1.0} ($r = 0,61$). Esses resultados sugerem que as medidas correspondentes à faixa inferior e à tendência central das medições contínuas do sensor são mais representativas do valor gravimétrico diário, indicando possível influência de episódios de pico ou de variabilidade ao longo do dia.

Em contrapartida, as estatísticas relacionadas à umidade relativa apresentam correlações negativas expressivas, com destaque para a mediana da umidade relativa

($r = -0,63$), a média da umidade relativa ($r = -0,61$) e o primeiro quartil da umidade relativa ($r = -0,55$). Isso reforça a influência da umidade relativa na dinâmica do $PM_{1.0}$ atmosférico e também nas respostas do sensor óptico, conforme apontado em diversos estudos da literatura (MALINGS et al., 2020a; CASTELL et al., 2017b; JAYARATNE, 2018). Valores elevados de umidade tendem a estar associados a menores concentrações gravimétricas, seja por deposição úmida ou por interferências ópticas nos sensores.

Para temperatura e pressão atmosférica, as correlações foram em geral baixas, com valores próximos de zero para a maioria das estatísticas, indicando influência marginal dessas variáveis no contexto das amostragens realizadas. A exceção foi o desvio padrão da temperatura, com correlação moderada ($r = 0,37$), sugerindo que a variabilidade térmica diária pode ter algum impacto sobre a formação e a dispersão das partículas finas.

Esses resultados ressaltam que, para modelos de calibração e análise preditiva envolvendo sensores ópticos de baixo custo, pode ser recomendável utilizar múltiplas estatísticas das variáveis ambientais e do próprio sensor, com ênfase nas medidas de tendência central e de dispersão da série temporal, além do controle rigoroso dos efeitos da umidade relativa.

4.2.3 Matriz de Dispersão das Variáveis

A Figura 19 apresenta a matriz de dispersão entre as principais variáveis do estudo: $PM_{1.0}$ medido pelo sensor PA-II-FLEX, $PM_{1.0}$ pelo método de referência HI MS&T, umidade relativa média, temperatura média, precipitação acumulada e pressão atmosférica média em cada período amostral.

Nota-se que a relação entre o sensor e o método de referência não é perfeitamente linear, mas evidencia uma tendência positiva, com alguns pontos dispersos indicando episódios de superestimação pelo sensor, especialmente em concentrações mais elevadas. Essa característica é relatada na literatura, que destaca que sensores ópticos de baixo custo frequentemente apresentam viés positivo e maior dispersão em valores altos de PM, demandando procedimentos de calibração adequados (MALINGS et al., 2020a; CASTELL et al., 2017b; JAYARATNE, 2018).

Observa-se ainda um padrão negativo marcante entre as variáveis $PM_{1.0}$ (de ambos os métodos) e a umidade relativa, com nuvem de pontos inclinada para baixo, reforçando a influência da umidade na resposta dos sensores ópticos, que tendem a superestimar concentrações sob alta umidade ambiental (CRILLEY, 2020; JAYARATNE, 2018; MALINGS et al., 2020a). Por outro lado, a relação com temperatura e pressão atmosférica é menos definida, com dispersão relativamente homogênea dos pontos, sugerindo que esses fatores desempenham papel secundário no comportamento das medições de $PM_{1.0}$ no contexto do presente estudo.

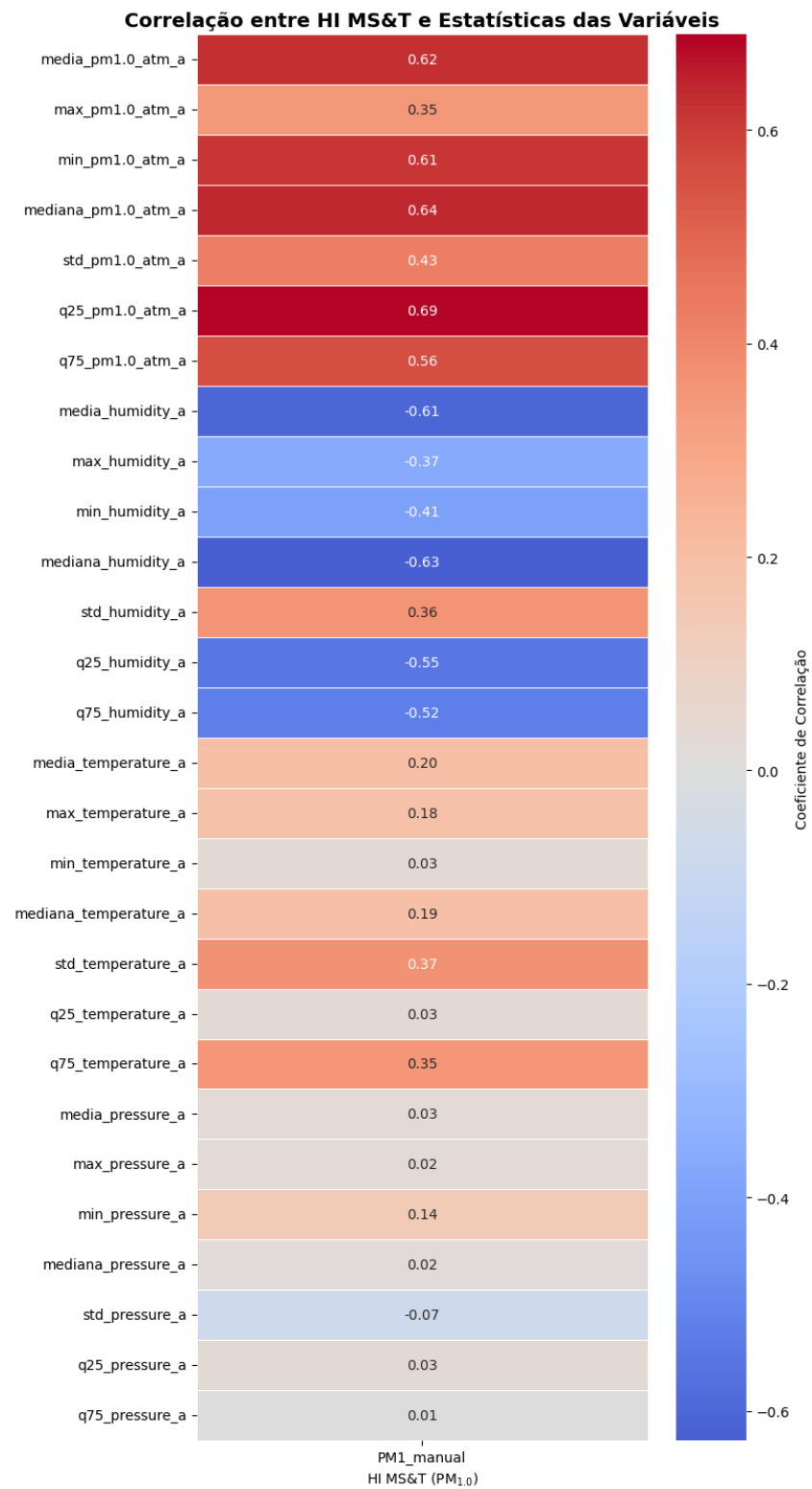


Figura 18 – Matriz de correlação entre HI MS&T (PM_{1.0} manual) e as estatísticas diárias das variáveis ambientais e do sensor PA-II-FLEX.

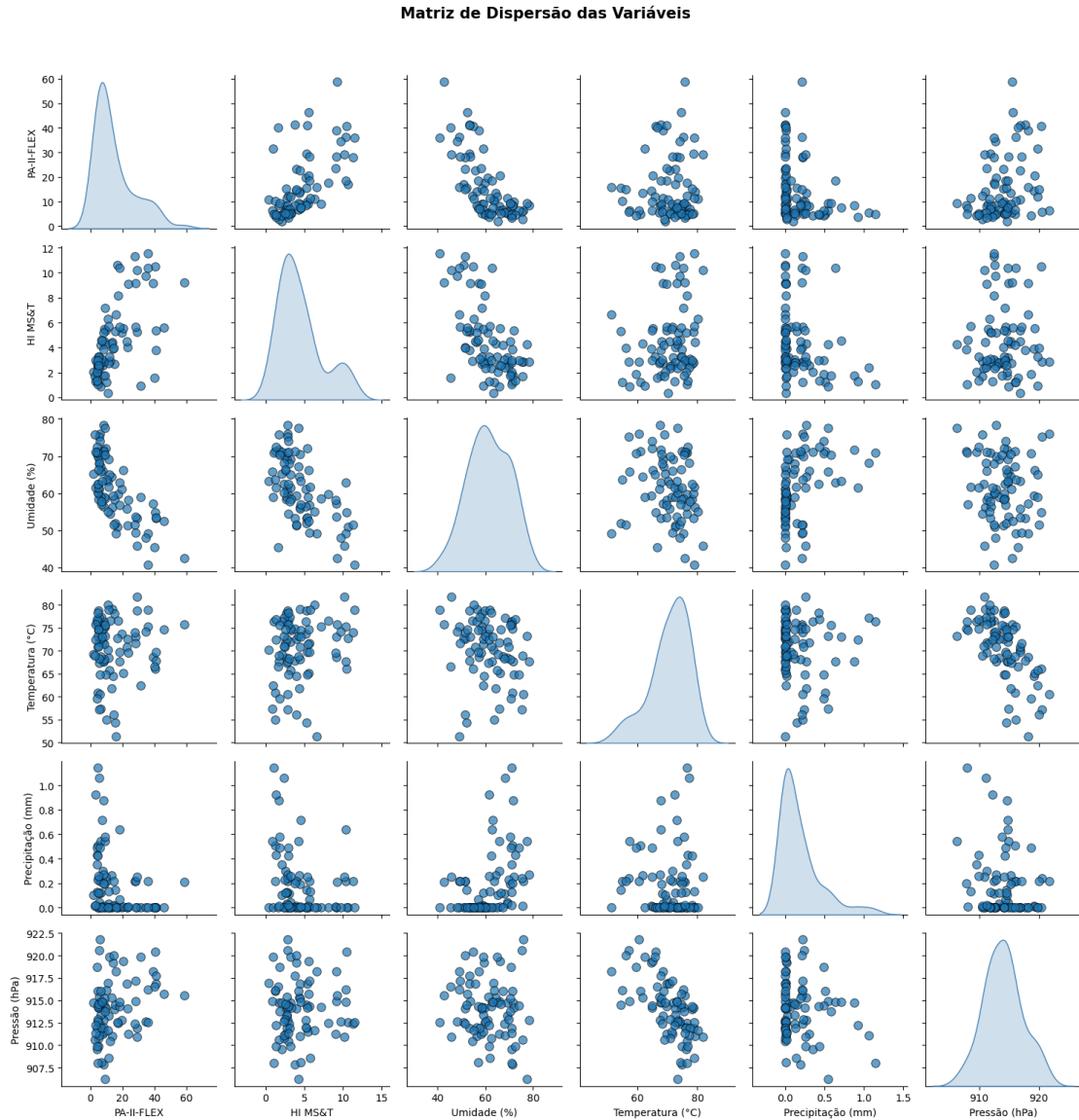


Figura 19 – Matriz de dispersão entre as concentrações de $PM_{1.0}$ (sensor e referência) e variáveis ambientais: umidade relativa, temperatura, precipitação e pressão atmosférica média.

A variável precipitação, como esperado, mostra distribuição fortemente assimétrica, concentrando a maioria dos pontos em valores próximos de zero, o que reflete a predominância de períodos secos durante a campanha de amostragem. Em eventos de chuva, observa-se tendência de redução nos valores de $PM_{1.0}$, conforme evidenciado em alguns dos gráficos de dispersão.

Os padrões visuais identificados nesta matriz sustentam os resultados da matriz de correlação e da literatura, indicando que a calibração dos sensores PA-II-FLEX deve, preferencialmente, incorporar a umidade relativa como variável explicativa principal, além de considerar possíveis efeitos da precipitação em cenários de chuva significativa (CRILLEY, 2020; MALINGS et al., 2020a; JAYARATNE, 2018; TRYNER et al., 2020b).

4.2.4 Análise das Relações Lineares entre as Variáveis

As Figuras 20, 21, 22, 23 e 24 apresentam os gráficos de dispersão e os ajustes lineares entre a concentração de $PM_{1.0}$ registrada pelo sensor PA-II-FLEX e as principais variáveis ambientais, além do método de referência HI MS&T.

No gráfico de dispersão entre o sensor PA-II-FLEX e o método de referência HI MS&T (Figura 20), observa-se uma tendência positiva, refletida pelo ajuste linear, embora com uma inclinação superior à linha 1:1. Isso evidencia que o sensor tende a superestimar os valores em relação ao método gravimétrico, especialmente em faixas de maior concentração. Este comportamento é consistente com as análises anteriores e encontra respaldo na literatura, que destaca a necessidade de calibração específica para sensores ópticos de baixo custo (MALINGS et al., 2020a; CASTELL et al., 2017b).

Na análise da relação entre $PM_{1.0}$ do sensor e umidade relativa (Figura 21), observa-se uma correlação negativa significativa, com o ajuste linear apresentando declive acentuado. Isso indica que valores mais elevados de umidade relativa tendem a estar associados a menores concentrações reportadas pelo sensor, ou ainda, podem evidenciar viés do sensor causado pela interferência da umidade, fenômeno amplamente documentado em estudos recentes (JAYARATNE, 2018; CRILLEY, 2020; TRYNER et al., 2020b).

Já o gráfico de dispersão entre o sensor e a temperatura média (Figura 22) revela uma relação linear muito fraca, praticamente nula, sugerindo que a temperatura não exerce influência significativa isolada sobre as medições do sensor no contexto estudado.

A dispersão entre $PM_{1.0}$ do sensor e a pressão atmosférica (Figura 23) indica uma correlação positiva discreta. Embora a relação seja menos acentuada que a observada com a umidade, há indícios de que a pressão pode ter um papel secundário na variação das medições do sensor, especialmente em condições de estabilidade atmosférica ou variações sazonais.

Por fim, a Figura 24 apresenta a relação entre a concentração de $PM_{1.0}$ e a precipitação acumulada em cada período amostral. Nota-se uma tendência negativa, com redução das concentrações de $PM_{1.0}$ conforme a precipitação aumenta, em linha com o fenômeno de deposição úmida. Embora a maioria dos pontos se concentre em baixos valores de precipitação, os episódios de chuva contribuem para quedas pontuais nas concentrações de partículas, reforçando a importância de considerar este parâmetro em análises de calibração e modelagem.

Esses resultados evidenciam a relevância de considerar, especialmente, a umidade relativa e, em menor escala, a precipitação acumulada e a pressão atmosférica como variáveis preditoras na calibração dos dados dos sensores ópticos. A temperatura, por sua vez, apresentou influência marginal nas condições observadas neste estudo.

Diante disso, a aplicação de procedimentos de calibração torna-se essencial para

assegurar a acurácia e a comparabilidade dos dados, atendendo às recomendações das principais revisões internacionais sobre monitoramento de material particulado com sensores de baixo custo (LIANG, 2021; SAYAHI et al., 2019). O próximo capítulo detalha as estratégias metodológicas adotadas para o desenvolvimento e avaliação dos modelos de correção.

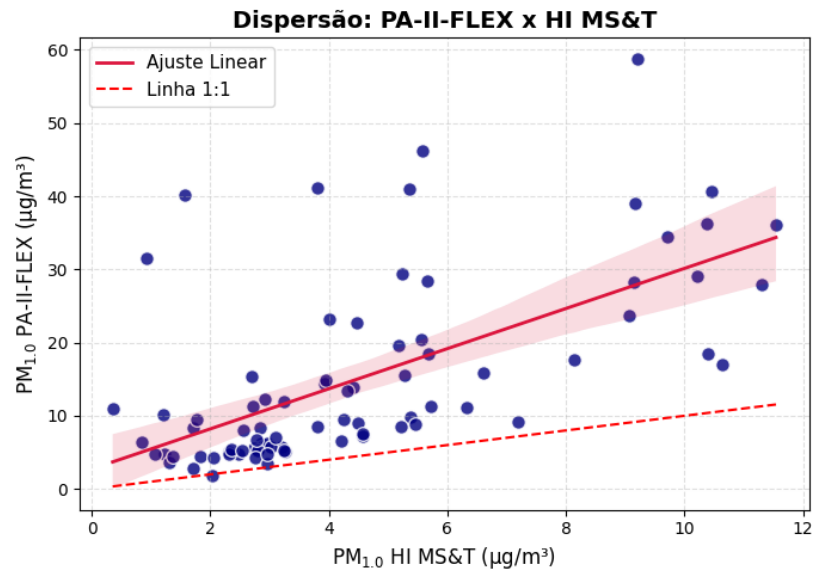


Figura 20 – Dispersão e ajuste linear entre $PM_{1.0}$ PA-II-FLEX e HI MS&T (linha 1:1 em vermelho tracejado).

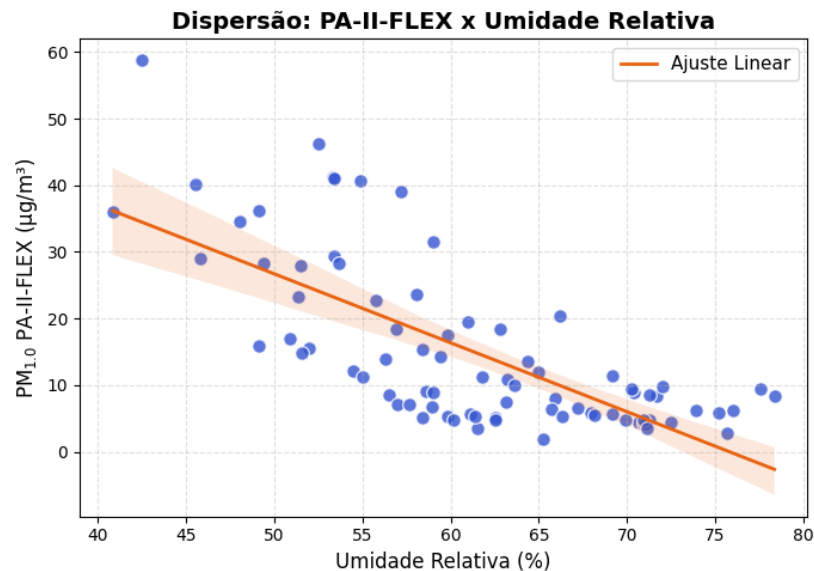


Figura 21 – Dispersão e ajuste linear entre $PM_{1.0}$ PA-II-FLEX e umidade relativa média (%).

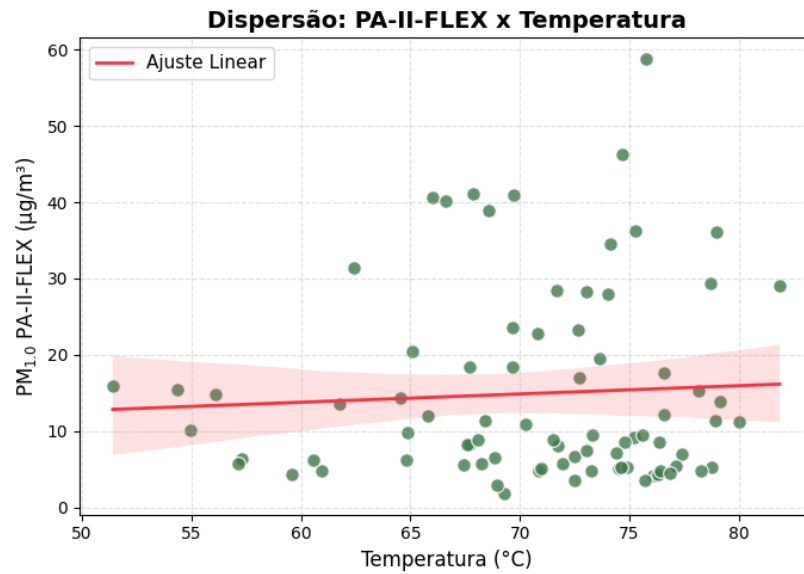


Figura 22 – Dispersão e ajuste linear entre PM_{1.0} PA-II-FLEX e temperatura média (°C).

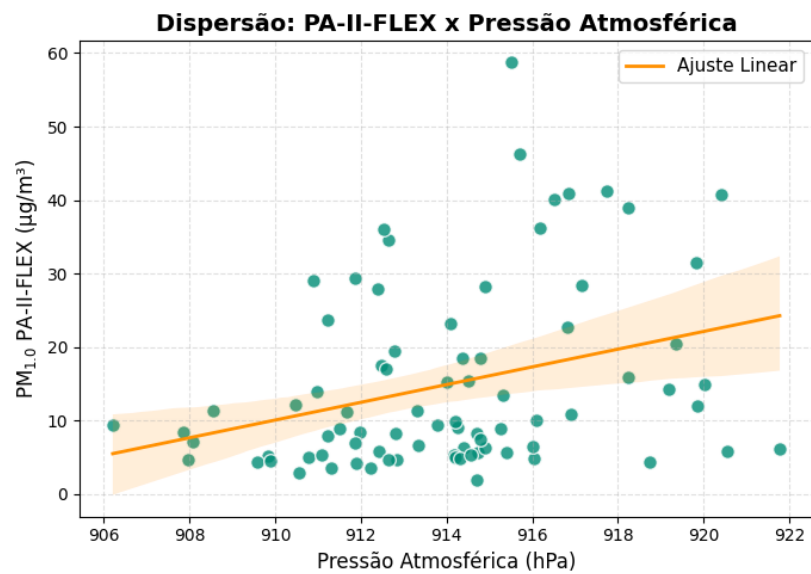


Figura 23 – Dispersão e ajuste linear entre PM_{1.0} PA-II-FLEX e pressão atmosférica média (hPa).

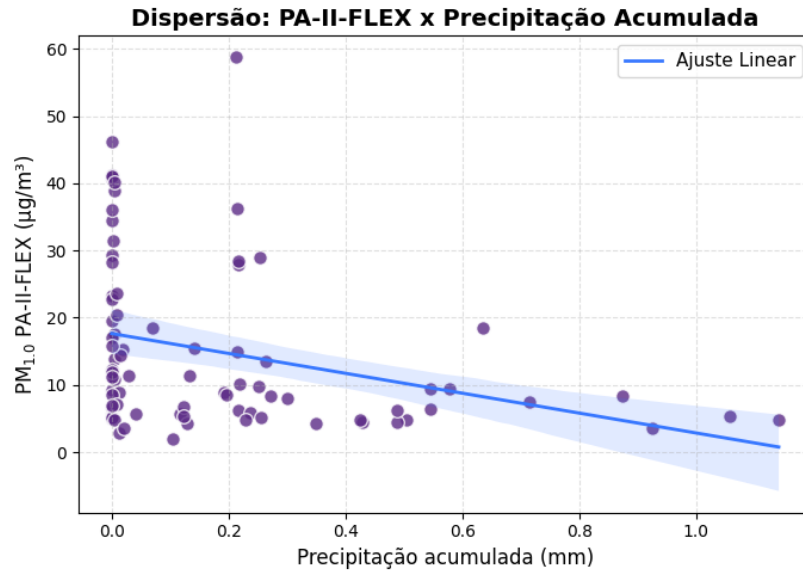


Figura 24 – Dispersão e ajuste linear entre $PM_{1.0}$ PA-II-FLEX e precipitação acumulada (mm).

4.3 Aplicação e Avaliação dos Modelos

4.3.1 Modelos Lineares Simples e Múltiplos

4.3.1.1 Regressão Linear Simples

Neste trabalho, inicialmente empregou-se o método *holdout* para avaliar o desempenho dos modelos de calibração, no qual o conjunto de $n = 79$ amostras foi segmentado utilizando a função *train_test_split*, destinando 80% dos dados para treinamento e 20% para teste. O modelo de regressão linear simples, que utiliza exclusivamente a média das medições do PA-II-FLEX como variável independente, resultou na seguinte equação de calibração:

$$PM_{1.0}^{HI} = 1,4665 + 0,2209 \times \overline{PM_{1.0}^{PA}} \quad (4.1)$$

No conjunto de dados de teste, esse modelo apresentou $R^2 = 0,7021$, RMSE de $1,7921 \mu\text{g}/\text{m}^3$, MAE de $1,4464 \mu\text{g}/\text{m}^3$ e MNB de 71,7611%, como pode ser observado na Figura 25A. Tais métricas são compatíveis com os resultados reportados na literatura para calibração linear simples de sensores ópticos de baixo custo para $PM_{1.0}$.

Nesse contexto, os estudos conduzidos pela *South Coast Air Quality Management District* (AQMD), agência regulatória responsável pela melhoria da qualidade do ar em importantes regiões dos Estados Unidos, constituem uma referência particularmente relevante. Em suas avaliações, foram testadas seis unidades do sensor PurpleAir PA-II-FLEX, equipado com o PMS6003, em comparação direta com o instrumento de referência

GRIMM, utilizando médias diárias de $PM_{1.0}$ — abordagem metodológica compatível com a empregada neste trabalho. Os resultados reportados pela AQMD incluem valores de R^2 entre 0,90 e 0,93 e RMSE variando de 1,2 a 1,7 $\mu\text{g}/\text{m}^3$, com inclinações das retas de calibração entre 0,74 e 0,81 e interseptos próximos de 1,5 (por exemplo, $y = 0,7885x + 1,676$; $R^2 = 0,93$). Ressalta-se, no entanto, que as métricas de desempenho podem variar consideravelmente em função do tipo de aerossol, condições ambientais (umidade, temperatura), escala temporal das medições e diferenças de fabricação entre sensores e, diferença entre equipamentos de referencia.

Portanto, os resultados obtidos neste trabalho para $PM_{1.0}$, considerando médias diárias e calibrando o PA-II-FLEX frente a método gravimétrico, são abaixo do encontrado em estudos semelhantes. Isso evidencia as limitações do uso exclusivo de modelos lineares simples. Ainda assim, tais resultados reforçam a necessidade do desenvolvimento de abordagens, como modelos de regressão múltipla ou métodos baseados em aprendizado de máquina, para aprimorar a precisão e reduzir o viés das estimativas em cenários mais complexos e variáveis.

4.3.1.2 Modelos de Regressão Linear Múltipla

Para aprimorar a calibração do sensor óptico PA-II-FLEX na estimativa de $PM_{1.0}$, foram avaliados diferentes modelos de regressão linear múltipla, combinando a concentração média de $PM_{1.0}$ registrada pelo sensor com médias diárias das variáveis meteorológicas — umidade relativa, temperatura e pressão atmosférica. O objetivo principal foi investigar se a incorporação dessas variáveis, reconhecidas por influenciar tanto a dinâmica do material particulado quanto a resposta dos sensores ópticos, poderia proporcionar ganhos significativos na acurácia dos modelos de calibração.

O melhor desempenho foi alcançado pelo modelo que utiliza, como preditores, a média da concentração de $PM_{1.0}$ do sensor e a umidade relativa média. A equação ajustada é:

$$PM1.0^{HI} = 1,6151 + 0,2193 \times \overline{PM1.0^{PA}} - 0,0021 \times \overline{UR} \quad (4.2)$$

onde $PM1.0^{HI}$ é a concentração gravimétrica de referência, $\overline{PM1.0^{PA}}$ representa a média das leituras do PA-II-FLEX e $\overline{UR}$ a umidade relativa média no período amostral.

Os resultados para o conjunto de teste demonstraram $R^2 = 0,7034$, RMSE de 1,7881 $\mu\text{g}/\text{m}^3$, MAE de 1,4446 $\mu\text{g}/\text{m}^3$ e MNB de 71,6936%. Esses valores, detalhados na Tabela 3, evidenciam que o ganho em relação ao modelo linear simples (com apenas a concentração média do sensor) foi marginal, tanto para as métricas de ajuste quanto para os erros médios.

A análise comparativa, ilustrada pelas Figuras de dispersão dos dados de teste para todas as combinações de variáveis da Figura 25, mostra que a inclusão de variáveis meteorológicas adicionais — seja umidade relativa, temperatura, pressão atmosférica ou a combinação de múltiplas preditoras — não trouxe avanços relevantes. Em alguns casos, modelos mais complexos apresentaram até piora nos indicadores, como redução do R^2 e aumento do erro absoluto médio (MAE) e do erro quadrático médio (RMSE), reforçando que, para este conjunto de dados, variáveis como temperatura e pressão atmosférica têm influência secundária ou nula no ajuste. Assim, a média da concentração de $PM_{1.0}$ do sensor PA-II-FLEX permanece como principal preditora para a calibração local do instrumento.

O coeficiente negativo associado à umidade relativa ($-0,0021$) está sugerindo que condições de maior umidade atmosférica levam à redução da concentração estimada. Isso pode ser explicado por dois mecanismos: a interferência ótica causada por gotículas de água aderidas às partículas, que altera a resposta do sensor, e o aumento da deposição úmida, que reduz a quantidade de material particulado em suspensão. Mesmo assim, a persistência de viés percentual médio (MNB) elevado indica uma tendência sistemática de superestimação das concentrações reais, sugerindo que a variabilidade residual está relacionada a outros fatores não contemplados pelo modelo, como flutuações diárias não capturadas pelas médias, efeitos físico-químicos específicos do aerossol local ou limitações do próprio sensor óptico.

Em síntese, a adição de variáveis ambientais, embora física e estatisticamente justificável, não resultou em ganhos práticos expressivos para o ajuste dos modelos de calibração testados neste estudo. Esses resultados, sistematizados na Tabela 3 e visualizados nas Figuras 25, reforçam a necessidade de uma calibração local criteriosa, além do entendimento aprofundado das limitações inerentes aos sensores de baixo custo, sobretudo em cenários de elevada umidade e dinâmica atmosférica variada.

Tabela 3 – Resumo das métricas de desempenho dos modelos de regressão para diferentes combinações de variáveis independentes (conjunto de teste).

Modelo	Variáveis In- dependentes	R^2	RMSE	MAE	MNB (%)	CvMAE	nRMSE
A	$\overline{PM_{1,0}^{PA}}$	0,7021	1,7921	1,4464	71,7611	0,2936	0,3638
B	$\overline{PM_{1,0}^{PA}}, \overline{UR}$	0,7034	1,7881	1,4446	71,6936	0,2933	0,3630
C	$\overline{PM_{1,0}^{PA}}, \overline{T}$	0,6555	1,9271	1,5945	61,7912	0,3237	0,3912
D	$\overline{PM_{1,0}^{PA}}, \overline{P}$	0,6798	1,8579	1,5044	62,2760	0,3054	0,3771
E	$\overline{PM_{1,0}^{PA}}, \overline{UR}, \overline{T}$	0,6136	2,0409	1,6822	60,9416	0,3415	0,4143
F	$\overline{PM_{1,0}^{PA}}, \overline{UR}, \overline{T}, \overline{P}$	0,6481	1,9477	1,6051	59,8679	0,3258	0,3954
G	$\overline{PM_{1,0}^{PA}}, \overline{UR}, \overline{P}$	0,6818	1,8522	1,5029	62,2078	0,3051	0,3760
H	$\overline{PM_{1,0}^{PA}}, \overline{T}, \overline{P}$	0,6657	1,8982	1,5627	60,0994	0,3172	0,3853

$\overline{PM_{1,0}^{PA}}$ = média do $PM_{1,0}$ do sensor PA (canal CF_1); $\overline{UR}$ = média da umidade relativa; $\overline{T}$ = média da temperatura; $\overline{P}$ = média da pressão atmosférica.

4.3.1.3 Regressão Linear Múltipla com Inclusão de Variáveis Estatísticas

A inclusão de variáveis estatísticas nos modelos de regressão múltipla teve como objetivo capturar aspectos adicionais da variabilidade dos dados do sensor PA-II-FLEX e das condições meteorológicas, em especial a umidade relativa, buscando aprimorar a calibração dos sensores ópticos para $PM_{1,0}$.

Nesta etapa do estudo, foram desenvolvidos onze modelos de regressão linear múltipla incorporando diferentes combinações de estatísticas descritivas das séries temporais durante a amostragem com o HI MS&T. Os modelos avaliados consideraram, individualmente ou em conjunto, medidas centrais (média e mediana), estatísticas de dispersão (desvio padrão), quartis (Q25 e Q75), além de valores extremos (máximos e mínimos) tanto para o $PM_{1,0}$ quanto para a umidade relativa. A Tabela 4 apresenta um resumo detalhado das métricas obtidas para cada modelo avaliado, enquanto a Figura 26 ilustra visualmente os desempenhos preditivos.

Dentre os modelos testados, o Modelo N apresentou o melhor desempenho global, incluindo a mediana da umidade relativa, a média do $PM_{1,0}$ do sensor, o desvio padrão, o percentil 75 e o valor máximo das concentrações. O modelo pode ser expresso pela seguinte equação:

$$\begin{aligned}
 PM_{1,0}^{HI} = & 2,3509 - 0,0107 \times \tilde{UR} - 0,0174 \times \overline{PM_{1,0}^{PA}} - 0,1855 \times \sigma_{PM_{1,0}^{PA}} \\
 & + 0,2082 \times Q_{75}(PM_{1,0}^{PA}) + 0,0158 \times \max(PM_{1,0}^{PA})
 \end{aligned} \quad (4.3)$$

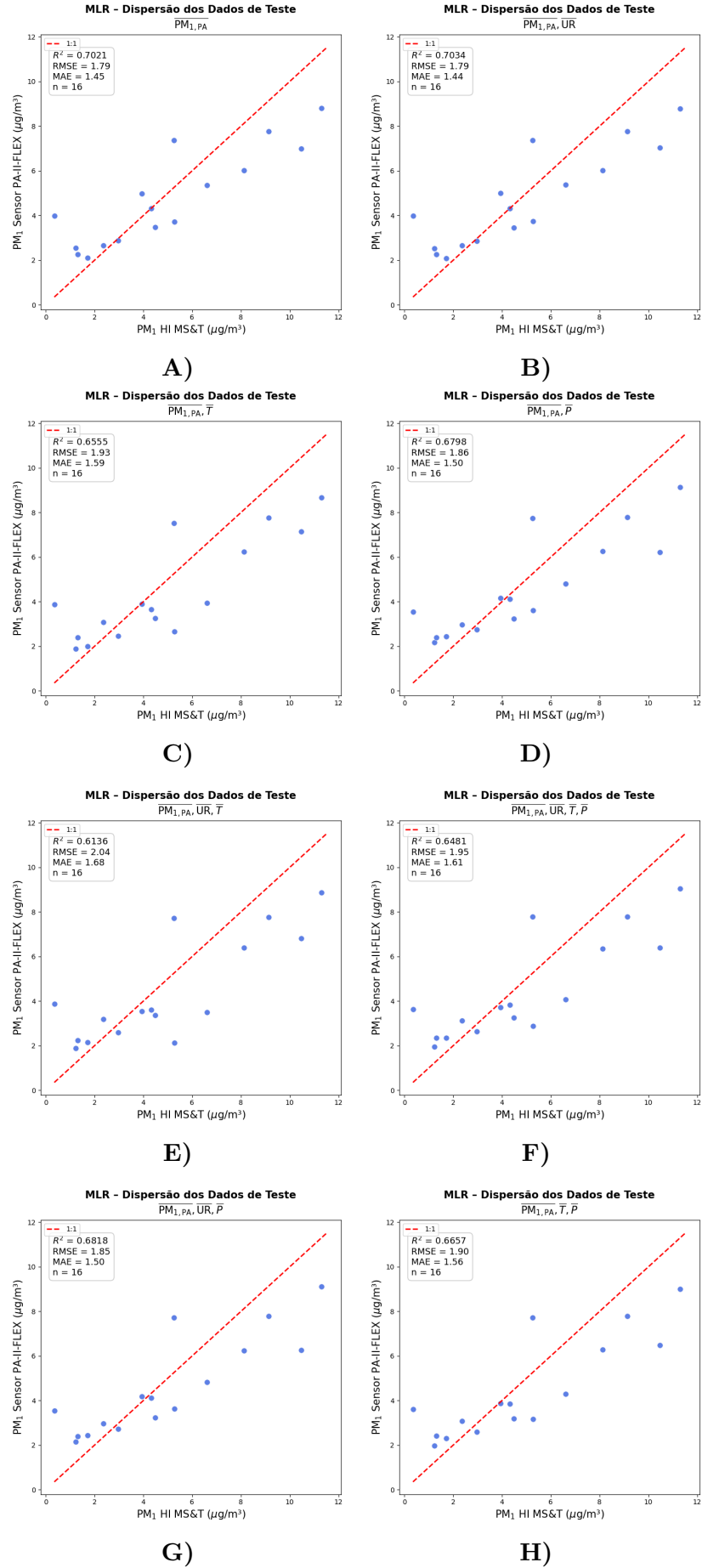


Figura 25 – Dispersão dos dados de teste para diferentes modelos MLR para as médias de amostragem.

No conjunto de teste, esse ajuste resultou em $R^2 = 0,7404$, RMSE de $1,6728 \mu\text{g}/\text{m}^3$, MAE de $1,3865 \mu\text{g}/\text{m}^3$ e MNB de 62,89%. Esses valores superam, de forma consistente, todos os demais modelos avaliados, evidenciando a importância de incorporar estatísticas que capturam a dispersão e a assimetria dos dados ao lado das medidas centrais.

Modelos que utilizam apenas médias ou medianas (Modelos I, K) apresentaram desempenho satisfatório, porém inferior, com R^2 entre 0,65 e 0,70. A utilização de estatísticas ligadas à cauda inferior da distribuição, como quartis inferiores ou valores mínimos (Modelo J), não contribuiu para um ajuste robusto, confirmando que os maiores desafios de predição se concentram em episódios de concentração elevada de partículas. Isso reforça a necessidade de modelos capazes de representar não só a tendência central, mas também os picos, geralmente ligados a eventos de poluição episódica.

No contexto da literatura, observa-se uma diferença marcante no comportamento dos modelos para $\text{PM}_{1,0}$ e $\text{PM}_{2,5}$. Para $\text{PM}_{2,5}$, a média das concentrações costuma ser suficiente para explicar boa parte da variabilidade, mas, no caso do $\text{PM}_{1,0}$, os dados deste estudo sugerem que a dispersão e a assimetria da distribuição têm peso mais significativo. Assim, a inclusão de estatísticas como o percentil 75 e o valor máximo mostrou-se especialmente relevante para captar episódios de concentração extrema. Embora a literatura revisada não aborde diretamente o impacto de picos ou extremos na calibração de sensores ópticos para $\text{PM}_{1,0}$, os resultados obtidos aqui indicam que a consideração dessas estatísticas pode aprimorar o ajuste dos modelos, tornando-os mais sensíveis a variações pontuais de concentração ao longo do período amostral.

Outro ponto relevante é a robustez dos modelos frente a *outliers* e valores extremos. Estatísticas como a mediana e o percentil 75 são menos sensíveis a picos anômalos, proporcionando ajustes mais estáveis em contextos de alta variabilidade atmosférica. Em contrapartida, modelos baseados apenas em médias podem ser fortemente impactados por esses eventos, levando a subestimções ou superestimções indesejadas, principalmente em séries temporais curtas ou com poucos episódios extremos.

Apesar da melhora proporcionada pela inclusão de múltiplas estatísticas, é importante observar que os ganhos permanecem relativamente modestos, especialmente devido ao número limitado de amostras disponíveis para calibração e validação. Esse fator, comum em estudos com sensores de baixo custo, limita o potencial de exploração das variáveis estatísticas e reforça a importância de séries temporais mais extensas e amostragens mais variadas em estudos futuros.

A Tabela 4 permite visualizar rapidamente o desempenho relativo de todos os modelos, facilitando a comparação entre os diferentes conjuntos de preditores. Observa-se que os melhores desempenhos estão sistematicamente associados à combinação de medidas centrais (média ou mediana da umidade relativa e do $\text{PM}_{1,0}$ do sensor) e medidas que capturam a cauda superior da distribuição (Q75, máximo).

Do ponto de vista operacional, embora modelos mais complexos possam oferecer ajustes estatisticamente superiores, a escolha do modelo final deve considerar, além dos ganhos quantitativos, aspectos práticos como simplicidade, facilidade de implementação e robustez frente a condições não previstas no conjunto de calibração. Essa abordagem integrada facilita a decisão sobre qual modelo utilizar em aplicações reais de monitoramento de partículas ultrafinas com sensores ópticos de baixo custo.

Em síntese, a análise evidencia que, para $PM_{1.0}$, a calibração de sensores ópticos baseada apenas em médias ou medianas é insuficiente para garantir precisão nos extremos da distribuição. A inclusão de estatísticas relacionadas à dispersão e picos de concentração é fundamental para capturar a variabilidade real do ambiente amostrado. Estudos futuros devem ampliar o espectro de amostras e explorar o potencial de métodos não lineares e aprendizado de máquina para aprimorar ainda mais o ajuste desses modelos, especialmente em cenários de alta complexidade atmosférica.

Tabela 4 – Resumo das métricas de desempenho dos modelos de regressão com variáveis estatísticas (conjunto de teste).

Modelo	Variáveis Independentes	R^2	RMSE	MAE	MNB (%)	CvMAE	nRMSE
I	$\tilde{PM}_{1.0}^{PA}$ $\tilde{UR}$	0,6494	1,9440	1,6051	72,1669	0,3258	0,3946
J	$Q_{25}(PM_{1,PA})$ $Q_{25}(UR)$	0,5929	2,0949	1,7712	80,7472	0,3596	0,4252
K	$\overline{PM}_{1.0}^{PA}$, $\overline{UR}$ $\sigma_{PM_{1,PA}}$, σ_{UR}	0,6989	1,8017	1,4618	72,9963	0,2967	0,3657
L	$\tilde{PM}_{1.0}^{PA}$, $\tilde{UR}$ $\sigma_{PM_{1,PA}}$	0,6803	1,8564	1,4780	68,7914	0,3000	0,3768
M	$\max(PM_{1,PA})$, $\min(PM_{1,PA})$ $\max(UR)$, $\min(UR)$	0,6473	1,9498	1,5340	75,0192	0,3114	0,3958
N	$\tilde{UR}$, $\overline{PM}_{1.0}^{PA}$ $\sigma_{PM_{1,PA}}$, $Q_{75}(PM_{1,PA})$ $\max(PM_{1,PA})$	0,7404	1,6728	1,3865	62,8888	0,2815	0,3396
O	$\overline{UR}$, $\overline{PM}_{1.0}^{PA}$ $\sigma_{PM_{1,PA}}$, $Q_{75}(PM_{1,PA})$ $\max(PM_{1,PA})$	0,7339	1,6935	1,3983	63,5470	0,2838	0,3438
P	$\overline{UR}$, $\overline{P}$ $\overline{PM}_{1.0}^{PA}$, $Q_{75}(PM_{1,PA})$ $\max(PM_{1,PA})$	0,6989	1,8017	1,4617	59,8124	0,2967	0,3657
Q	$\tilde{UR}$, $\overline{PM}_{1.0}^{PA}$ $Q_{75}(PM_{1,PA})$, $\max(PM_{1,PA})$	0,7219	1,7314	1,4267	70,5912	0,2896	0,3515
R	$\overline{UR}$, $\overline{PM}_{1.0}^{PA}$ $Q_{75}(PM_{1,PA})$, $\max(PM_{1,PA})$	0,7131	1,7586	1,4401	71,4393	0,2923	0,3570
S	$\tilde{UR}$, $\tilde{PM}_{1.0}^{PA}$ $Q_{75}(PM_{1,PA})$, $\max(PM_{1,PA})$	0,7248	1,7223	1,4110	67,8270	0,2864	0,3496

$(-)$ = média $(\tilde{-})$: mediana; $\sigma_{(-)}$: desvio padrão; $Q_{25}(-)$: percentil 25; $Q_{75}(-)$: percentil 75; $\max(-)$: máximo; $\min(-)$: mínimo.

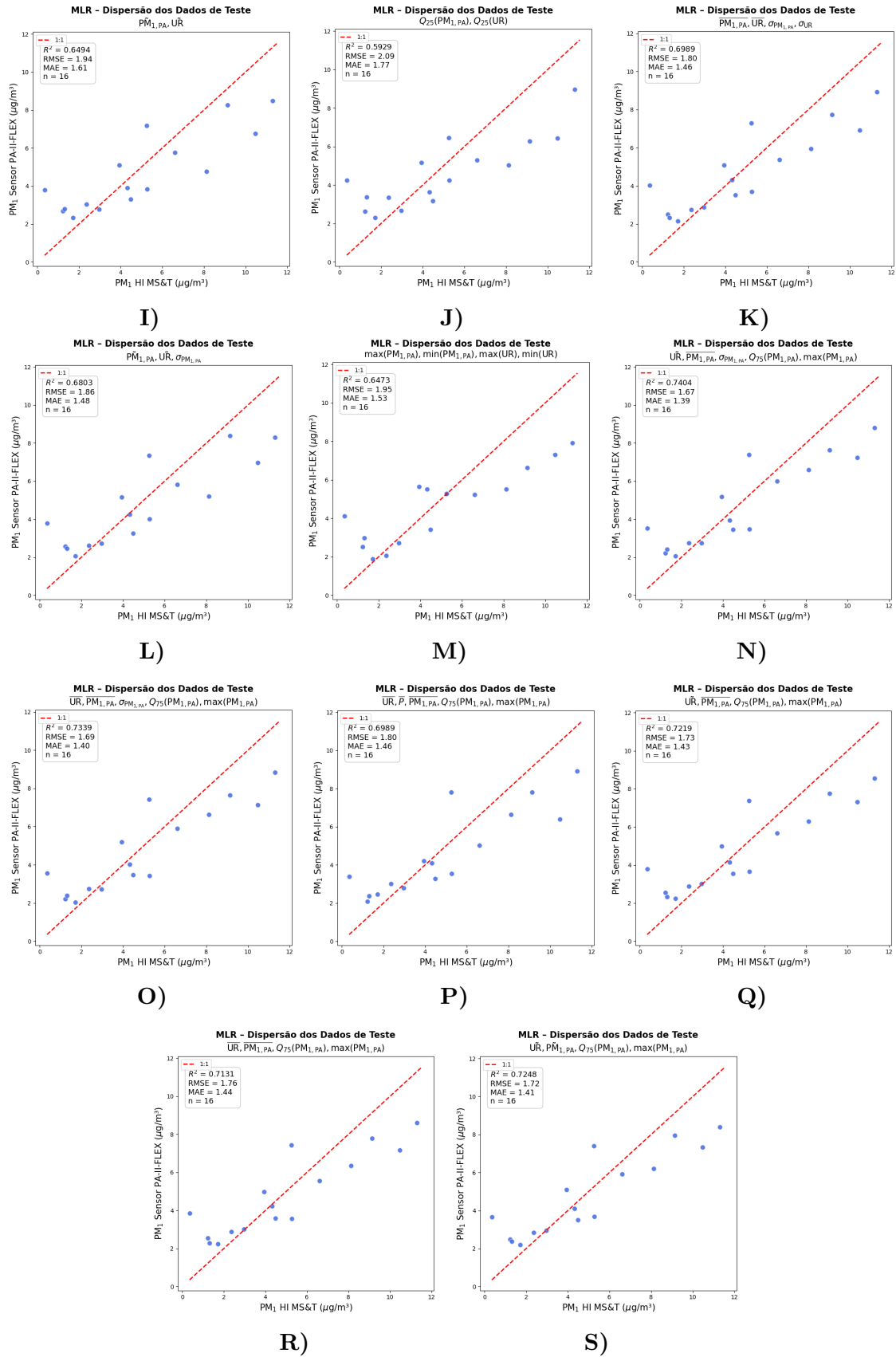


Figura 26 – Dispersão dos dados de teste para diferentes modelos MLR com inclusão de variáveis estatísticas.

4.3.2 Machine Learning

4.3.2.1 Random Forest Regressor

Foram avaliados 19 modelos de Random Forest, correspondentes às mesmas combinações de variáveis preditoras empregadas nas regressões lineares múltiplas (Modelos A a S). Cada modelo utilizou diferentes combinações de estatísticas descritivas de $PM_{1.0}$ (média, mediana, quartis, máximo, mínimo e desvio padrão) de variáveis meteorológicas (umidade relativa, temperatura e pressão), conforme detalhado na Tabela 5. Essa abordagem sistemática permitiu comparar, de forma direta, o desempenho do Random Forest frente às estratégias convencionais de regressão.

O desempenho dos modelos foi avaliado com base em métricas padronizadas: coeficiente de determinação (R^2), erro quadrático médio (RMSE), erro absoluto médio (MAE) e viés médio normalizado (MNB), calculados sobre o conjunto de teste. O gráfico de dispersão de todos os modelos é apresentado na Figura 31. Os melhores resultados foram observados nos modelos que incluíram medidas de dispersão e extremos superiores de $PM_{1.0}$, especialmente o Modelo Q, composto pelas preditoras $\tilde{U}R$, $\overline{PM}_{1.0}^{PA}$, $Q_{75}(PM_{1.0}^{PA})$ e $\max(PM_{1.0}^{PA})$, que apresentou $R^2 = 0,7800$, $RMSE = 1,54 \mu g/m^3$, $MAE = 1,07 \mu g/m^3$ e $MNB = 81\%$. Este desempenho é visualmente apresentado na Figura 27, com clara aderência dos valores preditos à linha 1:1.

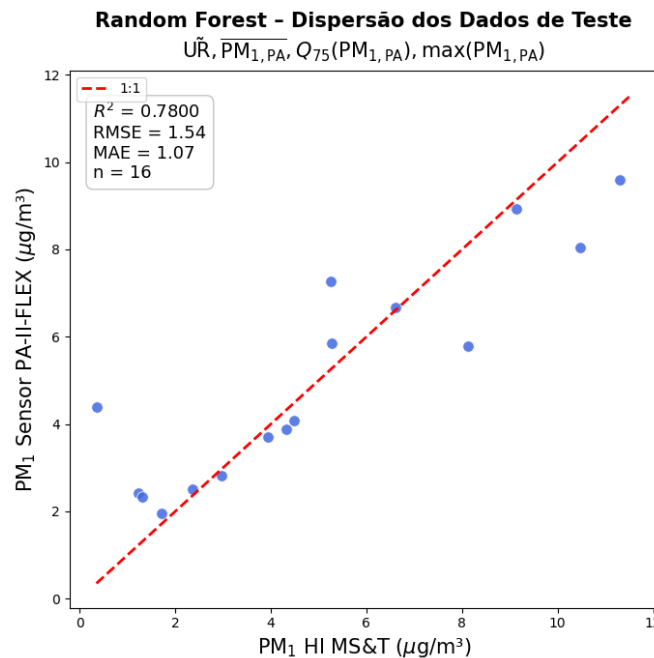


Figura 27 – Random Forest – Dispersão dos dados de teste para o modelo Q (40 árvores).

Além disso, a escolha do número de árvores no modelo foi fundamentada na análise do desempenho para diferentes valores de $n_estimators$, conforme apresentado na Figura 28. Observou-se que o uso de 40 árvores resultou nos melhores valores de R^2 e RMSE dentre

os cenários testados, indicando que esse parâmetro proporcionou o desempenho mais adequado para o conjunto de dados avaliado.

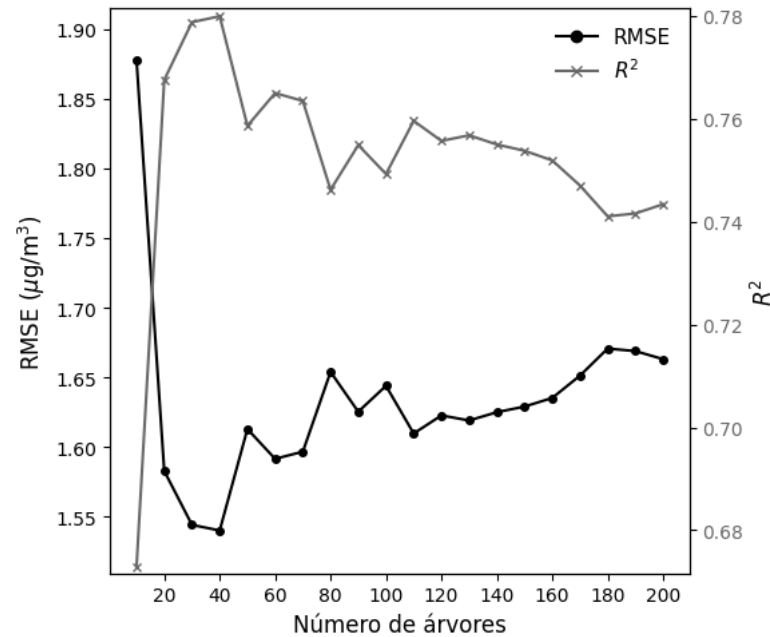


Figura 28 – Desempenho do Random Forest em função do número de árvores: evolução dos indicadores R^2 e RMSE para o modelo Q.

Por fim, a Figura 29 apresenta a importância relativa das variáveis no modelo de melhor desempenho, destacando que o percentil 75 de $PM_{1.0}$ (Q_{75}) e a média de $PM_{1.0}$ explicam a maior parte da variância preditiva, seguidos pela umidade relativa e pelo valor máximo de $PM_{1.0}$.

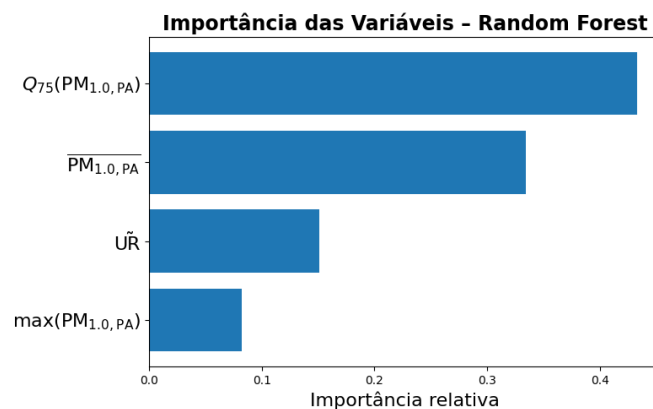


Figura 29 – Importância relativa das variáveis no modelo *Random Forest* de melhor desempenho.

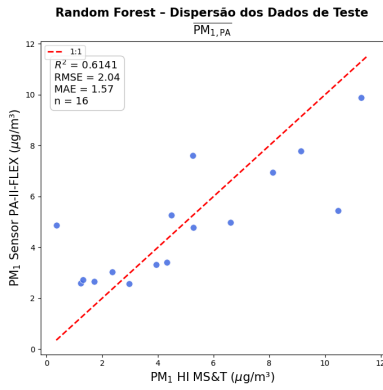
Os resultados obtidos demonstram ganhos substanciais de desempenho do *Random Forest* em relação aos modelos lineares tradicionais, principalmente na capacidade de capturar não linearidades e interações entre variáveis, além de efeitos extremos de $PM_{1.0}$. Os

valores de R^2 próximos a 0,78 e de RMSE na ordem de $1,54 \mu\text{g}/\text{m}^3$ demonstram uma relação consistente com resultados reportados na literatura para calibração de sensores ópticos de baixo custo para $\text{PM}_{1.0}$ em condições de campo. Este desempenho é especialmente relevante considerando que o equipamento de referência empregado neste estudo não é do tipo *Federal Equivalent Method* (FEM) e o número reduzido de dados, o que geralmente implica desafios adicionais em termos de precisão e comparabilidade dos dados.

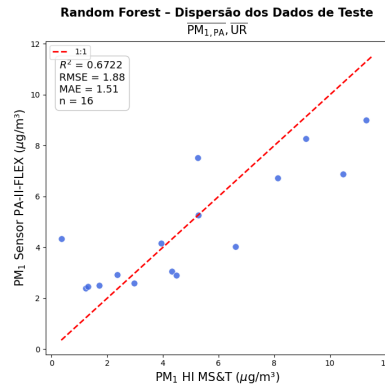
Estudos recentes destacam que o uso de modelos baseados em *machine learning*, especialmente *Random Forest*, tende a superar regressões lineares na calibração de sensores ópticos (HAYWARD et al., 2024; WANG et al., 2015; WALLACE, 2022; KIM et al., 2015b). Por exemplo, Wang et al. (2020) calibraram uma rede de sensores $\text{PM}_{2.5}$ em Taiwan e relataram reduções de RMSE da ordem de 18,4 para $5,9 \mu\text{g}/\text{m}^3$ (e MAE de 16,2 para $5,0 \mu\text{g}/\text{m}^3$) ao utilizar *Random Forest*, com R^2 entre 0,76 e 0,89 nos conjuntos de validação. Embora a maioria dos trabalhos foque em $\text{PM}_{2.5}$, o desafio para $\text{PM}_{1.0}$ é ainda maior, devido à maior variabilidade temporal e sensibilidade a eventos extremos (KIM et al., 2015b).

Destaca-se que, enquanto a literatura normalmente emprega variáveis meteorológicas (umidade, temperatura) e as médias dos sensores como preditores nos modelos *Random Forest* (WANG et al., 2015; HAYWARD et al., 2024), neste estudo a inclusão de estatísticas associadas à dispersão e aos extremos superiores das concentrações de $\text{PM}_{1.0}$ (como máximos e quartis) demonstrou ganhos significativos de desempenho. Esse diferencial metodológico reforça a noção de que a variabilidade e a ocorrência de picos de concentração desempenham papel central na calibração para $\text{PM}_{1.0}$.

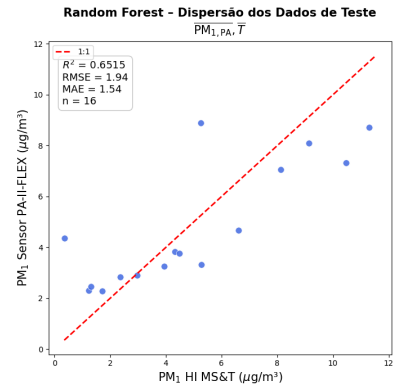
Em termos práticos, os resultados evidenciam que a adoção de *Random Forest* como método de calibração oferece ganhos claros de acurácia e robustez para sensores de baixo custo, especialmente em cenários onde a composição do material particulado e as condições meteorológicas variam significativamente ao longo do tempo (HAYWARD et al., 2024; WANG et al., 2015). Ressalta-se, contudo, que a seleção criteriosa de variáveis predictoras, aliando medidas centrais e de dispersão, é fundamental para a maximização do desempenho dos modelos.



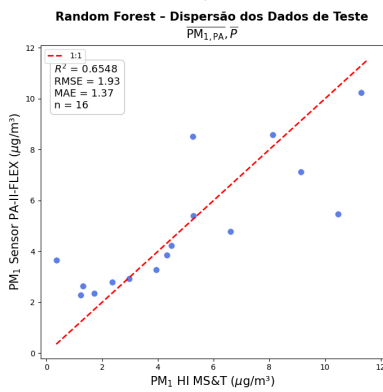
A)



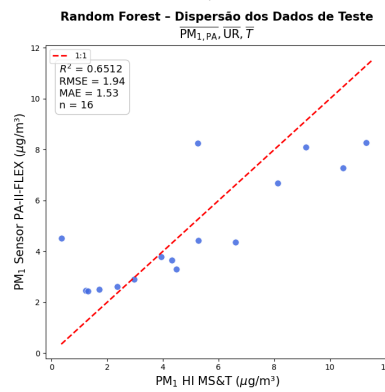
B)



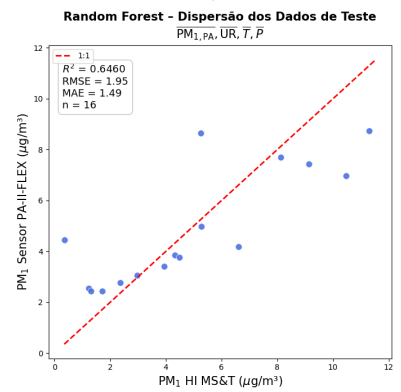
C)



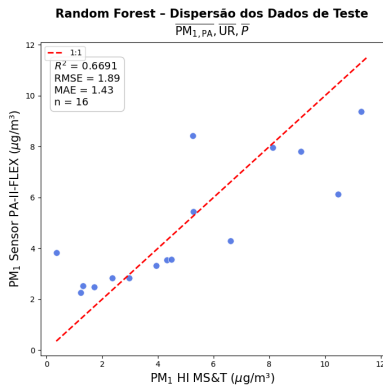
D)



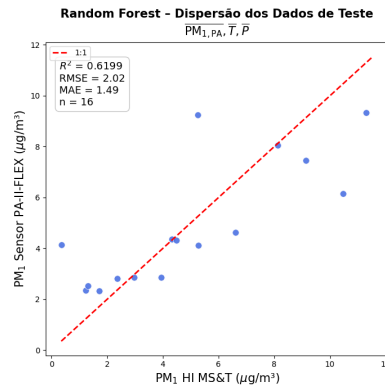
E)



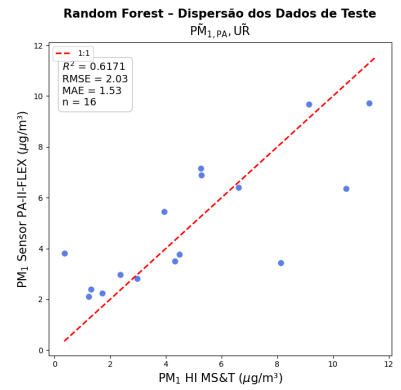
F)



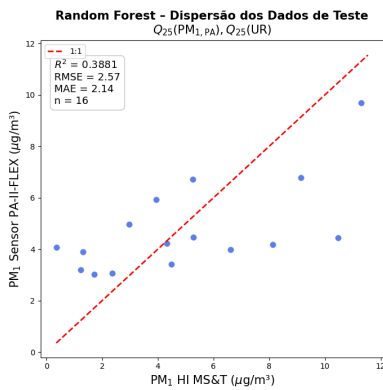
G)



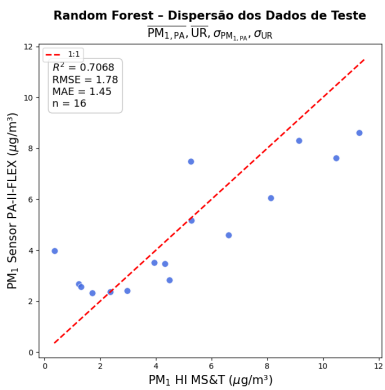
H)



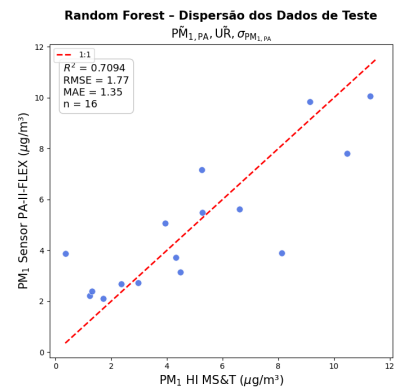
I)



J)



K)



L)

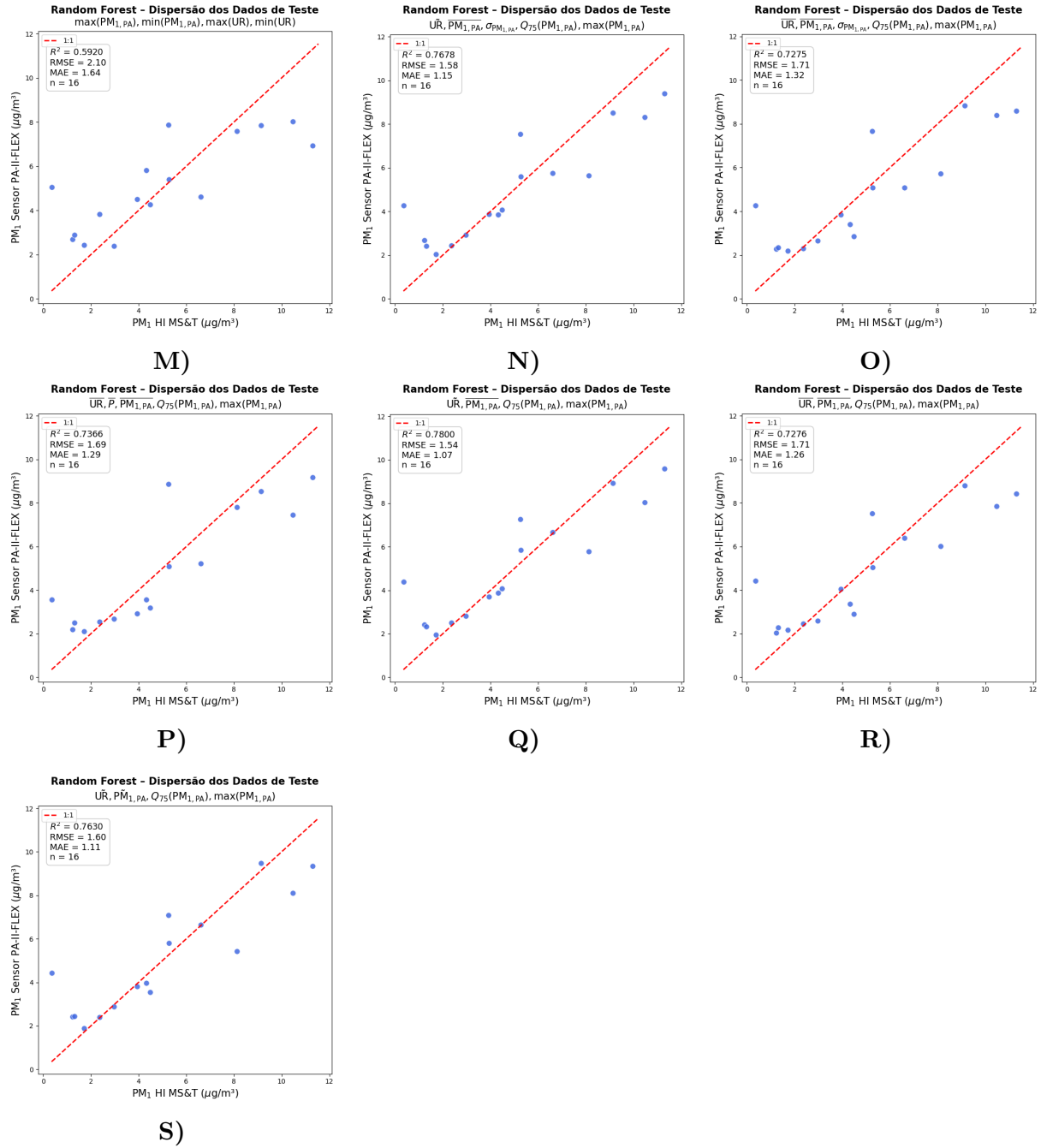


Figura 31 – Dispersão dos dados de teste para diferentes modelos Random Forest (continuação).

Tabela 5 – Resumo das métricas de desempenho dos modelos Random Forest para diferentes combinações de variáveis independentes (conjunto de teste).

Modelo	Variáveis Independentes	R^2	RMSE	MAE
A	$\overline{PM}_{1,0}^{PA}$	0,6141	2,0395	1,5680
B	$\overline{PM}_{1,0}^{PA}$ $\overline{UR}$	0,6722	1,8799	1,5082
C	$\overline{PM}_{1,0}^{PA}$ $\overline{T}$	0,6515	1,9383	1,5408
D	$\overline{PM}_{1,0}^{PA}$ $\overline{P}$	0,6548	1,9290	1,3713
E	$\overline{PM}_{1,0}^{PA}$ $\overline{UR}$ $\overline{T}$	0,6512	1,9391	1,5297
F	$\overline{PM}_{1,0}^{PA}$ $\overline{UR}$ $\overline{T}$ $\overline{P}$	0,6460	1,9535	1,4896
G	$\overline{PM}_{1,0}^{PA}$ $\overline{UR}$ $\overline{P}$	0,6691	1,8885	1,4259
H	$\overline{PM}_{1,0}^{PA}$ $\overline{T}$ $\overline{P}$	0,6199	2,0242	1,4884
I	$\tilde{PM}_{1,0}^{PA}$ $\tilde{UR}$	0,6171	2,0317	1,5274
J	$Q_{25}(PM_{1,0}^{PA})$ $Q_{25}(UR)$	0,3881	2,5683	2,1440
K	$\overline{PM}_{1,0}^{PA}$ $\overline{UR}$ $\sigma_{PM_{1,0}^{PA}}$ σ_{UR}	0,7068	1,7779	1,4532

Modelo	Variáveis Independentes	R^2	RMSE	MAE
L	$\tilde{PM}_{1,0}^{PA}$ $\overline{UR}$ $\sigma_{PM_{1,0}^{PA}}$	0,7094	1,7700	1,3501
M	$\max(PM_{1,0}^{PA})$ $\min(PM_{1,0}^{PA})$ $\max(UR)$ $\min(UR)$	0,5920	2,0971	1,6356
N	$\overline{UR}$ $\overline{PM}_{1,0}^{PA}$ $\sigma_{PM_{1,0}^{PA}}$ $Q_{75}(PM_{1,0}^{PA})$ $\max(PM_{1,0}^{PA})$	0,7678	1,5821	1,1548
O	$\overline{UR}$ $\overline{PM}_{1,0}^{PA}$ $\sigma_{PM_{1,0}^{PA}}$ $Q_{75}(PM_{1,0}^{PA})$ $\max(PM_{1,0}^{PA})$	0,7275	1,7140	1,3200
P	$\overline{UR}$ $\overline{P}$ $\overline{PM}_{1,0}^{PA}$ $Q_{75}(PM_{1,0}^{PA})$ $\max(PM_{1,0}^{PA})$	0,7366	1,6852	1,2873
Q	$\overline{UR}$ $\overline{PM}_{1,0}^{PA}$ $Q_{75}(PM_{1,0}^{PA})$ $\max(PM_{1,0}^{PA})$	0,7800	1,5401	1,0745
R	$\overline{UR}$ $\overline{PM}_{1,0}^{PA}$ $Q_{75}(PM_{1,0}^{PA})$ $\max(PM_{1,0}^{PA})$	0,7276	1,7135	1,2567
S	$\overline{UR}$ $\tilde{PM}_{1,0}^{PA}$ $Q_{75}(PM_{1,0}^{PA})$ $\max(PM_{1,0}^{PA})$	0,7630	1,5982	1,1145

$\overline{PM}_{1,0}^{PA}$: média do PM_1 do sensor PA; $\overline{UR}$: média da umidade relativa; $\overline{T}$: média da temperatura; $\overline{P}$: média da pressão atmosférica; $\tilde{PM}_{1,0}^{PA}$: mediana do PM_1 do sensor PA; $Q_{25/75}(-)$: percentil 25/75; $\sigma_{(-)}$: desvio padrão; $\max(-)$: máximo; $\min(-)$: mínimo.

4.3.3 Avaliação dos Modelos

4.3.3.1 Validação Cruzada (k -fold)

Para garantir a robustez e a generalização dos modelos desenvolvidos neste estudo, foi empregada a validação cruzada do tipo k -fold (com $k = 5$) para todos os 19 modelos de regressão múltipla linear (MLR) e Random Forest testados (Modelos A a S). Os resultados são apresentados nas Figuras 32 e 33, que trazem, para cada modelo, a média e o desvio-padrão das métricas R^2 e RMSE obtidas nos $folds$ de validação.

Observa-se que os modelos MLR apresentam desempenho relativamente consistente entre si, com valores médios de R^2 variando de aproximadamente 0,35 a 0,58 e RMSEs

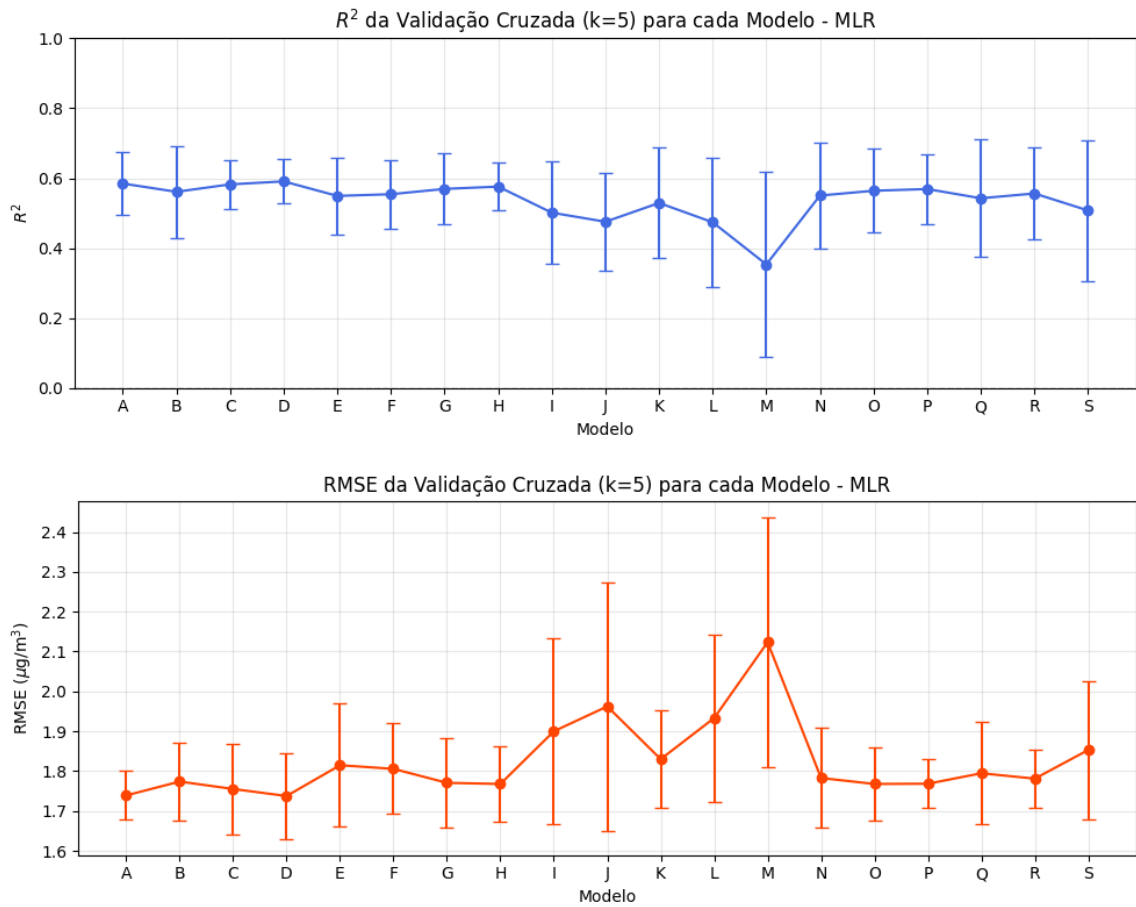


Figura 32 – R^2 e RMSE da validação cruzada ($k = 5$) para cada modelo de regressão múltipla linear (MLR).

entre 1,74 e 2,12 $\mu\text{g}/\text{m}^3$. A maioria dos modelos apresenta R^2 superior a 0,5, evidenciando capacidade moderada de explicação da variabilidade dos dados de $\text{PM}_{1.0}$. Os desvios-padrão das métricas, representados pelas barras de erro nos gráficos, indicam certa estabilidade entre os *folds*, embora haja aumento da variabilidade em modelos que incorporam medidas de extremos ou maior número de variáveis.

No caso dos modelos *Random Forest*, observa-se desempenho inferior ao dos modelos de regressão linear múltipla (MLR) nas condições do estudo. Os maiores valores médios de R^2 dos modelos *Random Forest* alcançaram até 0,54, levemente inferiores aos valores de MLR. Essa redução no desempenho pode ser atribuída, principalmente, ao número limitado de amostras disponíveis ($n = 79$) e à alta variabilidade das concentrações de $\text{PM}_{1.0}$, fatores que dificultam a generalização de modelos mais complexos, como *Random Forest*, sobre bases pequenas.

Além disso, ambos os métodos apresentaram desvios-padrão consideráveis nas métricas, Principalmente para o *Random Forest*, indicando que a divisão dos *folds* na validação cruzada pode impactar fortemente a avaliação de desempenho devido à heterogeneidade sazonal dos dados (com destaque para a maior variabilidade observada no

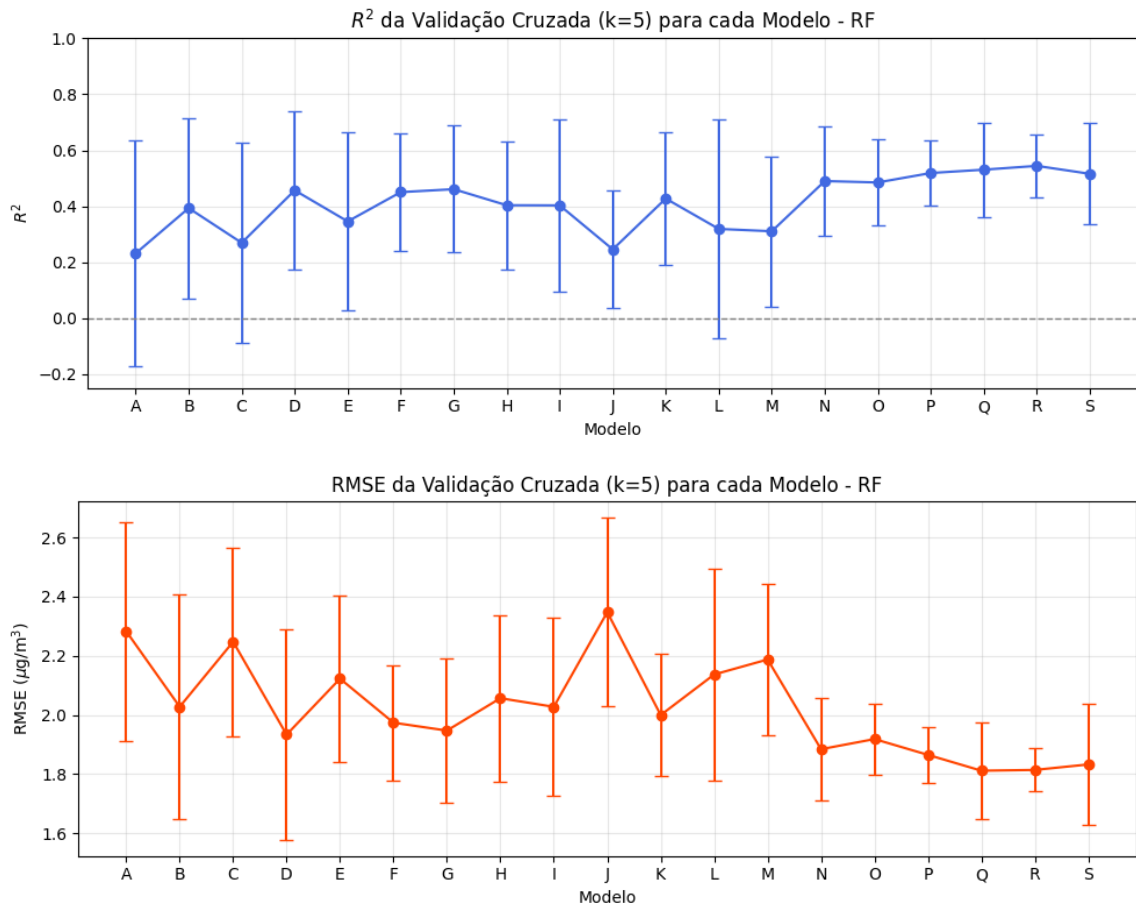


Figura 33 – R^2 e RMSE da validação cruzada ($k = 5$) para cada modelo *Random Forest* (RF).

inverno.

A literatura, normalmente aponta vantagens para *Random Forest* em conjuntos de dados maiores e/ou para $\text{PM}_{2.5}$, especialmente em cenários de elevada variabilidade meteorológica (HAYWARD et al., 2024; GIORDANO et al., 2021). Entretanto, os resultados deste estudo indicam que, em bases menores e para $\text{PM}_{1.0}$, ambos os métodos apresentam desempenhos próximos, sugerindo que a complexidade do modelo deve ser sempre ponderada em função do tamanho e da variabilidade amostral.

A variabilidade observada nas métricas de validação cruzada notadamente os desvios-padrão elevados de R^2 e RMSE em alguns modelos pode ser explicada por dois fatores centrais: (i) a elevada variabilidade sazonal das concentrações de $\text{PM}_{1.0}$, com destaque para o período de inverno, em que se observam flutuações mais intensas e ocorrência frequente de picos de concentração; e (ii) o número limitado de amostras disponíveis para calibração ($n = 79$), que torna os modelos mais sensíveis à divisão dos dados nos *folds* da validação cruzada.

Entre os modelos avaliados, destaca-se o Modelo A da Regressão Linear Simples, que obteve um valor médio de R^2 ($0,58 \pm 0,09$), aliado ao menor RMSE ($1,74 \pm 0,06 \mu\text{g}/\text{m}^3$)

e MAE ($1,31 \pm 0,10 \mu\text{g}/\text{m}^3$), além de apresentar o menor desvio-padrão das métricas entre todas as combinações testadas. Isso indica não apenas bom desempenho preditivo, mas também robustez e estabilidade em relação à divisão dos dados nos *folds* da validação cruzada. Já para o *Random Forest*, o Modelo R mostrou-se superior, com R^2 médio de $0,54 \pm 0,11$, RMSE de $1,83 \pm 0,20 \mu\text{g}/\text{m}^3$ e MAE de $1,35 \pm 0,19 \mu\text{g}/\text{m}^3$, além de apresentar, de modo similar ao MLR, baixa variabilidade nas métricas.

A equação média de calibração obtida para o Modelo A, considerando a validação cruzada ($k = 5$) pra a Regressão Linear Simples, é dada por:

$$\text{PM}_{1.0}^{\text{HI}} = 1,26 \pm 0,12 + (0,24 \pm 0,01) \cdot \overline{\text{PM}_{1.0}^{\text{PA}}} \quad (4.4)$$

Dessa forma, a seleção final dos modelos para análise comparativa e aplicação prática recaiu sobre o Modelo A para a regressão linear múltipla (MLR) e o Modelo R para o *Random Forest*, uma vez que ambos se destacaram não apenas pelos melhores valores médios de R^2 , RMSE e MAE, mas também pela menor dispersão dessas métricas ao longo dos *folds* de validação cruzada. Essa escolha assegura que as análises subsequentes estejam baseadas nos modelos com maior desempenho e estabilidade estatística entre todas as alternativas avaliadas.

Assim, recomenda-se a adoção destas configurações em aplicações futuras, ressaltando-se que a escolha do modelo ótimo pode ser sensível ao contexto amostral e à disponibilidade de variáveis preditoras.

Em síntese, os resultados da validação cruzada reforçam a necessidade de empregar técnicas robustas de avaliação em estudos de calibração de sensores, considerando tanto o desempenho médio quanto a incerteza associada a cada modelo. Recomendam-se estudos futuros com amostragens maiores e estratificadas por estação do ano, visando reduzir a influência da sazonalidade e aumentar a generalização dos modelos calibrados.

Tabela 6 – Comparação das métricas de desempenho (validação cruzada k -fold) para os modelos de regressão linear múltipla (MLR) e *Random Forest* (RF).

Modelo	MLR			RF		
	R^2	RMSE	MAE	R^2	RMSE	MAE
A	$0,58 \pm 0,09$	$1,74 \pm 0,06$	$1,31 \pm 0,10$	$0,23 \pm 0,40$	$2,28 \pm 0,37$	$1,57 \pm 0,16$
B	$0,56 \pm 0,13$	$1,77 \pm 0,10$	$1,32 \pm 0,08$	$0,39 \pm 0,32$	$2,03 \pm 0,38$	$1,42 \pm 0,19$
C	$0,58 \pm 0,07$	$1,76 \pm 0,11$	$1,33 \pm 0,16$	$0,27 \pm 0,36$	$2,25 \pm 0,32$	$1,56 \pm 0,15$
D	$0,59 \pm 0,06$	$1,74 \pm 0,11$	$1,32 \pm 0,18$	$0,46 \pm 0,28$	$1,93 \pm 0,36$	$1,36 \pm 0,16$
E	$0,55 \pm 0,11$	$1,81 \pm 0,15$	$1,36 \pm 0,17$	$0,35 \pm 0,32$	$2,12 \pm 0,28$	$1,48 \pm 0,10$
F	$0,55 \pm 0,10$	$1,81 \pm 0,11$	$1,36 \pm 0,19$	$0,45 \pm 0,21$	$1,97 \pm 0,19$	$1,43 \pm 0,09$
G	$0,57 \pm 0,10$	$1,77 \pm 0,11$	$1,33 \pm 0,16$	$0,46 \pm 0,23$	$1,95 \pm 0,24$	$1,42 \pm 0,13$
H	$0,58 \pm 0,07$	$1,77 \pm 0,10$	$1,35 \pm 0,18$	$0,40 \pm 0,23$	$2,06 \pm 0,28$	$1,47 \pm 0,15$
I	$0,50 \pm 0,15$	$1,90 \pm 0,23$	$1,37 \pm 0,14$	$0,40 \pm 0,31$	$2,03 \pm 0,30$	$1,47 \pm 0,14$
J	$0,48 \pm 0,14$	$1,96 \pm 0,31$	$1,49 \pm 0,24$	$0,25 \pm 0,21$	$2,35 \pm 0,32$	$1,77 \pm 0,30$
K	$0,53 \pm 0,16$	$1,83 \pm 0,12$	$1,37 \pm 0,05$	$0,43 \pm 0,24$	$2,00 \pm 0,21$	$1,49 \pm 0,11$
L	$0,47 \pm 0,19$	$1,93 \pm 0,21$	$1,39 \pm 0,10$	$0,32 \pm 0,39$	$2,14 \pm 0,36$	$1,54 \pm 0,15$
M	$0,35 \pm 0,27$	$2,12 \pm 0,31$	$1,56 \pm 0,20$	$0,31 \pm 0,27$	$2,19 \pm 0,26$	$1,61 \pm 0,20$
N	$0,55 \pm 0,15$	$1,78 \pm 0,12$	$1,36 \pm 0,12$	$0,49 \pm 0,19$	$1,88 \pm 0,17$	$1,39 \pm 0,15$
O	$0,56 \pm 0,12$	$1,77 \pm 0,09$	$1,37 \pm 0,13$	$0,48 \pm 0,15$	$1,92 \pm 0,12$	$1,44 \pm 0,09$
P	$0,57 \pm 0,10$	$1,77 \pm 0,06$	$1,37 \pm 0,17$	$0,52 \pm 0,12$	$1,87 \pm 0,09$	$1,44 \pm 0,11$
Q	$0,54 \pm 0,17$	$1,79 \pm 0,13$	$1,38 \pm 0,09$	$0,53 \pm 0,17$	$1,81 \pm 0,16$	$1,33 \pm 0,16$
R	$0,56 \pm 0,13$	$1,78 \pm 0,07$	$1,39 \pm 0,10$	$0,54 \pm 0,11$	$1,81 \pm 0,07$	$1,37 \pm 0,10$
S	$0,51 \pm 0,20$	$1,85 \pm 0,17$	$1,41 \pm 0,07$	$0,52 \pm 0,18$	$1,83 \pm 0,20$	$1,35 \pm 0,19$

4.3.3.2 Curvas de Aprendizado dos Modelos Selecionados

A análise das curvas de aprendizado é fundamental para avaliar a capacidade de generalização dos modelos de regressão, especialmente em contextos de conjunto de dados reduzido, como o deste estudo. As Figuras 34 e 35 apresentam as curvas de aprendizado para o modelo de Regressão Linear Múltipla (MLR, Modelo A) e para o Random Forest (RF, Modelo R), respectivamente. Em ambos os casos, são mostrados os valores médios e desvios-padrão do R^2 e do RMSE em função do número de amostras de treinamento.

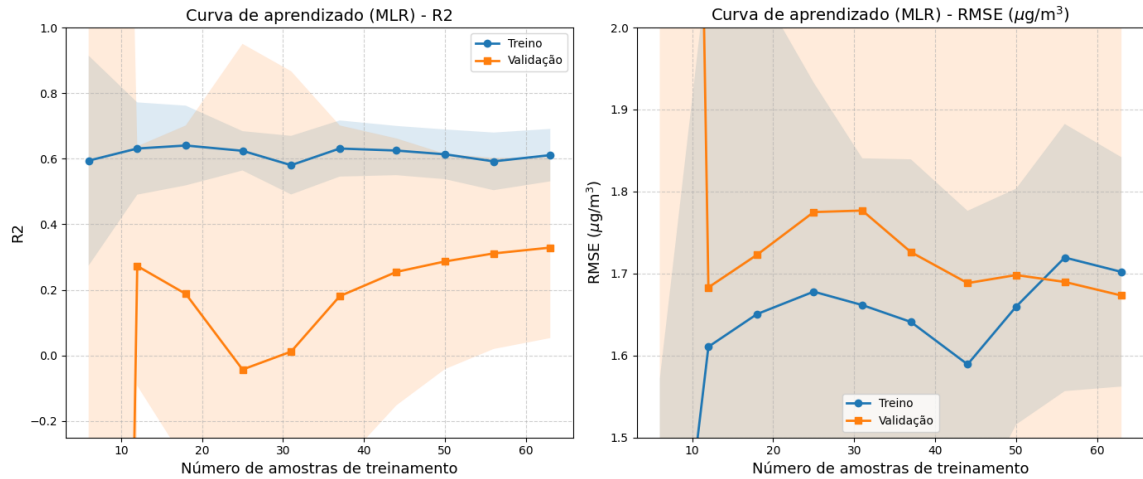


Figura 34 – Curva de aprendizado para o modelo MLR-A: evolução de R^2 e RMSE em função do tamanho do conjunto de treinamento.

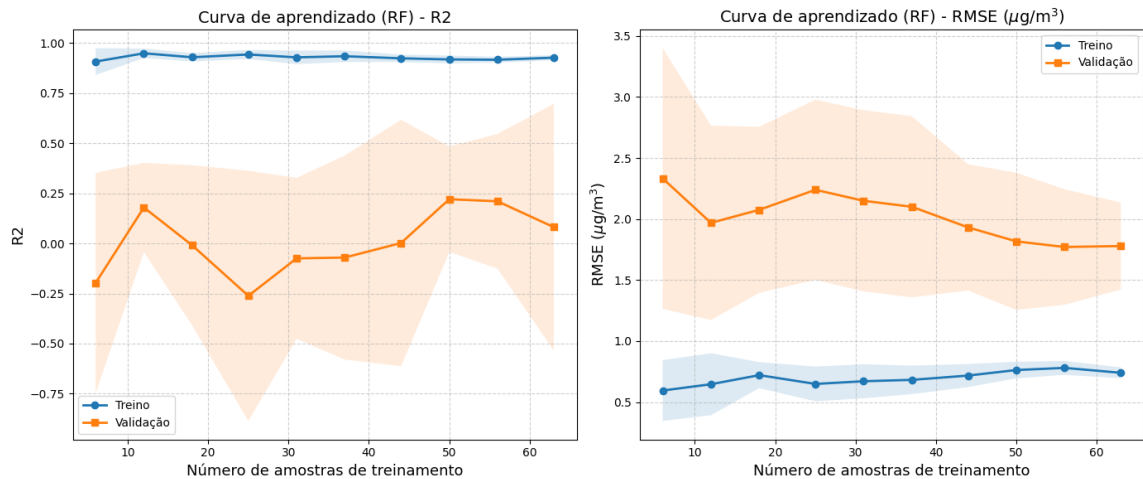


Figura 35 – Curva de aprendizado para o modelo RF-R: evolução de R^2 e RMSE em função do tamanho do conjunto de treinamento.

No caso do modelo MLR, observa-se que o desempenho no conjunto de treino ($R^2_{\text{treino}} \sim 0,7$) é consistentemente superior ao observado na validação, cuja média de R^2 cresce de valores negativos (para amostras pequenas) até valores positivos próximos de 0,3 a 0,4, à medida que aumenta o número de amostras de treinamento. O RMSE de validação, por sua vez, é consideravelmente maior do que o RMSE de treino para tamanhos reduzidos de amostra, convergindo gradualmente à medida que mais dados são utilizados. As amplas faixas de desvio padrão observadas para as métricas de validação indicam alta variância e instabilidade estatística nos resultados, fenômeno esperado em bases pequenas e com alta variabilidade intrínseca.

No caso do *Random Forest* (RF, Modelo R), o modelo exibe um R^2 de treino muito elevado (acima de 0,85), denotando forte capacidade de ajuste aos dados de treinamento. Entretanto, o R^2 de validação permanece, na maior parte dos tamanhos de amostra,

inferior ao do MLR, situando-se predominantemente entre 0 e 0,3, com intervalos de confiança ainda mais amplos, principalmente para conjuntos de treinamento reduzidos. O RMSE de validação segue o mesmo padrão: valores elevados e pouco estáveis, convergindo apenas parcialmente com o aumento do número de amostras. Esse comportamento é característico de *overfitting*, comum em modelos mais complexos aplicados a conjuntos de dados limitados,

Tanto MLR quanto RF exibem bandas de incerteza largas, reforçando a importância de conjuntos de dados mais extensos e representativos para se obter estimativas robustas de desempenho. Esse comportamento também pode ser atribuído à elevada variabilidade das concentrações de $PM_{1.0}$ ao longo da série temporal, em especial nos meses de inverno, aumentando o desafio para a generalização dos modelos.

Cabe ressaltar que as curvas de aprendizado tendem a subestimar o desempenho preditivo real dos modelos quando comparadas às métricas médias de validação cruzada, devido à elevada variabilidade entre os *splits* com poucos dados. Na validação cruzada, o uso mais eficiente do conjunto de dados resulta em métricas mais estáveis e valores de R^2 superiores ($R^2_{MLR} = 0.58 \pm 0.09$ e $R^2_{RF} = 0.54 \pm 0.11$), acompanhados de desvios padrão menores. Portanto, os valores obtidos via validação cruzada devem ser priorizados na avaliação da capacidade preditiva dos modelos, enquanto as curvas de aprendizado são fundamentais para indicar tendências de estabilidade, impacto do tamanho amostral e ocorrência de sobreajuste.

4.3.4 Análise dos Erros dos Modelos

A avaliação detalhada dos erros de predição é fundamental para caracterizar o desempenho e as limitações dos modelos ajustados. Neste estudo, as métricas *RMSE* (erro quadrático médio), *MAE* (erro absoluto médio), *MNB* (viés médio normalizado), *CvMAE* (erro absoluto médio relativo) e *nRMSE* (erro quadrático médio normalizado) foram empregadas para quantificar o desvio entre os valores preditos e observados de $PM_{1.0}$.

A Tabela 7 apresenta o resumo das métricas obtidas por validação cruzada (*k*-fold, *k* = 5) para os modelos selecionados: MLR-A e RF-R.

Tabela 7 – Métricas de erro médias e desvios-padrão dos modelos ajustados ($n = 79$, $k = 5$).

Modelo	<i>RMSE</i>	<i>MAE</i>	<i>MNB</i> (%)	<i>CvMAE</i>	<i>nRMSE</i>	R^2
MLR-A	1.74 ± 0.06	1.31 ± 0.10	32.5 ± 21.8	0.297 ± 0.017	0.178 ± 0.015	0.58 ± 0.09
RF-R	1.81 ± 0.07	1.37 ± 0.10	34.0 ± 22.2	0.307 ± 0.016	0.186 ± 0.013	0.54 ± 0.11

Nota-se que ambos os modelos apresentaram desempenho similar em termos de

erro absoluto (MAE) e erro quadrático médio ($RMSE$), com valores próximos de 1,3 a 1,8 $\mu\text{g}/\text{m}^3$, e coeficiente de determinação (R^2) entre 0,54 e 0,58. O viés normalizado (MNB) manteve-se ao redor de 33%, indicando tendência de superestimação nas predições.

4.3.4.1 Análise dos Resíduos

A Tabela 8 resume os principais parâmetros estatísticos dos resíduos para ambos os modelos. Em ambos os casos, as médias e medianas são próximas de zero, sugerindo ausência de viés sistemático significativo (ou seja, erros que tendem a se compensar). O desvio-padrão dos resíduos é similar (MLR-A: 1,74 $\mu\text{g}/\text{m}^3$; RF-R: 1,83 $\mu\text{g}/\text{m}^3$), evidenciando a variabilidade típica de dados ambientais de campo, especialmente em cenários com número de amostras reduzido ($n = 79$).

A assimetria (*skewness*) e a curtose dos resíduos também foram avaliadas. A assimetria indica a simetria da distribuição em torno da média; valores próximos de zero sugerem uma distribuição aproximadamente simétrica, como observado para ambos os modelos (MLR-A: -0,06; RF-R: 0,31). Já a curtose indica o grau de concentração dos resíduos ao redor da média; valores próximos de zero indicam proximidade à distribuição normal. Ambos os modelos apresentaram curtose levemente positiva (MLR-A: 0,71; RF-R: 0,55), sinalizando uma leve concentração de resíduos próximos ao zero, sem indícios de caudas longas ou excesso de *outliers*.

Tabela 8 – Estatísticas descritivas dos resíduos para os modelos MLR-A e RF-R da validação cruzada.

Métrica	MLR-A	RF-R
Média	-0,01	-0,05
Mediana	-0,05	-0,04
Desvio-padrão	1,74	1,83
Assimetria	-0,06	0,31
Curtose	0,71	0,55
n outliers	5	5
Proporção outliers	0,063	0,063

Para verificar se os resíduos seguem distribuição aproximadamente normal (requisito importante para a validade de diversos testes estatísticos), foi aplicado o teste de Shapiro-Wilk. Os valores de p -valor encontrados foram 0,562 para o MLR-A e 0,196 para o RF-R, ambos superiores ao limiar convencional de 0,05. Assim, não há evidências estatísticas para rejeitar a hipótese de normalidade dos resíduos em ambos os modelos, reforçando a adequação dos ajustes para uso em inferência e análise dos erros.

A Figura 36 exibe o boxplot dos resíduos para ambos os modelos. Observa-se que a maioria dos resíduos se concentra próxima de zero, com poucos outliers nos extremos. As

linhas tracejadas representam as médias dos resíduos, acompanhadas do valor numérico destacado sobre cada boxplot. Esse padrão sugere que ambos os modelos não apresentam viés sistemático (tendência de super ou subestimação global).

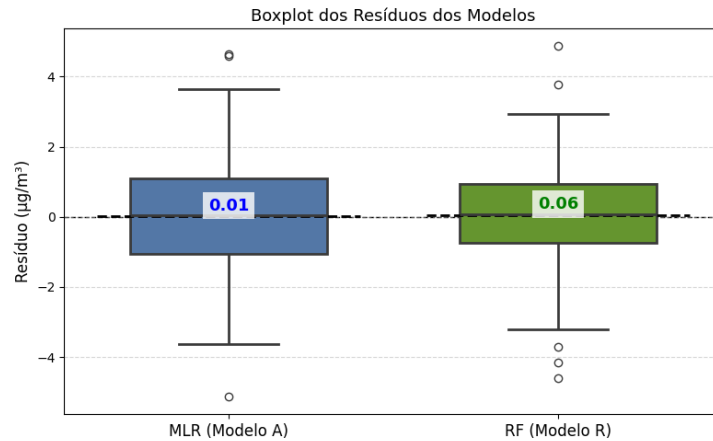


Figura 36 – Boxplot dos resíduos dos modelos MLR-A e RF-R.

A Figura 37 apresenta a densidade dos resíduos, evidenciando a aproximação à distribuição normal, como reforçado pelo teste de Shapiro-Wilk e pelos valores baixos de assimetria e curtose.

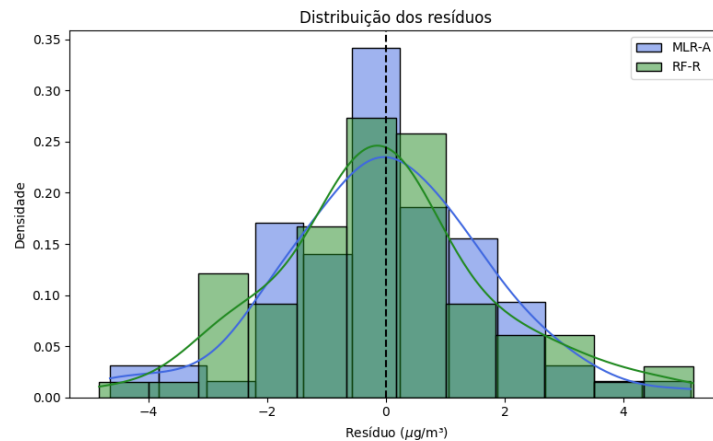


Figura 37 – Distribuição de densidade dos resíduos dos modelos MLR-A e RF-R

A autocorrelação dos resíduos foi analisada com gráficos de autocorrelação parcial na Figura 38, os quais demonstraram ausência de autocorrelação significativa (valores dentro das bandas de confiança em todos os *lags*), indicando independência temporal dos erros. Esse resultado é fundamental para garantir que não há efeitos de memória ou padrões não modelados ao longo das previsões.

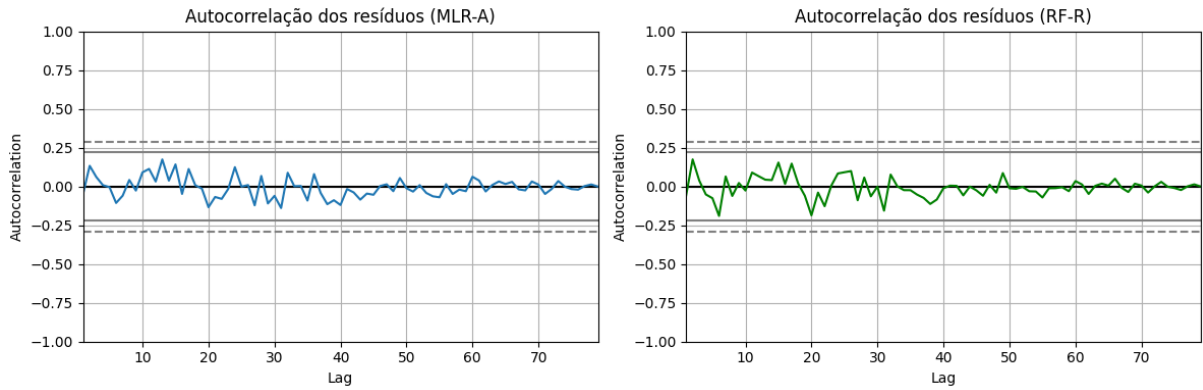


Figura 38 – Autocorrelação dos resíduos para MLR-A e RF-R.

A Figura 39 apresenta os gráficos de resíduos versus valores previstos. Nota-se leve tendência de aumento da dispersão dos resíduos para concentrações mais altas, principalmente para o modelo MLR-A, o que foi confirmado pelo teste de Breusch-Pagan ($p = 0,0126$), indicando presença de heterocedasticidade. Em aplicações ambientais, esse fenômeno geralmente é atribuído à maior dificuldade de previsão em cenários de alta variabilidade e concentração de $PM_{1.0}$.

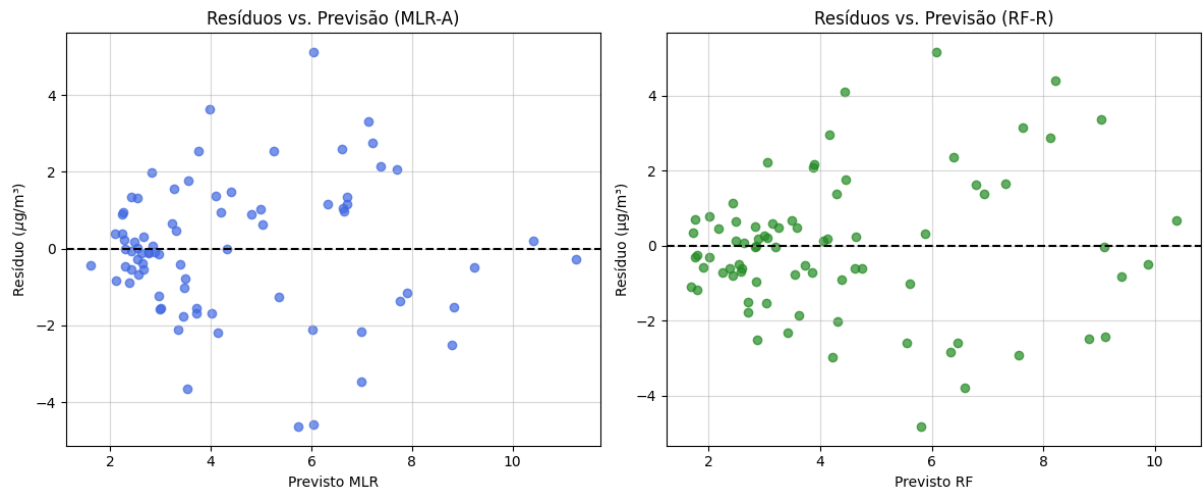


Figura 39 – Gráficos de resíduos versus valores previstos para os modelos MLR-A e RF-R.

A análise integrada dos resíduos mostra que ambos os modelos, MLR-A e RF-R, apresentam desempenho estatístico semelhante no contexto de calibração de $PM_{1.0}$ em campo. Os resíduos são centrados, com distribuição próxima da normal, baixa autocorrelação e proporção moderada de outliers. O RF-R apresenta leve assimetria positiva, mas ainda dentro de limites aceitáveis. Além disso, presença de resíduos extremos e a leve heterocedasticidade observada refletem o impacto da variabilidade ambiental¹ e dos eventos de $PM_{1.0}$ atípicos.

¹ Conforme observado no período de inverno da Série Temporal da Fig. 15

4.3.5 Influência das Condições Meteorológicas nos Modelos

A Figura 40 apresenta os gráficos de dispersão das previsões dos modelos MLR-A e RF-R, coloridos de acordo com a temperatura média observada durante o período amostral. De modo geral, observa-se que a distribuição dos resíduos em relação à linha 1:1 não revela um padrão claro de viés associado a faixas específicas de temperatura, indicando que ambos os modelos mantêm desempenho relativamente consistente ao longo do espectro térmico observado.

No caso da umidade relativa (Figura 41), observa-se que as maiores discrepâncias entre as previsões e a referência tendem a ocorrer nos extremos do intervalo de umidade. Os dados demonstram leve tendência à subestimação das concentrações reais, sobretudo quando a umidade relativa do ar atinge valores mais elevados, um efeito consistente com o que é relatado na literatura para sensores ópticos de baixo custo.

Quanto à pressão atmosférica (Figura 42), os gráficos de dispersão não evidenciam padrão sistemático de erro associado a diferentes faixas de pressão. A distribuição dos pontos não sugere relação direta entre desvios do modelo e variações na pressão.

De forma geral, a análise conjunta dos gráficos indica que a umidade relativa, podem impactar de maneira pontual a acurácia das estimativas, principalmente em concentrações mais elevadas.

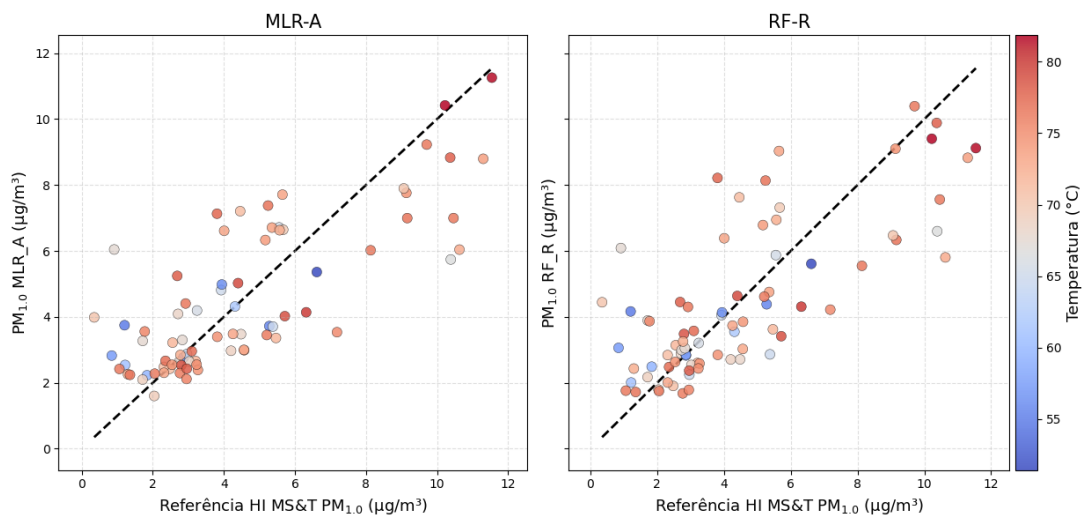


Figura 40 – Gráficos de dispersão das previsões calibradas dos modelos MLR-A e RF-R, coloridos conforme a temperatura média durante o período amostral. A linha tracejada indica a correspondência 1:1.

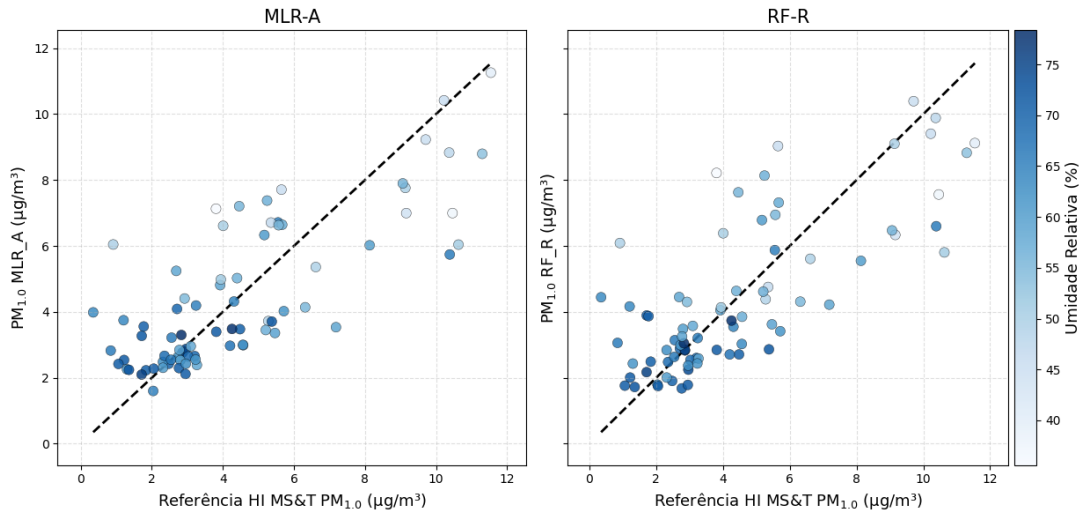


Figura 41 – Gráficos de dispersão das previsões calibradas dos modelos MLR-A e RF-R, coloridos conforme a umidade relativa média durante o período amostral. A linha tracejada indica a correspondência 1:1.

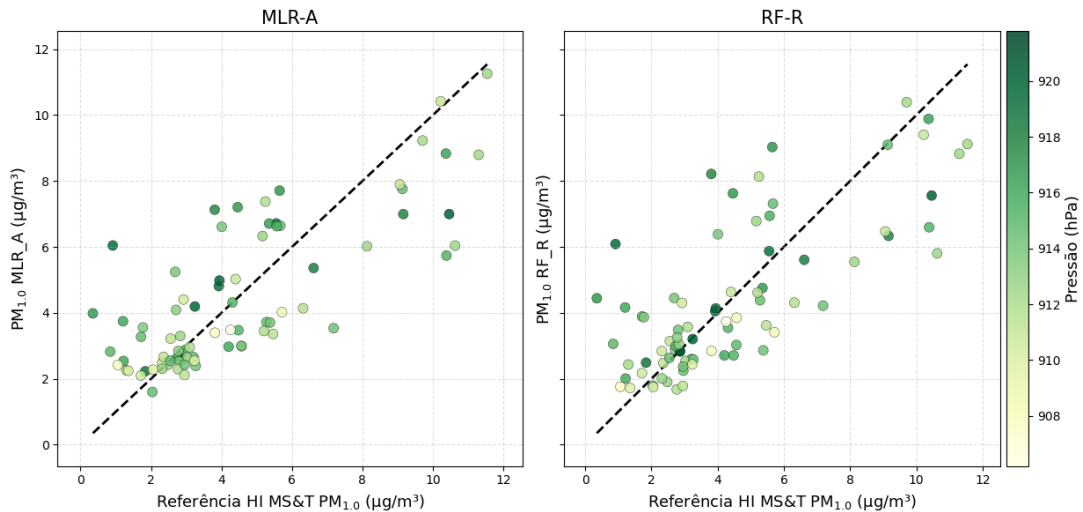


Figura 42 – Gráficos de dispersão das previsões calibradas dos modelos MLR-A e RF-R, coloridos conforme a pressão atmosférica média durante o período amostral. A linha tracejada indica a correspondência 1:1.

4.4 Concentração pelo método ALT

A determinação da concentração mássica de $PM_{1.0}$ pelo método alternativo (ALT), baseada na conversão das contagens diferenciais de partículas do sensor PA-II-FLEX, foi avaliada por meio de regressão linear simples frente ao método de referência HI MS&T. A Tabela 9 apresenta as métricas de desempenho das principais combinações de médias e densidades, tanto na validação *holdout* (20% teste) quanto por validação cruzada (k -Fold, $k = 5$).

Tabela 9 – Desempenho das combinações de médias e densidades para o cálculo alternativo de $PM_{1.0}$ (método Hanout), comparando o ajuste por holdout (20% teste) e validação cruzada ($K=5$).

Média	Densidade	holdout (20% teste)			K-Fold (k=5)		
		R^2	RMSE	MAE	R^2	RMSE	MAE
Geométrica	2,50	0,6886	1,8276	1,4169	$0,5537 \pm 0,1026$	$1,7985 \pm 0,0287$	$1,4085 \pm 0,0583$
Aritmética	2,50	0,6711	1,8783	1,3464	$0,4528 \pm 0,1797$	$1,9703 \pm 0,1036$	$1,4669 \pm 0,1407$
Harmônica	2,50	0,6001	2,0711	1,6875	$0,5055 \pm 0,0657$	$1,9117 \pm 0,0956$	$1,5065 \pm 0,0908$
Aritmética	1,65	0,4809	2,3598	1,9179	$0,3958 \pm 0,0564$	$2,1209 \pm 0,1519$	$1,6794 \pm 0,1148$
Geométrica	1,65	0,2943	2,7514	2,2245	$0,2058 \pm 0,0566$	$2,4364 \pm 0,2055$	$1,9411 \pm 0,1393$
Harmônica	1,65	0,0899	3,1246	2,5051	$-0,0149 \pm 0,0663$	$2,7558 \pm 0,2434$	$2,1946 \pm 0,1523$
Aritmética	1,00	-0,2011	3,5894	2,9229	$-0,3410 \pm 0,0906$	$3,1676 \pm 0,2811$	$2,5779 \pm 0,1873$
Geométrica	1,00	-0,4060	3,8836	3,1792	$-0,5751 \pm 0,1125$	$3,4323 \pm 0,3028$	$2,8217 \pm 0,2087$
Harmônica	1,00	-0,5976	4,1397	3,3981	$-0,7964 \pm 0,1358$	$3,6648 \pm 0,3206$	$3,0327 \pm 0,2265$

Observa-se que as melhores combinações, em particular a média geométrica com densidade $2,50 \text{ g/cm}^3$, atingiram valores de R^2 próximos a 0,68 no teste *holdout* e $0,55 \pm 0,10$ por validação cruzada, com RMSE entre 1,8 e $1,9 \text{ } \mu\text{g/m}^3$. Apesar desses resultados indicarem desempenho razoável para uma abordagem baseada apenas em dados primários de contagem, a precisão dos modelos alternativos foi inferior àquela observada para a concentração fornecida diretamente pelo sensor, que atingiu valores de R^2 superiores a 0,70 com menores erros absolutos médios (MAE) (*holdout*).

O gráfico de dispersão apresentado na Figura 43 ilustra o desempenho do modelo de regressão linear simples ajustado sobre os dados de teste (*holdout*), para a configuração de melhor desempenho (média geométrica, densidade $2,50 \text{ g/cm}^3$). Nota-se que, embora a correlação seja razoável ($R^2 = 0,68$), o número de amostras disponíveis na validação foi reduzido ($n = 16$), o que pode limitar a robustez da avaliação. Os pontos de dispersão mostram tendência de alinhamento com a linha 1:1, porém com maior dispersão para concentrações elevadas e certo viés subestimativo em valores extremos.

Esses resultados evidenciam as limitações do método alternativo de cálculo para $PM_{1.0}$, especialmente no contexto de sensores ópticos de baixo custo. A necessidade de suposições adicionais (densidade, formato das partículas, distribuição de tamanhos) e o aumento da propagação de incertezas tornam o desempenho final sensivelmente dependente da configuração adotada. Além disso, a calibração direta baseada na concentração total reportada pelo sensor tende a capturar melhor as características do aerossol e as singularidades do instrumento, resultando em métricas estatísticas superiores.

Portanto, embora o método alternativo de cálculo apresente desempenho compatível, ele se mostra menos competitivo em relação à estratégia de ajuste direto sobre a leitura bruta do sensor.

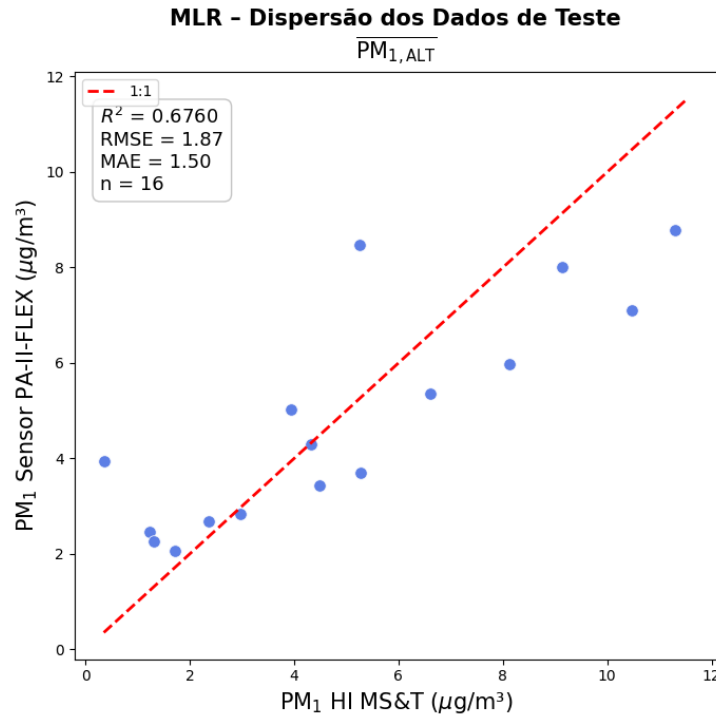


Figura 43 – Dispersão das concentrações $PM_{1,0}$ calculadas pelo método alternativo (média geométrica, densidade $2,50 \text{ g/cm}^3$) em relação ao método de referência HI MS&T nos dados de teste (*holdout*).

4.5 Análise Temporal dos Resultados

A análise temporal dos resultados obtidos pelos modelos MLR-A e RF-R permite avaliar não apenas a precisão global das correções, mas também sua estabilidade ao longo do tempo, considerando variações sazonais e possíveis tendências sistemáticas.

A Figura 44 apresenta o boxplot das concentrações de $PM_{1,0}$ estimadas pelos modelos, juntamente com os valores de referência. Observa-se que as medianas e médias dos modelos calibrados (MLR-A e RF-R) são semelhantes com a referência, indicando baixo viés sistemático relevante. Além disso, a dispersão das previsões corrigidas se aproxima da dispersão dos dados de referência, enquanto o sensor bruto tende a apresentar maior variabilidade e valores máximos mais elevados, como pode ser observado na Figura 14.

A Figura 45 mostra a evolução temporal das concentrações de $PM_{1,0}$ para cada campanha amostral, comparando os valores observados (referência) com as correções obtidas pelos modelos MLR-A, RF-R e o sensor bruto. É possível observar que ambos os modelos acompanharam de forma bastante próxima as variações sazonais e tendências temporais presentes nos dados de referência, com destaque para a redução dos picos extremos observados no sensor não corrigido.

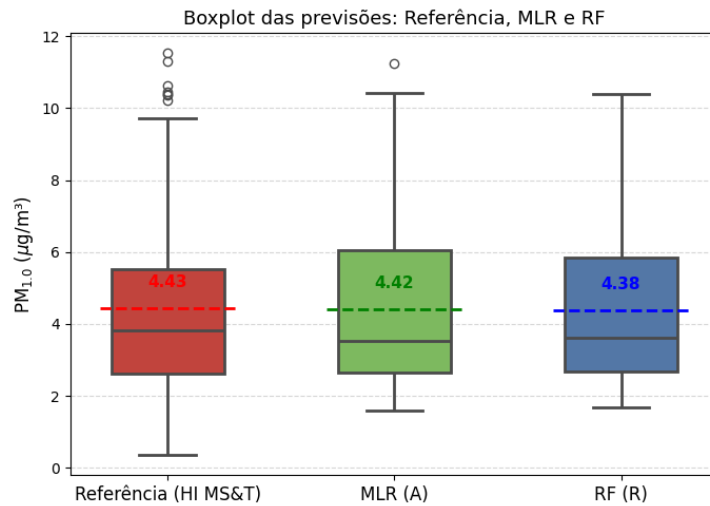


Figura 44 – Boxplot das concentrações de $PM_{1.0}$ da referência (HI MS&T), modelo MLR-A e modelo RF-R. As linhas tracejadas indicam as médias.

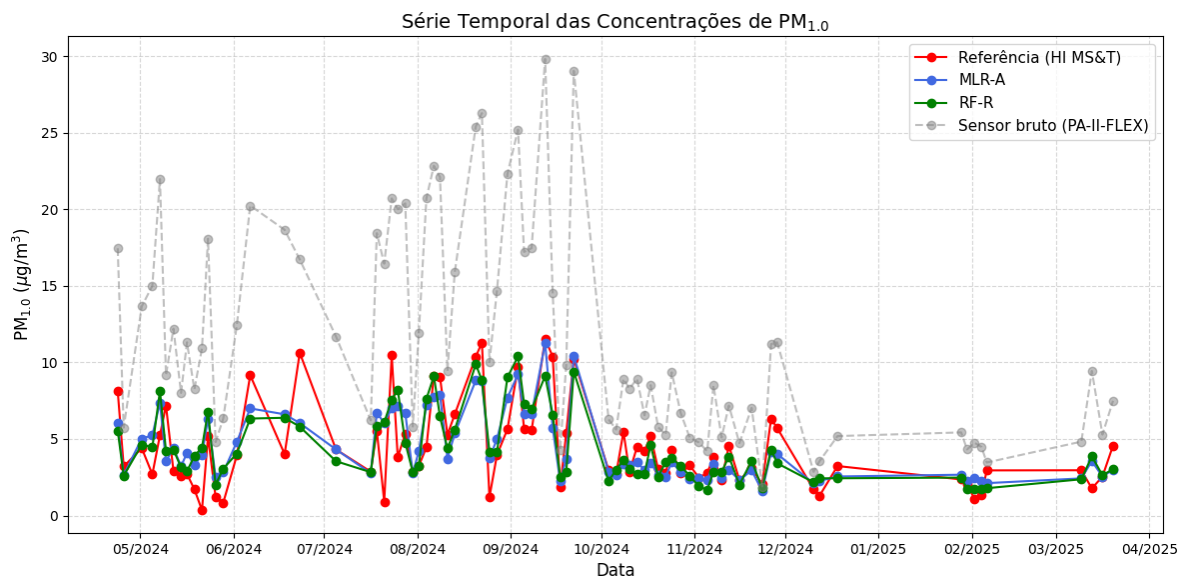


Figura 45 – Série temporal das concentrações de $PM_{1.0}$: referência, modelos calibrados (MLR-A e RF-R) e sensor bruto (PA-II-FLEX).

Os maiores picos de concentração de $PM_{1.0}$ foram registrados no intervalo entre 10/07/2024 e 15/08/2024, atingindo valores superiores a $10 \mu g/m^3$, período que coincide com o inverno na região de Curitiba. Estes eventos críticos podem ser atribuídos a condições meteorológicas adversas, como estabilidade atmosférica, inversões térmicas e, possivelmente, à ocorrência de queimadas regionais, que favorecem o acúmulo de material particulado na atmosfera.

Após setembro de 2024, verifica-se uma tendência de queda nas concentrações, refletindo a transição para períodos de maior precipitação e maior dispersão atmosférica durante a primavera e o verão. Durante os meses de dezembro de 2024 a fevereiro de 2025,

as concentrações de $PM_{1.0}$ permaneceram em níveis reduzidos, inferiores a $4 \mu g/m^3$, com poucas oscilações significativas.

Ambos os modelos (MLR e RF) mostraram boa capacidade de replicar as variações temporais observadas pela referência, destacando-se a acurácia do RF nas situações de maior variabilidade, com ajustes mais próximos dos picos máximos e mínimos. Pequenas divergências entre as curvas dos modelos e da referência foram mais evidentes nos eventos de pico, o que pode ser reflexo de limitações na resolução temporal dos métodos gravimétricos ou de variáveis ambientais não incorporadas ao modelo.

A estabilidade dos modelos ao longo do tempo também pode ser avaliada pela evolução dos resíduos (diferença entre previsão e referência) durante toda a série amostral, conforme apresentado na Figura 46. Ambos os modelos exibem resíduos oscilando em torno de zero.

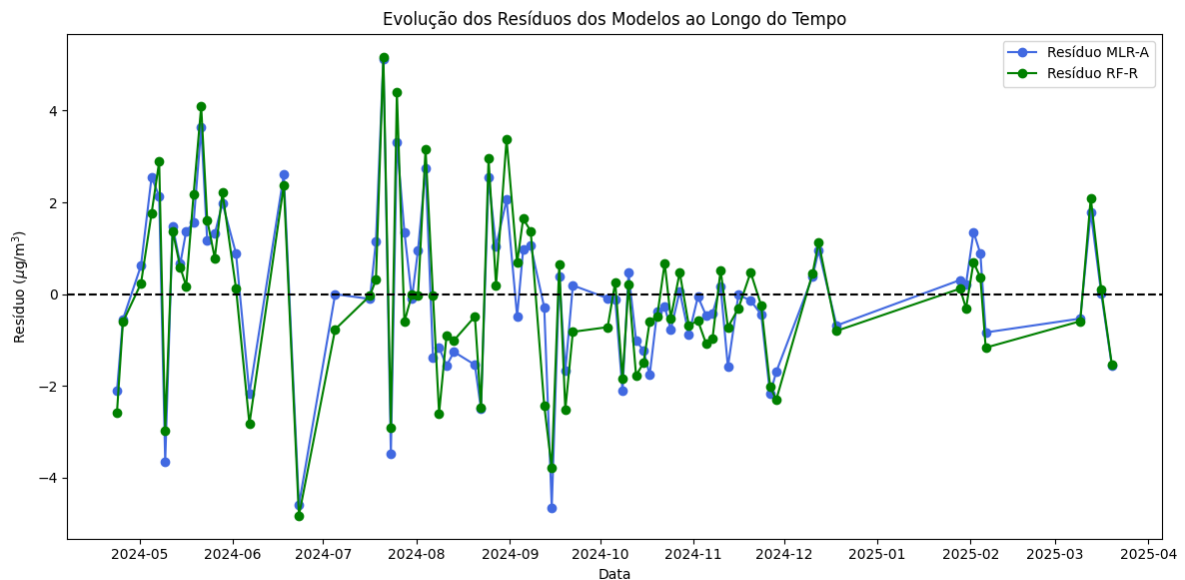


Figura 46 – Série temporal dos resíduos para os modelos MLR-A e RF-R (diferença entre previsão e referência em cada campanha).

Ressalta-se, contudo, que em alguns episódios isolados de alta concentração, ambos os modelos apresentaram resíduos mais acentuados, o que pode ser atribuído à maior incerteza nas condições extremas de operação e ao próprio desafio de calibrar sensores ópticos em cenários ambientais reais, com variabilidade meteorológica e de fontes de emissão.

De modo geral, a avaliação temporal demonstra que os modelos calibrados apresentaram desempenho consistente ao longo das diferentes campanhas e estações do ano, reduzindo vieses do sensor óptico e proporcionando previsões robustas para aplicações de monitoramento ambiental em campo. O acompanhamento da estabilidade das correções ao longo do tempo é fundamental para validar a aplicabilidade dos modelos em estudos de

longa duração ou redes de sensores de baixo custo.

4.6 Comparação dos Modelos

Durante o desenvolvimento deste estudo, foram analisadas 19 diferentes combinações de variáveis independentes, denominadas de Modelos A a S, com o intuito de calibrar de forma mais precisa o sensor óptico de $PM_{1.0}$. Cada configuração foi avaliada utilizando dois algoritmos principais: regressão linear (simples e múltipla) e *Random Forest*. O processo envolveu a seleção criteriosa das variáveis e a comparação do desempenho preditivo entre diferentes abordagens estatísticas.

Para garantir uma avaliação abrangente e confiável dos modelos, foram empregados dois métodos de validação. O primeiro, conhecido como separação *holdout*, consiste em dividir o conjunto de dados em 80% para treinamento e 20% para teste, permitindo medir diretamente o desempenho do modelo em um subconjunto reservado. Já a validação cruzada *k-Fold* ($k = 5$) subdivide os dados em cinco partes iguais e alterna o grupo de teste em cada iteração, reduzindo a influência de partições amostrais e proporcionando métricas de desempenho mais robustas, especialmente relevantes diante do número limitado de amostras ($n = 79$) disponível neste estudo.

Os resultados obtidos pelo método *holdout* mostraram, em geral, métricas superiores, em especial para o modelo RF-Q, como destacado na Tabela 10 e ilustrado na Figura 47. Observa-se, nesses gráficos, um melhor alinhamento dos pontos previstos à linha 1:1, indicando um ajuste eficaz aos dados de teste reservados. Contudo, é importante salientar que tal abordagem pode superestimar o desempenho real dos modelos, pois a escolha aleatória dos dados de teste pode não refletir adequadamente a variabilidade do conjunto completo, especialmente em amostras pequenas, facilitando a ocorrência de sobreajuste (*overfitting*).

Por esse motivo, a validação cruzada *k-Fold* foi adotada como referência principal para seleção dos modelos. Conforme mostra a Figura 48, tanto o MLR-A quanto o RF-R calibrados por validação cruzada apresentam um desempenho consistente, com previsões mais próximas da linha 1:1 em relação aos dados brutos do sensor (pontos cinza), refletindo o aumento de precisão e a efetividade da calibração. De modo geral, os pontos se concentram ao redor da linha de concordância, especialmente nas faixas de baixa e média concentração, o que indica um ajuste satisfatório dos modelos na maior parte do espectro amostral. Entretanto, verifica-se uma tendência de subestimação das concentrações mais elevadas de $PM_{1.0}$, mais pronunciada no modelo RF-R, cuja dispersão aumenta nos extremos superiores do eixo de referência. Esse comportamento revela a dificuldade dos modelos em extrapolar para valores pouco representados na amostra.

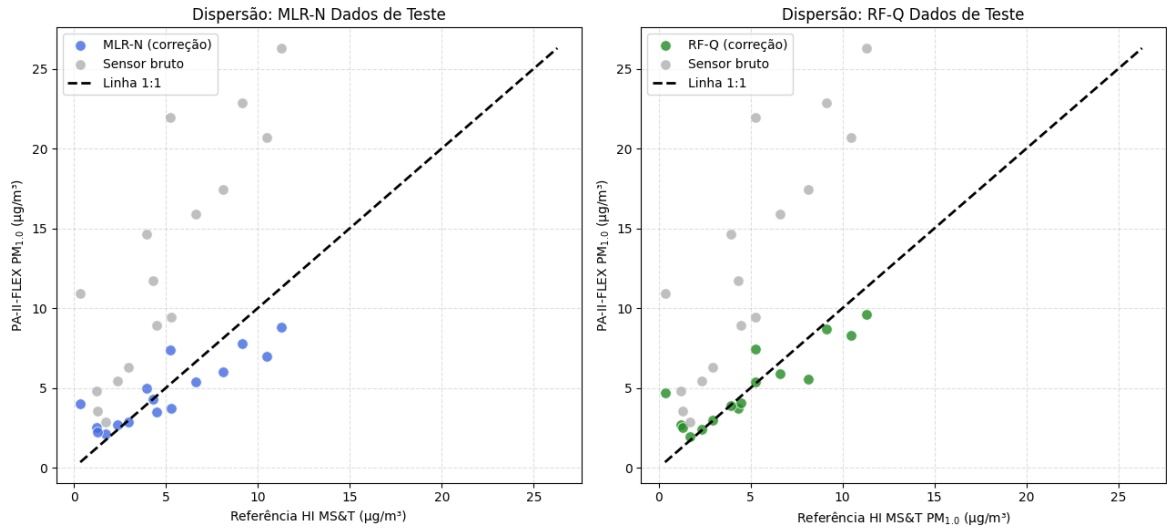


Figura 47 – Gráficos de dispersão das previsões calibradas dos modelos MLR-N e RF-Q nos dados de teste (*holdout*), em relação à referência HI MS&T.

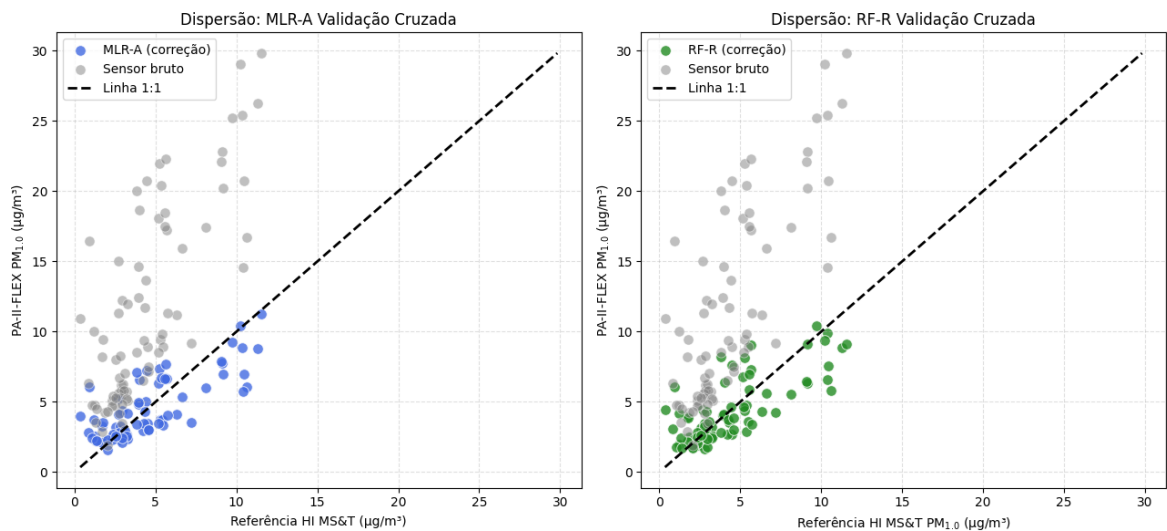


Figura 48 – Gráficos de dispersão das previsões calibradas dos modelos MLR-A (esquerda) e RF-R (direita) por validação cruzada, em relação à referência gravimétrica HI MS&T. A linha tracejada representa a correspondência 1:1.

Essa análise é aprofundada na Figura 49, que apresenta as médias e desvios padrão das previsões dos modelos MLR-A e RF-R em diferentes faixas de concentração gravimétrica de referência. Nota-se que ambos os modelos acompanham fielmente os valores de referência nas faixas intermediárias (2–4 e 4–6 $\mu\text{g}/\text{m}^3$), com pequenas diferenças entre si e baixa dispersão dos resultados. Por outro lado, tanto em faixas de concentração muito baixas quanto elevadas, há uma leve tendência de subestimação, acompanhada de aumento nos desvios padrão. Isso reforça a limitação dos modelos para capturar eventos menos frequentes ou extremos, mas demonstra que, na maior parte do tempo, a calibração proporciona estimativas robustas e alinhadas ao comportamento real das medições de referência.

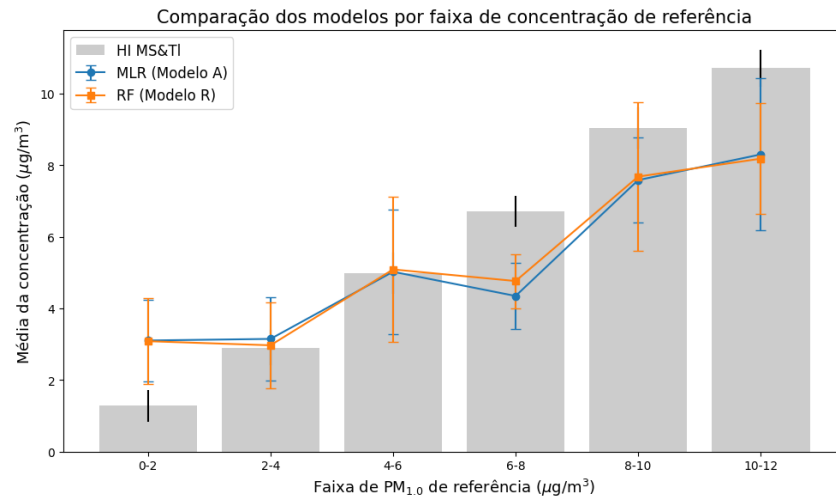


Figura 49 – Média das concentrações previstas pelos modelos MLR-A e RF-R em comparação com a referência gravimétrica HI MS&T, segmentadas por faixas de concentração de referência. As barras de erro representam o desvio padrão em cada faixa.

A Figura 50 ilustra o ganho de precisão obtido com a aplicação dos modelos calibrados sobre os dados brutos do sensor, evidenciando a aproximação dos pontos previstos à linha 1:1 em toda a faixa de concentração. Esse resultado destaca a importância do processo de calibração, que reduz vieses e amplia a confiabilidade dos dados do sensor óptico.

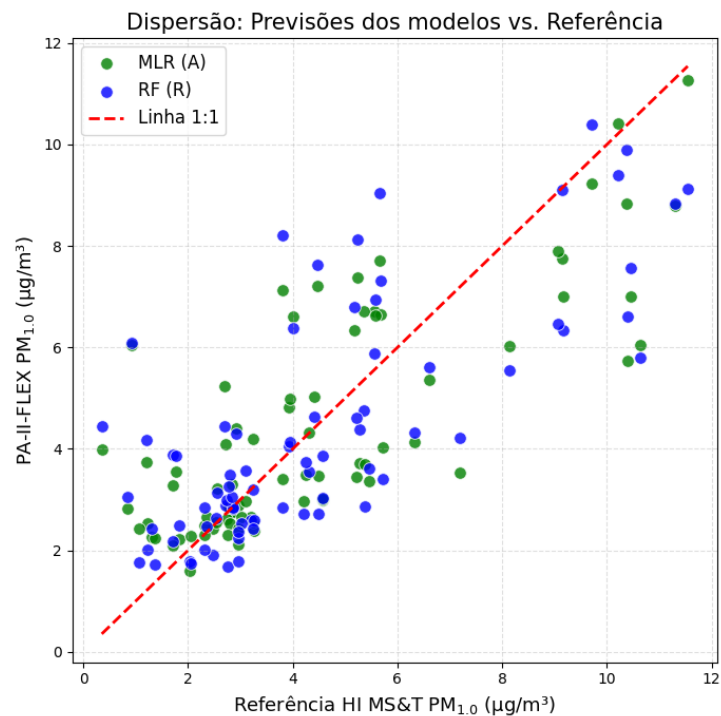


Figura 50 – Gráfico de dispersão das previsões dos modelos MLR-A (azul) e RF-R (verde) em relação à referência. A linha tracejada representa a correspondência 1:1.

Por fim, a Tabela 10 sintetiza os principais resultados estatísticos dos modelos avaliados pelos dois métodos de validação. Embora o método *holdout* apresente métricas mais otimistas, a escolha pelo **modelo MLR-A** (Equação 4.4) ajustado via validação cruzada é justificada por sua maior robustez, estabilidade das métricas ($R^2 = 0,58 \pm 0,09$) e menor risco de sobreajuste. Além disso, o modelo linear simples favorece a transparência interpretativa e a aplicabilidade prática no monitoramento, enquanto métodos mais complexos como o Random Forest tendem a exigir maior volume de dados para atingir seu potencial máximo, podendo sofrer perda de generalização em amostragens reduzidas.

Tabela 10 – Resumo dos modelos finais avaliados para MLR e RF, considerando validação cruzada (*k-Fold*, $k = 5$) e *holdout* (treino/teste 80/20).

Validação	Modelo	R^2	RMSE	MAE
<i>k-Fold</i>	MLR-A	$0,58 \pm 0,09$	$1,74 \pm 0,06$	$1,31 \pm 0,10$
<i>k-Fold</i>	RF-R	$0,54 \pm 0,11$	$1,81 \pm 0,07$	$1,37 \pm 0,10$
holdout	MLR-N	0,74	1,67	1,38
holdout	RF-Q	0,78	1,54	1,07

Em síntese, o processo comparativo demonstra que, mesmo com um conjunto de dados limitado, a estratégia de validação cruzada aliada à seleção de modelos parcimoniosos proporciona resultados mais realistas e confiáveis para a calibração do PurpleAir PA-II-FLEX, equipado com o sensor PMS6003, para $PM_{1.0}$.

5 Aplicação Prática dos Modelos

A calibração e aplicação de modelos de correção para sensores ópticos de baixo custo, como o PurpleAir PA-II-FLEX, são fundamentais para garantir a confiabilidade das medições de material particulado ultrafino ($PM_{1.0}$), especialmente em estudos ambientais e epidemiológicos. Após o processo de treinamento e validação dos modelos de Regressão Linear (MLR) e de *Random Forest* (RF), torna-se essencial orientar sua aplicação prática, considerando as especificidades do local de amostragem, das características do sensor e das condições ambientais.

Para situações em que o objetivo é monitorar o $PM_{1.0}$ no mesmo local onde foi realizada a calibração do sensor, ou em regiões muito próximas e com características ambientais similares, recomenda-se a utilização do modelo de regressão linear. A abordagem linear tem a vantagem de ser mais transparente, apresentar baixo risco de sobreajuste e, sobretudo, garantir robustez frente a pequenas variações do ambiente.

Além disso, para amostragem em campo, o modelo linear apresenta facilidade de aplicação, podendo ser implementado em planilhas ou dispositivos embarcados de forma direta, Pela Equação 4.4, sem exigir alta capacidade computacional.

No entanto, ao considerar a aplicação dos modelos em outras localidades, com características ambientais, fontes de poluição e até mesmo tipos de aerossóis potencialmente diferentes, a recomendação é empregar o modelo baseado em *Random Forest* (Modelo R). Esse algoritmo, por ser não-linear e mais flexível, tende a capturar interações complexas entre as variáveis de entrada, acomodando mudanças que não seriam representadas por um modelo linear simples. Assim, a utilização do *Random Forest* é mais indicada para cenários onde há expectativa de maior variabilidade ambiental ou para extrapolação para diferentes regiões.

Apesar das vantagens descritas, é importante destacar as limitações inerentes à aplicação de modelos de calibração fora do contexto original de treinamento. A performance dos modelos pode ser comprometida devido a fatores como diferenças climáticas, predominância de outras fontes de partículas ou até variações entre lotes de sensores, já que sensores ópticos de baixo custo podem apresentar diferenças de fabricação que afetam a resposta ótica.

Portanto, recomenda-se que a aplicação dos modelos seja sempre acompanhada de avaliações periódicas de desempenho, e, quando possível, seja feita uma nova calibração local. Dessa forma, a escolha entre os modelos deve considerar não apenas o desempenho estatístico, mas, principalmente, o contexto de aplicação, buscando sempre garantir a melhor representatividade possível das medições.

Para facilitar a aplicação prática do modelo de *Random Forest* e da Regressão Linear, foi desenvolvido um *notebook* interativo no Google Colab, disponível em: <<https://colab.research.google.com/drive/1GR36ONRbii2lv7fHFPJibAlsmyLtCYXa?usp=sharing>>. Nesse ambiente, o usuário pode carregar seus próprios dados do sensor PurpleAir e executar todas as etapas necessárias para aplicar o modelo de correção de forma automatizada. Ressalta-se que é necessário substituir o arquivo de entrada pelos dados obtidos no local de interesse, conforme orientações detalhadas no próprio *notebook*. Dessa forma, o procedimento torna-se acessível para os usuários, garantindo a flexibilidade de aplicação do modelo.

6 Conclusão

O presente estudo avaliou a calibração do PurpleAir PA-II-FLEX, equipado com o sensor óptico de baixo custo PMS6003, para a quantificação de material particulado ultrafino ($PM_{1.0}$) em ambiente urbano. A calibração foi realizada em comparação ao método gravimétrico de referência, utilizando o Impactador Harvard MS&T operando com corte aerodinâmico de $1,0 \mu m$ e fluxo de 23 l/min, conforme os protocolos internacionalmente adotados. A campanha amostral, conduzida ao longo de diferentes condições meteorológicas, permitiu a análise detalhada do desempenho do sensor frente às variações ambientais.

Foram testados diferentes métodos de correção e calibração dos dados do sensor, incluindo regressões lineares (MLR), modelos do tipo *Random Forest* (RF) e um método alternativo baseado na conversão de contagens diferenciais de partículas para massa (ALT). A análise estatística dos modelos, utilizando tanto a abordagem de separação holdout (80% treino e 20% teste) quanto validação cruzada *k-Fold* ($k=5$), demonstrou que os modelos calibrados apresentaram melhorias expressivas em relação às medições brutas do sensor, reduzindo o erro e o viés em relação ao método de referência.

O modelo de regressão linear simples (MLR-A), que utiliza apenas a concentração média do sensor como variável preditora, destacou-se pela robustez, simplicidade e boa capacidade de explicação, apresentando os melhores resultados médios de R^2 , RMSE e MAE na validação cruzada. Já o modelo Random Forest (RF-R), que incorpora múltiplas variáveis meteorológicas, apresentou desempenho similar, embora com maior variabilidade para concentrações mais elevadas ($R^2=0.54 \pm 0.11$, RMSE= $1.81 \pm 0.07 \mu g/m^3$ e MAE= $1.37 \pm 0.10 \mu g/m^3$). A avaliação dos ajustes indicou que, embora o método *holdout* resulte em métricas de R^2 acima de 0,77 (RF-Q), a validação cruzada oferece estimativas mais realistas da capacidade preditiva, sendo menos suscetível ao sobreajuste, especialmente diante do número limitado de amostras disponíveis ($n = 79$). A escolha do modelo MLR-A para correção local do sensor mostrou-se a mais adequada frente ao conjunto de dados utilizado, conciliando desempenho estatístico na validação cruzada e facilidade de aplicação ($R^2=0.58 \pm 0.09$, RMSE= $1.74 \pm 0.06 \mu g/m^3$ e MAE= $1.31 \pm 0.10 \mu g/m^3$).

A análise segmentada por faixas de concentração de referência demonstrou que ambos os modelos calibrados acompanham adequadamente os valores gravimétricos em faixas baixas e intermediárias, mas tendem a subestimar concentrações superiores a $6 \mu g/m^3$, efeito atribuído principalmente à menor frequência de dados nessas faixas e à variabilidade intrínseca das partículas ultrafinas. Além disso, a investigação da influência das variáveis meteorológicas evidenciou que parâmetros como umidade relativa e temperatura exercem papel relevante sobre as leituras do sensor óptico, justificando a inclusão dessas variáveis em

modelos de calibração mais avançados, principalmente em ambientes de alta variabilidade climática.

Entre as limitações do estudo, destaca-se o tamanho amostral relativamente reduzido, que limita a extrapolação dos resultados para outros contextos ambientais ou períodos mais longos. Ressalta-se também a necessidade de testes adicionais em diferentes ambientes, escalas temporais e perfis de aerossóis, a fim de consolidar a aplicabilidade dos sensores ópticos de baixo custo em programas de monitoramento ambiental de $PM_{1.0}$.

De forma geral, os resultados obtidos reforçam a importância da adoção de protocolos de calibração e desenvolvimento de modelos que incorporem a influência das condições ambientais. Por fim, recomenda-se que futuros estudos em diferentes regiões e condições ambientais, de modo a aumentar a robustez e a aplicabilidade dos resultados. Também é fundamental a utilização de FEMs como referência, o que possibilita uma avaliação mais completa e confiável do desempenho dos sensores para monitoramento do material particulado ultrafino.

Referências

- Air Diagnostics and Engineering, Inc. *MS&T Area Sampler (Harvard Impactor)*. Disponível em: <https://airdiagnostics.com/indoor_samp_equip.php>. Acesso em: jun. 2025.
- ALEXANDER, D. L.; TROPSHA, A.; WINKLER, D. A. Beware of r²: simple, unambiguous assessment of the prediction accuracy of qsar and qspr models. *Journal of chemical information and modeling*, ACS Publications, v. 55, n. 7, p. 1316–1322, 2015.
- ALVES, C. Aerossis atmosféricos: perspectiva histórica, fontes, processos químicos de formação e composição orgânica. *Química nova*, SciELO Brasil, v. 28, p. 859–870, 2005.
- ANTONIO, A. D.; POPOOLA, O. A.; OUYANG, B.; SAFFELL, J.; JONES, R. L. Developing a relative humidity correction for low-cost sensors measuring ambient particulate matter. *Sensors*, MDPI, v. 18, n. 9, p. 2790, 2018.
- AQMD, S. C. *Field Evaluation of the PurpleAir PA-II-FLEX Sensor*. Diamond Bar, CA, USA, 2022. Disponível em: <<http://www.aqmd.gov/aq-spec/evaluations/summary-pm>>.
- BADURA, M.; BATOG, P.; DRZENIECKA-OSIADACZ, A.; MODZEL, P. Evaluation of low-cost sensors for ambient pm_{2.5} monitoring. *Journal of Sensors*, Wiley Online Library, v. 2018, n. 1, p. 5096540, 2018.
- BREIMAN, L. Random forests. *Machine Learning*, Springer, v. 45, n. 1, p. 5–32, 2001.
- BREUSCH, T.; PAGAN, A. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, Wiley, v. 47, n. 5, p. 1287–1294, 1979.
- BRITO, A. O impacto do material particulado na qualidade do ar. *Revista Virtual de Química*, v. 10, n. 5, p. 1335–1354, 2018.
- BROOK, R. D.; RAJAGOPALAN, S.; III, C. A. P.; BROOK, J. R.; BHATNAGAR, A.; DIEZ-ROUX, A. V.; HOLGUIN, F.; HONG, Y.; LUEPKER, R. V.; MITTLEMAN, M. A. et al. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. *Circulation*, Am Heart Assoc, v. 121, n. 21, p. 2331–2378, 2010.
- BULOT, F. M.; JOHNSTON, S. J.; BASFORD, P. J.; EASTON, N. H.; APETROAIE-CRISTEA, M.; FOSTER, G. L.; MORRIS, A. K.; COX, S. J.; LOXHAM, M. Long-term field comparison of multiple low-cost particulate matter sensors in an outdoor urban environment. *Scientific reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 7497, 2019.
- CASTELL, N.; DAUGE, F. R.; SCHNEIDER, P.; VOGT, M.; LERNER, U.; FISHBAIN, B.; BRODAY, D.; BARTONOVA, A. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment international*, Elsevier, v. 99, p. 293–302, 2017.

CASTELL, N.; DAUGE, F. R.; SCHNEIDER, P.; VOGT, M.; LERNER, U.; FISHBAIN, B.; BRODAY, D.; BARTONOVA, A. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment international*, Elsevier, v. 99, p. 293–302, 2017.

CETESB. *Relatório de Qualidade do Ar no Estado de São Paulo 2022*. São Paulo, 2023. Disponível em: <<https://cetesb.sp.gov.br/ar/>>.

CONSIDINE, E.; FORDE, E.; REECE, S.; KLEANTHOUS, S.; BEEVERS, S.; GANI, S.; STETTLER, M. E. J.; FONT, A.; GRANGE, S. K.; POPOOLA, O.; JONES, R.; DUNMORE, R. E.; HAMILTON, J. F.; LEE, J. D.; LEIGH, R. J.; WILLIAMS, M. L.; BARRATT, B. Improving accuracy of air pollution exposure measurements—statistical correction of a municipal low-cost airborne particulate matter sensor network. *Atmospheric Environment*, Elsevier, v. 262, p. 118599, 2021.

CONSIDINE, E.; FORDE, E.; REECE, S.; KLEANTHOUS, S.; BEEVERS, S.; GANI, S.; STETTLER, M. E. J.; FONT, A.; GRANGE, S. K.; POPOOLA, O.; JONES, R.; DUNMORE, R. E.; HAMILTON, J. F.; LEE, J. D.; LEIGH, R. J.; WILLIAMS, M. L.; BARRATT, B. Application of gaussian mixture regression for the correction of low cost pm2.5 monitoring data in accra, ghana. *Atmosphere*, MDPI, v. 15, n. 4, p. 415, 2024.

CRILLEY, L. R. e. a. Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. *Atmospheric Measurement Techniques*, v. 13, p. 1181–1193, 2020.

(EEA), E. E. A. *Air quality in Europe — 2019 report*. Copenhagen, Denmark: [s.n.], 2019. Available at: <<https://www.eea.europa.eu/publications/air-quality-in-europe-2019>>.

EPA. *Report on the 2013 U.S. Environmental Protection Agency (EPA) International Decontamination Research and Development Conference*. Washington, DC, USA, 2013. Presented at Research Triangle Park, NC, November 05-07, 2013.

EPA. *Particulate Matter (PM) Basics*. [S.l.], 2024. Acesso em: 27 nov. 2024. Disponível em: <<https://www.epa.gov/pm-pollution/particulate-matter-pm-basics>>.

FAROOQUI, Z.; BISWAS, J.; SAHA, J. Long-term assessment of purpleair low-cost sensor for pm2. 5 in california, usa. *Pollutants*, MDPI, v. 3, n. 4, p. 477–493, 2023.

FERNANDES, J. S.; CARVALHO, A. M. d.; CAMPOS, J. F.; FILHO, L. C. Poluição atmosférica e efeitos respiratórios, cardiovasculares e reprodutivos na saúde humana. *Rev Med Minas Gerais*, v. 20, n. 1, p. 92–101, 2010.

GAUTAM, P.; RAMIREZ, A.; BAIR, S.; ARNOTT, W. P.; CHOW, J. C.; WATSON, J. G.; MOOSMÜLLER, H.; WANG, X. Sizing accuracy of low-cost optical particle sensors under controlled laboratory conditions. *Atmosphere*, MDPI, v. 14, n. 5, p. 502, 2023. Disponível em: <<https://www.mdpi.com/2073-4433/16/5/502>>.

GIORDANO, M. R.; MALINGS, C.; PANDIS, S. N.; PRESTO, A. A.; MCNEILL, V.; WESTERVELT, D. M.; BEEKMANN, M.; SUBRAMANIAN, R. From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. *Journal of Aerosol Science*, Elsevier, v. 158, p. 105833, 2021.

- GUARIEIRO, L. L.; VASCONCELLOS, P. C.; SOLCI, M. C. Poluentes atmosféricos provenientes da queima de combustíveis fósseis e biocombustíveis: uma breve revisão. 2011.
- HAGAN, K. Assessing the accuracy of low-cost optical particle sensors using a physics-based approach. *Atmospheric measurement techniques*, Copernicus GmbH, v. 13, n. 11, p. 6343–6355, 2020.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. ed. New York: Springer, 2009. ISBN 978-0-387-84857-0.
- HAYWARD, I.; MARTIN, N. A.; FERRACCI, V.; KAZEMIMANESH, M.; KUMAR, P. Low-cost air quality sensors: Biases, corrections and challenges in their comparability. *Atmosphere*, MDPI, v. 15, n. 12, p. 1523, 2024.
- HEI. *State of Global Air 2024: A Special Report on Global Exposure to Air Pollution and Its Health Impacts*. Boston, MA: [s.n.], 2024. Available at <<https://www.stateofglobalair.org>>.
- HINDS, W. C. *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles*. 2nd. ed. New York: Wiley-Interscience, 1999.
- HOLDER, A. L.; MEBUST, A. K.; MAGHRAN, L. A.; MCGOWN, M. R.; STEWART, K. E.; VALLANO, D. M.; ELLEMAN, R. A.; BAKER, K. R. Field evaluation of low-cost particulate matter sensors for measuring wildfire smoke. *Sensors*, MDPI, v. 20, n. 17, p. 4796, 2020.
- IHME. *Global Burden of Disease Study 2019 (GBD 2019) Results*. Seattle, WA, USA: [s.n.], 2020. Available at: <<http://ghdx.healthdata.org/gbd-results-tool>>.
- JACOBSON, L. d. S. V.; HACON, S. de S.; CASTRO, H. A. de; IGNOTTI, E.; ARTAXO, P.; LEON, A. C. M. P. de. Association between fine particulate matter and the peak expiratory flow of schoolchildren in the brazilian subequatorial amazon: a panel study. *Environmental research*, Elsevier, v. 117, p. 27–35, 2012.
- JAYARATNE, R. e. a. The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog. *Atmospheric Measurement Techniques*, v. 11, p. 4883–4890, 2018.
- JOHARI. Health effects of ultrafine particles (pm1. 0): A review. *Mater. Sci*, v. 3, p. 1–10, 2017.
- JOHNSON, E. R.; INAMDAR, S.; KAN, J.; VARGAS, R. Particulate organic matter composition in stream runoff following large storms: role of pom sources, particle size, and event characteristics. *Journal of Geophysical Research: Biogeosciences*, Wiley Online Library, v. 123, n. 2, p. 660–675, 2018.
- KIM, P. S.; JACOB, D. J.; MICKLEY, L. J.; KOPLITZ, S. N.; MARLIER, M. E.; DEFRIES, R. S.; MYERS, S. S.; CHEW, B. N.; MAO, Y. H. Sensitivity of population smoke exposure to fire locations in equatorial asia. *Atmospheric Environment*, Elsevier, v. 102, p. 11–17, 2015.

KIM, P. S.; JACOB, D. J.; MICKLEY, L. J.; KOPLITZ, S. N.; MARLIER, M. E.; DEFRIES, R. S.; MYERS, S. S.; CHEW, B. N.; MAO, Y. H. Sensitivity of population smoke exposure to fire locations in equatorial asia. *Atmospheric Environment*, Elsevier, v. 102, p. 11–17, 2015.

KULSHRESTHA, U. C. Pm1 is more important than pm2.5 for human health protection. *Current World Environment*, v. 13, n. 1, p. 01–02, 2018.

LIANG, L. Calibrating low-cost sensors for ambient air monitoring: Techniques, trends, and challenges. *Environmental Research*, Elsevier, v. 197, p. 111163, 2021.

LIANG, S.; ZHAO, T.; XU, Q.; DUAN, J.; SUN, Z. Evaluation of fine particulate matter on vascular endothelial function in vivo and in vitro. *Ecotoxicology and environmental safety*, Elsevier, v. 222, p. 112485, 2021.

LIU, C.; CHEN, R.; SERA, F.; VICEDO-CABRERA, A. M.; GUO, Y.; TONG, S.; COELHO, M. S.; SALDIVA, P. H.; LAVIGNE, E.; MATUS, P. et al. Ambient particulate air pollution and daily mortality in 652 cities. *New England Journal of Medicine*, Mass Medical Soc, v. 381, n. 8, p. 705–715, 2019.

MALINGS, C.; TANZER, R.; HAURYLIUK, A.; KUMAR, S.; SUBRAMANIAN, R.; PRESTO, A. A.; ROBINSON, A. L. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmospheric Measurement Techniques*, v. 13, n. 9, p. 4693–4709, 2020.

MALINGS, C.; TANZER, R.; HAURYLIUK, A.; SAHA, P. K.; ROBINSON, A. L.; PRESTO, A. A.; SUBRAMANIAN, R. Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. *Aerosol Science and Technology*, Taylor & Francis, v. 54, n. 2, p. 160–174, 2020.

MARKOWICZ, K. M.; CHILINSKI, M. T. Evaluation of two low-cost optical particle counters for the measurement of ambient aerosol scattering coefficient and ångström exponent. *Sensors*, MDPI, v. 20, n. 9, p. 2617, 2020. Disponível em: <<https://www.mdpi.com/1424-8220/20/9/2617>>.

Ministério do Meio Ambiente. *Plano Nacional de Qualidade do Ar (PNQA)*. Brasília, 2014. Disponível em: <https://www.gov.br/mma/pt-br/assuntos/qualidade-do-ar>. Disponível em: <<https://www.gov.br/mma/pt-br/assuntos/qualidade-do-ar>>.

MISKELL, G.; SALMOND, J. A.; WILLIAMS, D. E. Solution to the problem of calibration of low-cost air quality measurement sensors in networks. *ACS sensors*, ACS Publications, v. 3, n. 4, p. 832–843, 2018.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to Linear Regression Analysis*. 5th. ed. Hoboken, NJ: John Wiley & Sons, 2012.

MOREIRA, D. Meteorologia e poluição atmosférica. *Ambiente & Sociedade*, SciELO Brasil, v. 11, p. 1–13, 2008.

OBERDÖRSTER, G.; OBERDÖRSTER, E.; OBERDÖRSTER, J. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. *Environmental health perspectives*, National Institute of Environmental Health Sciences, v. 113, n. 7, p. 823–839, 2005.

OKAFOR, V. I. Accounting estimates and profitability of listed agricultural companies in nigeria. *Nigerian Journal of Management Sciences Vol*, v. 24, n. 1b, 2023.

PATERAKI, S.; ASIMAKOPOULOS, D.; MAGGOS, T.; ASSIMAKOPOULOS, V.; BOUGIATIOTI, A.; BAIRACHTARI, K.; VASILAKOS, C.; MIHALOPOULOS, N. Chemical characterization, sources and potential health risk of pm2.5 and pm1 pollution across the greater athens area. *Chemosphere*, Elsevier, v. 241, p. 125026, 2020.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

POPE, C. A.; EZZATI, M.; DOCKERY, D. W. Tradeoffs between income, air pollution and life expectancy: Brief report on the us experience, 1980–2000. *Environmental Research*, Elsevier, v. 142, p. 591–593, 2015.

PSOTKA, J.; TRACEY, E.; SICA, R. Retrieval of bulk hygroscopicity from purpleair pm2. 5 sensor measurements. *EGUsphere*, Copernicus Publications Göttingen, Germany, v. 2024, p. 1–14, 2024.

SAYAHI, T.; KAUFMAN, D.; BECNEL, T.; KAUR, K.; BUTTERFIELD, A.; COLLINGWOOD, S.; ZHANG, Y.; GAILLARDON, P.-E.; KELLY, K. Development of a calibration chamber to evaluate the performance of low-cost particulate matter sensors. *Environmental Pollution*, Elsevier, v. 255, p. 113131, 2019.

SCHRAUFNAGEL, D. E.; BALMES, J. R.; COWL, C. T.; MATTEIS, S. D.; JUNG, S.-H.; MORTIMER, K.; PEREZ-PADILLA, R.; RICE, M. B.; RIOJAS-RODRIGUEZ, H.; SOOD, A. et al. Air pollution and noncommunicable diseases: A review by the forum of international respiratory societies' environmental committee, part 2: Air pollution and organ systems. *Chest*, Elsevier, v. 155, n. 2, p. 417–426, 2019.

SEINFELD, J. H.; PANDIS, S. N. *Atmospheric chemistry and physics: from air pollution to climate change*. [S.l.]: John Wiley & Sons, 2016.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, Oxford University Press, v. 52, n. 3/4, p. 591–611, 1965.

SPINELLE, L.; GERBOLES, M.; VILLANI, M. G.; ALEIXANDRE, M.; BONAVIDACOLA, F. Calibration of a cluster of low-cost sensors for the measurement of air pollution in ambient air. In: IEEE. *SENSORS, 2014 IEEE*. [S.l.], 2014. p. 21–24.

STOCK, J. H.; WATSON, M. W. Indicators for dating business cycles: Cross-history selection and comparisons. *American Economic Review*, American Economic Association, v. 100, n. 2, p. 16–19, 2010.

TRYNER, J.; L'ORANGE, C.; MEHAFFY, J.; MILLER-LIONBERG, D.; HOFSTETTER, J. C.; WILSON, A.; VOLCKENS, J. Laboratory evaluation of low-cost purpleair pm monitors and in-field correction using co-located portable filter samplers. *Atmospheric Environment*, Elsevier, v. 220, p. 117067, 2020.

- TRYNER, J.; L'ORANGE, C.; MEHAFFY, J.; MILLER-LIONBERG, D.; HOFSTETTER, J. C.; WILSON, A.; VOLCKENS, J. Laboratory evaluation of low-cost purpleair pm monitors and in-field correction using co-located portable filter samplers. *Atmospheric Environment*, Elsevier, v. 220, p. 117067, 2020.
- TURNER, W.; OLSON, B.; ALLEN, G. Calibration of sharp cut impactors for indoor and outdoor particle sampling. *Journal of the Air & Waste Management Association*, v. 50, n. 4, p. 484–487, 2000. Disponível em: <https://www.researchgate.net/publication/12531354_Calibration_of_Sharp_Cut_Impactors_for_Indoor_and_Outdoor_Particle_Sampling>.
- VANLOON, G.; DUFFY, S. Environmental chemistry, first printed. *New York*, 2000.
- WALLACE, L. Intercomparison of purpleair sensor performance over three years indoors and outdoors at a home: bias, precision, and limit of detection using an improved algorithm for calculating pm_{2.5}. *Sensors*, MDPI, v. 22, n. 7, p. 2755, 2022.
- WANG, W.-C. V.; LUNG, S.-C. C.; LIU, C.-H. Application of machine learning for the in-field correction of a pm_{2.5} low-cost sensor network. *Sensors*, MDPI, v. 20, n. 17, p. 5002, 2020.
- WANG, Y.; LI, J.; JING, H.; ZHANG, Q.; JIANG, J.; BISWAS, P. Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement. *Aerosol science and technology*, Taylor & Francis, v. 49, n. 11, p. 1063–1077, 2015.
- WHO. *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*. [S.l.]: World Health Organization, 2006.
- WHO. *2018 Health SDG Profile: Democratic People's*. [S.l.]: World Health Organization. Regional Office for South-East Asia, 2018.
- WHO. *WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Geneva, Switzerland: World Health Organization, 2021. Available at: <<https://www.who.int/publications/i/item/9789240034228>>.
- XU, J.; CHEN, Y.; LU, F.; CHEN, L.; DONG, Z. The association between short-term exposure to pm₁ and daily hospital admission and related expenditures in beijing. *Toxics*, MDPI, v. 12, n. 6, p. 393, 2024.
- YAN, M.; XU, J.; LI, C.; GUO, P.; YANG, X.; TANG, N.-j. Associations between ambient air pollutants and blood pressure among children and adolescents: a systemic review and meta-analysis. *Science of The Total Environment*, Elsevier, v. 785, p. 147279, 2021.
- YANG, M.; LIU, X.; XU, S.-L.; WANG, X.-F.; ROPONEN, M.; JIN, N.-X.; HUANG, J.-W.; WU, Q.-Z.; CHU, C.; SUN, M.-K.; BAO, W.-W.; LIN, L.-Z.; GUI, Z.-H.; LIU, R.-Q.; HU, L.-W.; XU, B.; DONG, G.-H.; JALAVA, P. I. Concentrations, sources and health risks of per- and polyfluoroalkyl substances in ambient pm₁ in the pearl river delta region, china. *Environment International*, Elsevier, v. 198, p. 109439, 2025.
- YEE, J.; CHO, Y. A.; YOO, H. J.; YUN, H.; GWAK, H. S. Short-term exposure to air pollution and hospital admission for pneumonia: a systematic review and meta-analysis. *Environmental Health*, Springer, v. 20, p. 1–10, 2021.

YIN, P.; BRAUER, M.; COHEN, A.; BURNETT, R. T.; LIU, J.; LIU, Y.; LIANG, R.; WANG, W.; QI, J.; WANG, L. et al. Long-term fine particulate matter exposure and nonaccidental and cause-specific mortality in a large national cohort of chinese men. *Environmental health perspectives*, v. 125, n. 11, p. 117002, 2017.

ZHANG, Y.; CHEN, S.; WEI, J.; JIANG, J.; LIN, X.; WANG, Y.; HAO, C.; WU, W.; YUAN, Z.; SUN, J. et al. Long-term pm₁ exposure and hypertension hospitalization: A causal inference study on a large community-based cohort in south china. *Science Bulletin*, Elsevier, v. 69, n. 9, p. 1313–1322, 2024.