Modelagem de Estações Virtuais de Qualidade do Ar com Redes Neurais Artificiais para Monitoramento de PM10

Jonathan da Silva Saldanha[1]; 0009-0003-1096-7426
Italo Pinto Rodrigues[1]; 0000-0002-6832-8358

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
jonathans@id.uff.br (contato principal)

Resumo: O monitoramento da qualidade do ar enfrenta desafios relacionados ao alto custo e à manutenção de estações físicas. Como alternativa, este estudo investigou a viabilidade de construir uma estação virtual de PM10 utilizando redes neurais artificiais do tipo Multilayer Perceptron (MLP). Foram utilizados dados de duas estações reais para prever concentrações em uma terceira, adotando diferentes arquiteturas de rede. Os resultados mostraram que arquiteturas de complexidade intermediária alcançaram melhor desempenho, com destaque para a configuração [2 – 30 – 30 – 30 – 1], que obteve Média Relativa do Erro Quadrático (MRSE) de aproximadamente 63,98%. Apesar de os resultados ainda não serem competitivos frente a estudos de referência, eles confirmam o potencial de RNAs para aplicações de baixo custo em estações virtuais. Trabalhos futuros deverão explorar estratégias de regularização, variações arquiteturais, uso de múltiplas estações e inclusão de variáveis ambientais adicionais, de modo a reduzir erros e aumentar a competitividade das predições.

Palavras-chave: Estações virtuais. PM10. Redes neurais artificiais. Multilayer Perceptron. Qualidade do ar..

INTRODUÇÃO

O monitoramento da qualidade do ar é estratégico para proteção da saúde pública, mas a implantação e a manutenção de estações físicas elevam custos e limitam a cobertura espacial, sobretudo em áreas urbanas dinâmicas e em países com redes pouco densas. Nesse cenário, estações virtuais — modelos preditivos que estimam concentrações de poluentes a partir de dados ambientais, meteorológicos e de estações vizinhas — surgem como alternativa tecnicamente viável e economicamente atraente. Estudos recentes mostram que modelos de aprendizado de máquina conseguem prever concentrações de PM10/PM2.5 (material particulado, que são partículas sólidas e líquidas no ar, onde o número indica o diâmetro das partículas em micrômetros (μm)) com com desempenho competitivo, e que informações de estações próximas figuram entre os preditores mais eficazes para estimar os níveis em um ponto sem medições diretas, abrindo caminho para substituir ou complementar estações físicas por “estações virtuais” em diferentes cidades europeias (Samad *et al.*, 2023).

A literatura internacional também registra ganhos substanciais quando técnicas de ensemble empilham modelos heterogêneos, como Random Forest (RF), Support Vector Machine (SVM) e Extreme Gradient Boosting (XGBoost), capturando relações não lineares e variações espaço-temporais nas séries de PM. Em estudo nacional de alta resolução (1 km) na Itália, um *deep ensemble* de três níveis atingiu $R^2 \approx 0{,}87$ e RMSE≈5,38 μg/m³, superando cada algoritmo isolado e apresentando boa capacidade de generalização espacial e temporal (Yu *et al.*, 2022).

Em regiões do Sudeste Asiático, trabalhos de escala nacional demonstram o Random Forest (RF) e o Support Vector Regression (SVR estimam PM2.5 com acurácia moderada-alta ($R^2 \approx 0{,}64$–$0{,}76$ na validação por estação), sobretudo em ambientes urbano-industriais com maior carga de aerossóis e gases traço, reforçando o potencial de estações virtuais para cobrir vazios de monitoramento (Zaman *et al.*, 2021).

Do ponto de vista de arquitetura, há evidências de que modelos simples e bem projetados, como a *Multilayer Perceptron (MLP)*, podem igualar ou superar arquiteturas mais profundas, favorecendo implementação enxuta e robusta em cenários com dados e recursos computacionais limitados — situação comum em aplicações de engenharia ambiental (Hasanpour *et al.*, 2016). Ao mesmo tempo, a literatura alerta para o viés de simplicidade (*simplicity bias*): durante o treinamento, redes neurais tendem a apoiar-se em padrões mais simples e negligenciar sinais preditivos mais complexos, o que pode reduzir robustez sob mudanças de distribuição. Esse viés justifica uma validação rigorosa e uma seleção criteriosa de entradas e metas de regularização quando se busca extrapolar para pontos sem sensores (Shah *et al.*, 2020).

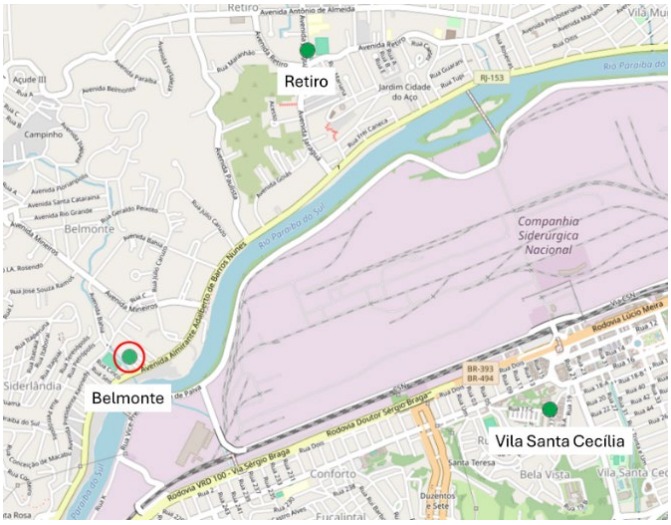
Neste trabalho, propõe-se o uso de uma Rede Neural Artificial MLP para treinar um modelo capaz de estimar indiretamente PM10 em Volta Redonda, caracterizando uma estação virtual a partir de duas estações físicas e prevendo o que ocorreria em uma terceira. O problema de pesquisa é motivado pelo alto custo de estações e pela necessidade de ampliar cobertura de monitoramento com software que triangule informações. Portanto, o objetivo principal é demonstrar a viabilidade de uma estação virtual de PM10 baseada em rede neural artificial do tipo *Multilayer Perceptron (MLP)* alimentada por duas estações reais, quantificando ganhos e limites dessa abordagem e discutindo diretrizes de projeto (entradas,

regularização e topologia) à luz das melhores práticas reportadas para estações virtuais e ensembles em qualidade do ar (Samad *et al.*, 2023; Yu *et al.*, 2022; Zaman *et al.*, 2021).

MÉTODOS

Neste estudo, foram utilizadas três estações de monitoramento da qualidade do ar localizadas na cidade de Volta Redonda (RJ): Belmonte, Retiro e Vila Santa Cecília (Figura 1). A escolha dessas estações deve-se à proximidade geográfica e à cobertura urbana homogênea, fatores que reduzem a influência de variáveis topográficas ou climáticas diferenciadas.

Figura 1 – Posição das estações.

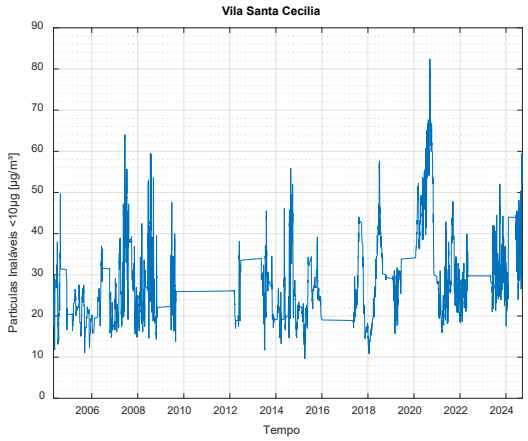


Fonte: Adaptada de INEA (2025).

O objetivo principal foi estimar as concentrações de PM10 na estação Vila Santa Cecília utilizando como preditores os valores históricos registrados simultaneamente nas estações Belmonte e Retiro. Essa abordagem caracteriza uma estação virtual, substituindo medições físicas por inferências derivadas de padrões de dados ambientais.

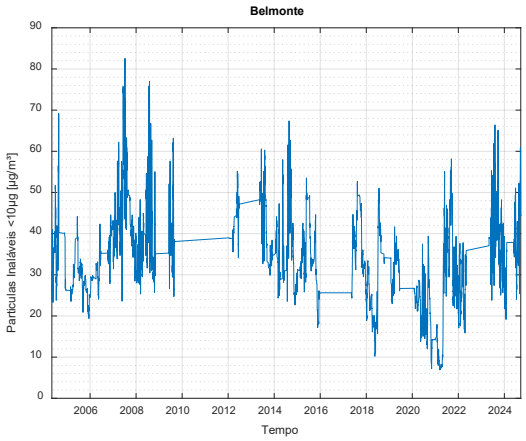
Foram utilizados dados diários do período entre 2004 e 2024, extraídos do banco de dados do Instituto Estadual do Ambiente (INEA/RJ). Apenas registros completos e sincronizados nas três estações foram considerados, resultando em 36.315 amostras válidas. As séries temporais de cada estação foram analisadas em termos de tendência geral, sendo ilustradas pelas médias móveis a cada 100 amostras (Figuras 2, 3 e 4).

Figura 2 – Medições de PM10 na Vila Santa Cecília.



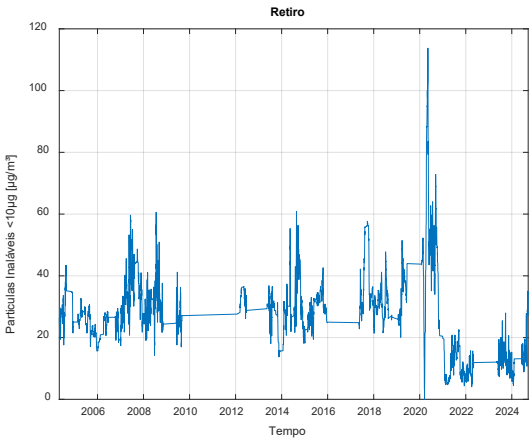
Fonte: Elaborada pelos autores (2025).

Figura 3 – Medições de PM10 no Belmonte.



Fonte: Elaborada pelos autores (2025).

Figura 4 – Medições de PM10 no Retiro.



Fonte: Elaborada pelos autores (2025).

A rede neural artificial utilizada foi do tipo *Multilayer Perceptron* (MLP), estruturada em cinco camadas: Camada de entrada: 2 neurônios, correspondentes às estações Belmonte e Retiro (valores normalizados); Três camadas ocultas: variando o número de neurônios empiricamente, com função de ativação tangente hiperbólica; Camada de saída: 1 neurônio, representando a concentração prevista de PM10 na estação Vila Santa Cecília.

O algoritmo de treinamento empregado foi o Levenberg-Marquardt (LM), conhecido por equilibrar rapidez de convergência e estabilidade. O conjunto de dados de cerca de 12000 amostras foi dividido em três subconjuntos: Treinamento (70%): ajuste dos pesos e vieses

da rede; Validação (15%): monitoramento do desempenho e prevenção de sobreajuste; Teste (15%): estimativa do erro de generalização com dados não vistos.

A métrica adotada para avaliar o desempenho foi o Mean Relative Squared Error (MRSE), que mede a média relativa do erro quadrático entre os valores previstos e observados, conforme a Equação 1. Valores mais próximos de zero indicam maior precisão. O MRSE utilizou cerca de 24210 amostras de PM10 de cada estação, que não foram apresentadas à rede durante o treinamento.

$$MRSE(\%) = \frac{1}{n_o}\sum_{o=1}^{n_o}\left[\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(\hat{y}_i)^2}}\right] \quad (1)$$

Onde y_i representa a saída medida (telemetria de VBAT1); $\hat{y}_i$ a saída simulada; no o número total de saídas, e N a quantidade de dados (amostras) utilizada para medir o desempenho do modelo.

RESULTADOS E DISCUSSÃO

As diferentes arquiteturas da rede neural artificial do tipo Multilayer Perceptron (MLP) foram testadas com o objetivo de identificar a configuração mais adequada para estimar os valores de PM10 na estação Vila Santa Cecília a partir das entradas Belmonte e Retiro. A métrica de avaliação utilizada foi a média relativa do erro quadrático (MRSE) em porcentagem.

Os experimentos realizados com diferentes arquiteturas de redes neurais artificiais do tipo Multilayer Perceptron (MLP) indicaram que a topologia [2 30 30 30 1] apresentou o melhor desempenho, alcançando um erro médio relativo quadrático (Mean Relative Squared Error – MRSE) de aproximadamente 63,98%. Esse resultado foi superior tanto às arquiteturas muito profundas, como a configuração [2 100 100 100 1] ($MRSE \approx 64{,}94\%$), quanto às arquiteturas excessivamente simplificadas, como [2 2 2 2 1] ($MRSE \approx 66{,}72\%$). Esses achados sugerem que a complexidade moderada da rede favorece o equilíbrio entre viés e variância, evitando tanto o subajuste quanto o sobreajuste.

Tabela 1 — Arquiteturas testadas e desempenho (MRSE em %)

Arquitetura (neurônios por camada)	MRSE (%)
[2 20 40 60 1]	69,5609
[2 10 10 10 1]	67,7606

[2 30 30 30 1]	63,9789
[2 50 50 50 1]	65,6214
[2 100 100 100 1]	64,9377
[2 10 20 40 1]	66,5688
[2 2 2 2 1]	66,7208
[2 2 10 2 1]	65,5198
[2 10 2 10 1]	67,7045
[2 2 50 2]	66,7506

Fonte: Elaborada pelos autores (2025).

Esse comportamento está em consonância com evidências da literatura que apontam para a eficácia de modelos simples e bem projetados, capazes de superar arquiteturas mais profundas em determinados contextos, sobretudo em cenários de limitação de dados ou ruído acentuado nas séries temporais (Hasanpour *et al.*, 2016). Por outro lado, o aumento indiscriminado do número de neurônios não se traduziu em melhor desempenho, o que reforça a ideia de que redes neurais podem ser afetadas pelo chamado viés de simplicidade (simplicity bias), ou seja, a tendência de priorizar padrões superficiais em detrimento de relações mais complexas presentes nos dados (Shah *et al.*, 2020).

Quando comparados a outras abordagens reportadas na literatura para a construção de estações virtuais, os resultados obtidos aqui confirmam a viabilidade da RNA tipo MLP para predição de PM10, ainda que o erro percentual obtido seja relativamente elevado. Estudos anteriores baseados em algoritmos como *Random Forest* (RF), *Support Vector Machine* (SVM) e *Support Vector Regression* (SVR) demonstraram desempenhos satisfatórios na predição de PM2.5, atingindo coeficientes de determinação (R^2) entre 0,64 e 0,76 em cenários urbanos-industriais (Zaman *et al.*, 2021). Nessa comparação, o presente trabalho mostra que, mesmo sem recorrer a técnicas de ensemble, uma rede neural de arquitetura moderada consegue capturar padrões relevantes de correlação espacial entre estações vizinhas.

A literatura também mostra que abordagens baseadas em ensembles profundos alcançaram erros significativamente menores, com valores de Root Mean Square Error (RMSE) próximos a 5,38 µg/m³ em estudos de larga escala na Europa (Yu *et al.*, 2022). Isso evidencia que o uso de combinações de modelos heterogêneos pode aumentar substancialmente a capacidade de generalização, sobretudo quando aplicado a grandes bases de dados com

diversidade espacial e temporal. Embora o presente estudo não tenha explorado ensembles, seus resultados indicam um ponto de partida promissor, em que a simplicidade arquitetural e o baixo custo computacional podem ser explorados em aplicações práticas de estações virtuais.

Em termos de implicações, os resultados reforçam que arquiteturas intermediárias de RNA oferecem melhor compromisso entre eficiência computacional e desempenho preditivo. Esse equilíbrio é particularmente relevante para aplicações em tempo real ou em regiões com infraestrutura limitada, onde soluções leves e escaláveis são preferíveis. Além disso, a confirmação da viabilidade de uma estação virtual de PM10 baseada em duas estações reais abre caminho para estudos que explorem dados adicionais, como variáveis meteorológicas, ou que avancem para poluentes mais críticos, como o PM2.5, frequentemente associado a maiores riscos à saúde.

CONCLUSÕES

Este trabalho demonstrou a viabilidade da utilização de redes neurais artificiais MLP para a construção de uma estação virtual de monitoramento de PM10, a partir da combinação de dados provenientes de duas estações reais vizinhas. Os resultados indicaram que arquiteturas de complexidade intermediária apresentaram melhor desempenho, com destaque para a configuração [2 30 30 30 1], que obteve um MRSE de aproximadamente 63,98%.

A análise revelou que modelos demasiadamente simplificados apresentaram limitações na capacidade preditiva, enquanto arquiteturas excessivamente profundas não proporcionaram ganhos significativos. Isso confirma que a eficácia de um modelo não depende apenas do número de camadas ou neurônios, mas da sua adequação ao problema e ao conjunto de dados. Embora os erros obtidos ainda não sejam competitivos em relação a trabalhos de referência, os resultados reforçam que uma rede neural MLP devidamente calibrada pode servir como solução inicial para ampliar a cobertura espacial do monitoramento atmosférico sem a necessidade de instalação de novas estações físicas.

Como trabalhos futuros, recomenda-se explorar diferentes arquiteturas de RNA, incluindo variações assimétricas no número de neurônios ocultos e funções de ativação alternativas,

além de aplicar estratégias de regularização, como *dropout*, penalização e normalização por batch, para reduzir o risco de sobreajuste e aumentar a capacidade de generalização. Também se destaca a importância de investigar diferentes janelas temporais e médias móveis como variáveis de entrada, de modo a capturar tendências de médio e longo prazo, bem como ampliar a triangulação para múltiplas estações, verificando se o acréscimo de uma terceira ou quarta entrada melhora a estimativa da estação virtual. Além disso, a inclusão de outras variáveis ambientais e meteorológicas no modelo deverá ser considerada, pois pode contribuir para reduzir os erros e tornar as previsões mais competitivas.

AGRADECIMENTOS

Os autores agradecem ao Centro Universitário de Volta Redonda (UniFOA) pelo apoio institucional e financeiro, por meio do Programa Institucional de Bolsas de Iniciação Científica (PIBIC) (93978/17/RPE).

REFERÊNCIAS

HASANPOUR, Seyyed Hossein; ROUHANI, Mohammad; FAYYAZ, Mohsen; SABOKROU, Mohammad. Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures. [*s. l.*], 2016. DOI 10.48550/ARXIV.1608.06037. Disponível em: https://arxiv.org/abs/1608.06037. Acesso em: 14 set. 2025.

INEA. AQMIS. 2025. **SIGQAr - Sistema Integrado de Gestão da Qualidade do Ar**. Disponível em: https://ei.weblakes.com/INEAPublico/ViewerMap. Acesso em: 19 jul. 2025.

SAMAD, A.; GARUDA, S.; VOGT, U.; YANG, B. Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations. **Atmospheric Environment**, [*s. l.*], v. 310, p. 119987, out. 2023. https://doi.org/10.1016/j.atmosenv.2023.119987.

SHAH, Harshay; TAMULY, Kaustav; RAGHUNATHAN, Aditi; JAIN, Prateek; NETRAPALLI, Praneeth. The Pitfalls of Simplicity Bias in Neural Networks. [*s. l.*], 2020. DOI 10.48550/ARXIV.2006.07710. Disponível em: https://arxiv.org/abs/2006.07710. Acesso em: 14 set. 2025.

YU, Wenhua; LI, Shanshan; YE, Tingting; XU, Rongbin; SONG, Jiangning; GUO, Yuming. Deep Ensemble Machine Learning Framework for the Estimation of PM2.5 Concentrations. **Environmental Health Perspectives**, [*s. l.*], v. 130, n. 3, mar. 2022. DOI 10.1289/ehp9752. Disponível em: https://ehp.niehs.nih.gov/doi/10.1289/EHP9752. Acesso em: 1 maio 2025.

ZAMAN, Nurul Amalin Fatihah Kamarul; KANNIAH, Kasturi Devi; KASKAOUTIS, Dimitris G.; LATIF, Mohd Talib. Evaluation of Machine Learning Models for Estimating PM2.5 Concentrations across Malaysia. **Applied Sciences**, [*s. l.*], v. 11, n. 16, p. 7326, 9 ago. 2021. https://doi.org/10.3390/app11167326.