Built to Breathe? Modeling Air Quality with Features of the Built and Natural Environment

Lara S. Furtado[1,5], Nayara O. Gurjão[1,2], Nicolas C. Monteiro[3,5], Edilson Filho[3], Carlos Matheus Ferreira[3], Jarbas A. Nunes[3], Jorge B. Soares[1,2], José A. Macedo[4]

[1]Grupo de Inteligência de Dados (GrID), INCT Infra, Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brazil

[2]Programa de Pós-Graduação em Engenharia de Transportes (PETRAN) – UFC

[3]Departamento de Engenharia de Teleinformática (DETI) - UFC

[4]Programa de Pós-Graduação em Ciência da Computação (MDCC) – UFC

[5]Programa de Pós-Graduação em Engenharia de Teleinformática (PPGETI) – UFC

lara.furtado@insightlab.ufc.br, nayara.gurjao@det.ufc.br, nicolas.eng.comp@gmail.com, edilsonfilho@lesc.ufc.br, cmatheuslf@alu.ufc.br,jarbas@lesc.ufc.br, jsoares@det.ufc.br, jose.macedo@insightlab.ufc.br

Abstract. *Although air quality data is often limited by the cost and complexity of sensor networks, open geospatial data provides detailed information on the built environment, which can be used to estimate concentrations of pollutants. Using point-based sensor data and urban features from a pilot city, the research presented herein has trained and validated multiple supervised regression models finding that features such as tree density, building height, street connectivity, and infrastructure coverage can effectively predict spatial variation in Particulate Matter size 2.5µm, even in areas without direct measurements. This scalable and data-driven solution supports environmental monitoring and sustainable planning in cities worldwide with minimal reliance on primary sensor data.*

1. Introduction

Air quality plays a fundamental role in public health and urban sustainability. Among the many pollutants that affect urban populations, fine particulate matter—particularly PM2.5 (*Particulate Matter of size 2.5µm*)—stands out due to its ability to penetrate deep into the respiratory system and its strong association with cardiovascular and respiratory diseases, as well as cancer (Dapper et al., 2016; WHO, 2021; Liang et al., 2023). Despite growing recognition of its health impacts, access to high-resolution, localized air quality data remains a significant challenge, especially in cities of the Global South. The reliance on sparse official monitoring stations often limits the ability to fully capture environmental inequalities and localized pollution patterns.

In response to these limitations, recent advances in urban data availability and spatial analytics have opened new pathways for modeling air quality using secondary data sources. The proliferation of georeferenced information about the built environment—such as land use, vegetation cover, street networks, and building characteristics—offers an alternative means to estimate pollution levels with greater

spatial granularity. These data, when analyzed through data science and machine learning techniques, enable predictive modeling even in the absence of dense sensor networks (Liu et al., 2020; Tella & Balogun, 2022).

This paper presents a methodology to model PM2.5 concentrations using urban and environmental data from Fortaleza, the fourth largest city in Brazil. While the model is trained and validated in this specific context, the framework is scalable and adaptable to other cities. Our goal is twofold: to demonstrate how urban form and infrastructure contribute to local air quality outcomes, and present a data-driven approach in promoting environmentally informed urban planning.

2. Literature Review

Understanding the urban environment presents several challenges, particularly due to the presence of multiple interdependent and often uncontrollable variables. To support interventions and investments in pollution mitigation actions, the development of predictive models and the simulation of future scenarios becomes essential.

Historical data on air pollutant emissions have been the main source for analyzing this phenomenon (Xie et al., 2022; Ji et al., 2022). However, most cities do not have continuous pollution monitoring networks and data collection stations can be expensive to purchase and maintain (Li et al., 2023). This limitation makes it difficult to assess pollution patterns at a local scale, which reinforces the need for predictive models based on urban and environmental data that can be replicated in other regions.

Computational modeling is increasingly used in environmental research to simulate air pollution in areas with limited data (Lou et al., 2023). Techniques such as regression models, random forests, neural networks, and deep learning help identify patterns and predict pollutant concentrations. Many models integrate machine learning with spatial data to increase accuracy and interpretability, offering a scalable solution for assessing urban air quality (Tella & Balogun, 2022). Air quality prediction can be improved by incorporating additional variables, such as meteorological data (Faraji et al., 2022), traffic volume estimates, microclimate variations, and urban thermal comfort data (Yang et al., 2020). Similarly, models have been studied for large-scale pollution forecasting using point data collected using low-cost sensors and georeferenced information (Liu et al., 2020; Gurjão et al., 2024).

Human activities are widely recognized as the main drivers of change in urban environments, especially due to the rapid expansion of built-up areas, the increase in impervious surfaces, and the reduction of green spaces. In this context, urban planning and land use optimization are fundamental for analyzing pollution in cities (Che et al., 2023). The relationship between building height and street width (including sidewalks) defines the so-called urban canyons, which are characterized as confined spaces between buildings with limited ventilation, high traffic intensity, and often elevated pollutant concentrations, suggesting an increase in surface temperature (Vardoulakis et al., 2003; Seaton et al., 2022). Gurjão et al. (2024) analyzed multiple urban factors and found that the height of buildings and the presence of urban canyons stood out as significant variables in predictive models of particulate matter concentration, especially when integrated with meteorological variables. Studies also indicate higher pollutant concentrations in areas near high-traffic roads or in regions with dense road networks

that accommodate heavy vehicles, particularly where adjacent buildings hinder pollutant dispersion (Silva & Mendes, 2006; Che et al., 2023).

Vegetation can also play an important role in pollutant mitigation and thermal regulation in urban environments. Wooded environments help absorb solar radiation, tree canopies provide shade and contribute to reducing surface temperatures through evapotranspiration (Stache et al., 2022). In addition, urban vegetation acts as a natural filter, capturing pollutant particles, releasing oxygen, increasing humidity and reducing air temperature (Zheng et al., 2021). These mechanisms also interact with the built environment, influencing microclimatic conditions and pollutant dispersion dynamics.

The literature reveals that several variables act differently on the dispersion and concentration of pollutants, contributing to different effects on air quality. These findings demonstrate the importance of connecting air quality with built environment indicators, using data science tools such as machine learning, remote sensing and geospatial analysis, as well as direct field measurements as baseline data for prediction (Sakti et al., 2023).

Therefore, this study focuses on the construction of predictive models on the concentration of particulate matter. The emphasis is on urban variables given their availability in public databases in the investigated area, in addition to the possibility of being collected and analyzed with spatial scope in other regions.

3. Methodology

This study was conducted in Fortaleza, Brazil, with a focus on two urban neighborhoods: Meireles and Aldeota. These areas were selected due to their high population density, urban complexity, and the availability of primary air quality data, which enabled testing and training of the proposed air quality model (Gurjão, 2024).

3.1. Data Collection and Preparation

Primary pollution data were collected using a portable sensor capable of detecting concentrations of PM2.5 (see sensor specifications in Furtado et al., 2024). The sensor was installed on tripods approximately 1.8 m above ground level, near major roads and canyon-shaped street sections, where the study area ensured the safety of the team and equipment (Figure 1). Measurements were conducted at 11 georeferenced locations within the study area between August and September 2023, from Tuesday to Thursday, with data recorded every minute over continuous 4-hour sampling windows.

The procedures performed are described in Gurjão (2024) and Gurjão et al. (2024), where details on sensor calibration, measurement protocol, and point selection are reported in full. Briefly, missing values were imputed by linear interpolation, and all series were aggregated into one-minute averages to match the resolution used for model training.

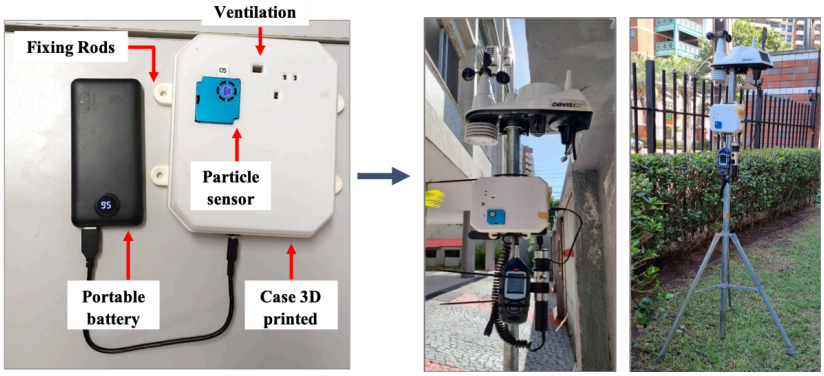

Figure 1. Structure and positioning of sensors in outdoor city locations.

Environmental and infrastructure data about the built environment were obtained from the Fortaleza Municipal Secretary of Finances (SEFIN) via the city's official spatial data infrastructure portal (SEFIN, 2016). The datasets consist of: (i) Point data on individual tree locations; (ii) Polygon data of building footprints with associated attributes; (iii) Line data representing the street network, including infrastructure indicators. In this way, the variables considered in the models were: 2.5 Particulate Matter; Buildings Count; Street Connectivity; Average Building Area; Average Canyon Width; Average Building Height; Total Built Area; Sewage Coverage Rate; Drainage Coverage Rate; Street Lighting Rate; Pavement Coverage Rate; Gutter Coverage Rate; Tree Count; Water Coverage Rate.

3.2. Feature Engineering

To estimate PM2.5 concentrations across the urban area, we developed a supervised regression modeling pipeline based on point-level sensor data enriched with contextual urban features. The original air quality measurements were collected at discrete geographic points, each corresponding to a sensor location. To incorporate surrounding urban conditions, these points were spatially joined to a uniform hexagonal grid with 200-meter resolution. This allowed each observation to inherit built environment characteristics from the hexagon it fell within, such as tree density, street connectivity, building height, built-up area, infrastructure coverage, and the canyon effect (defined as the height-to-width ratio of surrounding structures).

The predictor variables used in the regression models capture key aspects of the built environment. Vegetation, represented by tree density, can act to mitigate pollutants and regulate the microclimate (Zheng et al., 2021; Stache et al., 2022), while building height and the urban canyon effect can influence pollutant accumulation and ventilation restrictions in urban areas (Vardoulakis et al., 2003; Seaton et al., 2022). Traffic-related effects were considered through road connectivity, given the established relationship between dense road networks and higher pollutant concentrations (Silva & Mendes, 2006; Che et al., 2023). Finally, infrastructure coverage was incorporated to reflect broader urbanization processes that shape air quality dynamics (Che et al., 2023; Gurjão et al., 2024).

Figure 2 illustrates the spatial data formats used in this study. Red points mark sensor locations, while the base map includes building footprints (gray), street network

(yellow), neighborhood boundaries (red lines), and the hexagonal aggregation grid (light blue).

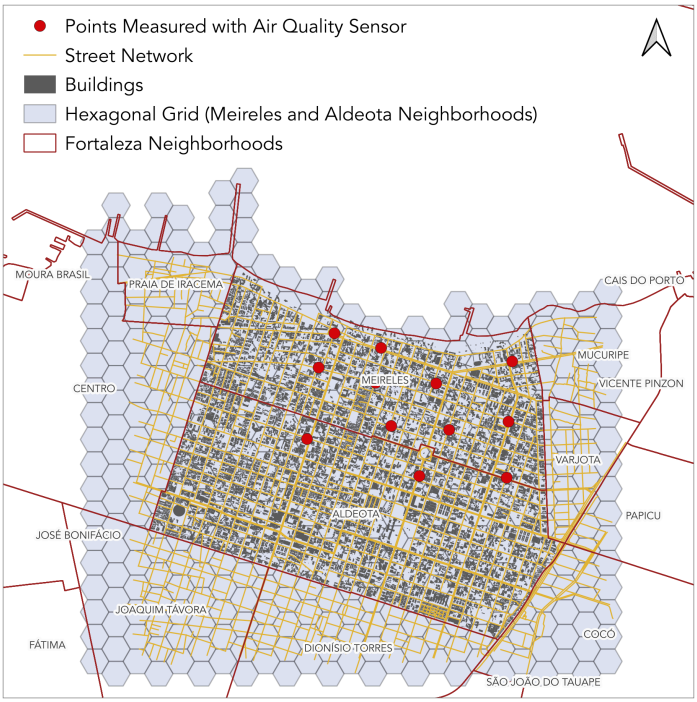

Figure 2. Study area in Fortaleza, Brazil, showing the neighborhoods where PM2.5 concentrations were measured.

For illustration purposes, Figure 3 details some of the additional urban features aggregated in the hexagonal spatial grid:

- Street Connectivity: computed as both the number of unique street intersections (3a) and the total number of street segments within each hexagon (3b).
- Tree Density: calculated as the number of geolocated trees falling within each hexagon (3c).
- Building characteristics: average height of buildings per hexagon (3d), average building area and urban canyon effect; derived per building as the height-to-width ratio to neighboring structures; values are averaged across all buildings in each hexagon.
- Urban Infrastructure: represented by the proportion of buildings served by pavement, water, sewage, drainage, guttering, and public lighting infrastructure.

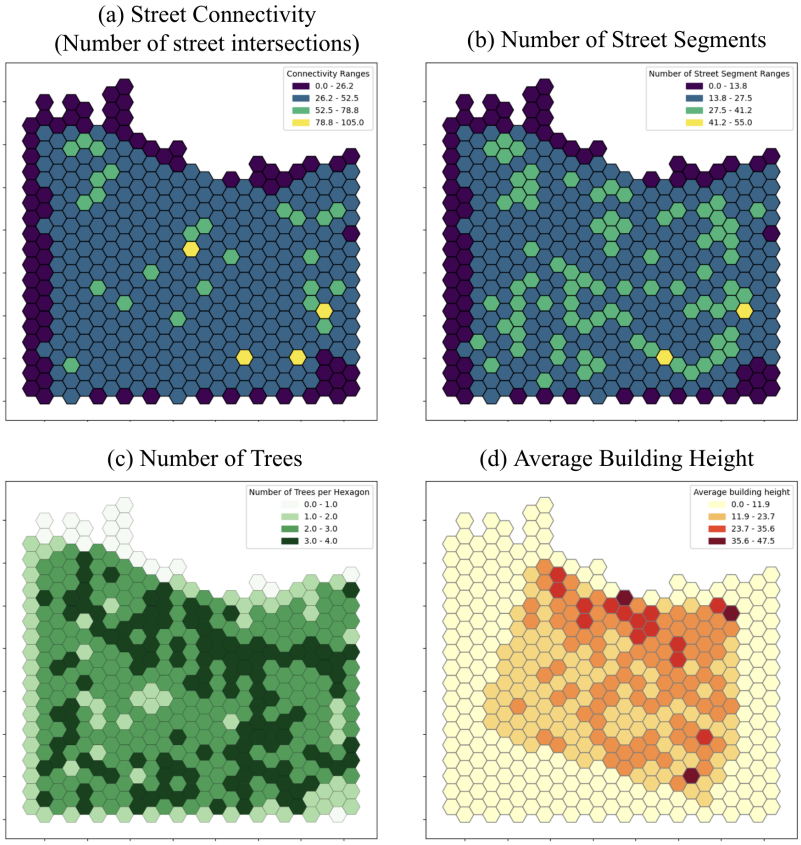

Figure 3. Spatial datasets and aggregation framework used in the analysis.

3.3. Predictive Modeling

The enriched dataset composed of point-level PM2.5 values and hexagon-derived urban indicators was used to train and compare multiple machine learning models. To evaluate the predictive capacity of urban features on fine particulate matter concentration (PM2.5), five regression models were tested: Linear Regression, Random Forest, Extra Trees, AdaBoost, and Gradient Boosting. All models were evaluated using five-fold cross-validation on the training dataset, and their performance was compared using three metrics: the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

4. Model Results

Table 1 summarizes the performance results of the investigated models. The three tree-based ensemble models (Gradient Boosting, Extra Trees, and Random Forest) presented the highest performance, each achieving an R^2 of approximately 0.805 and RMSE values around 2.25. AdaBoost performed less effectively, with the lowest R^2 (0.7567) and the highest RMSE and MAE, indicating its limitations in this specific context.

Table 1. Variables to be considered on the evaluation of interaction techniques

Model	R^2	RMSE	MAE
Gradient Boosting	0.805372	2.2471	1.6143
Extra Trees	0.805371	2.2471	1.6142
RandomForest	0.805357	2.2472	1.6138
Linear Regression	0.801671	2.2683	1.6474
AdaBoost	0.756733	2.5122	1.9647

Among them, Gradient Boosting yielded the best overall performance, with R^2 = 0.8054, RMSE = 2.25, and MAE = 1.61, and was selected as the final model for spatial prediction across the study area. These findings support the value of ensemble learning methods to model air pollution in urban settings based solely on built environment indicators.

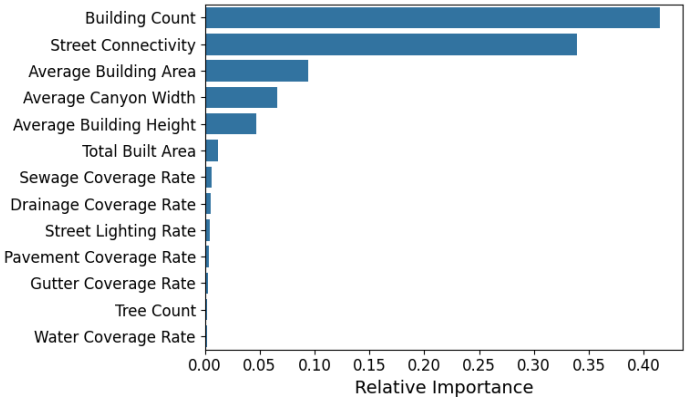

Figure 4. Importance of urban features employed in the Gradient Boosting Regressor Model.

The feature importance plot (Figure 4) reveals that Building Count and Street Connectivity are the most influential predictors in the Gradient Boosting model for estimating PM2.5 concentrations, together accounting for the majority of model relevance. This suggests that denser and more connected urban areas tend to exhibit stronger associations with air pollution levels, likely due to increased built surfaces and traffic-related emissions. Average Building Area, Average Canyon Width, and Average Building Height also contribute meaningfully, indicating that morphological characteristics of the built environment, such as enclosed urban canyons or larger structures, may affect air flow and pollutant dispersion. In contrast, variables related to infrastructure coverage (e.g., drainage, lighting, pavement, water) and Tree Count show very low importance.

4.1. Spatial Application of the Air Quality Model

After validating model performance, the best-performing model (Gradient Boosting Regressor) was applied to all hexagons in the study area, including those without sensor coverage (Figure 5). This final prediction step enabled spatial extrapolation of PM2.5 concentrations based solely on the urban fabric. By linking sensor data to contextual features through the hexagonal framework, the method combines the granularity of point measurements with the scalability of spatial modeling over the full urban area.

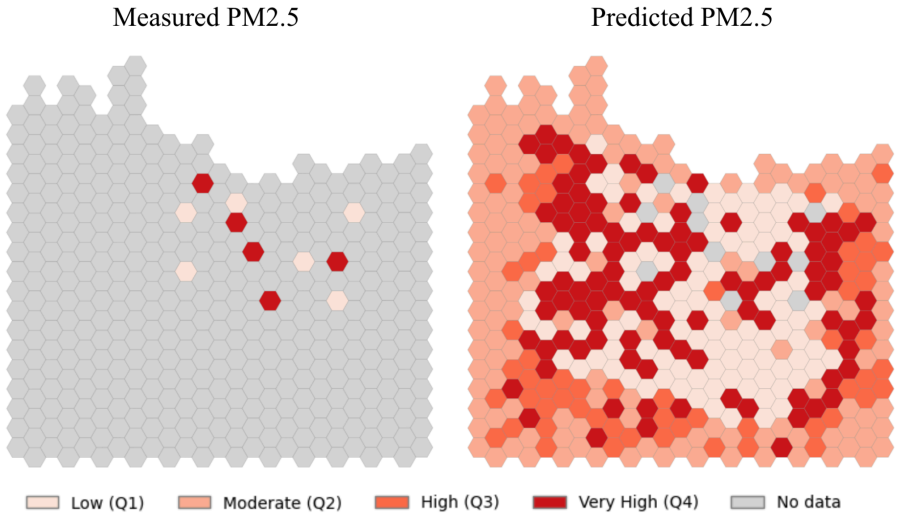

Figure 5. Comparison of measured and predicted PM2.5 concentrations across hexagons.

The 3D visualization (Figure 6) presents a detailed perspective of how predicted PM2.5 concentrations vary across the urban fabric. Each building is assigned a color based on the predicted air quality value of the hexagon it falls within, allowing a block-level interpretation of environmental quality. The classification follows the quantile-based ranges depicted in the hexagon-level prediction map (Figure 5), which divides the entire distribution of predicted values into four equal groups (Q1–Q4). Buildings located in hexagons categorized as "Very High" (Q4) are rendered in deep red, signaling higher exposure to air pollution.

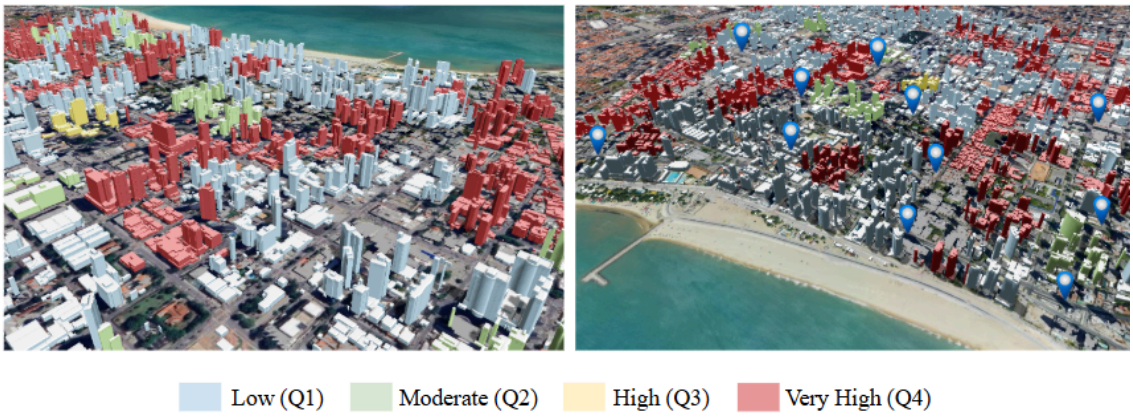

Figure 6. 3D building visualization colored by predicted PM2.5 concentration levels and the location of primary data collection with the sensor.

This approach bridges the aggregated spatial prediction and the physical form of the city, highlighting how built density and configuration relate to local air quality. It enhances interpretation by grounding pollution estimates in recognizable urban morphology.

5. Conclusions

The primary goal of this research was to investigate how features of the built environment such as density, connectivity, and morphology are structurally associated with pollution levels. Data analysis showed that urban form plays a critical role in environmental outcomes, calling for an integrated planning approach in which sustainable urban development becomes a key vector for improving air quality and, consequently, public health and well-being.

It is important to note that the primary air quality data used in this study represent a static snapshot in time, capturing PM2.5 concentrations during a specific measurement period. While air pollution levels are known to fluctuate due to temporal factors such as weather conditions, traffic cycles, and seasonal variations, this dynamic aspect was not captured in the current analysis. Future model improvements could benefit from integrating time series data to enhance the temporal resolution of predictions.

Additionally, the analysis of feature importance revealed that Average Building Area, Average Canyon Width, and Average Building Height all contributed modestly to the model, and are likely to be strongly correlated. Similarly, variables related to infrastructure, such as drainage, lighting, pavement, and water coverage, showed individually low importance but may collectively represent an underlying dimension of urban service provision. To reduce redundancy and improve model parsimony, future work could explore dimensionality reduction techniques such as Principal Component Analysis to group correlated features into composite indicators.

References

Che, W., Zhang, Y., Lin, C., Fung, Y. H., Fung, J. C. H., & Lau, A. K. H. (2023). Impacts of pollution heterogeneity on population exposure in dense urban areas using ultra-fine resolution air quality data. Journal of Environmental Sciences, 125, 513–523. https://doi.org/10.1016/j.jes.2022.02.041.

Dapper, S. N., Spohr, C., & Zanini, R. R. (2016). Poluição do ar como fator de risco para a saúde: Uma revisão sistemática no estado de São Paulo. Estudos Avançados, 30, 83–97. https://doi.org/10.1590/S0103-40142016.00100006.

Faraji, M., Nadi, S., Ghaffarpasand, O., Homayoni, S., & Downey, K. (2022). An integrated 3D CNN-GRU deep learning method for short-term prediction of PM2.5 concentration in urban environment. Science of the Total Environment, 834, 155324. https://doi.org/10.1016/j.scitotenv.2022.155324.

Furtado, L. S., Monteiro, N., Gurjão, N., Cavalcante, R. M., Silva Filho, J. E., da Silveira, J. A. N., ... & de Macedo, J. A. F. (2024). Low-Cost Smart Sensing Pipeline:

Assembly, Calibration, and Interpretation of Air Quality Data. In 2024 IEEE International Smart Cities Conference (ISC2) (pp. 1-6). IEEE.

Gurjão, N. O. (2024). Estudo sobre a contribuição do tráfego e de fatores urbanos na poluição atmosférica de Fortaleza/CE utilizando um equipamento de baixo custo. Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia de Transportes, Universidade Federal do Ceará, Fortaleza, Brasil. Available at: https://repositorio.ufc.br/handle/riufc/80631.

Gurjão, N. O., Lucas Júnior, J. L. O., Furtado, L. S., & Soares, J. B. (2024). Air pollution dynamics in Fortaleza, Brazil: Exploring the interplay of traffic and high-rise development. Urban Climate, 58, 102176. http://dx.doi.org/10.1016/j.uclim.2024.102176.

Ji, C., Zhang, C., Hua, L., Ma, H., Nazir, M. S., & Peng, T. (2022). A multi-scale evolutionary deep learning model based on CEEMDAN, improved whale optimization algorithm, regularized extreme learning machine and LSTM for AQI prediction. Environmental Research, 215, 114228. https://doi.org/10.1016/j.envres.2022.114228.

Li, G., Wu, Z., Liu, N., Liu, X., Wang, Y., & Zhang, L. (2023). A variational Bayesian blind calibration approach for air quality sensor deployments. IEEE Sensors Journal, 23(7), 7129–7141. https://doi.org/10.1109/JSEN.2022.3212009.

Liang, H., Zhou, X., Zhu, Y., Li, D., Jing, D., Su, X., & Zhang, Y. (2023). Association of outdoor air pollution, lifestyle, genetic factors with the risk of lung cancer: A prospective cohort study. Environmental Research, 218, 114996. https://doi.org/10.1016/j.envres.2022.114996.

Liu, X., Jayaratne, R., Thai, P., Kuhn, T., Zing, I., ChristenseN, B., Lamont, R., Dunbabin, M., Zhu, S., & Gao, J. (2020). Low-cost sensors as an alternative for long-term air quality monitoring. Environmental Research, 185, 109438. https://doi.org/10.1016/j.envres.2020.109438.

Lou, C., Jiang, F., Tian, X., Zou, Q., Zheng, Y., Shen, Y., Feng, S., Chen, J., Zhang, L., & Jia, M. (2023). Modeling the biogenic isoprene emission and its impact on ozone pollution in Zhejiang Province, China. Science of the Total Environment, 865, 161212. https://doi.org/10.1016/j.scitotenv.2022.161212.

Sakti, A. D., Anggraini, T. S., Ihsan, K. Y., Misra, P., Trang, N. T. Q., Pradhan, B., Wenten, I. G., Hadi, P. O., & Wikantika, K. (2023). Multi-air pollution risk assessment in Southeast Asia region using integrated remote sensing and socio-economic data products. Science of the Total Environment, 854, 158825. https://doi.org/10.1016/j.scitotenv.2022.158825.

Seaton, M., O'Neill, J., Bien, B., Hood, C., Jackson, M., Jackson, R., Johnson, K., Oades, M., Stidworthy, A., & Stocker, J. (2022). A multi-model air quality system for health research: Road model development and evaluation. Environmental Modelling & Software, 155, 105455. https://doi.org/10.1016/j.envsoft.2022.105455.

Secretaria Municipal das Finanças de Fortaleza. (2016). GeoServiços - IDE Fortaleza. Infraestrutura de Dados Espaciais (IDE). Available at: https://ide.sefin.fortaleza.ce.gov.br/geoservicos.

Silva, L. T., & Mendes, J. F. G. (2006). Determinação do índice de qualidade do ar numa cidade de média dimensão. In Anais do 2º Congresso Luso-Brasileiro de Planeamento Urbano Regional Integrado Sustentável, Braga. [CD-ROM]. ISBN 85-85205-67-9. https://hdl.handle.net/1822/7177.

Stache, E., Schilperoort, B., Ottelé, M., & Jonkers, H. M. (2022). Comparative analysis in thermal behaviour of common urban building materials and vegetation and consequences for urban heat island effect. Building and Environment, 213, 108489. http://dx.doi.org/10.1016/j.buildenv.2021.108489.

Tella, A., & Balogun, A. (2021). GIS-based air quality modelling: Spatial prediction of PM10 for Selangor state, Malaysia using machine learning algorithms. Environmental Science and Pollution Research, 29(57), 86109–86125. https://doi.org/10.1007/s11356-021-16150-0.

Vardoulakis, S., Fisher, B. E. A., Pericleous, K., & Gonzalez-Flesca, N. (2003). Modelling air quality in street canyons: A review. Atmospheric Environment, 37(2), 155–182. http://dx.doi.org/10.1016/S1352-2310(02)00857-9.

World Health Organization (WHO). (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. https://apps.who.int/iris/handle/10665/345329.

Xie, M., Lu, X., Ding, F., Cui, W., Zhang, Y., & Feng, W. (2022). Evaluating the influence of constant source profile presumption on PMF analysis of PM2.5 by comparing long- and short-term hourly observation-based modeling. Environmental Pollution, 314, 120273. https://doi.org/10.1016/j.envpol.2022.120273.

Yang, L., Zhang, L., Stettler, M. E. J., Sukitpaneenit, M., Xiao, D., & Dam, K. H. (2020). Supporting an integrated transportation infrastructure and public space design: A coupled simulation method for evaluating traffic pollution and microclimate. Sustainable Cities and Society, 52, 101796. https://doi.org/10.1016/j.scs.2019.101796.

Zheng, T., Jia, Y., Zhang, S., Li, X., Wu, Y., Wu, C., He, H., & Peng, Z. (2021). Impacts of vegetation on particle concentrations in roadside environments. Environmental Pollution, 282, 117067. http://dx.doi.org/10.1016/j.envpol.2021.117067.