Global Journal of Earth and Environmental Science
Volume 10(4), pages 126-139, August 2025
Article Number: 43BCF7831
ISSN: 2636-6002
https://doi.org/10.31248/GJEES2024.191
https://integrityresjournals.org/journal/GJEES
Full Length Research

Spatial estimation of particulate matter (PM2.5) in selected African cities using machine learning-based models for improved air quality assessment

Kaothar Ayomide ABUDULAWAL* and Ismaheel Olajide BELLO

Innovative Analytics Technology Limited, Lagos State, Nigeria.

*Corresponding author. Email: abudulawalkaothar@gmail.com; Co-author: belloismaheel33@gmail.com

Received 15th April 2025; Accepted 19th July 2025

ABSTRACT: Nine (9) out of ten (10) people around the world breathe air that does not meet WHO's recommended air quality standards. This study aims to create an accurate and scalable machine learning model using satellite-derived Aerosol Optical Depth (AOD) data, temporal and seasonal derived features to estimate Particulate Matter (PM2.5) levels, enhance air quality monitoring and inform interventions for public health and environmental sustainability. The geographical locations considered in this study include Lagos (Nigeria), Bujumbura (Burundi), Nairobi (Kenya), and Kampala (Uganda). Predictive Regression models used in this study are XGBoost, LightGBM, Ridge, Polynomial, and Feedforward Neural Network (FNN). XGBoost emerged as the best-performing model, which achieves an RMSE of 12.01 μg/m³ and an R^2 of 0.76. Spatial analysis using Local Indicators of Spatial Association (LISA) and Global Moran's I statistic revealed varying degrees of spatial clustering of PM2.5 concentrations across the cities. Lagos, which exhibits the strongest positive spatial autocorrelation with Moran's I statistic of 0.686 and Nairobi the weakest with Moran's I statistic of 0.046. This study shows the effectiveness of combining satellite AOD data with temporal and seasonal variables in enhancing the predictive accuracy of PM2.5 estimation models. It provides critical insights for air quality management and highlights the importance of spatially informed models. This is important to identify localised pollution hotspots for more effective environmental health interventions.

Keywords: Air quality management, environmental health, environmental sustainability, machine learning, particulate matter, Sub-Saharan Africa.

INTRODUCTION

Air pollution is mainly caused by a variety of human and natural activities, such as the burning of fossil fuels for energy and transportation, industrial emissions and agricultural activities that release harmful substances into the atmosphere. According to Nguyen *et al.* (2024), air pollution is the introduction of hazardous or excessive levels of substances like gases, particles, and biological molecules into the Earth's atmosphere. This is detrimental to our health and poses a serious environmental and public health concern.

Air pollutants such as ground-level ozone and PM2.5, according to the National Oceanic and Atmospheric Administration (2025) have been observed to be influenced by climate change and the cause of respiratory and cardiovascular ailments. One (1) in (8) deaths is recorded globally per year, and this has been attributed to air pollution (Locke *et al*., 2022), with an estimated 4.2 million people dying annually from exposure to outdoor air pollution (World Health Organisation, 2021). The detrimental effects of air pollutants emphasise the need for effective monitoring and prediction mechanisms, particularly in densely populated areas (Panaite *et al,* 2024).

Unfortunately, despite the variety of in-depth studies on air-quality assessment in developed cities, there is presently little or no study that has assessed the air quality

estimation of more than one city or country at the same time in Africa. Recent urban studies do not provide insights into the air quality assessment of the studied regions: Lagos in Nigeria, Bujumbura in Burundi, Kampala in Uganda and Nairobi in Kenya. This research, therefore, aims to bridge the gap in air quality assessment in sub-Saharan African cities by predicting the air quality of the study areas and providing a robust solution for estimating PM2.5 levels across these regions, which is a major source of concern for public health among various pollutants. Hence, this study is unique for its focus on African urban areas and its inclusion of satellite-derived AOD data, thus offering both novelty and significance. This research, therefore, attempts to develop predictive models specifically for these cities, which are crucial for mitigating air pollution and improving air quality in sub-Saharan Africa.

This study is expected to enhance the understanding and application of air-quality assessment estimation, mitigation of air pollution, and preparation for future trends, given our predicted air quality models for the study areas.

Furthermore, this work reinforces the initial different initiatives to mitigate environmental challenges, most especially improving air quality, thereby providing useful findings applicable to similar settings in Africa. Also, it adds to the body of knowledge by focusing more in-depth on air quality improvements within unique settings of sub-Saharan Africa. This will serve as a practical tool for academia, policymakers, government, and stakeholders to ensure and improve sustainable development in those regions and beyond.

MATERIALS AND METHODS

Data and material description

The data used in this research was obtained from a public data science competition hosted on Zindi Africa by AirQo (2024). The dataset is Sentinel-5p data extracted from Google Earth Engine. The dataset consists of observations from four cities in four African Countries. This dataset contains 80 feature variables that cover various aspects, including Location and Time Features, Pollutant Measurements and Atmospheric Features, Carbon Monoxide (CO) variables, Nitrogen Dioxide (NO_2) variables, Formaldehyde (HCHO) variables, Ozone (O_3) variables, and Aerosol and Cloud variables. In total, the data contains 8071 observations covering the period from the 1st of January 2023 to the 26th of February 2024.

Methods

This research leverages Statistical models, Machine Learning models, and Deep learning models to determine the approach that best performs in estimating the value of PM2.5. It is believed that this can help reduce environmental health risks in the regions of Africa. The statistical methods employed include Ridge regression and Polynomial Ridge regression. The machine learning models employed are XGBoost and LightGBM, while a Feedforward Neural Network was employed for deep learning techniques.

Data preprocessing

The data preprocessing started from the inspection of missing values within the variables. Upon inspection, it was observed that some variables have missing values up to 94% of the total sample. Variables with missing values of more than 40% of the total sample were dropped from the data. Also, variables like id and site id were also dropped, after which the total number of variables decreased from 80 to 35. Then, the dataset was partitioned into train and test datasets using a ratio of 80:20, respectively. Stratified sampling was employed for the partitioning, with the site ID variable used as the stratification criterion. After partitioning, the missing values for each variable in both the train and test sets were replaced using the mean imputation method derived from the Train set. To evaluate the performance of the models, several metrics were used, including R-squared Goodness of Fit, Adjusted R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Feature engineering

To enhance model performance and capture underlying patterns in the datasets, several feature engineering strategies were implemented, which include Cyclical Encoding of temporal features, Derived Cloud Features, and Interaction terms.

The temporal features were derived as follows:

$$hour\_\sin = \sin\left(\frac{2\pi * hour}{24}\right), hour\_\cos = \cos\left(\frac{2\pi * hour}{24}\right)$$

$$month\_\sin = \sin\left(\frac{2\pi * month}{12}\right), month\_\cos = \cos\left(\frac{2\pi * month}{12}\right)$$

$$is\_weekend = \begin{cases} 1, & if\ day\ is\ saturday\ or\ sunday \\ 0, & otherwise \end{cases}$$

The additional Cloud features which were derived are:

$$cloud\_height\_diff = cloud\_top\_height - cloud\_base\_height$$

$$cloud\_pressure\_diff = cloud\_top\_pressure - cloud\_base\_pressure$$

Interaction terms were also derived using:

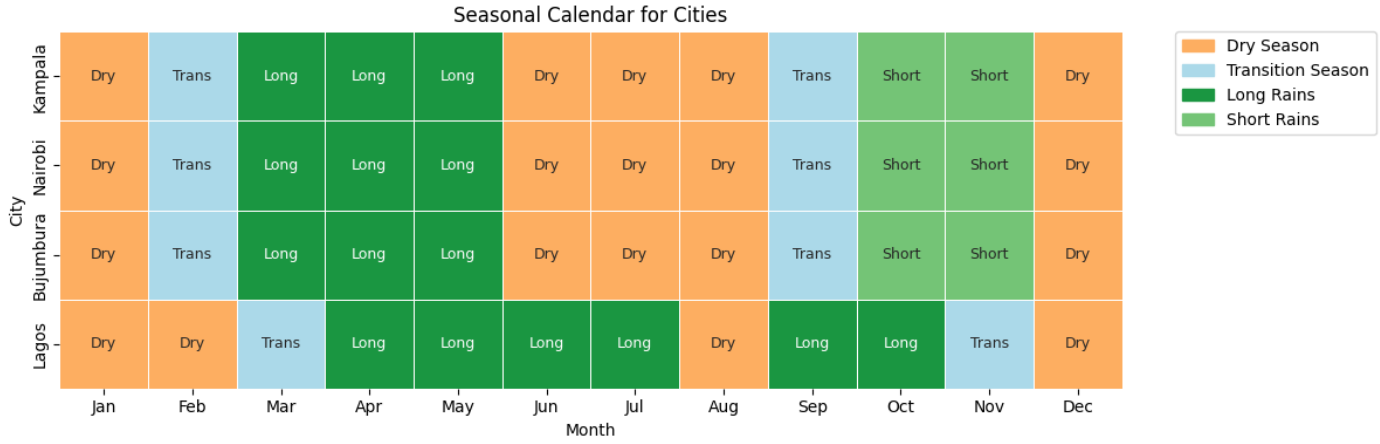

Figure 1. Seasonal classification calendar for the four cities.

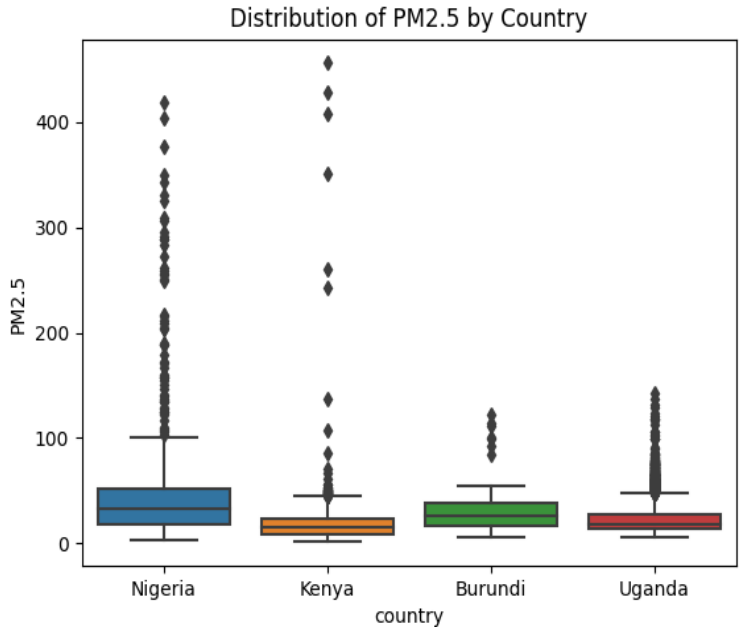

Figure 2. Distribution of PM2.5 by country.

$$aerosol\_solar\_interaction = absorbing\_aerosol\_index \times solar\_zenith\_angleozone\_temp\_interaction = ozone\_column\_density \times ozone\_effective\_temperature$$

Additionally, a seasonal feature was derived for each city. Figure 1 below illustrates how the seasonal classifications were assigned across the months for different cities.

Data exploration

Figure 2 above shows the distribution of PM2.5 across the four cities in the selected countries. There is a wide disparity in the concentration of PM2.5 across the cities, which can be attributed to the varying degrees of industrialisation, urbanisation, and environmental regulation in these countries. Nigeria has the highest concentration of PM2.5, with numerous data points reaching up to 400, indicating severe air pollution. Kenya follows with a significantly lower concentration, although there are a few outliers higher than 400. Burundi and Uganda have the lowest concentration levels, with Uganda slightly higher than Burundi, but both are below 200, indicating that there might be relatively better air quality in these countries compared to Nigeria and Kenya. The linear relationships between PM2.5 and the feature variables were conducted using pair-wise Pearson correlations and visualised to reveal the strength and direction of association.

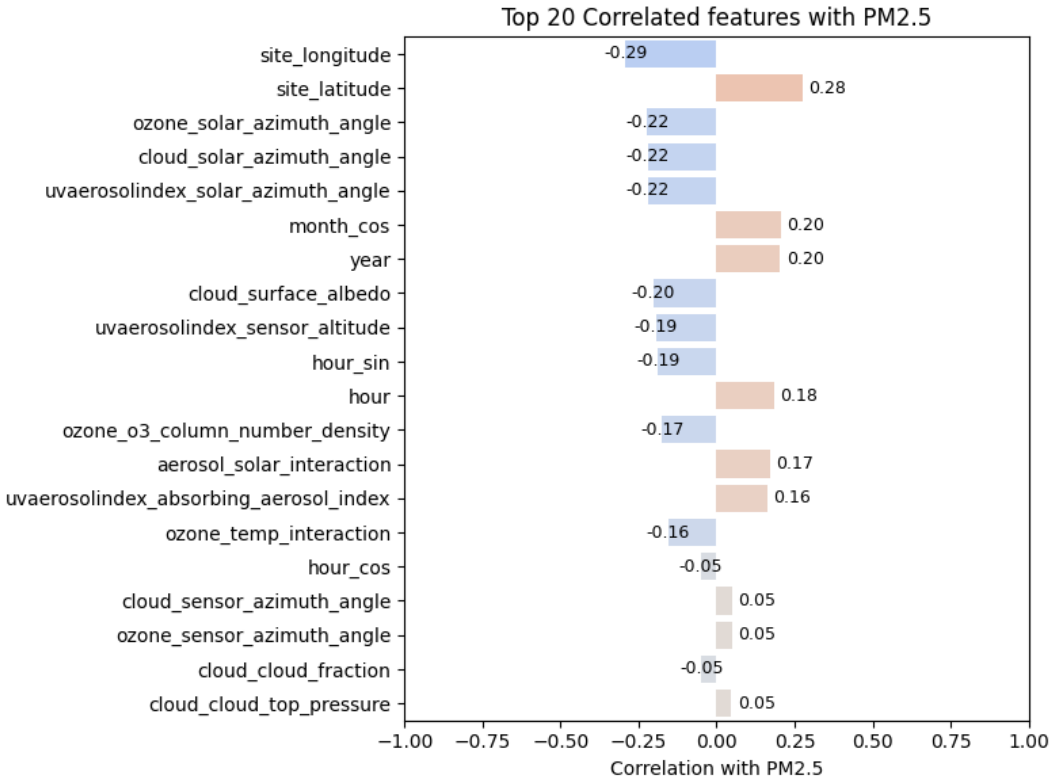

Figure 3. Top 20 Features with their Correlations with PM2.5.

Figure 3 shows the top 20 features that are most positively and negatively correlated with PM2.5. It shows that spatial, solar, and atmospheric features stand out in their relationship with PM2.5. Longitude has a correlation of −0.29, while latitude is 0.28. This implies that PM2.5 tends to increase as we move north and decrease as we move east. Also, cloud surface albedo (−0.20), uvaerosolindex sensor altitude (−0.19), hour_sin (−0.19), and ozone column number density (−0.17) show negative relationships, meaning PM2.5 reduces as these increase. On the positive side, month and year both have a correlation of 0.20, while hour, aerosol solar interaction, and absorbing aerosol index also show moderate positive values. This suggests that time-related patterns and aerosol properties contribute to PM2.5 levels. A few features, like cloud fraction and cloud top pressure, have very weak correlations. They may still matter in non-linear ways, but are not strongly linearly related to PM2.5. Other features with strong negative correlation include ozone solar azimuth angle, cloud solar azimuth angle, and uvaerosolindex solar azimuth angle, each with about −0.22. These show that solar angles have a role in reducing PM2.5, likely because of their influence on sunlight and how it interacts with aerosols. Also, cloud surface albedo (−0.20), uvaerosolindex sensor altitude (−0.19), hour_sin (−0.19), and ozone column number density (−0.17) show negative correlations, meaning PM2.5 reduces as these increase.

Evaluation metrics

To evaluate the performance of the models used in this study, four widely used error and goodness-of-fit metrics, which are Mean Squared Error (MSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2) and Adjusted R^2. were employed. These metrics will help us capture both the accuracy of the predictions and the explanatory power of the models for a balanced assessment of performance from different perspectives.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where y_i is the actual PM2.5 value and y_i is the predicted PM2.5 values, and n is the total number of observations.

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

Where $\bar{y}$ is the mean of the actual PM2.5 values

$$Adjusted\ R^2 = 1 - (1 - R^2) \times \frac{n-1}{n-p-1}$$

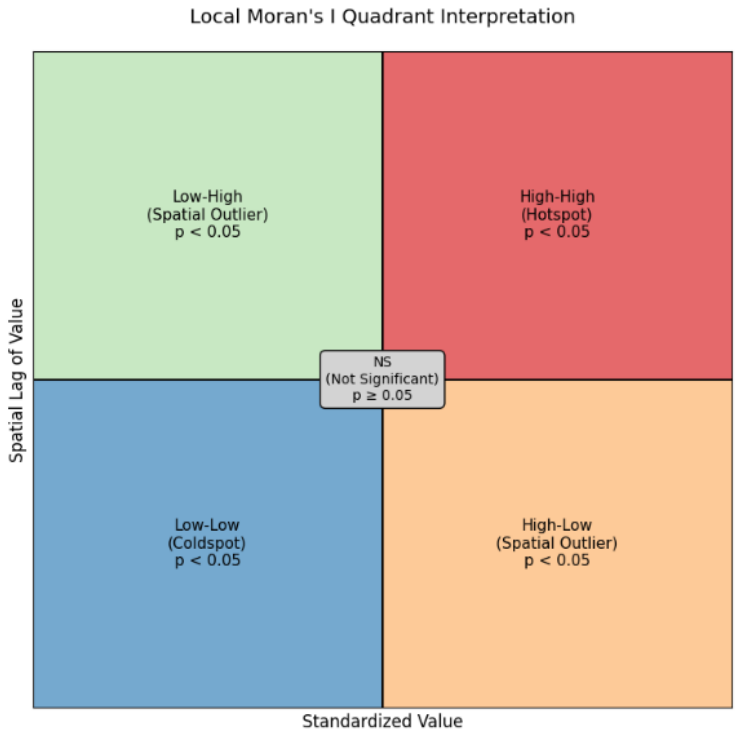

Figure 4. LISA Classification scheme.

Where n is the number of observations and p is the number of independent variables.

Spatial autocorrelation analysis

Global and local spatial autocorrelation of PM2.5 concentrations were evaluated using Moran's I and Local Indicators of Spatial Association (LISA). Spatial weights matrices were constructed using k-nearest neighbours. Global Moran's I was computed to assess overall clustering; it was calculated using:

$$I = \frac{N}{W} \times \frac{\sum_i \sum_{jwij}(x_i - \bar{x})(x_j - \overline{x})}{\sum_i (x_i - \bar{x})^2}$$

Where: N = total number of spatial units (sites), W = the sum of all spatial weights W_{ij}, x_i *and* x_j = The PM2.5 concentration values at locations i and j, w_{ij} = is the spatial weight between locations i and j based on proximity (k-nearest neighbors).

The value of Moran's I typically ranges between -1 and +1:

I > 0 indicates spatial clustering (similar values are near each other),
I < 0 suggests spatial dispersion (dissimilar values are near each other), and
I ≈ 0 signifies a random spatial distribution.

This research also uses LISA to identify local patterns, including High-High (HH), Low-Low (LL), High-Low (HL), and Low-High (LH) clusters. Statistical significance was assessed at p<0.05.

Figure 4 illustrates the classification scheme used in Local Indicators of Spatial Association (LISA) analysis for interpreting spatial clustering patterns. The High-High (HH) quadrant represents areas with high concentration values surrounded by other high values (Hotspots) and is considered statistically significant at p<0.05. The Low-Low (LL) quadrant depicts locations with low PM2.5 concentrations, also surrounded by low values (Cold spots), significant at p < 0.05. High-Low (HL) areas are those where a high concentration site is surrounded by low concentration sites, typically reflecting spatial outliers, significant at p<0.05, while the non-significant quadrant depicts other areas which are not statistically significant at p<0.05.

Table 1 shows the revised air quality standards for particle pollution and updates to the air quality index (AQI) according to NAAQS. This standard categorisation is used in this research for a better understanding and interpretation of the different levels of PM2.5.

RESULTS

Table 2 presents the comparative performance of the five predictive models used in estimating PM2.5, which was evaluated using RMSE, MAE, R^2, Adj R^2, and model fitting

Table 1. The National Ambient Air Quality Standards for Particle Pollution (NAAQS, United States: Environmental and Protection Agency (EPA, 2024).

AQI Category	Index Values	Previous Breakpoints (1999 AQI) (µg/m³, 24-hour average)	Revised Breakpoints (µg/m³, 24-hour average)
Good	0 – 50	0.0 – 15.0	0.0 – 12.0
Moderate	51 – 100	>15.0 – 40	12.1 – 35.4
Unhealthy for Sensitivity Groups	101 – 150	>40 – 65	35.5 – 55.4
Unhealthy	151 – 200	>65 – 150	55.5 – 150.4
Very Unhealthy	201 – 300	>150 – 250	150.5 – 250.4
Hazardous	301 – 400	>250 – 350	250.5 – 350.4
	401 – 500	>350 – 500	350.5 – 500

Table 2. Performance of models in estimating PM2.5.

Metrics	Ridge	PolyRegression_order_2	XGBoost	LightGBM	Neural Network
RMSE	21.901584	21.327934	12.014032	12.866516	20.575495
MAE	11.226098	10.144270	5.314148	6.705398	9.062059
Rsquared	0.206685	0.247698	0.761289	0.726211	0.299843
Adj Rsquared	0.183932	0.226122	0.754443	0.718358	0.279762
Fit Time (s)	0.320522	7.062807	23.057269	2.182495	15.967955

time. From the results, it is evident that XGBoost outperformed all other models across all evaluation metrics, achieving the lowest RMSE of 12.01 µg/m³ and MAE of 5.31 µg/m³, which indicate superior accuracy in predicting PM2.5 concentrations. The model also achieved the highest R^2 value of 0.76, which implies that approximately 76% of the variability in PM2.5 concentration was explained by the features used in the model. Its corresponding Adjusted R^2 of 0.75 further confirms the model's robustness after accounting for the number of predictors in the model. LightGBM followed closely, with an RMSE of 12.87 µg/m³ and MAE of 6.71 µg/m³. The R^2 and Adjusted R^2 values for LightGBM are 0.73 and 0.72, respectively, which indicate strong explanatory power, though slightly less than XGBoost. The Feedforward Neural Network model demonstrated moderate performance, with an RMSE of 20.58 µg/m³ and an R^2 of 0.30. Ridge Regression and Polynomial Regression (order 2), on the other hand, exhibited the weakest predictive abilities, with higher errors of RMSE 21.90 µg/m³ and 21.33 µg/m³ and lower R^2 values 0.21 and 0.25, respectively. In terms of computational efficiency, Ridge Regression was found to be the fastest model, which trained at 0.32 seconds, while XGBoost, despite being the most accurate, required the longest fitting time at 23.06 seconds, which may be due to its iterative boosting process.

Spatial distribution of actual and predicted PM2.5 concentrations for various cities

Figure 5a (left) and 5b (right) above present the spatial distribution of actual and predicted PM2.5 concentrations for Bujumbura (Burundi), as estimated by the XGBoost model, which was identified as the best-performing model in this study. A comparison of the two maps reveals a close alignment between the predicted and actual PM2.5 values across most locations. Although there are minor discrepancies which are evident in specific areas. In the Ntahangwa region, the model predicted PM2.5 concentrations to fall within the 35.5–55.4 µg/m³ range, which corresponds to the Unhealthy for Sensitive Groups category based on the revised 24-hour AQI breakpoints. However, the actual measurements in this area predominantly range between 12.1–35.4 µg/m³, which is classified as Moderate. This indicates that there is a slight overestimation of the model in this locality. Generally, the model performed well in predicting PM2.5 in Bujumbura (Burundi).

Figure 6a (left) and 6b (right) above illustrate the spatial distribution of actual and predicted PM2.5 concentrations for Kampala (Uganda) as estimated by the XGBoost model. From the results, it is evident that the XGBoost model generally performs well in capturing the spatial distribution of PM2.5 across the city, except for some discrepancies in some localities. For example, in the Kesenko area, the XGBoost predicted that the value lies between 12.1–35.4 µg/m³ range, which corresponds to the Moderate AQI category, while the actual data shows that it lies between 55.5–150.4 µg/m³ range, which is classified as unhealthy. Similarly, the model estimated the concentration of PM2.5 around Yoka locality to be between 12.1–35.4 µg/m³, while the original concentration is between 0.0–12.0 µg/m³, which falls under the good category. The XGBoost model accurately captured the

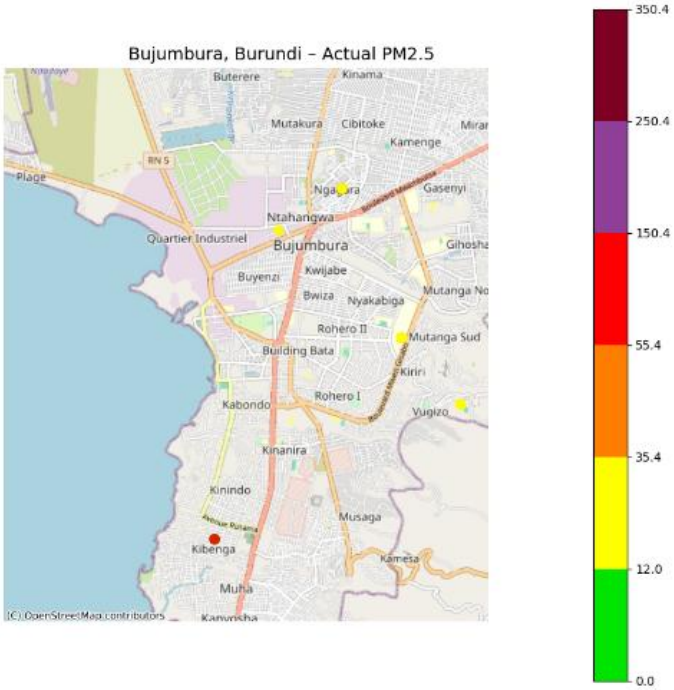

Figure 5a. Bujumbura, Burundi Actual Data

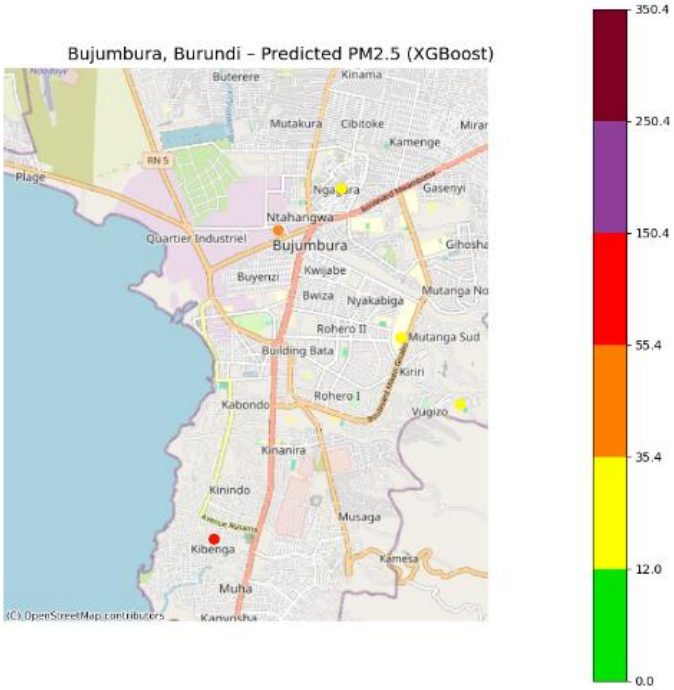

Figure 5b. Bujumbura, Burundi Predicted Data

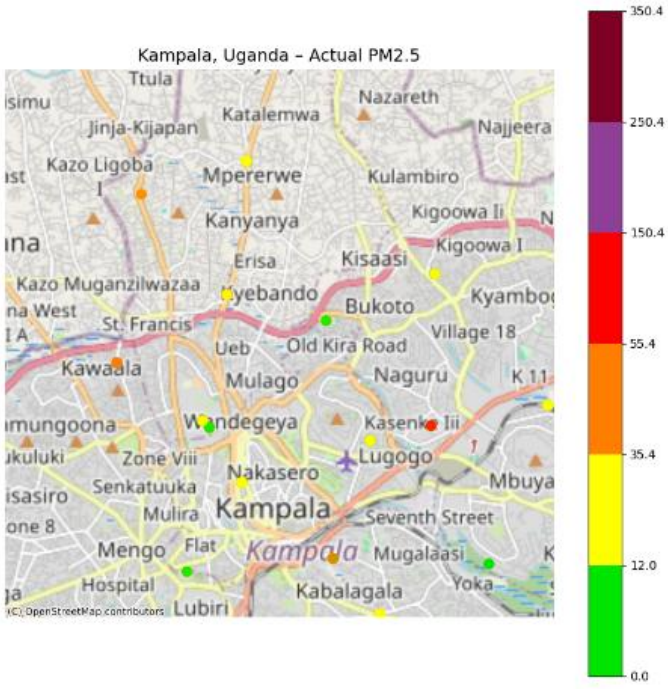

Figure 6a. Kampala, Uganda Actual Data.

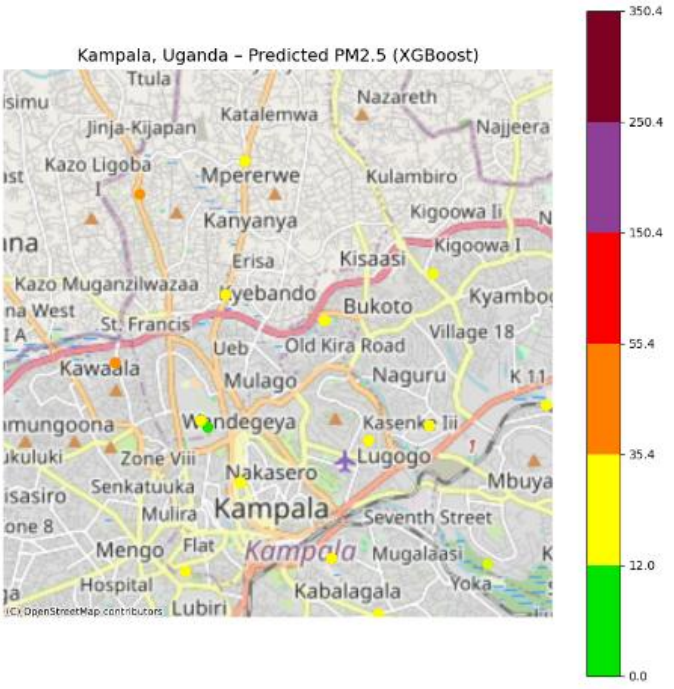

Figure 6b. Kampala, Uganda Predicted Data.

broader spatial distribution of PM2.5 concentrations in Kampala.

Figure 7a (left) and 7b (right) depict the actual and predicted spatial distribution of PM2.5 concentrations for Lagos (Nigeria) as estimated by the XGBoost model. From the figures, it can be observed that even though Lagos recorded the highest PM2.5 concentrations among all the cities considered in this study, the model generally

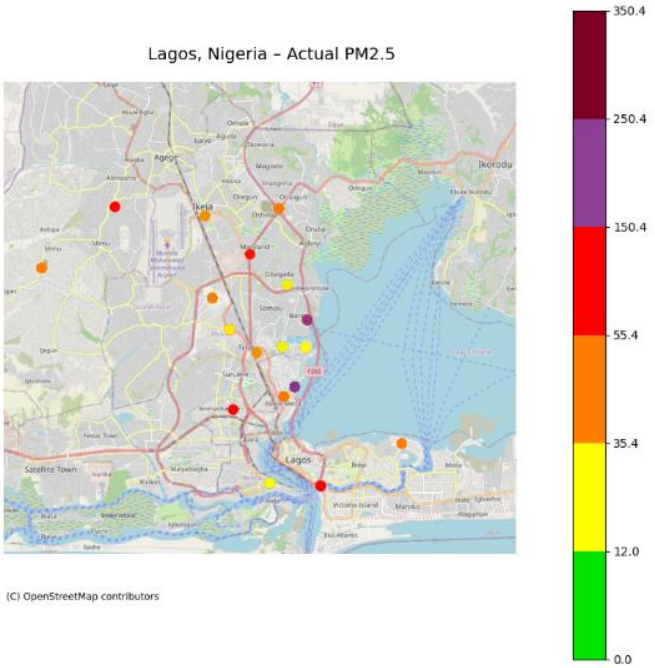

Figure 7a. Lagos, Nigeria Actual Data

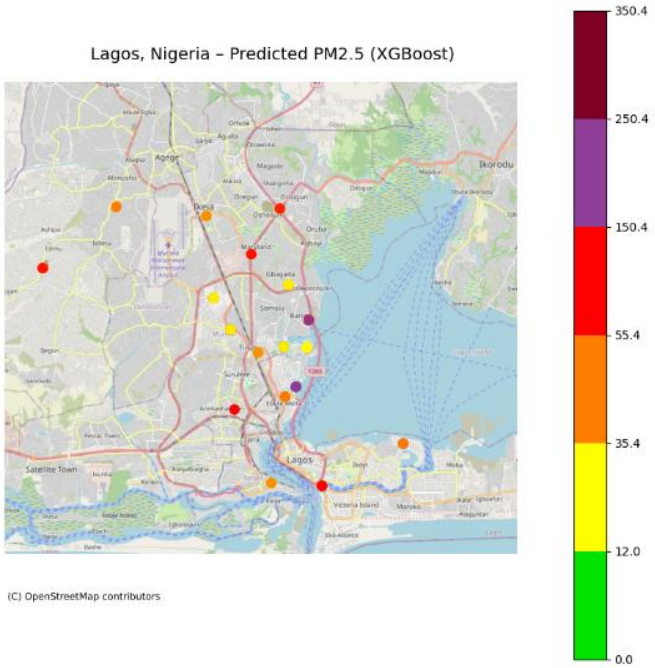

Figure 7b. Lagos, Nigeria Predicted Data.

performed well in estimating the spatial distribution of PM2.5 across the city. Most concentration ranges for different localities were accurately predicted by the model. However, there are noticeable discrepancies in specific areas. For instance, the model estimated PM2.5 concentrations in the Idimu area to fall within the 55.5–150.4 µg/m³ range, classified as Unhealthy, whereas the actual data indicates a concentration range of 35.5–55.4 µg/m³, which corresponds to the Unhealthy for Sensitive Groups category. Similarly, the model predicted PM2.5 concentrations in Illado locality to be between 35.5–55.4 µg/m³, while the actual spatial concentration was within 12.1–35.4 µg/m³, which is categorised as Moderate. Overall, while these discrepancies exist, the model was able to capture the broader pattern and intensity of PM2.5 concentrations across Lagos with reasonable accuracy.

Figure 8a (left) and 8b (right) show the spatial distribution of actual and predicted PM2.5 concentrations for Nairobi (Kenya), which was estimated by the XGBoost model. These visualisations show that the model generally performs well in capturing the spatial patterns of PM2.5 concentrations across the city, with only minor discrepancies observed in specific areas. For instance, the model estimated the concentration of PM2.5 in Karen C area to fall within the 12.1–35.4 µg/m³ range, which corresponds to being Moderate according to the AQI category, whereas the actual observed values fall within the 0.0–12.0 µg/m³ range, categorised as Good. Also, the model estimated concentrations in Tassia locality within the range of 0.0–12.0 µg/m³, while the actual measured concentrations were within 12.1–35.4 µg/m³. Despite these minor misestimations, the model accurately captured the broader spatial distribution and intensity of PM2.5 concentrations in Nairobi, which effectively identifies most areas of higher and lower air pollution levels across the city.

Spatial concentration and analysis

Figure 9a (left) and 9b (right) above present the actual PM2.5 concentration of the data used for training and the LISA cluster map for Bujumbura (Burundi). From the concentration map, it can be observed that Bujumbura exhibits a spatial distribution pattern ranging from Good to Moderate, Unhealthy for Sensitive Groups, and Unhealthy AQI categories. Although the spatial distribution map reveals that most of the study area comprises higher PM2.5 concentration values, with visible clusters of data points. A closer inspection of the LISA cluster map indicates that most of these apparent clusters are statistically non-significant at $p \geq 0.05$. One significant cluster identified is within Bujumbura itself, which is classified as a High-High (HH) cluster. This means that this area has a high PM2.5 concentration value and is surrounded by neighbouring sites with similarly high concentration values, forming a significant hotspot at $p < 0.05$. The computed Global Moran's I statistic for the PM2.5 concentrations in Bujumbura is 0.082, which suggests a weak but positive spatial autocorrelation. This value implies that while there is a slight tendency for similar PM2.5 concentration values to cluster together spatially,

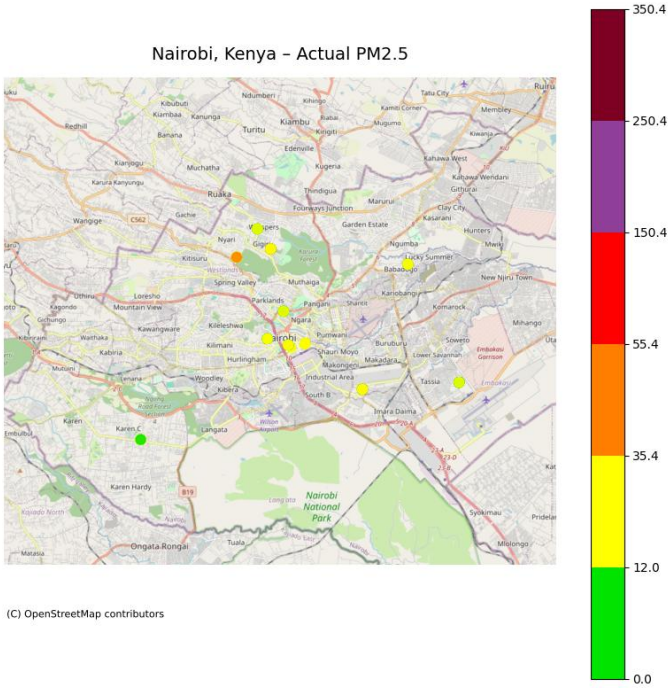

Figure 8a. Nairobi, Kenya Actual Data.

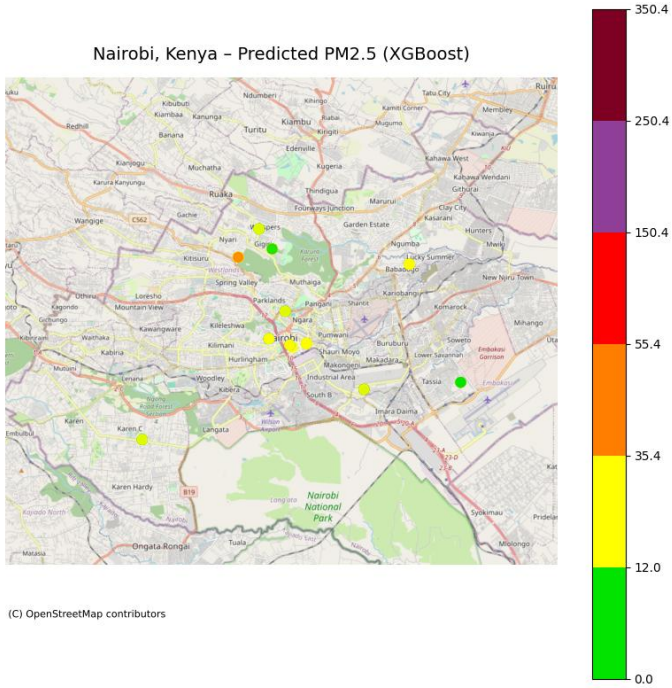

Figure 8b. Nairobi, Kenya Predicted Data.

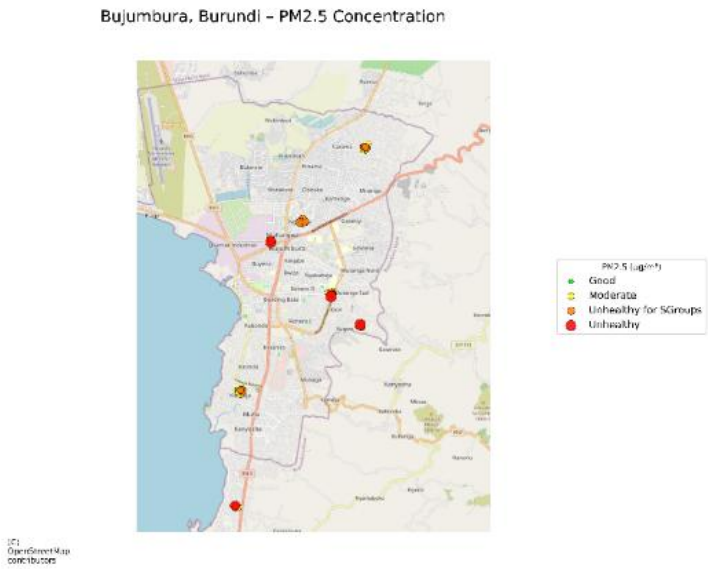

Figure 9a. Bujumbura Burundi PM2.5 Concentration

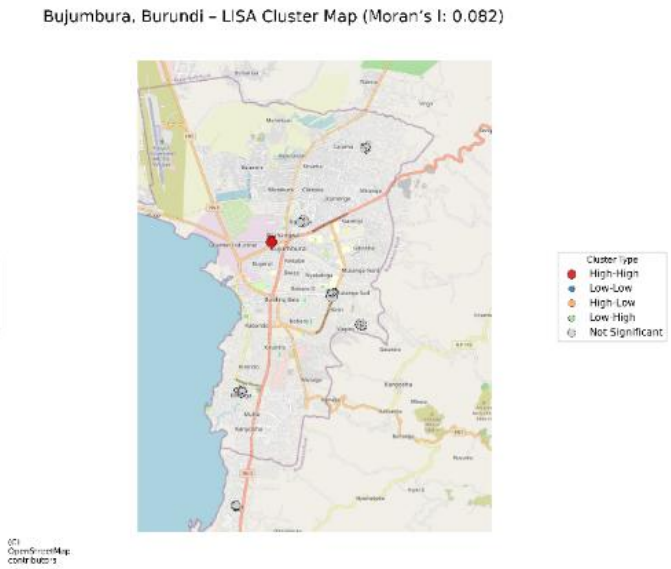

Figure 9b. Bujumbura Burundi LISA Cluster Map

the overall clustering pattern across the study area is not strongly pronounced.

Figure 10a (left) and 10b (right) present the spatial distribution of actual PM2.5 concentrations and the LISA cluster map for Kampala (Uganda). From the concentration map, it can be observed that a large proportion of PM2.5 data points across Kampala fall within the Unhealthy category according to the revised 24-hour AQI breakpoints. While a few locations recorded values within the Good, Moderate, or Unhealthy for Sensitive Groups categories, these points tend to cluster visually with areas of higher PM2.5 concentrations. The LISA cluster map reveals that most of these observed spatial groupings are statistically non-significant at $p \geq 0.05$, except for the Kyebando area, which emerges as a statistically significant High-High (HH) cluster at $p < 0.05$.

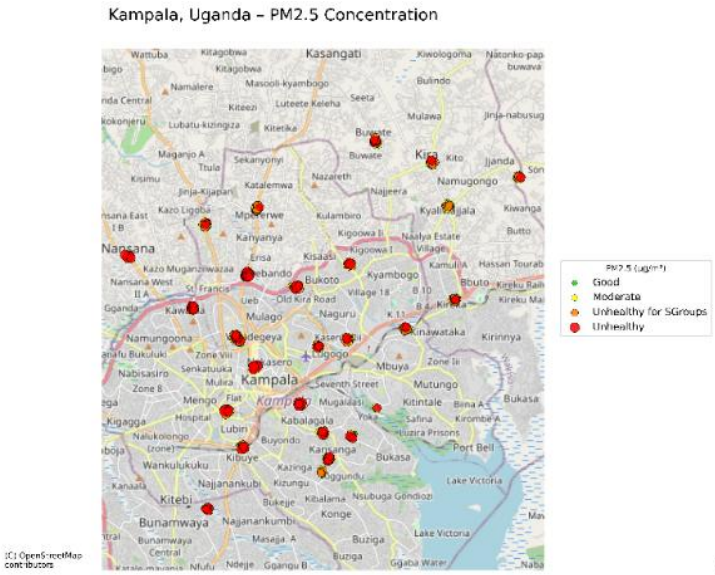

Figure 10a. Kampala, Uganda PM2.5 Concentration

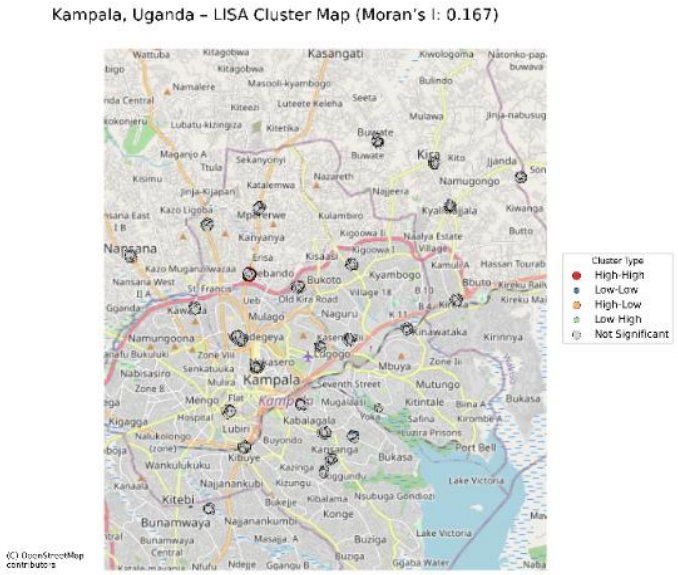

Figure 10b. Kampala, Uganda LISA Cluster Map.

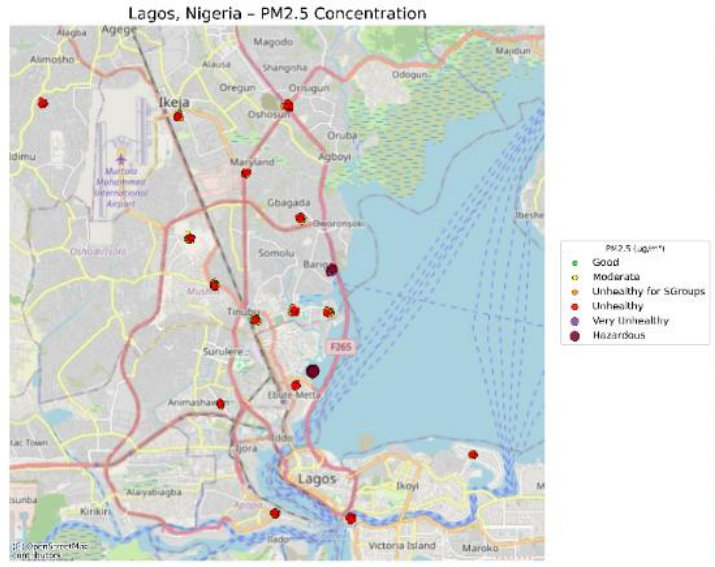

Figure 11a. Lagos, Nigeria PM2.5 Concentration

Figure 11b. Lagos, Nigeria LISA Cluster Map

The possible causes of elevated PM2.5 concentrations in this locality may likely include intense human and vehicular activities, the presence of small-scale informal industries, open waste burning, among others. The computed Global Moran's I statistic for PM2.5 concentrations in Kampala was 0.167, which indicates a weak to moderate positive spatial autocorrelation. This implies that while there is a tendency for similar PM2.5 concentration values to cluster spatially, the overall pattern of clustering is not strongly pronounced across the entire study area.

Figure 11a (left) and 11b (right) show the spatial distribution of actual PM2.5 concentrations and the LISA cluster map for Lagos (Nigeria), based on the observations used in this study. From Figure 11a, it is evident that Lagos exhibits a diverse spatial pattern, with concentration values cutting across all AQI categories. A closer examination of the spatial distribution reveals that areas with particularly high PM2.5 concentrations are concentrated around Ebute-Metta and parts of Bariga. These places are both densely populated and highly industrialised urban neighbourhoods within the Lagos Mainland. These areas are known to be characterised by a combination of intense vehicular traffic, proximity to industrial facilities, port-related activities, and a high

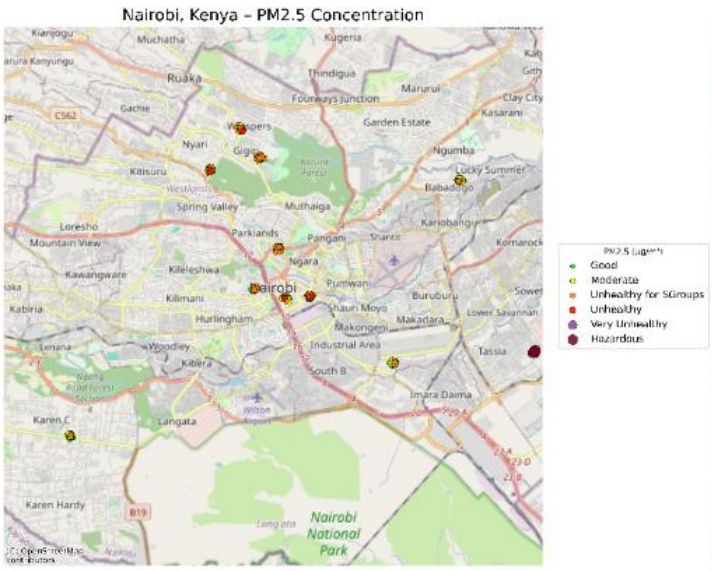

Figure 12a. Nairobi, Kenya PM2.5 Concentration.

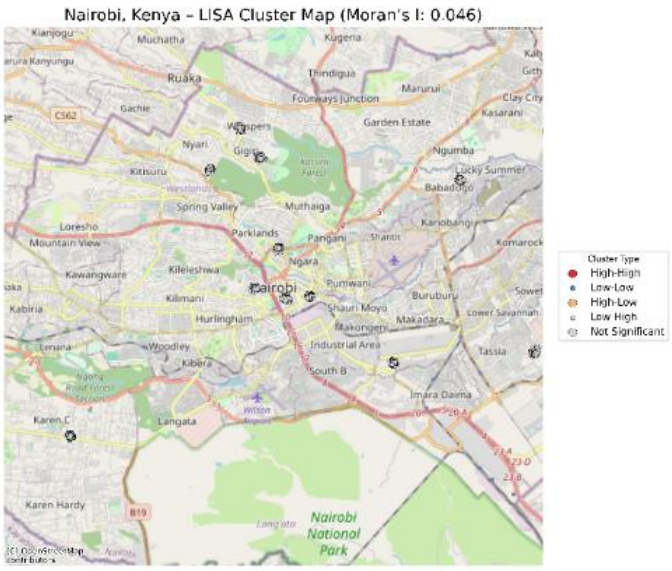

Figure 12b. Nairobi, Kenya LISA Cluster Map.

prevalence of biomass fuel used for domestic and informal industrial purposes, which contributes significantly to PM2.5 emissions. The LISA map indicates that the neighbourhood near Ebute-Metta is identified as a statistically significant High-High (HH) cluster at $p < 0.05$. This implies that this area not only records high PM2.5 concentrations but is also surrounded by other sites with similarly high concentration levels, forming a distinct air pollution hotspot. Global Moran's I statistic for Lagos is computed as 0.686, which implies a strong positive spatial autocorrelation. This relatively high Moran's I value suggests a pronounced tendency for similar PM2.5 concentration values to cluster together spatially across the city. In other words, high PM2.5 values are not randomly dispersed but are spatially concentrated in certain neighbourhoods, most notably around Ebute Metta and adjacent areas.

Figure 12a (left) and 12b (right) present the spatial distribution of actual PM2.5 concentrations and the corresponding LISA cluster map for Nairobi (Kenya). Figure 12a records a full range of AQI categories for Nairobi. A closer inspection of the LISA cluster map reveals that most of the observed data points are not statistically significant at $p \geq 0.05$, which suggests that apparent clusters of similar PM2.5 values on the concentration map may largely occur by chance. However, a notable exception is the Tassia area, which stands out as a statistically significant High-High (HH) cluster. This shows that Tassia records high PM2.5 concentrations and is surrounded by neighbouring sites with similarly high values, significant at $p < 0.05$. Interestingly, this HH cluster appears relatively isolated, as the immediate surrounding areas predominantly fall within the Non-Significant (NS) category on the LISA map. Global Moran's I statistic of 0.046 for PM2.5 concentrations in Nairobi confirms a very weak but positive spatial autocorrelation. This low value suggests that, overall, similar PM2.5 concentrations are only weakly clustered across the city, and most of the observed clustering patterns, apart from areas like Tassia, are statistically indistinguishable from random spatial patterns.

Actual vs predicted PM2.5 by models and feature importance comparison (XGBoost top 20 features)

Figure 13 above presents the relationship between the actual and predicted PM2.5 concentrations for each predictive model used in this study. The scatter plots reveal that the predictions generated by the XGBoost model exhibit a strong positive relationship with the actual values, with a high R-squared value of 0.762. This is closely followed by LightGBM with 0.731, then, Feedforward Neural Network (FNN) model with a moderate relationship value of 0.320. In comparison, the Polynomial Ridge (order 2) and Ridge Regression (order 1) models recorded lower R-squared values of 0.254 and 0.209, respectively. These results show the superior performance of ensemble-based models, particularly XGBoost and LightGBM, in accurately capturing the relationship between actual and predicted PM2.5 concentrations in this multi-city study.

Figure 14 displays the top 20 features influencing PM2.5 prediction as determined by the XGBoost model, with its importance in percentage values. For context, the corresponding importance of these features in the LightGBM model is also shown alongside XGBoost. This

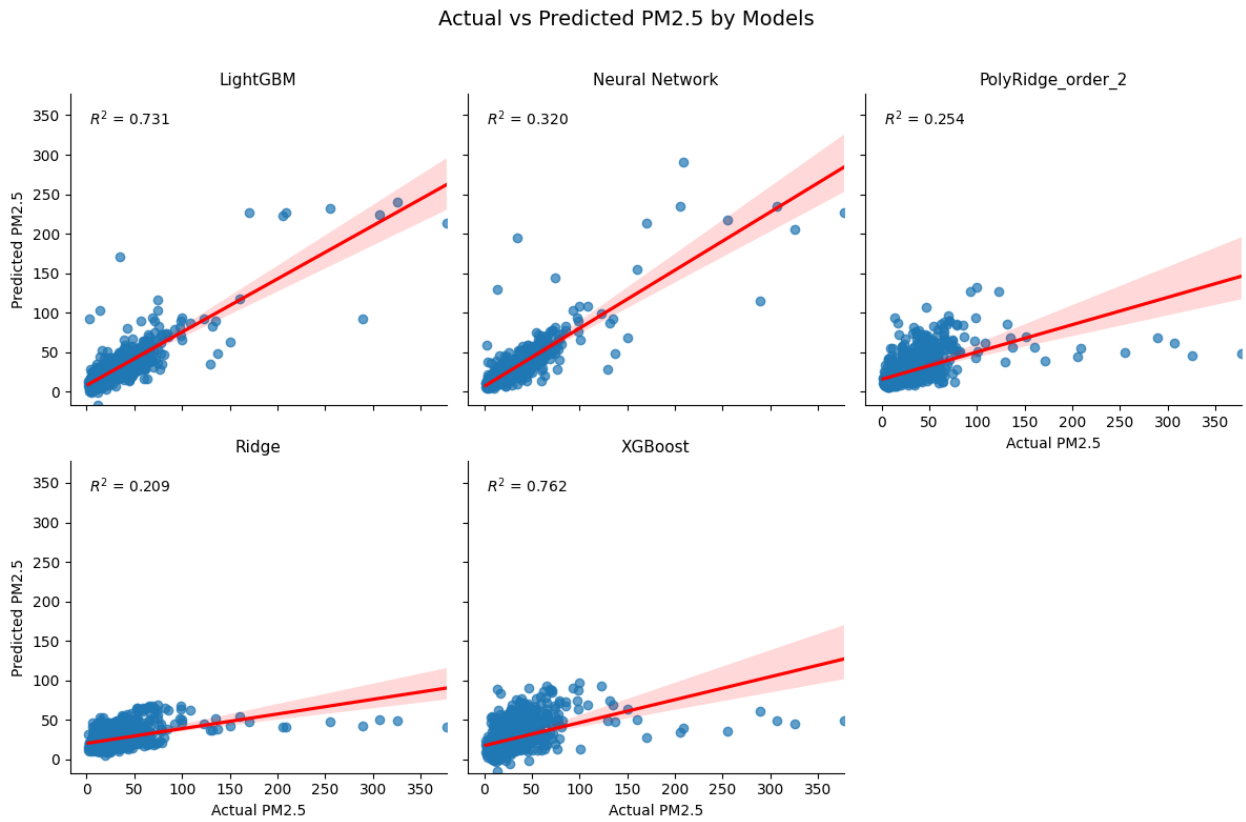

Figure 13. Actual vs predicted PM2.5 by Models.

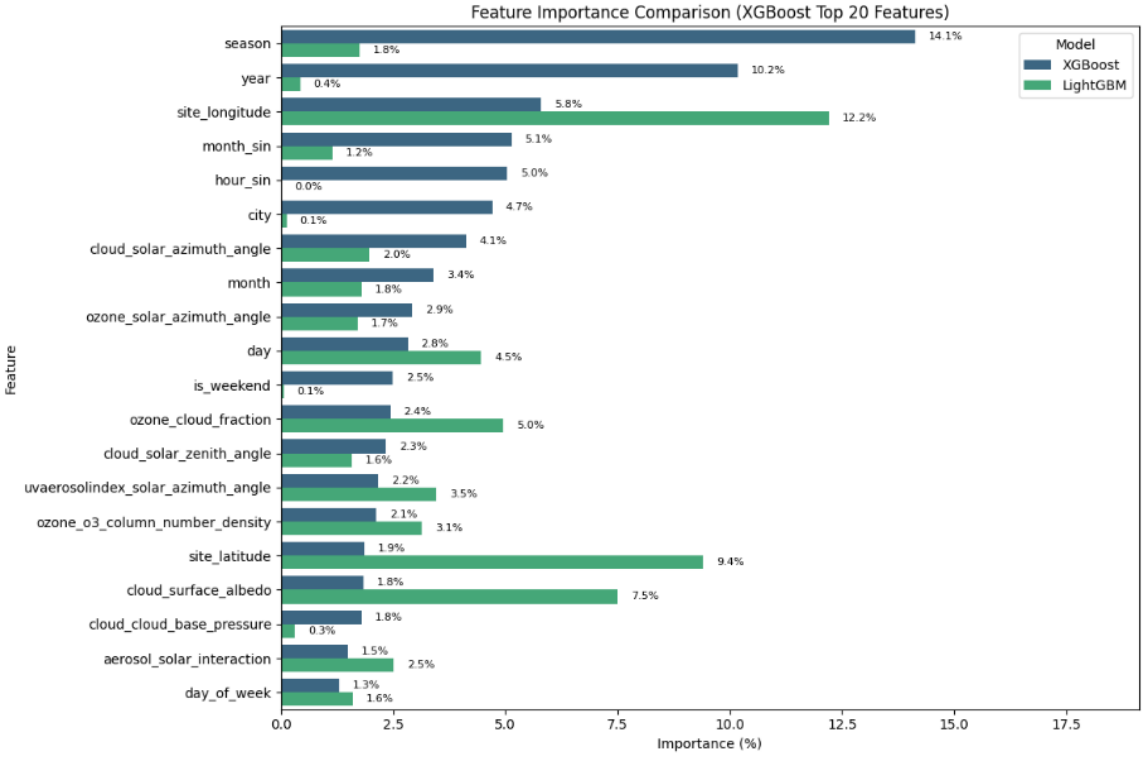

Figure 14. Feature Importance Comparison (XGBoost Top 20 Features).

is done to highlight both models' reliance on temporal, geospatial, and atmospheric indicators, while also revealing differences in their sensitivity to specific environmental factors. XGBoost identifies temporal and seasonal features as the most critical, with season (14.1%) and year (10.2%) ranked highest. These results show the strong seasonal variation in PM2.5 levels, which are likely driven by changes in meteorological conditions, biomass burning, and human activity patterns. Other notable temporal features include month_sin (5.1%), hour_sin (5.0%), and day (2.8%), which together highlight the relevance of periodic cycles such as diurnal and monthly variations. Geospatial features like site_longitude (5.8%) and site_latitude (1.9%) rank slightly lower in XGBoost but are given higher importance in LightGBM as 12.2% and 9.4% respectively. This suggests that LightGBM may be more sensitive to spatial heterogeneity in air pollution sources, such as urban layout and emission hotspots. Among the cloud-related and solar geometry features, variables like cloud_solar_azimuth_angle (4.1%), cloud_surface_albedo (1.8%), and cloud_solar_zenith_ angle (2.3%) are moderately important. These factors influence solar radiation reaching the surface and atmospheric photochemistry, thereby indirectly affecting PM2.5 formation and dispersion. The presence of ozone_solar_azimuth_angle and uvaerosolindex_solar_ azimuth_angle further indicates the role of atmospheric composition and aerosol dynamics, though their importance is comparatively lower. Also, features derived from atmospheric trace gases, such as ozone_o3_column_number_density (2.1%) and ozone_cloud_fraction (2.4%), suggest a weaker but non-negligible relationship between PM2.5 and columnar ozone levels.

DISCUSSION

It is worth noting that a significant factor which contributes to the strong performance of this study's XGBoost model is the extensive feature engineering process undertaken. The inclusion of a custom-derived seasonal feature, which accounted for the unique rainfall and dry season cycles specific to each city, appears to have enhanced the model's ability to capture temporal variations in PM2.5 concentrations. Unlike many existing studies such as Jia *et al*. (2023), Musa *et al.* (2024), Parra *et al.* (2024), and Wang *et al*. (2023) that rely solely on standard meteorological, socio-economic and pollutant variables, this study tailored the feature space to reflect region-specific environmental dynamics to increase predictive accuracy in a heterogeneous multi-city context. The XGBoost model achieved an R^2 of 0.76 and RMSE of 12.01 μg/m³ across four cities. This result aligns closely with the research by Zaman *et al.* (2021), which shows that the Random Forest model achieved an R^2 value which ranges from 0.53–0.76 in estimating PM2.5 concentrations across 65 sites in Malaysia. Moursi *et al*. (2019) found that Extra Trees slightly performed better than Random Forest with an R^2 value above 0.9, while deep learning models like LSTM achieved competitive results. This study's R^2 value is comparatively lower, which is due to the diverse urban settings and complex pollution dynamics across the four cities in this study. A study by Zhang *et al.* (2021) in South Africa's industrialised Highveld region shows that the Random Forest model achieved an R^2 of 0.80 and an RMSE of 9.40 μg/m³, which is marginally higher than the one in this study. This is because Zhang *et al.* (2021) study focuses on a single region with higher data density, while this study covers four cities across four different countries with varying urban, peri-urban, and rural characteristics. A study by Patel *et al*. (2025) across five cities in Maharashtra (India), found that LSTM models performed better than traditional ML models, with an R^2 value ranging from 0.99 to 0.998. This level of accuracy is higher than the one obtained in this study because of a larger dataset of five years used. This is mostly suitable for deep learning models, which remain a limitation in many African contexts.

Conclusion

In this study, scalable machine learning models were employed to estimate PM2.5 concentrations across four African cities using satellite-derived Aerosol Optical Depth (AOD) data, temporal and seasonally engineered features. The results show that the ensemble-based models used performed better than other methods in accurately predicting PM2.5 levels across the cities, with XGBoost achieving the highest performance metrics across all evaluation criteria. The importance of including temporal cycles and localized seasonal characteristics in air quality modeling was also highlighted in this study. Some certain limitations were encountered in this study despite the encouraging performance. Important cloud-related and solar geometry variables were excluded due to excessive missing values, which might have affected the performance of the models. Future studies are encouraged to address this by improving data collection, including more comprehensive atmospheric parameters from Sentinel-5P or other high-resolution earth observation platforms. They can also consider including some socio-economic variables such as population density, traffic volume, industrial activity, and land use characteristics. While this study focused on four cities across four countries, future studies can improve upon this by increasing the number of cities within each country and consider including more countries to strengthen the spatial representativeness and robustness of the models. Models developed in this study are lightweight and computationally efficient, which makes them well-suited for deployment in real-time air quality monitoring systems with an appropriate real-time data pipeline.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ACKNOWLEDGEMENT

The authors are sincerely grateful to Mr. Lukuman Abudulawal for his invaluable assistance, guidance, and mentorship throughout this research work, especially for spearheading the finances of this study. The authors also acknowledge and sincerely appreciate the contributions of peer reviewers for their insightful comments and suggestions that contributed to the quality of this article.

REFERENCES

Jia, H., Zang, S., Zhang, L., Yakovleva, E., Sun, H., & Sun, L. (2023). Spatiotemporal characteristics and socioeconomic factors of PM2.5 heterogeneity in mainland China during the COVID-19 epidemic. *Chemosphere*, *331*, 138785.

Locke, A. V., Heffernan, R. C., McDonagh, G., Yassa, J., & Flaherty, G. T. (2022). Clearing the air: a global health perspective on air pollution. *International Journal of Travel Medicine and Global Health*, *10*(2), 46-49.

Moursi, A. S. A. E. A., Shouman, M., Hemdan, E. E., & El-Fishawy, N. (2019). PM2.5 Concentration Prediction for Air Pollution using Machine Learning Algorithms. *Menoufia Journal of Electronic Engineering Research*, *28*(1), 349-354.

Musa, M., Rahman, P., Saha, S. K., Chen, Z., Ali, M. A. S., & Gao, Y. (2024). Cross-sectional analysis of socioeconomic drivers of PM2.5 pollution in emerging SAARC economies. *Scientific Reports*, *14*(1), 16357.

National Oceanic and Atmospheric Administration (NOAA), United State Department of Commerce (2025). *State of the Science FACT SHEET Air Quality*. Retrieved 7th July, 2025 from https://councilonstrategicrisks.org/wp-content/uploads/2025/03/NOAA-State-of-the-Science-Fact-Sheet-Air-Quality-January-2025.pdf.

Nguyen, A. T., Pham, D. H., Oo, B. L., Ahn, Y., & Lim, B. T. H. (2024). Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. *Journal of Big Data*, *11*(1), 71.

Panaite, F. A., Rus, C., Leba, M., Ionica, A. C., & Windisch, M. (2024). Enhancing air-quality predictions on university campuses: A machine-learning approach to PM2.5 forecasting at the University of Petroșani. *Sustainability*, *16*(17), 7854.

Parra, J. C., Gómez, M., Salas, H. D., Botero, B. A., Piñeros, J. G., Tavera, J., & Velásquez, M. P. (2024). Linking meteorological variables and particulate matter PM2.5 in the Aburrá Valley, Colombia. *Sustainability*, *16*(23), 10250.

Patel, P., Patel, S., Shah, K., Desai, K., Patel, S., Shah, M., & Patel, S. (2025). A systematic study on PM2.5 and PM10 concentration prediction in air pollution using machine learning and deep learning model. *Environmental Chemistry and Ecotoxicology*, *7*, 1401-1415.

United States Environmental Protection Agency (EPA) (2024). *National Ambient Air Quality Standards (NAAQS) for Particle Pollution*. Retrieved from https://www.orcaa.org/epa-updates-particulate-pollution-standards/

Wang, J., Han, J., Li, T., Wu, T., & Fang, C. (2023). Impact analysis of meteorological variables on PM2.5 pollution in the most polluted cities in China. *Heliyon*, *9*(7), e17609.

World Health Organization (WHO) (2021). *Air Pollution*. Retrieved 7th July, 2025 from https://www.who.int/health-topics/air-pollution#tab=tab_1.

Zaman, N. A. F. K., Kanniah, K. D., Kaskaoutis, D. G., & Latif, M. T. (2021). Evaluation of machine learning models for estimating PM2.5 concentrations across Malaysia. *Applied Sciences*, *11*(16), 7326.

Zhang, D., Du, L., Wang, W., Zhu, Q., Bi, J., Scovronick, N., Naidoo, M., Garland, R. M., & Liu, Y. (2021). A machine learning model to estimate ambient PM2.5 concentrations in industrialized highveld region of South Africa. *Remote Sensing of Environment*, *266*, 112713.

Zindi AirQo (2024). Africa Air Quality Competition Dataset. Retrieved 10th July 2024 from https://zindi.africa/competitions/airqo-african-air-quality-prediction-challenge.