Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Do Traffic Congestion Indicators Improve PM2.5 Forecasting? Evidence from Machine Learning Models in Kampala

Hedwig Orieba

oriehedwig@gmail.com

Victoria University

Deborah Mackenzi

Victoria University

Research Article

Keywords: Machine learning, AirQo, PM2.5 Forecasting, Air Pollution, Traffic Congestion, Google Routes API

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11720518/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Do Traffic Congestion Indicators Improve $PM_{2.5}$ Forecasting? Evidence from Machine Learning Models in Kampala

Hedwig Orieba
Faculty of Science and Technology
Victoria University Kampala, Uganda
oriehedwig@gmail.com
ORCID: 0009-0006-3335-6689

Ms. Mackenzi Deborah
Lecturer, Faculty of Science and Technology
Victoria University Kampala, Uganda
dmackenzie@vu.ac.ug
ORCID: 0009-0006-8348-6908

Abstract

Air pollution remains a major environmental and public health challenge in Kampala, Uganda, where $PM_{2.5}$ concentrations frequently exceed World Health Organization (WHO) guidelines values. Although traffic emissions are widely recognized as contributors to urban air pollution, limited empirical evidence exists on whether traffic-congestion indicators improve $PM_{2.5}$ forecasting beyond information provided by meteorological and temporal variables. This study evaluated the predictive contribution of traffic-congestion indicators to $PM_{2.5}$ using machine learning models. $PM_{2.5}$, temperature, and humidity observations were obtained from the AirQo monitoring network, while traffic data were collected along the Jinja Road transport using the Google Routes API. After preprocessing and temporal synchronization, a dataset consisting of 696 hourly observations was used to implement Random Forest and HistGradientBoosting forecasting models. Environmental baseline models were compared with traffic-enhanced models incorporating mean speed, travel duration, and Traffic Performance Index (TPI). Model

performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

Results showed that the best environmental baseline model, HistGradientBoosting, achieved superior performance (MAE=15.28, RMSE=21.40, R^2=0.5752) compared with the best traffic-enhanced model (MAE=23.16, RMSE=31.70, R^2=0.4753). Correlation and feature-importance analysis further indicated that traffic-congestion indicators exhibited weak relationships with PM2.5 concentrations and contributed substantially to less predictive information than temporal and meteorological variables. An additional analysis incorporating lagged $PM_{2.5}$ variables achieved the highest explanatory performance (R^2=0.6245), highlighting the importance of historical pollution measurements in forecasting future $PM_{2.5}$ concentrations.

The findings suggest that, within the study context, corridor-level congestion indicators provided limited incremental forecasting value, while temporal patterns, meteorological conditions, and historical $PM_{2.5}$ measurements were more influential predictors. The study contributes empirical evidence for evaluating predictor relevance in machine-learning based air-quality forecasting and provides insights for the design of efficient forecasting systems in rapidly urbanizing cities.

Keywords:

Machine learning, AirQo, $PM_{2.5}$ Forecasting, Air Pollution, Traffic Congestion, Google Routes API

1.0 Introduction

Worldwide, air pollution remains a major environmental and public health challenge. Fine particulate matter ($PM_{2.5}$) is among the most harmful air pollutants because its small aerodynamic diameter enables it to penetrate deep into the respiratory system and bloodstream, increasing the risk of respiratory diseases, cardiovascular complications, and premature mortality (WHO, 2024; Coker et al.,2021). The burden of $PM_{2.5}$ exposure is particularly severe in rapidly urbanizing cities of low and middle-income countries where population growth, motorization, and urban development often outpace environmental management efforts (Wang et al., 2019).

In Uganda, Kampala continues to experience increasing traffic congestion and elevated $PM_{2.5}$ concentrations. Studies have reported $PM_{2.5}$ levels that frequently exceed WHO guidelines values, while recent evidence indicates that long-term exposure to particulate matter pollution contributes

significantly to mortality and disease burden in the city (Atuyambe et al.,2025; WHO, 2024). These concerns have strengthened interest in data-driven approaches capable of improving air-quality monitoring and forecasting.

Machine-learning approaches have emerged as effective tools for air-quality forecasting because of their ability to model complex non-linear relationships among environmental variables and generate predictions from large datasets (Jordan & Mitchell, 2015; Alaran et al.,2024). Several studies have successfully applied machine-learning techniques to predict $PM_{2.5}$ concentrations using a combination of meteorological variables, temporal indicators, and historical pollution measurement (Alaran et al., 2024; Wabinyai et al.,2025). At the same time, advances in digital navigation platforms and location-based services have increased access to traffic-congestion data, creating new opportunities to integrate transportation information into environment forecasting systems.

Traffic emissions are widely recognized as an important contributor to urban air pollution. Consequently, variables such as vehicle speed, travel duration, and congestion indices are increasingly being incorporated into forecasting models. However, despite the growing availability of traffic data and their theoretical relationship with particulate emissions, limited evidence exists regarding whether traffic-congestion indicators improve $PM_{2.5}$ forecasting performance beyond what was already achieved using meteorological and temporal variables, particularly within African urban environments.

This study evaluates the contribution of traffic-congestion indicators collected along the Jinja Road transport corridor to $PM_{2.5}$ forecasting in Kampala, Uganda. Traffic data obtained through the Google Routes API were integrated with $PM_{2.5}$, temperature, and humidity observations collected through the AirQo monitoring network. Random Forest and HistGradientBoosting regressors were subsequently used to compare environmental baseline models with traffic-enhanced forecasting models and assess the predictive value of congestion indicators within machine-learning-based air-quality forecasting systems.

2.0 Literature Review

Air pollution and traffic congestion continue to affect the quality of life in many rapidly growing cities. As environmental monitoring systems and machine learning techniques become more

widely used, researchers are exploring ways of predicting air-quality conditions. This section reviews previous studies related to $PM_{2.5}$ pollution, traffic activity, and air-quality forecasting to provide the context for the present study.

2.1 $PM_{2.5}$ Pollution and Urban Air Quality

Fine particulate matter ($PM_{2.5}$) is widely recognized as one of the most harmful air pollutants because its small aerodynamic diameter allows it to penetrate deep into the respiratory system and enter the bloodstream, contributing to respiratory illness, cardiovascular diseases, and premature mortality (WHO, 2024; Coker et al., 2021). Globally, $PM_{2.5}$ exposure remains a leading environmental risk factor, particularly in rapidly urbanizing regions where industrialization, transportation activities, and population growth continue to intensify pressure on urban environments (Wang et al., 2019).

Urban transportation systems represent an important source of $PM_{2.5}$ emissions due to fuel combustion, vehicle exhaust, brake wear, and road dust resuspension. In many developing cities, increasing vehicle ownership, aging vehicles fleets, and persistent traffic congestion further worsens air-quality challenges (Sulaimon et al., 2024; Alaran et al., 2024). Consequently, understanding the factors that influence $PM_{2.5}$ concentrations has become an important research priority for environmental monitoring, policy development, and public-health protection.

In Uganda, studies conducted through the AirQo network have shown that $PM_{2.5}$ concentrations frequently exceed World Health Organization guideline values, raising concerns regarding long-term public exposure to air pollution (Atuyambe et al., 2025; Bainomugisha et al., 2023). These findings have increased the interest in air-quality forecasting systems capable of providing early warning information and supporting evidence-based environmental management.

2.2 Traffic Congestion and $PM_{2.5}$ Concentrations

Traffic emissions are widely recognized as an important contributor to urban air pollution, particularly in rapidly urbanizing cities where increasing vehicle ownership places significant pressure on transportation infrastructure. Previous studies have reported associations between traffic activity and elevated particulate matter concentrations, suggesting that transportation systems play a significant role in shaping urban air quality (Sulaimon et al.,2024).

Advances in geospatial technologies and navigation services have increased access to traffic-related information such as travel duration, vehicle speed and congestion levels. These indicators

have attracted growing interest in air-quality forecasting research because transportation emissions are widely recognized as contributors to urban air pollution. Previous studies have investigated the effect of incorporating traffic datasets into machine-learning based air-quality prediction models, highlighting the need to better understand the contribution of traffic information to forecasting performance (Sulaimon et al.,2024).

However, the relationship between traffic activity and $PM_{2.5}$ concentrations is influenced by additional factors, including meteorological conditions such as temperature, humidity, and wind patterns, which affect pollutant formation, transport, and dispersion (Alaran et al.,2024). Consequently, the contribution of traffic related variables to observed $PM_{2.5}$ concentrations may vary across locations and forecasting contexts. While traffic activity is widely recognized as an important contributor to urban air pollution, the extent to which traffic congestion indicators provide additional predictive information beyond meteorological and temporal variables remains less clearly understood.

2.3 Machine Learning Models for $PM_{2.5}$ Forecasting

Machine learning techniques have become increasingly important in air-quality forecasting because they can model complex and nonlinear relationships among environmental variables. Unlike conventional statistical approaches, machine learning algorithms are capable of learning hidden patterns from large datasets, enabling the use of multiple predictors to forecast $PM_{2.5}$ concentrations (Jordan & Mitchell, 2015).

Several machine learning approaches have been applied to $PM_{2.5}$ forecasting, including Support Vector Machines, Random Forests, Gradient Boosting Machines, and Neural Network models (Alaran et al.,2024). These approaches have demonstrated promising performance in predicting $PM_{2.5}$ concentrations using meteorological observations, temporal indicators, and historical pollution measurements.

2.4 Limitations of Existing Forecasting

Existing $PM_{2.5}$ forecasting studies have demonstrated the value of meteorological variables, temporal indicators and historical pollution measurements in predicting air-quality conditions (Alaran et al.,2024; Wabinyai et al., 2025).

The growing availability of environmental observations from monitoring initiatives such as AirQo, together with traffic information obtained through digital navigation platforms like the google

maps services, has created new opportunities for integrating transportation and environmental datasets within air-quality forecasting systems. However much of the existing literature has focused on forecasting accuracy, sensor calibration, land-use modelling and environmental monitoring (Bainomugisha et al.,2023; Coker et al.,2021). Comparatively less attention has been directed toward evaluating the incremental predictive contribution of traffic-congestion indicators. As a result, limited evidence exists regarding whether congestion-related variable provide meaningful improvements in $PM_{2.5}$ forecasting beyond what is already achieved using meteorological and temporal predictors, particularly within African urban environments where transportation systems, environmental conditions, and data availability differ substantially from many of the contexts represented in existing forecasting literature.

2.5 Positioning of the Present Study

Previous studies have demonstrated the importance of the meteorological conditions, historical pollution measurements, and transportation activity in understanding urban air quality. Machine learning approaches have also shown considerable promise in forecasting $PM_{2.5}$ concentrations across diverse urban environments. However, relatively limited attention has been directed towards evaluating whether traffic-congestion indicators provide meaningful additional predictive information once meteorological and temporal variability have already been incorporated into forecasting models.

In contrast to studies primarily focused on developing or comparing forecasting algorithms, this study adopts a predictor-evaluation perspective. The study integrates traffic-congestion data obtained through the Google Routes API with environmental observations collected through the AirQo monitoring network and compares environmental baseline models with traffic-enhanced machine learning models. By accessing the incremental forecasting value of traffic-congestion indicators within Kampala's urban context, the study contributes evidence that can inform the design of efficient, scalable, and data-driven air quality forecasting systems.

3.0 Methodology

The structured framework adopted for this study to evaluate the contribution of traffic congestion indicators to $PM_{2.5}$ forecasting in Kampala, Uganda identifies an understanding of the study area, data sources, preprocessing procedures, feature engineering process, model development strategy,

and the evaluation metrics used to compare environmental-baseline models with traffic-enhanced forecasting models.

3.1 Study Area and Data Sources

The study was conducted in Kampala, Uganda, with a particular focus on the Jinja Road transport corridor. Jinja Road is one of the city's major transportation routes that experiences substantial traffic congestion during peak travel periods. The corridor was selected because of its high traffic volumes and its proximity to locations covered by the AirQo air-quality monitoring network, thereby facilitating the integration of transportation and environmental observations.

The study utilized two primary data sources. Air-quality and meteorological data were obtained from the AirQo monitoring network, which provides continuous measurements of $PM_{2.5}$ concentrations, temperature, and humidity across Kampala. Traffic congestion data were collected using the Google Routes API, which provides route-based transportation information including travel duration, vehicle speed, and congestion characteristics.

The integration of these datasets enabled the development of a merged dataset containing both environmental and transportation-related variables. The geographical context of this study is illustrated in Fig.1 below:

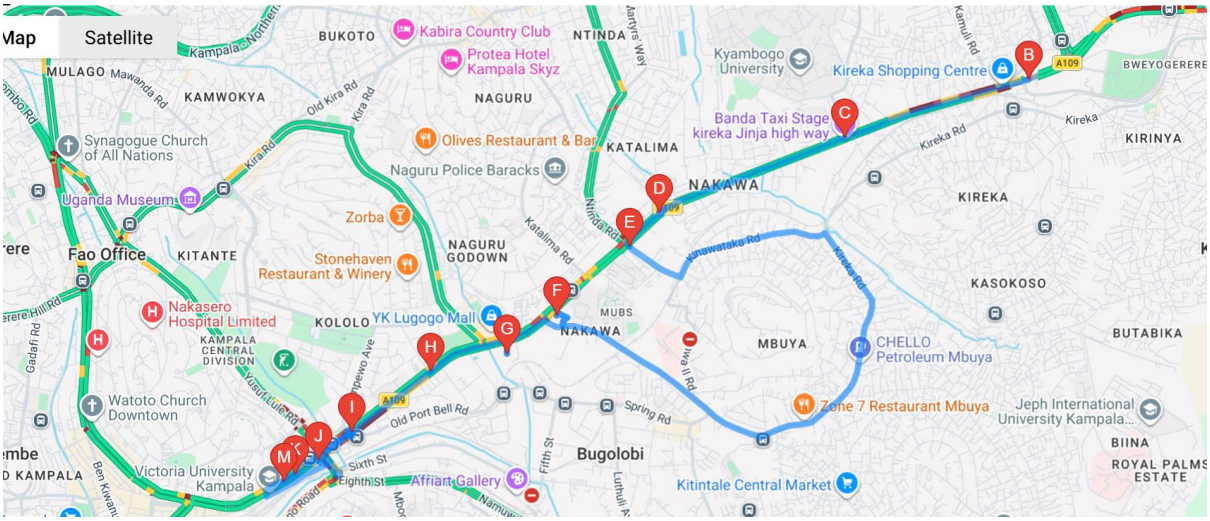


Fig. 1 Jinja Road Corridor of Study

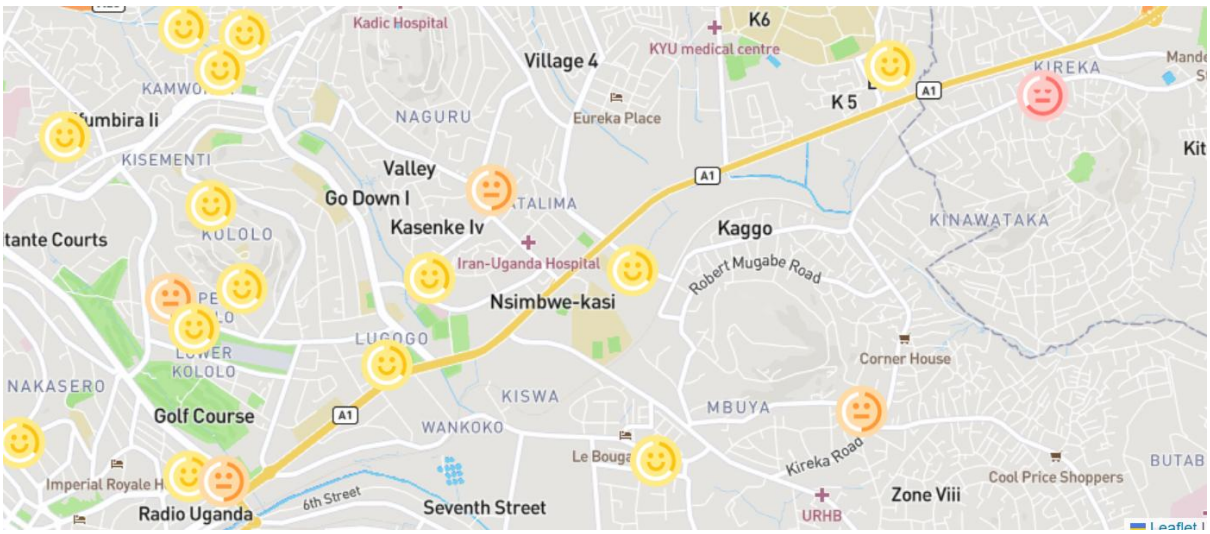


Fig. 2 AirQo Sensor Distribution Network around study corridor

3.2 Data Preparation and Integration

Data preparation involved the collection, synchronization, and integration of environmental and traffic observations obtained from the AirQo monitoring network and Google Routes API respectively. The collected datasets were first subjected to quality checks to identify missing observations, inconsistencies and any duplicate records.

Environmental and traffic data were subsequently synchronized using timestamp information to ensure temporal alignment between observations. Following synchronization, the datasets were aggregated at an hourly rate resolution to establish a common temporal framework suitable for forecasting analysis. Records containing incomplete observations were removed to improve data consistency and model reliability. The resulting integrated dataset combined $PM_{2.5}$ concentrations, meteorological variables, temporal attributes, and traffic-congestion indicators within a single analytical framework. Following preprocessing, a final dataset consisting of 696 complete hourly observations was obtained for model development and evaluation. The overall data integration workflow adopted in the study is illustrated in Fig. 3

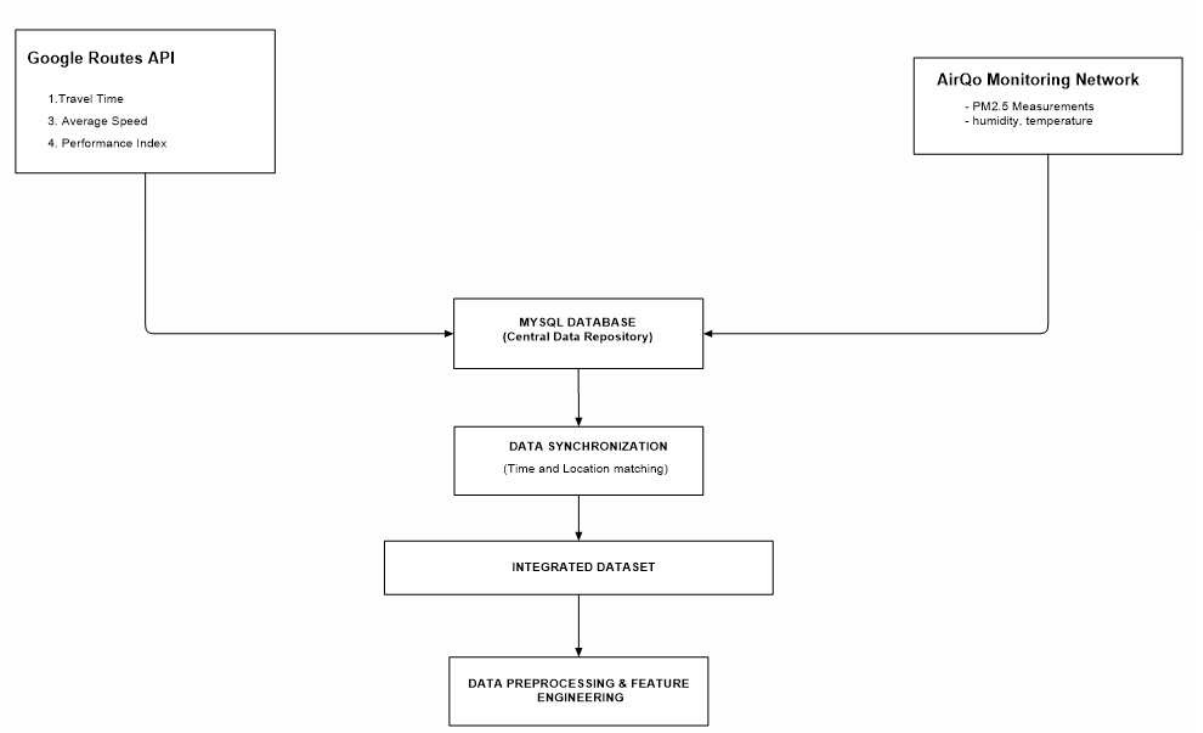


Fig. 3 Data Collection and Integration Workflow

The real-time data ingestion pipeline for traffic data using the google routes API service is as shown in Fig. 4:

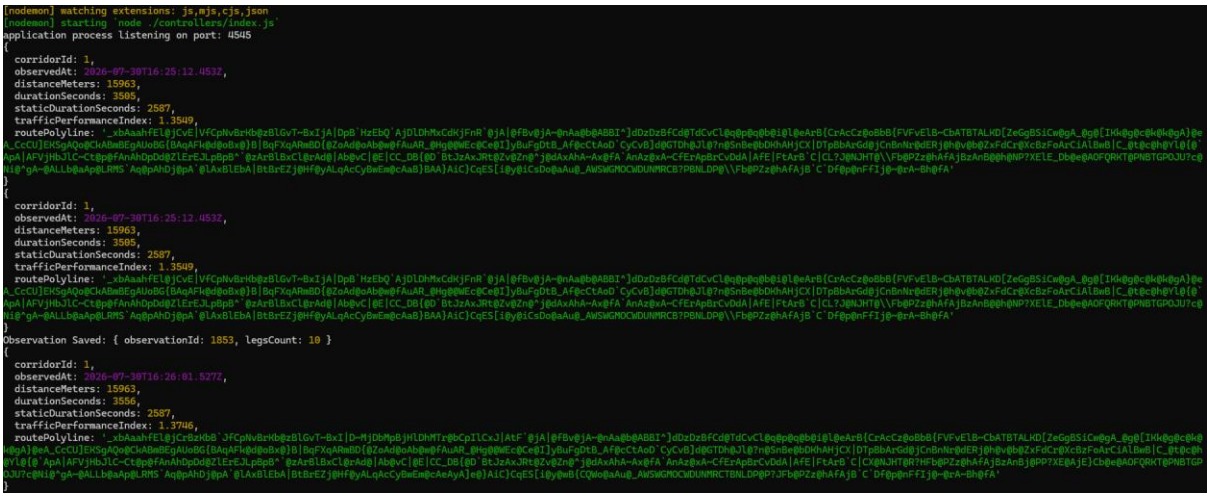


Fig. 4 Corridor Traffic Data Ingestion

3.3 Feature Engineering

Feature engineering was performed to transform the collected observations into predictor variables suitable for machine learning modelling. Three categories of features were developed: environmental variables, temporal variables, and traffic-congestion variables.

Environmental variables included $PM_{2.5}$ concentration measurements together with temperature and humidity observations obtained from AirQo sensors. Temporal variables were derived from timestamps and included hour of day, day of week, and weekend indicators to capture recurring temporal patterns associated with human activities and environmental conditions.

Traffic-congestion variables were derived from Google Routes API observations and included mean speed, travel duration, and the Traffic Performance Index (TPI). These indicators were selected because they provide quantitative representation of traffic conditions along the study corridor. The resultant feature set enabled the evaluation of whether traffic conditions indicators contribute additional forecasting information beyond environmental and temporal variables. Feature engineering was further extended through the creation of lagged $PM_{2.5}$ variables to capture temporal persistence in air-quality patterns. Three lagged variables were generated representing $PM_{2.5}$ concentrations observed one hour, three hours, and twenty-four hours before the prediction period. These variables were included in an additional exploratory analysis to evaluate the predictive contribution of historical pollutant concentrations when combined with environmental, temporal, and traffic-related predictors.

3.4 Model Development

The study adopted a comparative modelling approach using Random Forest and HistGradientBoosting regressors. To evaluate the contribution of traffic-congestion indicators, four forecasting models were implemented and pairwise grouped into environmental baseline models and traffic-enhanced models. Environmental baseline models were implemented using meteorological variables and temporal attributes. Historical $PM_{2.5}$ measurements represented by lagged $PM_{2.5}$ features were evaluated separately in an additional predictor-evaluation experiment. These models established a baseline for comparison against traffic-enhanced models

a) A1: Random Forest (Environmental and Temporal Variables)
b) A2: HistGradientBoosting (Environmental and Temporal Variables)

Traffic-enhanced models incorporated traffic-congestion indicators in addition to environmental and temporal variables.

a) B1: Random Forest (Environmental, Temporal Variables and Traffic Variables)
b) B2: HistGradientBoosting (Environmental, Temporal Variables and Traffic Variables)

This experimental design enabled a direct comparison between models constructed without traffic information and models incorporating congestion indicators. The objective was to determine whether the inclusion of traffic-congestion variables improves forecasting performance.

Category	Algorithm	Variable Used	Purpose
A1	Random Forest Regressor	Temperature, Humidity, Hour of Day, Day of Week, Weekend Indicator	Environmental Baseline
A2	HistGradientBoostingRegressor	Temperature, Humidity, Hour of Day, Day of Week, Weekend Indicator	Environmental Baseline
B1	Random Forest Regressor	Environmental Variables, mean speed, mean duration, and mean TPI	Traffic-enhanced model
B2	HistGradientBoostingRegressor	Environmental Variables, mean speed, mean duration, and mean TPI	Traffic-enhanced model.

Table 1 Modelling Framework used to evaluate the contribution of traffic congestion indicators to $PM_{2.5}$ forecasting

3.4.1 Additional Predictor Evaluation Analysis

Following the primary modelling experiments, an additional HistGradientBoosting model was implemented to investigate the combined contribution of traffic congestion indicators and historical $PM_{2.5}$ measurements. This model incorporated meteorological variables, temporal attributes, lagged $PM_{2.5}$ measurements (PM2.5 Lag1, PM2.5 Lag3 and PM2.5 Lag24), and traffic-congestion indicators within a single forecasting framework. A correlation analysis was also performed to examine the relationship among $PM_{2.5}$ concentrations, meteorological variables, temporal attributes, traffic-congestion indicators, and lagged $PM_{2.5}$ measurements. The objective of this additional analysis was to provide further insight into the relative importance of the predictor groups and help explain differences observed in forecasting performance.

3.5 Model Evaluation

Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics were selected because they provide complementary assessments of forecasting accuracy and explanatory performance. MAE measures the average magnitude of prediction errors, while RMSE places greater emphasis penalizing any large forecasting errors. The coefficient of determination (R^2) measures the proportion of variation in $PM_{2.5}$ concentrations explained by a forecasting model.

The performance of the environmental baseline models was compared with that of the traffic-enhanced models to evaluate the incremental forecasting contribution of traffic-congestion indicators. Lower MAE and RMSE values indicate better predictive accuracy, whereas higher ^{R2} values indicate stronger explanatory performance. The methodological workflow adopted in this study therefore consisted of five sequential stages: data acquisition, data integration, feature engineering, model development, and performance evaluation. This framework enabled a systematic assessment of whether traffic-congestion indicators provide meaningful predictive value beyond meteorological and temporal variables in $PM_{2.5}$ forecasting.

4.0 Results

This section presents the results of the forecasting experiments and supporting analyses. Model performance is first evaluated using environmental baseline and traffic enhanced forecasting models, followed by comparative performance analysis, correlation analysis, feature importance assessment, and an additional predictor-evaluation experiment incorporating historical $PM_{2.5}$ measurements.

4.1 Dataset Characteristics

Environmental observations obtained from AirQo monitoring network were integrated with traffic-congestion data collected through the Google Routes API. Following preprocessing, synchronization, and quality-control procedures, a final dataset consisting of 696 complete hourly observations was obtained for model development and evaluation.

The dataset contained $PM_{2.5}$ concentrations together with meteorological variables, temporal attributes and traffic-congestion indicators. The resulting observations captured variations in pollution levels, weather conditions, and traffic patterns across the study period, providing

sufficient information for evaluating the contribution of traffic-congestion indicators to $PM_{2.5}$ forecasting.

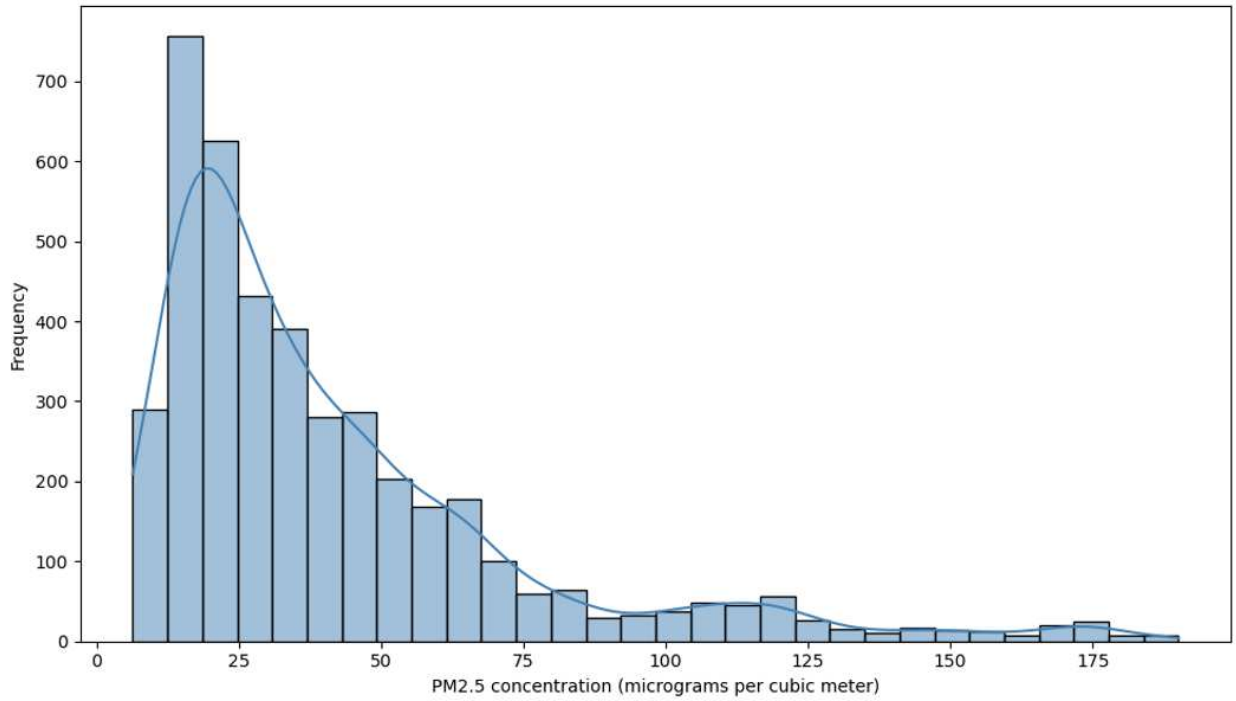


Fig. 5 Distribution of $PM_{2.5}$ Concentrations within the dataset

The temporal behavior of $PM_{2.5}$ concentrations through the study period is as illustrated below:

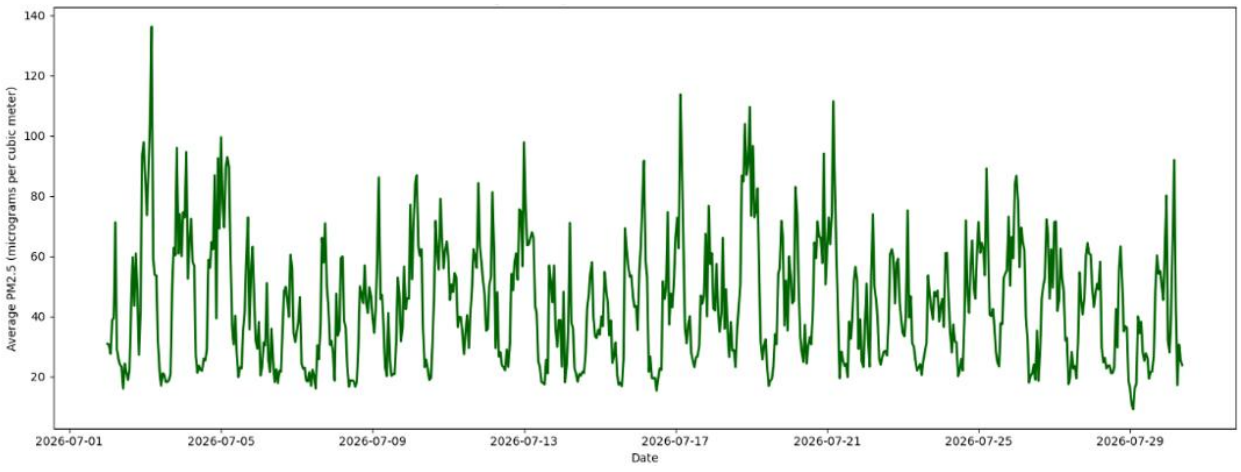


Fig. 6 Average Hourly $PM_{2.5}$ Concentrations Over Time

4.2 Performance of Environmental Baseline Models

Environmental baseline models were implemented using meteorological variables, temporal attributes, and historical $PM_{2.5}$ measurements. Random Forest (A1) and HistGradientBoosting (A2) were trained using identical datasets and performance metrics.

Model	MAE	RMSE	R^2
A1 Random Forest	17.13	24.36	0.4493
A2 HistGradientBoosting	15.28	21.40	0.5752

Table 2 Performance of Environmental Baseline Models

Among the baseline models, HistGradientBoosting achieved strongest overall performance with an MAE of 15.28, RMSE of 21.40, and R^2 value of 0.5752. These results indicate that meteorological, temporal, and historical $PM_{2.5}$ variables provided substantial predictive information for forecasting $PM_{2.5}$ concentrations.

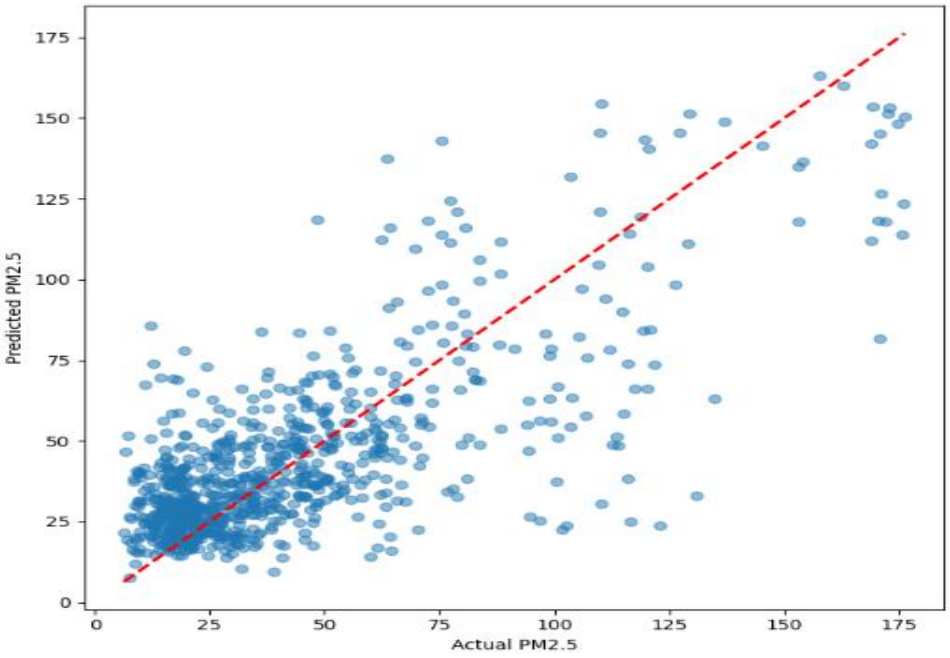


Fig. 7 Actual vs Predicted $PM_{2.5}$ concentrations for the HistGradientBoosting environmental baseline model

4.3 Performance of Traffic-Enhanced Models

Traffic enhanced models were developed by incorporating Mean Speed, Mean Duration, and Traffic Performance Index (TPI) alongside meteorological and temporal variables. Random Forest (B1) and HistGradientBoosting(B2) were subsequently trained and evaluated.

Model	**MAE**	**RMSE**	**R^2**
B1 Random Forest	24.97	35.29	0.3669
B2 HistGradientBoosting	23.16	31.70	0.4753

Table 3 Performance of Traffic-Enhanced Models

Among the traffic-enhanced models, HistGradientBoosting achieved strongest performance. However, both traffic-enhanced models produced higher prediction errors and lower explanatory power than their corresponding environmental baseline models.

These results suggest that the evaluated traffic-congestion indicators provided limited additional predictive information with the study dataset.

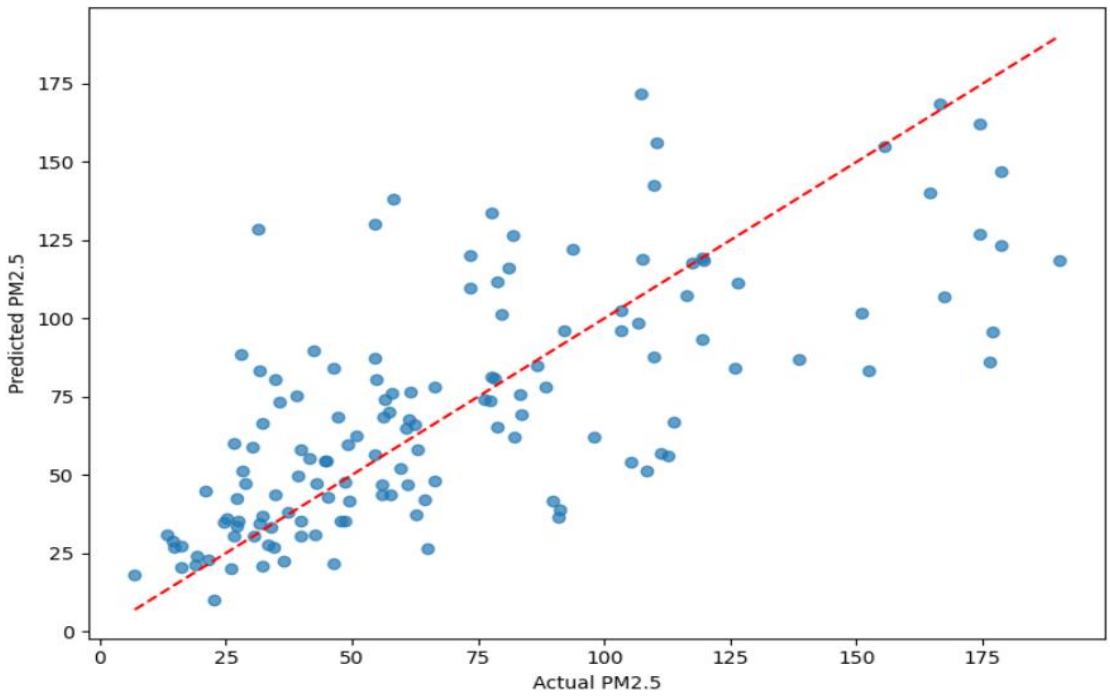


Fig. 8 Actual vs Predicted $PM_{2.5}$ concentrations for the HistGradientBoosting Traffic-enhanced model

4.4 Comparative Analysis

To assess the contribution of traffic-congestion indicators, the performance of environmental baseline models was compared directly with that of traffic-enhanced models.

Model	Variable Category	MAE	RMSE	R^2
A1	Environmental Baseline	17.13	24.36	0.4493
A2	Environmental Baseline	15.28	21.40	0.5752
B1	Traffic Enhanced	24.97	35.29	0.3669
B2	Traffic Enhanced	23.16	31.70	0.4753

Table 4 Comparative Model Performance

The best performing environmental baseline model (A2) outperformed the best performing traffic enhanced model (B2) across all evaluation metrics. The inclusion of traffic-congestion indicators increased prediction errors while reducing explanatory performance. Specifically, R^2 declined from 0.5752 to 0.4753, while both MAE and RMSE increased substantially.

These findings suggest that the evaluated traffic variables did not improve forecasting performance beyond the information already captured by meteorological and temporal variables.

4.5 Correlation Analysis

To better understand the contribution of individual predictor groups, a Pearson correlation analysis was conducted using $PM_{2.5}$ concentrations, meteorological variables, temporal attributes, and traffic congestion indicators.

#	Variable	Correlation with $PM_{2.5}$
1	Hour of Day	0.344
2	Temperature	-0,214
3	Day of Week	0.168
4	Weekend Indicator	0.124
5	Humidity	0.108
6	Mean Speed	0.023
7	Mean Duration	-0.013
8	Traffic Performance Index	-0.013

Table 5 Correlation of Predictor Variables with $PM_{2.5}$

The results revealed relatively weak associations between $PM_{2.5}$ concentrations and the evaluated traffic congestion indicators. Mean Speed exhibited a correlation coefficient of only 0.023, while

Mean Duration and Traffic Performance Index recorded correlation coefficient of approximately -0.013. These values indicate negligible linear relationships between traffic-congestion indicators and $PM_{2.5}$ concentrations within the study dataset. By contrast, temporal and meteorological variables demonstrated comparatively stronger relationships with $PM_{2.5}$, particularly Hour of Day (0.344) and temperature (-0.214).

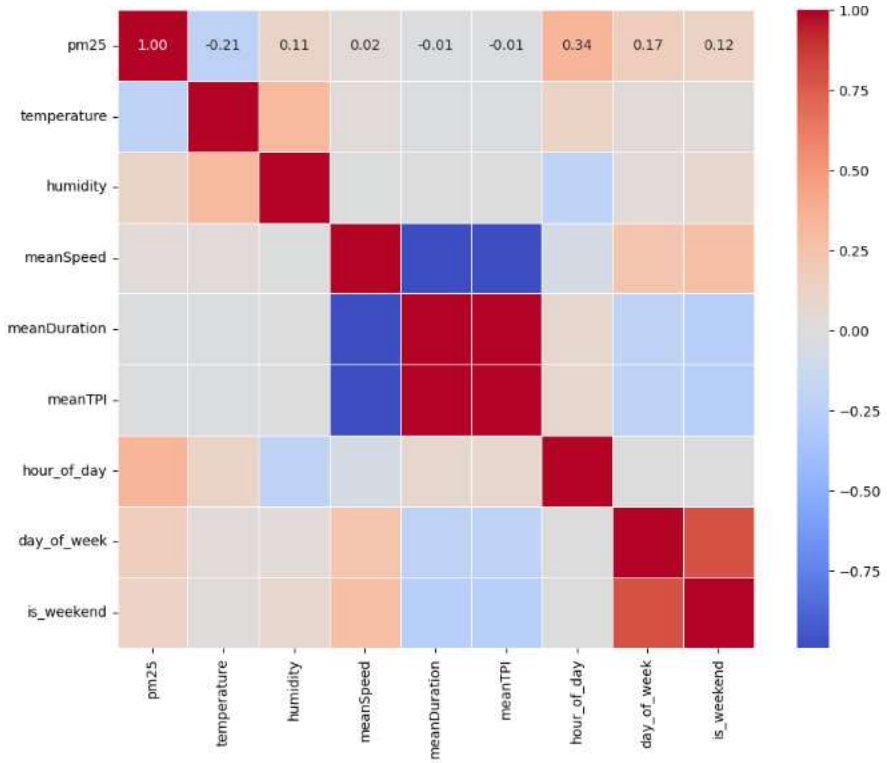


Fig. 9 Correlation Matrix of $PM_{2.5}$ and Predictor Variables

The analysis further revealed extremely strong correlations among the traffic variables themselves. Mean Duration and traffic Performance Index exhibited an almost perfect positive correlation (r=1.000), while both variables displayed near-perfect negative correlations with mean speed (r approx. -0.991). This suggests that the evaluated traffic indicators captured highly similar traffic conditions and therefore contributed limited unique information to the forecasting models.

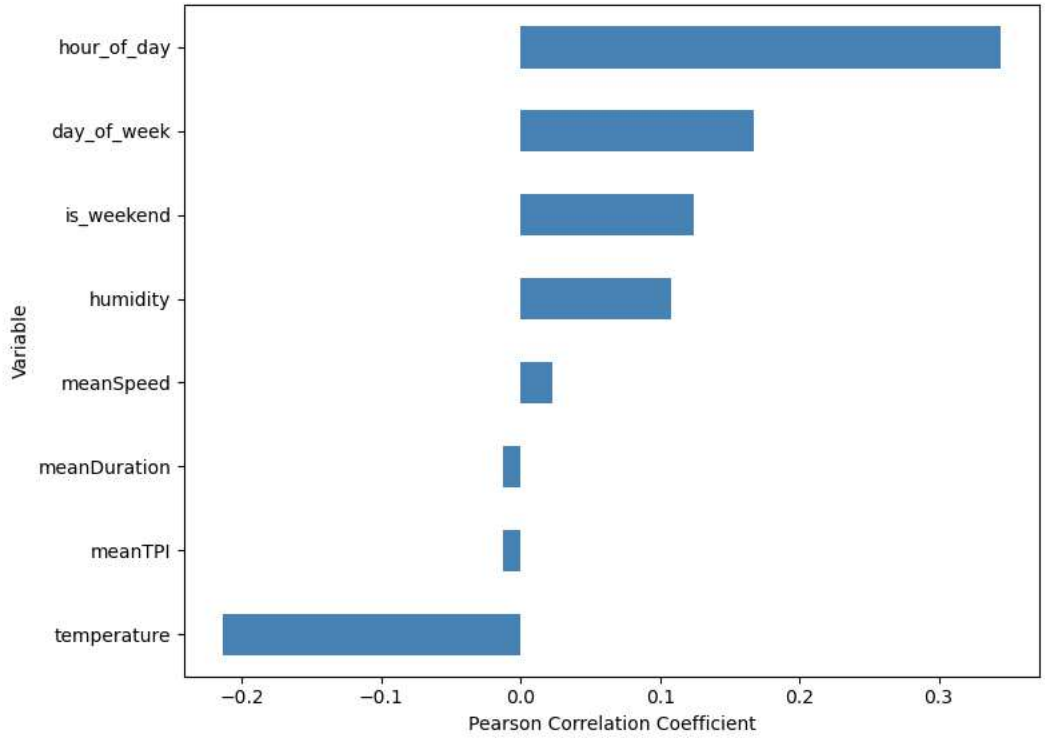


Fig. 10 Correlation Coefficients between $PM_{2.5}$ Concentrations and Predictor Variables

4.6 Feature Importance Analysis

Feature importance analysis was conducted to examine the relative contribution of individual variables to forecasting performance. The analysis revealed that temporal and meteorological variables contributed more significantly to forecasting performance than the evaluated traffic-congestion indicators. Hour of day consistently emerged among the most influential predictors, while humidity and temperature also contributed meaningfully to model predictions.

By contrast, Mean Speed, Mean Duration, and Traffic Performance Index demonstrated comparatively low importance scores, reinforcing the results in the correlation analysis.

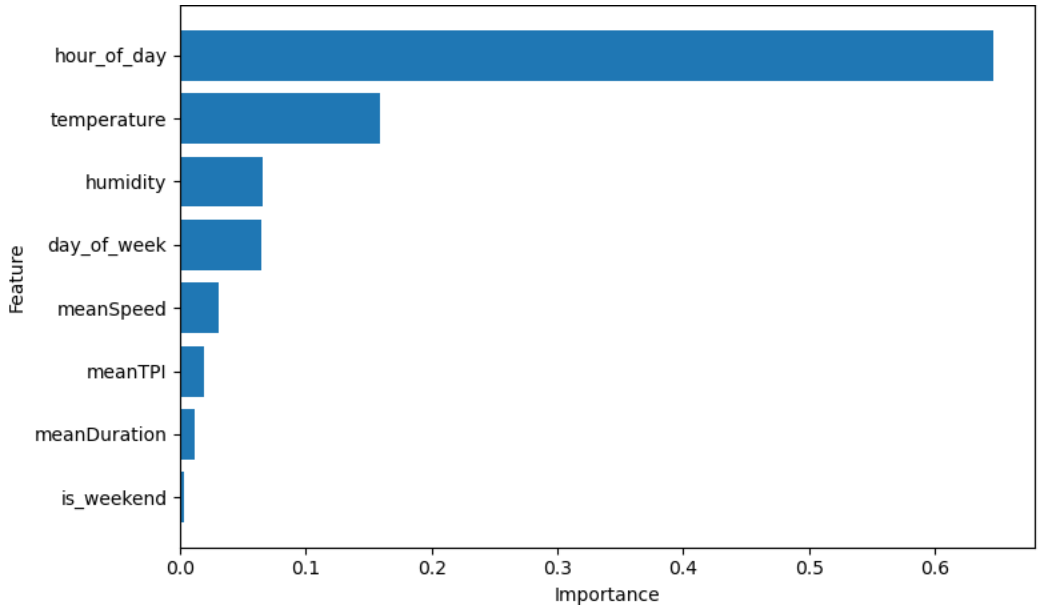


Fig. 11 Relative importance of environmental, temporal, and traffic-congestion variables.

The findings indicate that temporal activity patterns and meteorological conditions provided substantially more predictive information than the evaluated traffic-congestion indicators.

4.7 Additional Analysis Using Historical $PM_{2.5}$ Predictors

Environmental observations obtained from AirQo monitoring network were integrated with traffic-congestion data collected through Google Routes API.

Model	**MAE**	**RMSE**	**R^2**
HistGradientBoosting + LaggedPM2.5 +Traffic Variables (AllCombinedPredictorEvaluation)	22.10	31.19	0.6245

Table 6 Performance of Combined Forecasting Model

The combined model achieved an R^2 value of 0.6245, exceeding the performance of both the environmental baseline model (R^2=0.5752) and the original traffic-enhanced model (R^2=0.4753). This improvement indicates that historical $PM_{2.5}$ measurements contributed substantial predictive information.

Correlation analysis support this observation. $PM_{2.5}$ Lag1 exhibited the strongest relationship between with current $PM_{2.5}$ concentrations (r = 0.744), followed by $PM_{2.5}$ Lag24 (r = 0.570) and

$PM_{2.5}$ Lag3 (r = 0.379). These values were considerably higher than those observed for any of the evaluated traffic-congestion indicators.

Variable	**Correlation with** $PM_{2.5}$
PM2.5 Lag1	0.744
PM2.5 Lag24	0.570
PM2.5 Lag3	0.379

Table 7 Correlations Between $PM_{2.5}$ and Lagged $PM_{2.5}$ variables

Feature-importance analysis similarly showed Hour of Day and PM2.5 Lag1 were the dominant predictors, while traffic-congestion indicators contributed very little to forecasting performance.

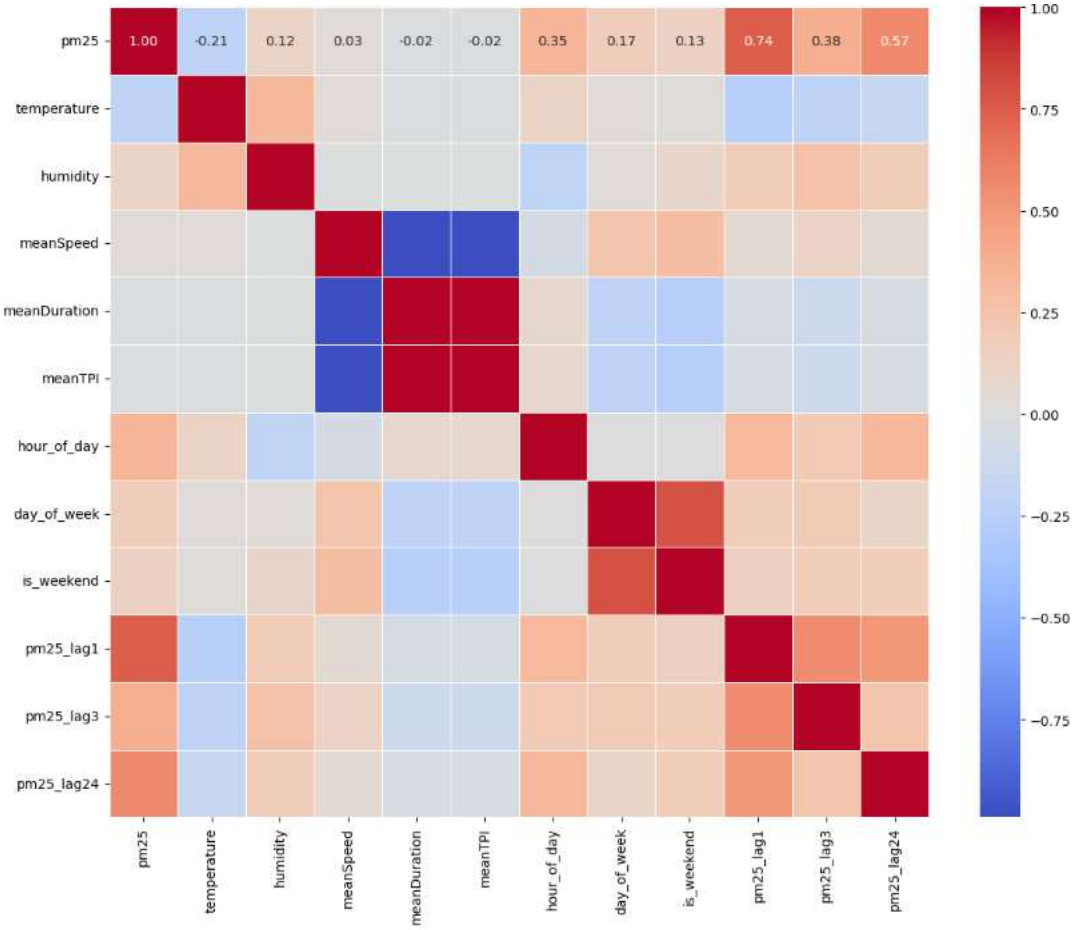


Fig. 12 Correlation matrix including lagged $PM_{2.5}$, meteorological, temporal, and traffic variables

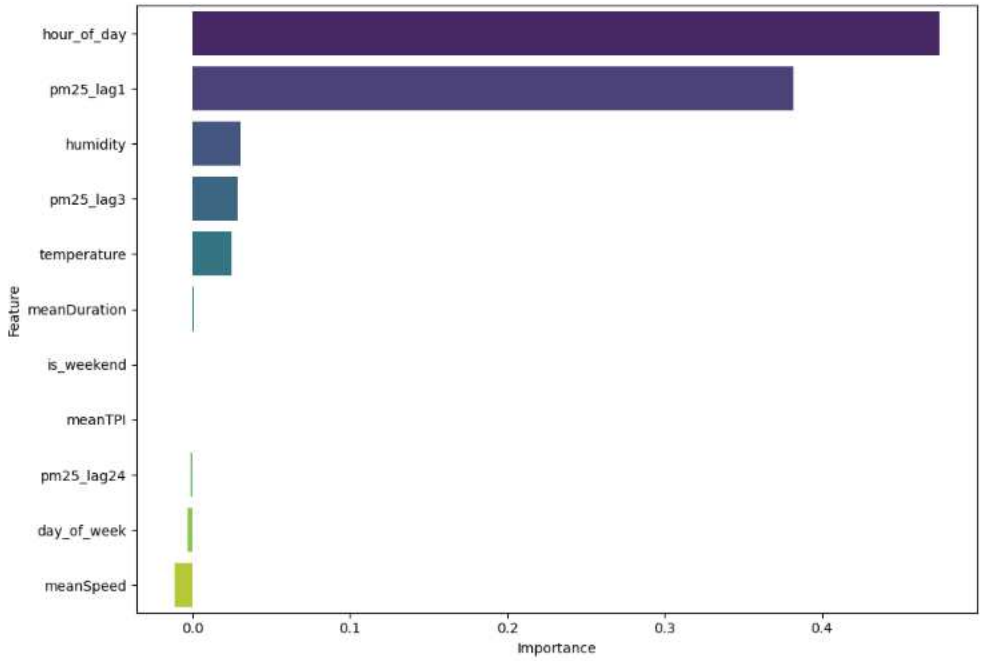


Fig. 13 Feature importance results for the combined forecasting model

Overall, the results indicate that historical $PM_{2.5}$ measurements and temporal variables contributed substantially more predictive information than the evaluated traffic-congestion indicators. While traffic activity is theoretically associated with urban air pollution, the corridor-level congestion metrics considered in this study provided limited incremental forecasting value within the evaluated dataset.

5.0 Discussion

This section discusses the findings of the study and its implications for $PM_{2.5}$ forecasting in Kampala.

5.1 Contribution of Traffic-Congestion Indicators to $PM_{2.5}$ Forecasting

The primary objective of this study was to evaluate whether traffic-congestion indicators improve $PM_{2.5}$ forecasting beyond information already provided by meteorological and temporal variables. The results showed that the environmental baseline models outperformed the traffic enhanced models across all evaluation metrics. The best performing environmental model achieved an R^2 value of 0.5752, compared with 0.4753 for the best performing environmental traffic-enhanced model.

These findings suggest that the evaluated traffic-congestion indicators provided limited additional predictive information within the study context. Although traffic emissions are widely recognized as contributors to urban air pollution, corridor-level traffic indicators considered in this study did not improve forecasting performance beyond that achieved using environmental and temporal predictors.

5.2 Relationship Between Traffic Indicators and $PM_{2.5}$ Concentrations

The correlation analysis provides a possible explanation for these results. Mean Speed, Mean Duration, and Traffic Performance Index (TPI) exhibited negligible correlation with $PM_{2.5}$ concentrations, with correlation coefficients ranging between -0.019 and 0.027. In contrast, Hour of Day (r=0.345), temperature (r = -0.213), and Humidity (r = 0.116) demonstrated comparatively stronger relationships with $PM_{2.5}$ concentrations. This analysis also revealed very strong correlation among traffic variables themselves. Mean Duration and TPI were almost perfectly correlated (r=1.000), while both variables exhibited near-perfect inverse relationships with mean Speed (r approx.-0.991). This suggests that the traffic variables largely represented the same traffic conditions and therefore contributed limited independent information to the forecasting models.

5.3 Importance of Historical PM2.5 Measurements

Additional analysis incorporating lagged $PM_{2.5}$ variables produced the highest forecasting performance (R^2 = 0.6245). Correlation analysis further showed strong relationship between current PM2.5 concentrations and Historical $PM_{2.5}$ measurements, particularly $PM_{2.5}$ Lag1(r = 0.744), PM2.5 Lag24 (r = 0.570), and PM2.5 Lag3 (r = 0.379).

Feature-importance analysis similarly identified Hour of Day and $PM_{2.5}$ Lag1 as the dominant predictors, while traffic-congestion indicators contributed little additional information. These findings suggest that temporal persistence and historical pollution levels play more important role in $PM_{2.5}$ forecasting than the evaluated traffic-congestion indicators within the study area.

5.4 Implications and Future Research

The findings highlight the importance of evaluating predictor relevance rather than simply increasing the number of variables include in forecasting models. While traffic data are increasingly available through digital navigation platforms, their inclusion does not automatically improve forecasting performance.

For Kampala, the results suggest that environmental monitoring data, temporal patterns and Historical $PM_{2.5}$ measurements currently provide the most valuable information for forecasting applications. Future studies should investigate traffic indicators derived from multiple transport corridors, examine alternative traffic metrics, and evaluate forecasting frameworks that combine traffic, environmental, land-use and historical pollution data within a unified modelling framework.

6.0 Conclusion

This study evaluated the contribution of traffic congestion indicators to $PM_{2.5}$ forecasting in Kampala, Uganda, using Random Forest and HistGradientBoosting machine-learning models. The results showed that environmental baseline models outperformed the traffic-enhanced models, indicating that the evaluated traffic-congestion indicators provided limited additional predictive value beyond meteorological and temporal variables.

Correlation analysis further revealed negligible relationship between $PM_{2.5}$ concentrations and the traffic indicators considered in the study, while temporal and meteorological variables exhibited comparatively strong er associations. Additional analysis incorporating lagged $PM_{2.5}$ measurements achieved the highest forecasting performance, demonstrating the importance of historical pollution levels in predicting future $PM_{2.5}$ concentrations.

Overall, the findings suggest that within the study context, historical $PM_{2.5}$ measurements, temporal patterns, and meteorological conditions contribute more to forecasting performance than the evaluated corridor-level traffic-congestion indicators. The study highlights the importance of assessing predictor relevance before incorporating additional datasets into forecasting frameworks and contributes evidence to support the design of efficient, data-driven air-quality forecasting systems in Kampala and similar urban environments.

7.0 Acknowledgment

The author acknowledges the Google Routes API and AirQo API for respective access to organized traffic congestion and $PM_{2.5}$ measurements. Foundational to the training of the models used in this study, the author also acknowledges the use of the Zenodo datasets: https://doi.org/10.5281/zenodo.24153153

REFERENCES

Bainomugisha, E., Ssematimba, J., & Okure, D. (2023). Design considerations for a distributed low-cost air quality sensing system for urban environments in low-resource settings. Atmosphere, 14(2), 354. https://doi.org/10.3390/atmos14020354

Atuyambe, L. M., Etajak, S., Walyawula, F., Kasasa, S., Nyabigambo, A., Bazeyo, W., Wipfli, H., Samet, J. M., & Berhane, K. T. (2025). Air quality and attributable mortality among city dwellers in Kampala, Uganda: Results from 4 years of continuous PM2.5 monitoring using a BAM 1022 reference instrument. Journal of Exposure Science & Environmental Epidemiology, 35(2), 288-293. https://doi.org/10.1038/s41370-024-00684-9

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Wang, Q., Kwan, M.-P., Zhou, K., Fan, J., Wang, Y., & Zhan, D. (2019). The impacts of urbanization on fine particulate matter (PM2.5) concentrations: Empirical evidence from 135 countries worldwide. Environmental Pollution, 247, 989–998. https://doi.org/10.1016/j.envpol.2019.01.086

Alaran, A. J., O'Sullivan, N., Tatah, L., Sserunjogi, R., & Okello, G. (2024). Air pollution (PM2.5) and its meteorology predictors in Kampala and Jinja cities, Uganda. Environmental Science: Atmospheres, 4(10), 1145-1156. https://doi.org/10.1039/D4EA00074A

AirQo. (2024). Air quality monitoring and data platform. AirQo Initiative, Makerere University. https://airqo.net

Coker, E. S., Amegah, A. K., Mwebaze, E., Ssematimba, J., & Bainomugisha, E. (2021). A land use regression model using machine learning and locally developed low-cost particulate matter sensors in Uganda. Environmental Research, 199, 111352. https://doi.org/10.1016/j.envres.2021.111352

Schwander, S., Okello, C. D., Freers, J., Chow, J. C., Watson, J. G., Corry, M., & Bennett, J. (2014). Ambient particulate matter concentrations in Kampala, Uganda: A pilot study. Air Quality, Atmosphere & Health, 7(4), 449-457.

Wabinyai, F. R., Sserunjogi, R., Lubisia, G., Okure, D., Akugizibwe, E., Kutesakwe, J., Nshimye, A., & Ndyabakira, A. (2025). Multilinear regression analysis of PM2.5 in Kampala and Fort Portal cities: Effects of meteorological factors and lagged pollution. Atmospheric Environment: X, 29, 100411.https://doi.org/10.1016/j.aeaoa.2025.100411

World Health Organization. (2024). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization.

Sulaimon, I. A., Alaka, H., Olu-Ajayi, R., Ahmad, M., Ajayi, S, & Hye, A. (2024). Effect of traffic data set on various machine-learning algorithms when forecasting air quality. Journal of Engineering, Design and Technology, 22(3), 1030-1056. https://doi.org/10.1108/JEDT-10-2021-0554

Statements and Declarations

Author Contributions

Hedwig Orieba conceived the study, collected and processed the data, developed the machine learning models, conducted the analysis, and drafted the manuscript. Mackenzi Deborah supervised the study, provided methodological guidance, reviewed the manuscript, and contributed to all technical revisions. Both authors approved the final manuscript.

Funding

The authors received no external funding for this research.

Data Availability

The dataset supporting this study is available through Zenodo at https://doi.org/10.5281/zenodo.24153153 .

Source code supporting the analysis is available at https://github.com/HedwigOrieba/pm2_5_congestion_ml_engine.git

Competing Interests

The Authors declare no competing interests.