Air Quality Forecasting in London Using Random Forest Regression

Xiaohan Yang*

Department of Information Science & Engineering, Lanzhou University, Lanzhou, 730000, China

*Corresponding author: yxiaohan2024@lzu.edu.cn

Abstract. This paper explores short-term forecasting of daily PM2.5 and AQI levels in London during 2024, employing Random Forest (RF) regression as the primary modeling approach. Air quality records were utilized to compile a set of predicted values consisting of the 6 pollutants recorded (PM10, NO_2, SO_2, CO, and O_3) and daily average values obtained via data pre-processing. Temporal variables such as day-of-year and weekend indicators were incorporated to account for seasonal patterns and human activity effects. The RF model, trained on multivariate inputs, demonstrates strong predictive accuracy, with results showing close alignment between predicted and observed values. Feature importance analysis reveals PM10 as the dominant predictor for PM2.5, supported by known emission and dispersion dynamics. For future work, RF's performance is also extended to benchmark with other regression approaches such as AutoRegressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM) and XGBoost in terms of its differences of prediction accuracy, model explainability and computation speed. The findings contribute to data-driven urban pollution monitoring, offering practical insights for short-term forecasting and public health early warning systems.

Keywords: Air quality forecasting; PM2.5; Random Forest.

1. Introduction

Air Quality Prediction spatial and temporal air pollution dynamics. Specific pollutants including PM2.5, nitrogen dioxide and so on, can be taken as indicators to quantify the health risks. Modeling methods of air quality prediction are studied widely. However, single method is rarely applied in practice, the selection is mainly based on the nature of data, prediction tasks, or computing resources at that moment. This still presents a point of interest for reasons not only related to the health effects of the pollution but also related to the time‐varying and to the fact that the pollutant levels are subjected to small variations of human activities or weather conditions.

The ARIMA model is a classical statistical approach that has been widely applied to time-series analysis of air pollution data. Kumar and Jain applied ARIMA for short-term forecasting of pollutant concentrations in Delhi. For one-day-ahead predictions, the reported MAPE values were 13.6% for CO, 12.1% for NO_2, 21.8% for NO, and 24.1% for O_3, indicating acceptable but pollutant-dependent performance [1]. In one of the other research projects, ARIMA was applied to the forecaster AQI in Shenzhen and was proven usable to grasp the time-varied nature of urban AQI data [2]. In order to have a good performance on analyzing periodicity, the Seasonal ARIMA(SARIMA) model was integrated with Factor Analysis in order to forecast PM2.5, proving that the methods have enhanced the ability to capture both seasonal patterns and weather disturbances [3].

Along with the development of machine learning, RF and LSTM network have been introduced to this task to predict future air pollution. Iskandaryan et al. have summarized the implementations of machine learning models including LSTM for predicting air quality and indicated that there are significant improvements compared with the traditional approaches [4]. Similar, a study examined air pollution coupling using both the RF model and LSTM in Bejing and concluded that LSTM needs much larger sample sizes for training, i.e. more time demanding [5]. Danesh Yazdi et al. in context of London studied and tested both the LSTM and RBF models for short-term predictions. They reported that the computational cost of training the RBF model on their dataset is quite expensive, designed an ensemble RF, Gradient Boosting Machine, and k-nearest neighbor algorithms ensemble

with satellite AOD, land-use, and meteorology data has been developed to obtain the high temporal predictive accuracy results and demonstrated the potential of ML for fine-scale estimation of PM2.5 at urban area [6]. RF is a simpler choice and also applicable to small data, compared with LSTM that needs huge amount of data. Furthermore, it is more interpretable and more computationally efficient.

To achieve acceptable trade-off between model interpretability and prediction ability, some works have been done to hybrid the conventional model with a ML technique. A CEEMDAN-LSTM-ARIMA model was presented by Li et al. to decompose the air pollution time series into several subcomponents and utilize for enhancing the forecast accuracy of PM2.5 [7]. Similarly, research also proposed a SARIMA-LSTM-BP neural network model for AQI prediction in Beijing to improve the accuracy and reliability with complex models and implementation difficulties [8]. Furthermore, Kang et al. compared multiple single models (RF, XGBoost, LightGBM, Adaboost, Decision Tree) and a stacking ensemble across different data types and time scales for PM2.5 prediction, finding stacking models generally outperformed individual models, which underscores the importance of comparative evaluation when selecting prediction approaches [9].

Although these models have contributed to some improvements in PM2.5 prediction, there are still some limitations. Most state-of-the-art models demand big data and complicated training, which do not apply to cities without enough data. The previous works of these studies focus on Beijing, Shenzhen and London very much [2,5,6,10]. The contribution is a clean RF baseline on recent London data with reproducible evaluation. It should be noted that the researches of long-term behaviour and seasonality of air pollution measurements of London have been carried out in the past, but they do not conduct predictive modeling [10].

Based on these research deficiencies, it is hypothesized that applying the RF model for estimating daily PM2.5 concentrations in London could offer valuable insights. RF is chosen for its interpretability, its general small data set fit, and its ability to accommodate complex relationships between the variables without data preparation. The study utilizes publicly available air quality data from London and assesses forecasting effectiveness through statistical metrics, including RMSE, MAPE, and R^2. The results are expected to provide empirical insights to support early warning systems, urban health management, and evidence-based policymaking, especially in cities with limited access to comprehensive datasets.

2. Method

2.1. Data Collection and Preprocessing

A public air quality dataset was obtained from Kaggle, originally sourced via the Open-Meteo API. The 2024 dataset was selected as it represents the most recent complete year and captures post-pandemic urban activity trends. It contains 8,760 hourly measurements (24 hours × 365 days) of major atmospheric pollutants across London, including PM2.5, PM10, NO_2, SO_2, CO, CO_2, O_3, and a calculated Air Quality Index (AQI).

Hourly data were aggregated into daily averages, reducing the dataset to 365 daily entries. The CO_2 column was excluded because of high missing values. Values missing less than half per year for each variable were replaced through linear interpolation. A tabular overview of the number of missing values prior to cleaning is given in Table 1.

Table 1. Missing Data Statistics Before Preprocessing

Column	Missing Count	Missing Percentage
CO	15	4.1%
CO_2	280	76.7%
PM2.5	7	1.9%
PM10	0	0.0%
O_3	0	0.0%
NO_2	3	0.8%
SO_2	12	3.3%
AQI	0	0.0%

After cleaning, a final dataset of 365 daily records was obtained, providing a complete basis for model training and evaluation.

2.2. Feature Selection

The explanatory variables considered in this work are chosen based on the published research and preliminary statistical analysis. The response variable is PM2.5, a vital sign of respiratory health and urban air quality.

These predictors can be grouped into three categories. The first category consists of major air pollutants, including gaseous compounds such as CO, NO_2, SO_2, and O_3, together with particulate matter (PM10), all of which are known to influence PM2.5 formation and dispersion. Second, the Air Quality Index (AQI) was incorporated as a composite metric reflecting overall pollution severity. Lastly, two temporal features—the day of the year and a binary indicator for whether the day falls on a weekend—were added to capture seasonal trends and human activity patterns.

A full description of all input features is provided in Table 2.

Table 2. Variable Descriptions

Feature	Unit	Description
PM2.5	μg/m³	Fine particulate matter
PM10	μg/m³	Coarse particulate matter
NO_2	ppb	Nitrogen dioxide
SO_2	ppb	Sulfur dioxide
CO	ppm	Carbon monoxide
O_3	ppb	Ozone
AQI	Index	Air Quality Index (composite)
Day of year	–	Integer from 1 to 365
Weekend	0 or 1	Weekday (0) or Weekend (1)

2.3. Model Description: Random Forest

Unlike ARIMA, RF does not need the input data to be stationary, and is trained much faster and performs better with less samples than deep learning approaches (e.g. LSTM). To model the nonlinear and multi-factorial nature of the PM2.5 dynamics, a RF regressor was applied—a powerful ensemble learning algorithm based on decision trees. RF uses a bagging strategy which builds numerous decision trees on bootstrapped subsets of the data and averages their outputs to reduce variance and increase the model's generalization performance. For regression the output is calculated as:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} h_i(x) \quad (1)$$

Where $h_i(x)$ is the prediction from the i-th tree and n is the total number of trees.
Each decision tree performs node splits by minimizing Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}(y_j - \bar{y})^2 \quad (2)$$

Figure 1 shows a simplified representation of how the RF regression model operates. As illustrated, the algorithm begins with input data comprising multiple features and a target variable. Through bootstrap sampling, random subsets of samples are produced, and a stochastic selection of candidate features is examined at every node split. These are then used to train multiple decision trees in parallel. By averaging the predictions from individual trees, the overall model output becomes more stable and less prone to overfitting.

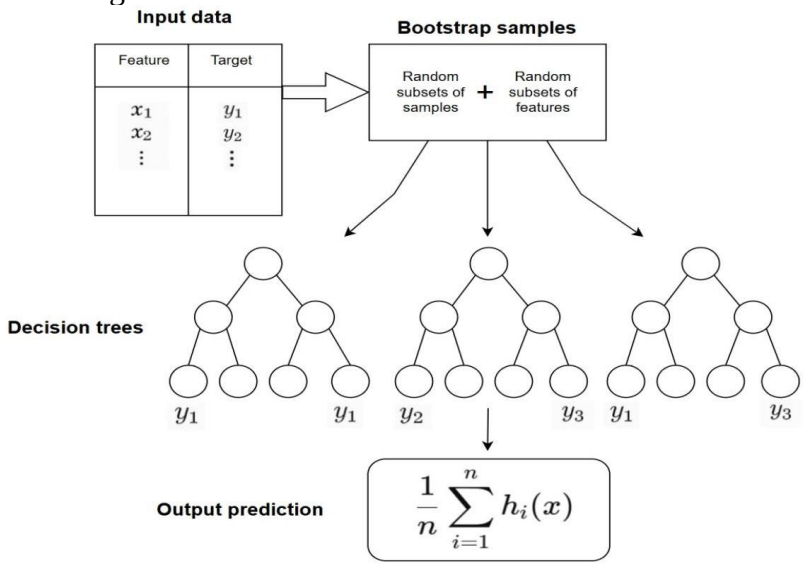

Figure 1. Structural Representation of the Random Forest Regression Framework

The implementation of the model is based on the RF from the Scikit-learn library. The data were split based on the time to a set of training (70%) and testing (30%). A grid-search procedure with five-fold cross-validation was employed to tune key settings, notably the number of estimators and the maximum tree depth. To quantify predictive effectiveness on the test dataset, three statistical indices—RMSE, MAPE, and R^2—were applied.

Table 3 presents the detailed parameter configurations used for the RF model.

Table 3. Random Forest Model Parameters

Parameter	PM2.5 Model	AQI Model
n_estimators	200	100
max_depth	None	15
min_samples_split	default (=2)	3
min_samples_leaf	default (=1)	2
max_features	default	'sqrt'
bootstrap	True	True
random_state	42	42

For benchmarking purposes, two additional models—ARIMA and XGBoost—were also implemented, enabling a direct comparison between RF and both a classical statistical approach and a state-of-the-art boosting method.

3. Results and Discussion

In this part, empirical evidence is provided for the RF model's application to daily PM2.5 and AQI forecasting in London, based on data gathered throughout 2024.

3.1. Descriptive Analysis

Before modeling, a descriptive analysis was conducted to understand the seasonal behavior and interdependencies among pollutants in London throughout 2024.

According to Figure 2, the average concentration of PM2.5 was notably seasonal, in that it was consistently low from mid-May through to September, followed by a steep rise toward late autumn and winter, thought to be partly as a result of decreased dispersion in the atmosphere and also to some increased emissions associated with central heating. The highest daily average exceeded 30 μg/m³, while the lowest fell below 5 μg/m³.

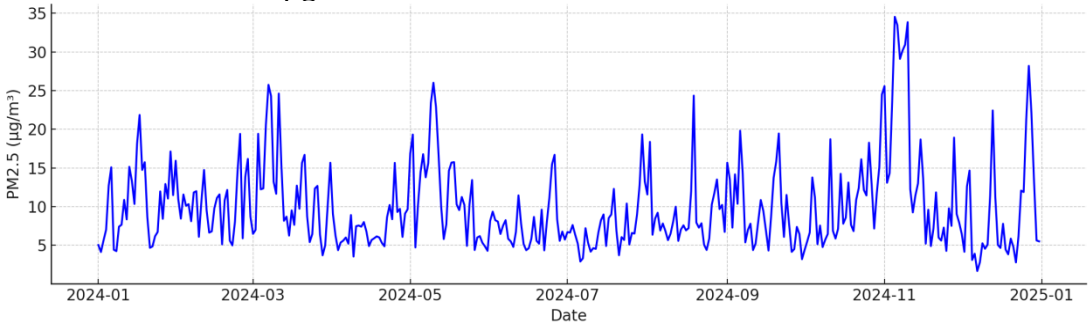

Figure 2. Daily average PM2.5 concentration in London, 2024

In a similar fashion, the variation of the Air Quality Index (AQI) is shown in Figure 3, and had the same seasonal pattern as PM2.5, reaffirming that fine particulate matter makes a significant contribution to composite pollution. As above, AQI also displayed greater values in colder months, suggesting a contribution from seasonal patterns of energy use and weather.

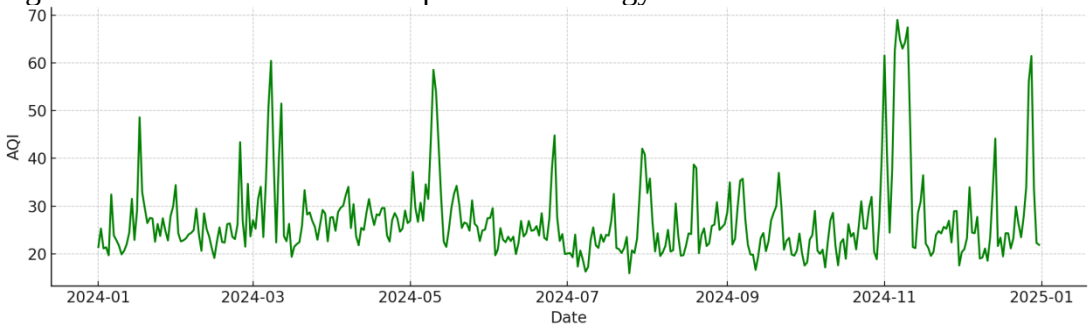

Figure 3. Daily average Air Quality Index (AQI) in London, 2024

In order to get the understanding of relationship between pollutants, the paper constructed correlation heatmap using hourly-mean data, as shown in Figure 4. It can be seen that PM2.5 was correlated with PM10(r = 0.97) in highly correlation, and in correlation was positive, especially in moderate and strong positive correlation with NO_2 and SO_2. However, O_3 was negatively correlated with nearly all other pollutants in most high PM2.5 time period.

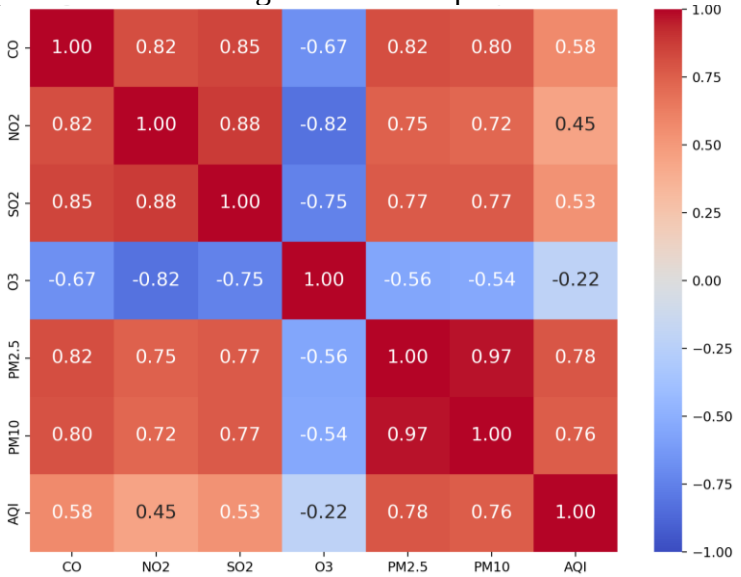

Figure 4. Feature Correlation Heatmap

These results suggest that multi-pollutant interactions are not only present but statistically significant. The high correlation between PM2.5 and PM10 justifies the inclusion of co-pollutants as model features and supports the feasibility of multivariate air quality forecasting.

3.2. Model Performance and Evaluation

Using a 70–30 split, the RF model was trained on the larger portion of the dataset and evaluated on the smaller held-out portion. The models predictive performance on both PM2.5 and AQI are shown in Table 4 and Table 5. In general the RF model delivered good accuracy and reliable explanatory power with strong consistency on the PM2.5 predictions and low average error on the AQI predictions.

RF has superior capability of capturing nonlinearities compared with traditional time-series models (e.g. ARIMA) as it doesn't require too many assumptions. The method works well in multivariate air quality data sets of moderate size.

Table 4. PM2.5 Prediction Results (Test Set)

Metric	PM2.5 (Train)	PM2.5 (Test)	AQI (Train)	AQI (Test)
RMSE	0.41	2.15	2.42	3.95
MAPE	3.61%	14.32%	5.05%	9.21%
R^2	0.993	0.916	0.918	0.810

To better understand the model's internal mechanism, feature importance was analyzed. Table 5 illustrates that PM10 explained more than 95% of the information contained in the model, while the other pollutants'(NO_2, SO_2, and CO) contributions are negligible. This is consistent with the known emission origins and physical behavior of fine particulate matter and provides helpful and practical directions for targeted pollution control policies.

Table 5. Feature Importance Scores in PM2.5 Prediction (Random Forest Model)

Feature	PM2.5 Importance	AQI Importance	Description
PM10	0.958	0.726	Coarse particulate
NO_2	0.013	0.046	Nitrogen dioxide
CO	0.011	0.053	Carbon monoxide
SO_2	0.011	0.038	Sulfur dioxide
O_3	0.007	0.137	Ozone

The other positive aspect of this RF model is its high reliability when used in real and messy urban conditions: RF works also on degraded and biased input (noisy, incomplete and smallsized samples, the current characteristic of sparse and underdeveloped monitoring networks in cities), despite the available number of environmental metrics in a city, and it can be easily updated with additional meteorological or temporal-related features, therefore allowing being deployed in large-scale for real-time predictions. This robustness, coupled with the model's interpretability, reinforces its practical relevance in supporting environmental decision-making and public health interventions.

To further contextualize these results, the RF model's performance was evaluated against ARIMA and XGBoost using identical training and test sets. This facilitates an objective assessment of each model's predictive accuracy, interpretability, and computational efficiency.

3.3. Model Comparison: RF vs ARIMA vs XGBoost

To evaluate the relative performance of different modeling approaches for daily PM2.5 and AQI forecasting in London, the RF model was compared with a traditional statistical method, ARIMA, and a modern gradient boosting method, XGBoost.

Table 6 and 7 show the results on test set for PM2.5 prediction and AQI prediction respectively. RF consistently demonstrates high predictive accuracy with balanced interpretability, while XGBoost offers comparable performance in PM2.5 forecasting but falls behind in AQI prediction. ARIMA, as

a univariate model, shows much lower accuracy in both tasks, reflecting its limited capacity to capture complex, nonlinear, multivariate relationships.

Table 6. PM2.5 Test Set Performance Comparison

Model	RMSE	MAPE	R^2
RF	2.15	14.32%	0.916
ARIMA	5.17	44.60%	0.426
XGBoost	2.18	16.68%	0.898

Table 7. AQI Test Set Performance Comparison

Model	RMSE	MAPE	R^2
RF	3.95	9.21%	0.810
ARIMA	7.42	17.66%	0.525
XGBoost	7.03	13.68%	0.573

Overall, the results highlight RF's ability to deliver accurate, stable, and interpretable forecasts under multivariate conditions. XGBoost performs competitively for PM2.5 but is less effective for AQI, while ARIMA's lower accuracy underscores the limitations of univariate time-series models in capturing the complex pollutant interactions present in urban air quality datasets.

3.4. Visualization and Interpretation

Although the comparative evaluation in Section 3.3 considered ARIMA, XGBoost, and RF, the following visualization focuses on the RF model due to its superior overall performance and interpretability. For brevity, only RF's prediction trends are illustrated, as it serves as the primary model in this study.

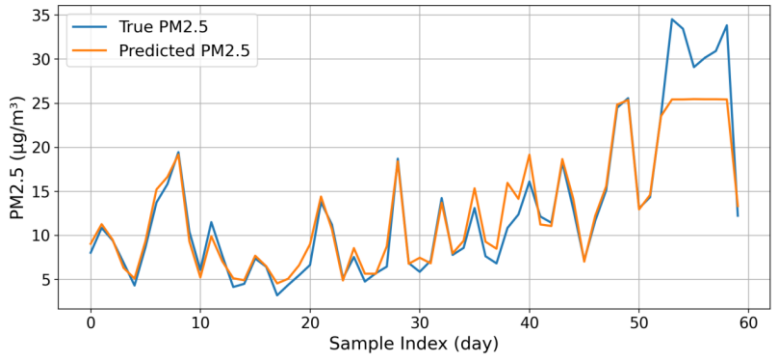

Figure 5. Predicted and observed PM2.5 levels (test data)

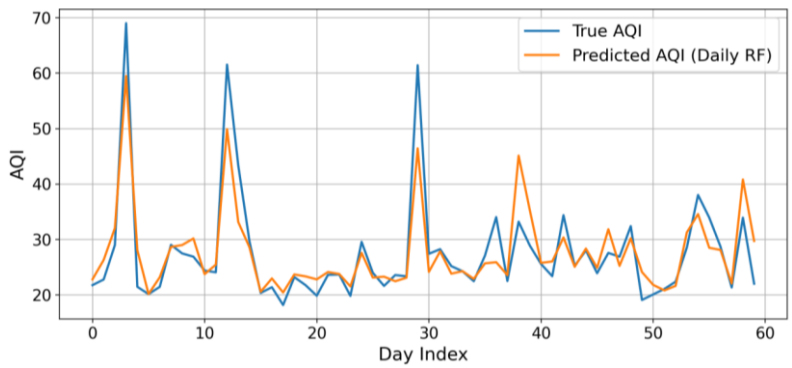

Figure 6. Predicted and observed AQI outcomes (test partition)

Figure 5 depicts how the predicted PM2.5 values align with the actual measurements in the test dataset. The model demonstrates a strong ability to track overall daily fluctuations, with good

alignment across most of the prediction horizon ($R^2 = 0.916$). While small misestimations are visible at the local peak levels of the particulate, overall the model generally performs well in capturing daily PM 2.5 estimates.

Figure 6 presents the true values vs. predicted AQI values. While the overall fit is slightly lower ($R^2 = 0.810$), the model achieves a test MAPE of 9.21%, indicating stable performance. The predicted AQI values follow the seasonal trend and magnitude of the observed values, suggesting that the model is effective for broad short-term air quality assessments, even if not perfectly accurate at finer levels.

In terms of performance, the model on both the two tasks of PM2.5 and AQI shows that it can effectively perform multi-variables air quality prediction in urban scenarios. Yet the modeling, e.g. introducing lagged information, temperature, wind speed, or adjusting calendar effects (e.g. intensity of holiday, time clustering, etc.), has the potential to be improved, especially when facing the sudden transition on pollution level. In the future, the results may be compared with deep learning models as well in diverse urban environments to trade-off accuracy and interpretability.

4. Conclusion

This paper implemented a RF regression model for the daily prediction of PM2.5 concentrations and AQI values in a city such as London making use of multivariate open-access data. A robust and reliable predictive performance has been obtained for the RF model, where PM10 is the prominent PM2.5 predictor, validating the predictive capability of the model to encapsulate the nonlinear relationships of pollutants in a parsimonious manner.

To examine the stability of RF, supplementary tests were carried out against ARIMA and XGBoost. The outcomes indicated that RF achieved substantially higher performance than ARIMA across both PM2.5 and AQI forecasting, and delivered results on par with or superior to XGBoost, especially in the AQI setting. This evidence highlights the appropriateness of RF for short-horizon, multivariable air quality prediction in urban contexts.

However, there are also some limitations in this study. This study solely uses one year's data and does not use the meteorological and event variables which decreases the accurate rate for some exceptional high pollution episodes. Future improvements could involve integrating lagged pollutant variables, meteorological factors, and refined calendar effects to enhance responsiveness to sudden environmental changes. Extending the evaluation to other cities and comparing with advanced neural networks such as LSTM could further improve robustness and generalizability.

In spite of these drawbacks, RF remains a practical and interpretable approach for air quality estimation in data limited environments, with real-world applications for local day-to-day air quality alerts, and policy interventions.

References

[1] Kumar U and Jain V K. ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). Stochastic Environmental Research and Risk Assessment, Stoch Environ Res Risk Assess. 2010, vol. 24, no. 5, pp. 751–760.

[2] Mou J, Zhao X, Fan J, Yan Z, Yan Y, Zeng D, Luo W, and Fan Z. Time series prediction of AQI in Shenzhen based on ARIMA model. Journal of Environmental Hygiene, 2017, vol. 7, no. 2, pp. 102–107.

[3] Bhatti U A, Yan Y, Zhou M, Ali S, Hussain A, Huo Q, Yu Z, and Yuan L. Time series analysis and forecasting of air pollution particulate matter (PM2.5): An SARIMA and factor analysis approach. IEEE Access, 2021, vol. 9, pp. 29222–29235.

[4] Iskandaryan D, Ramos F, and Trilles S. Air quality prediction in smart cities using machine learning technologies based on sensor data: A review. Applied Sciences, 2020, vol. 10, no. 7, p. 2401.

[5] Ren W, Liu J, Deng Y, Zou Y, Liu J, and Wang S. Research and prediction on air pollution coupling based on RF regression and LSTM neural networks: A case study of Beijing. Science and Technology Innovation, 2025, no. 10, pp. 22–26.

[6] Danesh Yazdi M, Kuang Z, Dimakopoulou K, Barratt B, Suel E, Amini H, Lyapustin A, Katsouyanni K, and Schwartz J. Predicting fine particulate matter (PM2.5) in the greater London area: An ensemble approach using machine learning methods. Remote Sensing, 2020, vol. 12, no. 6, p. 914.

[7] Li H. Prediction of air quality based on CEEMDAN-LSTM-ARIMA model. M.S. thesis, Chongqing University, 2023.

[8] Wang J. Prediction of Beijing air quality index based on SARIMA-LSTM-BP neural network. M.S. thesis, Shandong University of Finance and Economics, 2024.

[9] Kang J, Zou X, Tan J, Li J, and Karimian H. Short-term PM2.5 concentration changes prediction: a comparison of meteorological and historical data. Sustainability, 2023, vol. 15, no. 14, p. 11408.

[10] Gil-Alana L A, Yaya O S, and Carmona-González N. Air quality in London: evidence of persistence, seasonality and trends. Theoretical and Applied Climatology, 2020, vol. 142, pp. 103–115.