Prediction of Urban $PM_{2.5}$ Concentrations Using Machine Learning–Based Multiple Linear Regression: A Case Study of Jaipur, India

Shubham Goyal[1] and Ruchi Sharma[2*]

[1]Research Scholar, Department of Civil Engineering, Malaviya National Institute of Technology, Jaipur, Pin 302017, India

[2]Assistant Professor, Department of Civil Engineering, Malaviya National Institute of Technology, Jaipur, Pin 302017, India

Dr. Ruchi Sharma: ORCID ID: 0000-0002-8945-2788

Abstract

Fine particulate matter ($PM_{2.5}$) (aerodynamic dimension (≤2.5 µm)) poses a serious risk to public health and urban environmental quality, particularly in rapidly growing cities of developing countries. Reliable prediction of $PM_{2.5}$ concentrations is therefore essential for effective air quality management and policy formulation. This study develops a machine learning–based multiple linear regression (MLR) model to predict $PM_{2.5}$ concentrations using hourly air quality data from three monitoring stations in Jaipur, India, for the year 2019. Based on Pearson correlation analysis, PM_{10}, NO_x, and benzene were selected as input variables. Initial MLR models showed moderate predictive performance (R^2 = 0.27–0.71), which was significantly improved through systematic data refinement techniques, including outlier removal, logarithmic transformation, and bootstrapping. The optimized models achieved R^2 values ranging from 0.77 to 0.80 across the three sites, demonstrating strong agreement between predicted and observed $PM_{2.5}$ concentrations. The results highlight the effectiveness of combining simple machine learning techniques with robust data preprocessing to enhance air quality prediction accuracy. The proposed approach provides a practical and transferable framework for $PM_{2.5}$ forecasting in data-rich urban environments and can support decision-making for air pollution mitigation in Indian cities.

Keywords: Fine particulate matter ($PM_{2.5}$) prediction, Machine learning, Multiple linear regression (MLR), Outliers, Data transformation, Urban air pollution

1. Introduction

Air pollution has emerged as one of the most critical environmental challenges affecting human health and well-being, particularly in rapidly urbanizing regions. In India, outdoor air pollution is responsible for approximately 670,000 premature deaths annually, primarily due to inadequate enforcement of emission control regulations and rapid industrial and vehicular growth [1–3]. Several air pollutants, including particulate matter (PM), carbon monoxide (CO), nitrogen oxides (NO_x), ozone (O_3), and sulfur oxides (SO_x), frequently exceed the National Ambient Air Quality Standards in many Indian cities [4].

Among these pollutants, particulate matter is of particular concern due to its strong association with adverse cardiovascular and respiratory health effects [5–7]. The effects mainly rely upon the particle's atmospheric concentration, size, and chemical configuration [8]. PM can be categorized into "coarse" (with an aerodynamic diameter of ≤ 10µm, PM_{10}), “fine” (with an aerodynamic diameter of ≤ 2.5µm, $PM_{2.5}$), and “ultra-fine” (with an aerodynamic diameter of ≤ 100nm, PM_1) particles [9]. Fine particulate matter ($PM_{2.5}$) is especially hazardous as it can penetrate deep into the alveolar region of the lungs and enter the bloodstream, thereby affecting multiple organ systems [10,11]. Exposure to $PM_{2.5}$ has been linked to asthma, cardiovascular diseases, cancer, and increased vulnerability to respiratory infections, including COVID-19 [12–16].

Accurate measurement of $PM_{2.5}$ requires advanced monitoring equipment, which is often unavailable in many developing regions. However, continuous ambient air quality monitoring stations routinely record other pollutants that can be utilized for $PM_{2.5}$ estimation through predictive modeling [17–20]. Such predictive capability can support early warning systems, exposure reduction strategies, and air quality management planning.

Recent advances in machine learning have enabled efficient handling of large environmental datasets and improved predictive accuracy compared to conventional statistical approaches [21–24]. Despite this progress, limited studies have focused on machine learning–based $PM_{2.5}$ prediction in Indian urban environments, particularly using large hourly datasets and systematic data refinement strategies [25].

In this context, the present study aims to develop a machine learning–based multiple linear regression (MLR) model to predict $PM_{2.5}$ concentrations for Jaipur, India. Hourly air quality data for the year 2019 from three monitoring locations were used. Key input variables were selected using Pearson correlation analysis, and model performance was enhanced through outlier removal, logarithmic transformation, and bootstrapping. The study provides a practical and transferable framework for $PM_{2.5}$ prediction in urban Indian cities.

2. Methodology

2.1. Study area and data collection

Jaipur, the capital city of Rajasthan (26°25′N, 74°55′E), is a rapidly urbanizing and landlocked metropolitan area with limited atmospheric dispersion potential. The city covers approximately 11,061 km² and has experienced significant population growth and urban expansion, contributing to deteriorating air quality conditions [26].

Hourly air quality data for the year 2019 were obtained from the Rajasthan State Pollution Control Board (RSPCB) for three continuous ambient air quality monitoring stations: Jaipur Psychiatric Center (site-1), Jaipur Police Commissionerate Office (site-2), and Jaipur Science Park (site-3). The dataset included $PM_{2.5}$, PM_{10}, NO_x, and benzene concentrations. Location of the study area and the three air quality monitoring sites in Jaipur has been illustrated in Figure 1.

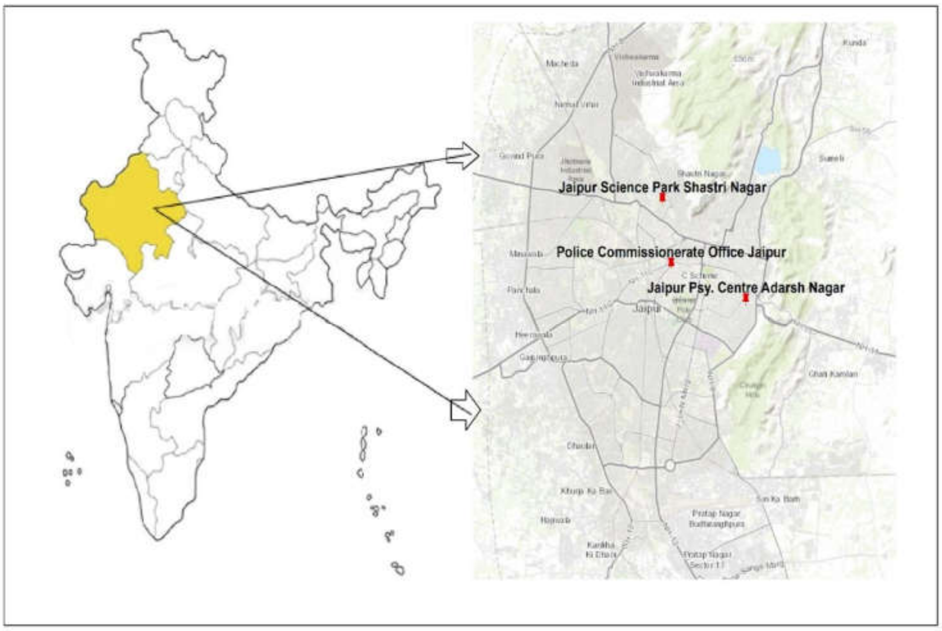

Jaipur Science Park Shastri Nagar
Police Commissionerate Office Jaipur
Jaipur Psy. Centre Adarsh Nagar
Jaipur

Figure 1: Location of the study area and the three air quality monitoring stations in Jaipur, India.

2.2. Data pre-processing

The dataset used in the model was hourly from January 01, 2019, to December 31, 2019. Missing data constituted only 2–3% of the total dataset and were removed to ensure data integrity. Descriptive statistics, including mean, range, and standard deviation, were computed for each pollutant at all three sites, as shown in Table 1. The hourly dataset was randomly divided into training (80%) and testing (20%) subsets, and the process was repeated multiple times to ensure robust model performance.

Table 1: Descriptive statistics of hourly air quality parameters used for model development at the three monitoring sites (all units are in µg/m^3)

	Site-1			**Site-2**			**Site-3**		
Parameter	**Range**	**Mean value**	**Standard deviation**	**Range**	**Mean value**	**Standard deviation**	**Range**	**Mean value**	**Standard deviation**
$PM_{2.5}$	0-341.6	36.8	19.7	0-1053.8	61.7	43.2	0-1759.7	49.5	41.9
PM_{10}	0-2924.2	99.6	77.8	0-1432.5	120.8	64.6	0-7883.1	11.3	134.1
NO_x	3.4-358.1	45.0	36.9	0-607.2	45.8	40.7	6-273.5	35.3	22.0
Benzene	0-34.8	1.5	2.8	0-40	1.6	2.1	0-20.4	1.1	1.7

2.3. Input variable selection

Due to the complexity and amount of input parameters, the selection of input variables is an important step for model development and forecasting. The correlation analysis technique provides a good estimate for input parameter selection; therefore, the Pearson correlation method was employed to identify the most influential predictors of $PM_{2.5}$ concentration in the present study. Pearson correlation coefficient (r) is one of the easiest and quickest methods for the selection of input parameters and helps in categorizing the best impactful input values for model forecasting [27, 28]. It is defined as the ratio among the covariance of the two variables and the standard deviation of them as indicated in the following Equation 1 [29]:

$$\mathrm{r} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{1}$$

here σ_{xy} denotes the covariance among the x and y variables which is determined by Equation (2):

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y}) \tag{2}$$

where σ_x, σ_y denote the standard deviation for the individual variable as determined by Equation 3 and Equation 4, respectively.

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{3}$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} \tag{4}$$

The value of r signifies the linear relationship between the two evaluated variables, and its value lies between $-1 < r < 1$. Here, the value 1 indicates a direct positive linear correlation; value -1 corresponds to a total indirect negative linear correlation, and value 0 describes no relationship among the variables. An empirical threshold of correlation coefficient ($r > 0.25$) was applied, resulting in the selection of PM_{10}, NO_x, and benzene as input variables for model development.

2.4. Machine learning-based MLR modeling

Multiple linear regression was selected due to its simplicity, interpretability, and proven effectiveness in air quality prediction studies [30–32]. This method determines the relationship between several independent variables and a dependent variable; therefore, MLR has been selected as a suitable technique for this study. The most common form of the MLR model can be represented by Equation (5) [32]:

$$y_n = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots + b_nx_n \tag{5}$$

Where,

y_n denotes the nth dependent variable value; x_n denotes the nth independent variable value; b_0 denotes the intercept of the equation, and b_n denotes the nth regression coefficient.

Due to the large amount of dataset, machine learning technique has been used for the development of the MLR model in this study. Moreover, Python programming language and Jupyter Notebook environment were used to implement the machine learning workflow due to their computational efficiency and extensive library support [33]. To enhance the accuracy of the model, the data were randomly distributed into two parts, i.e., training data and testing data, with proportions of 80% and 20%, respectively. The random distribution was conducted at least three times, and the best result was reported as the final model equation.

2.5. Model evaluation and refinement

Model performance was evaluated using the coefficient of determination (R^2), which represents the model's prediction accuracy [34]. The R^2 values range from 0 to 1, with a higher R^2 value showing that the model can make more accurate predictions. The following Equation 6 can be used to determine the indicator, R^2 [32]:

$$R^2 = \frac{\left(\sum_{i=1}^{N}(y_p^i - \overline{y_p})(y_o^i - \overline{y_o})\right)^2}{\sum_{i=1}^{N}(y_p^i - \overline{y^\iota})^2 \sum_{i=1}^{N}(y_o^i - \overline{y_0})} \tag{6}$$

Here,

y_p^i and y_o^i denote the ith predicted and observed values; $\overline{y_p}$ and $\overline{y_o}$ denote the average of the predicted and observed values, and n indicates the number of samples.

To improve prediction accuracy, three data refinement strategies were sequentially applied as discussed in the Section 2.5.1-2.5.3. It should be noted that the technical process of converting data from one format, standard, or structure to another without impacting the dataset's content is known as data transformation [35, 36].

2.6.1. Removal of extreme outliers

Removing outliers is essential and cannot be neglected because their presence in the dataset may affect the performance of machine-learning approaches and developed models [37]. Data outliers are usually the result of inaccurate measurements. Outliers in the dataset may lead to mistakes in model evaluation and may cause issues with model fitting [38]. The box plot method, the Z-score method, and applying upper and lower limit thresholds are examples of outlier removal techniques [39]. The upper and lower threshold limit method has been applied in the present study to exclude outliers from the dataset. Concentration with more than three times the standards was taken as the cutoff criterion for higher extreme values, while concentration values below 10 μg/m^3 for the pollutant were considered as the cutoff criterion for lower extreme values.

2.6.2. Logarithmic transformation of data

Some machine learning or deep learning models, such as MLR, artificial neural network (ANN), and logistic regression, assume that parameters or features are normally distributed. If the parameters used during modeling are normally distributed, they can perform significantly better [40]. The general equation for the Gaussian distribution function can be given by Equation (7) [41]:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2} \tag{7}$$

Here, σ denotes the standard deviation, and μ denotes the mean value.

The above equation is based on the mathematical assumption that the selected input parameters follow a Gaussian distribution while performing regression analysis [42]. Therefore, in this study, the data were log-normalized to reduce data skewness and to make the data distribution closer to Gaussian in order to enhance the model performance.

2.6.3. Bootstrapping technique

Bootstrapping is a resampling technique that involves drawing samples from the source data over and over with replacement, usually to estimate a population parameter in statistics and machine learning methods [43, 44]. With replacement means that the same data point may be included in the resampled dataset more than once. The performance of any machine learning model with a limited number of data points can be improved through the bootstrapping method [45]. In this study, after log normalizing the dataset, the bootstrapping technique was used to enhance model robustness with limited effective samples. During bootstrapping, more data points were generated, and thereafter, the performance of the MLR model was evaluated.

3. Results and Discussion

3.1. Input variable selection using Pearson correlation coefficient

After computing the Pearson correlation coefficient among dependent and independent variables, an empirical selection benchmark with values of r more than 0.25 has been applied, and considering this criterion, PM_{10}, NO_x, and benzene have been selected as input variables to predict the concentration of $PM_{2.5}$. The values of correlation coefficients between $PM_{2.5}$ and the three selected parameters (PM_{10}, NO_x, and benzene) have been shown in Figure 2 for the three monitoring sites. Correlation analysis revealed that PM_{10} exhibited the strongest association with $PM_{2.5}$ across all three sites, followed by benzene and NO_x. These findings indicate the significant contribution of resuspended dust and combustion-related emissions to fine particulate pollution in Jaipur.

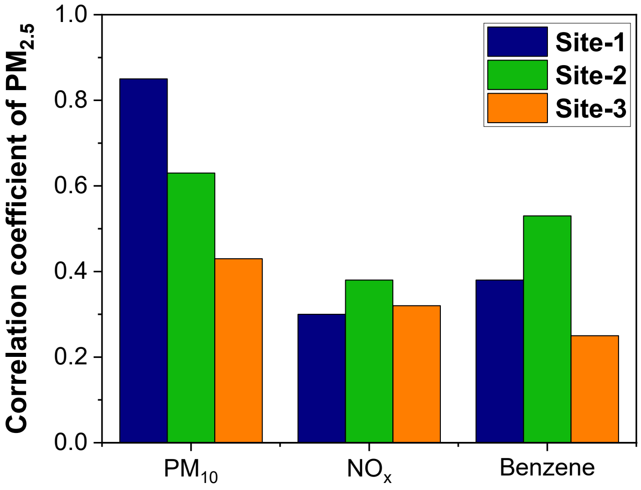


Figure 2: Pearson correlation coefficients between $PM_{2.5}$ and selected predictor variables (PM_{10}, NO_x, and benzene) at the three monitoring sites.

3.2. Development and evaluation of MLR models

Based on the three input parameters, i.e., PM_{10}, NO_x, and benzene, and output parameters as $PM_{2.5}$, the MLR model was developed using a machine learning technique. Three MLR model equations, i.e., Equation 8, Equation 9, and Equation 10, were developed for the prediction of $PM_{2.5}$ concentration for site-1, site-2, and site-3, respectively. Equations 8-10 reflected relatively higher coefficients for benzene and PM_{10}, indicating that they were significant contributors in the prediction of $PM_{2.5}$ concentrations, as also supported by their relatively higher Pearson correlation coefficients. However, at site-3, the contribution from NO_x was found to be relatively higher than PM_{10}.

$$\mathrm{PM_{2.5}} = 14.58 + 0.20\ \mathrm{PM_{10}} + 0.72\ \mathrm{Benzene} + 0.02\ \mathrm{NO_x} \tag{8}$$

$$\mathrm{PM_{2.5}} = 7.54 + 0.37\ \mathrm{PM_{10}} + 4.48\ \mathrm{Benzene} + 0.06\ \mathrm{NO_x} \tag{9}$$

$$\mathrm{PM_{2.5}} = 19.01 + 0.11\ \mathrm{PM_{10}} + 0.75\ \mathrm{Benzene} + 0.02\ \mathrm{NO_x} \tag{10}$$

Initial MLR models produced R^2 values as as 0.71, 0.31, and 0.27 for site-1, site-2, and site-3, respectively, indicating lower performance observed at site-2 and site-3. This reduced accuracy was attributed to data skewness, extreme pollution events, and potential measurement anomalies [46]. The presence of these factors may affect the accuracy of the model; thereby, the model was further improved using various techniques as discussed in Section 3.3.

3.3. Improvement of model using data refinement techniques

3.3.1. Removal of outliers

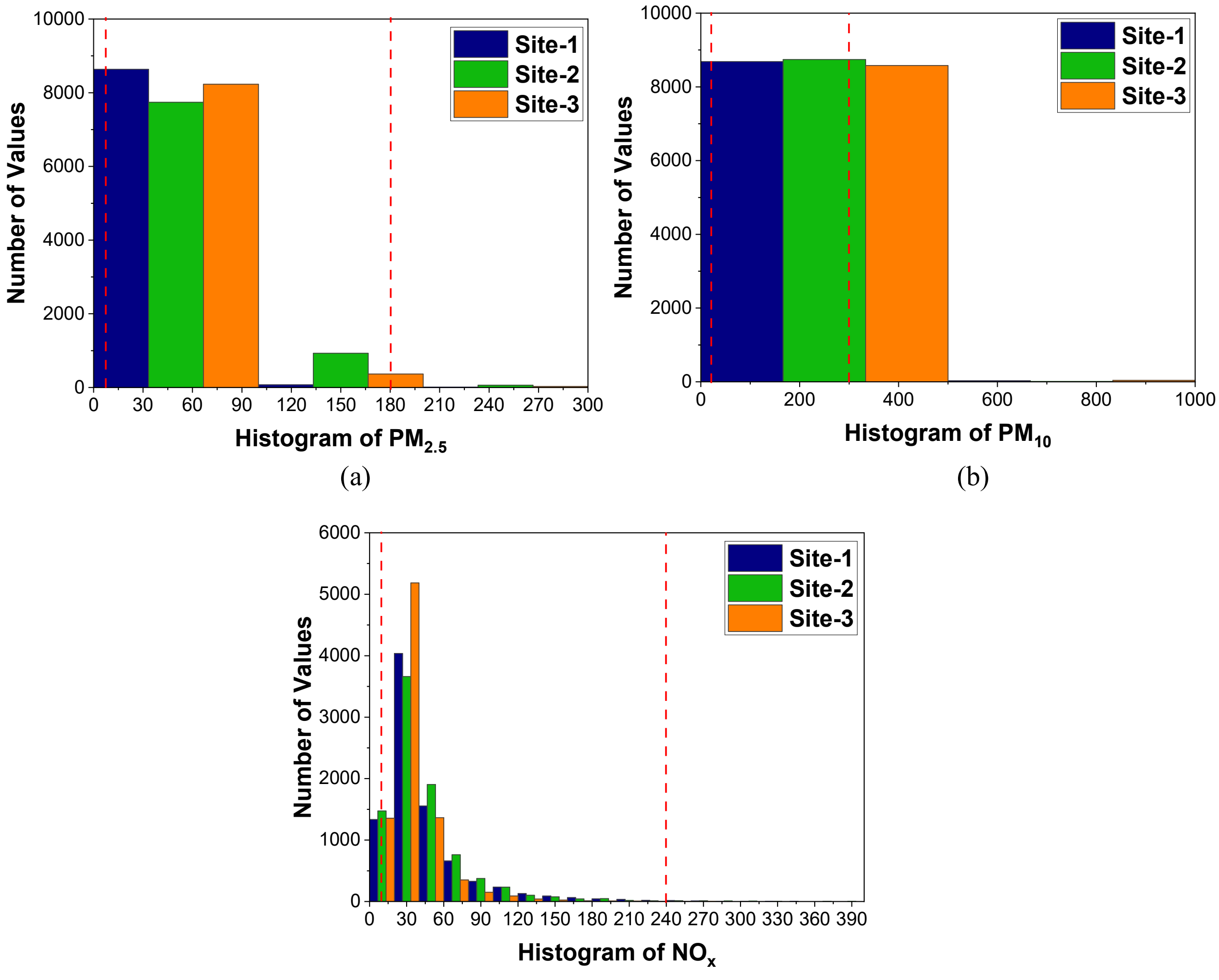
After eliminating the missing quantities from the dataset, the values of outliers were also removed to improve the model performance. After closely studying the dataset and considering the national ambient air quality standards [47], $PM_{2.5}$, PM_{10}, and NO_x concentrations with more than three times the standards were taken as the cutoff criterion for higher extreme values, while concentration values below 10 µg/m^3 for the pollutant were considered as a cutoff criterion for lower extreme values as shown in Figure 3. It is noteworthy that since the 24-hour average value for $PM_{2.5}$ is 60 µg/m^3, for PM_{10} is 100 µg/m^3, and for NO_x is 80 µg/m^3 as per the standards, therefore, 180 µg/m^3 for $PM_{2.5}$, 300 µg/m^3 for PM_{10}, and 240 µg/m^3 for NO_x were considered as the outliers on the higher extreme end and removed from the dataset. However, for benzene, the pollutant concentration was not varying significantly from its standard value, therefore, no outliers were detected.

The general statistics of R^2 values before and after the outlier's removal for $PM_{2.5}$ prediction are shown in Table 2. It can be observed from Table 2 that after the outlier's removal, the values of R^2 significantly improved for site-2 from 0.31 to 0.61 and for site-3 from 0.27 to 0.51, respectively. The new and improved MLR models were obtained after the outlier's removal, as represented in Equations 11-13 for site-1, site-2, and site-3, respectively.

$$\mathrm{PM_{2.5}} = 11.50 + 0.25\ \mathrm{PM_{10}} + 0.32\ \mathrm{Benzene} + 0.02\ \mathrm{NO_x} \tag{11}$$

$$\mathrm{PM_{2.5}} = 7.94 + 0.39\ \mathrm{PM_{10}} + 2.83\ \mathrm{Benzene} + 0.01\ \mathrm{NO_x} \tag{12}$$

$$\mathrm{PM_{2.5}} = 13.05 + 0.31\ \mathrm{PM_{10}} + 0.94\ \mathrm{Benzene} + 0.04\ \mathrm{NO_x} \tag{13}$$



(a)

(b)

(c)

Figure 3: Frequency distribution of for (a) $PM_{2.5}$, (b) PM_{10}, and (c) NO_x concentrations at the three monitoring sites showing upper and lower outlier thresholds with red-dashed line.

Table 2: Comparison of MLR model performance (R^2) before and after outlier removal for $PM_{2.5}$ prediction

Location	R^2 value before outliers removal	R^2 value after outliers removal
Site-1	0.71	0.71
Site-2	0.31	0.61
Site-3	0.27	0.51

3.3.2. Logarithmic data transformation

A second approach employing data transformation and data generation methods while including extreme values was also investigated for improving the MLR model's performance. In this case, the data was initially checked for skewness in the distribution, and in order to match it with the Gaussian distribution, logarithmic transformation was applied. After reviewing the skewness, the dataset was logarithmically normalized three times, and model efficiency was checked.

The model results were found to improve significantly after 3 times logarithmic normalization, and R^2 values were in the range of 0.62-0.74 for the three respective sites, as shown in Figure 4. It is noteworthy that after the dataset was logarithmically normalized three times, the value of R^2 significantly improved, especially for site-2 from 0.31 to 0.62 and for site-3 from 0.27 to 0.73, thereby reflecting the effectiveness of the method for enhancing model performance. This performance enhancement may be attributed to reduced skewness and improved model linearity, The improved MLR models were obtained after the logarithmic normalization as given in Equations 14-16 for site-1, site-2, and site-3, respectively.

$$\mathrm{PM_{2.5}} = 0.37 + 0.67\ \mathrm{PM_{10}} + 0.07\ \mathrm{Benzene} + 0.04\ \mathrm{NO_x} \tag{14}$$

$$\mathrm{PM_{2.5}} = 1.02 + 0.54\ \mathrm{PM_{10}} + 0.34\ \mathrm{Benzene} + 0.06\ \mathrm{NO_x} \tag{15}$$

$$\mathrm{PM_{2.5}} = 0.46 + 0.68\ \mathrm{PM_{10}} + 0.12\ \mathrm{Benzene} + 0.06\ \mathrm{NO_x} \tag{16}$$

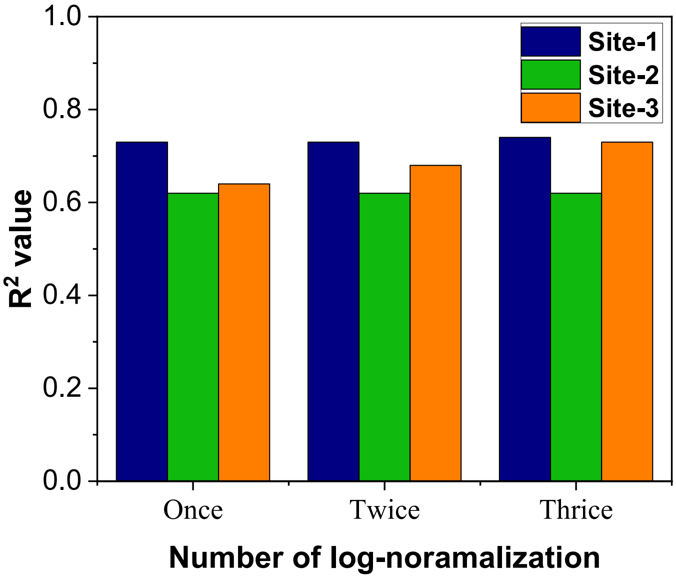


Figure 4: Variation in model performance (R^2) with successive logarithmic data transformations at the three monitoring sites.

3.3.3. Logarithmic data transformation followed by bootstrapping technique

After logarithmic normalization, use of bootstrapping technique, a resampling technique for data generation which involves drawing samples from the source data over and over with replacement, was also explored to further improve the performance of the MLR models. The general statistics of the R^2 values after the application of the three times logarithmic normalization followed by the bootstrapping technique are given in Figure 5. The results revealed that the highest model performance was obtained when logarithmic transformation was combined with bootstrapping, resulting in R^2 values of 0.77, 0.77, and 0.80 for site-1, site-2, and site-3, respectively. The improved and optimized MLR models, after applying the three times logarithmic data transformations followed by bootstrapping technique, are given in Equations 17, 18, and 19 for site-1, site-2, and site-3, respectively.

$$\mathrm{PM_{2.5}} = 0.35 + 0.69\ \mathrm{PM_{10}} + 0.06\ \mathrm{Benzene} + 0.04\ \mathrm{NO_x} \tag{17}$$

$$\mathrm{PM_{2.5}} = 0.98 + 0.56\ \mathrm{PM_{10}} + 0.34\ \mathrm{Benzene} + 0.04\ \mathrm{NO_x} \tag{18}$$

$$\mathrm{PM_{2.5}} = 0.49 + 0.68\ \mathrm{PM_{10}} + 0.12\ \mathrm{Benzene} + 0.05\ \mathrm{NO_x} \tag{19}$$

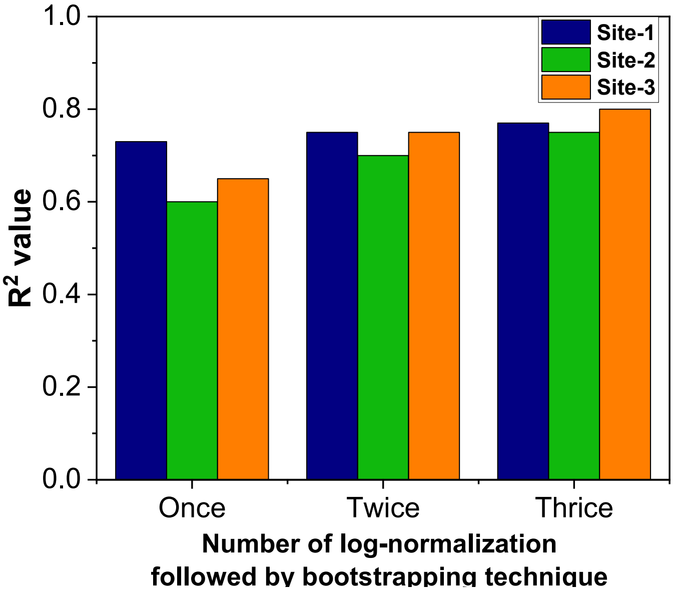


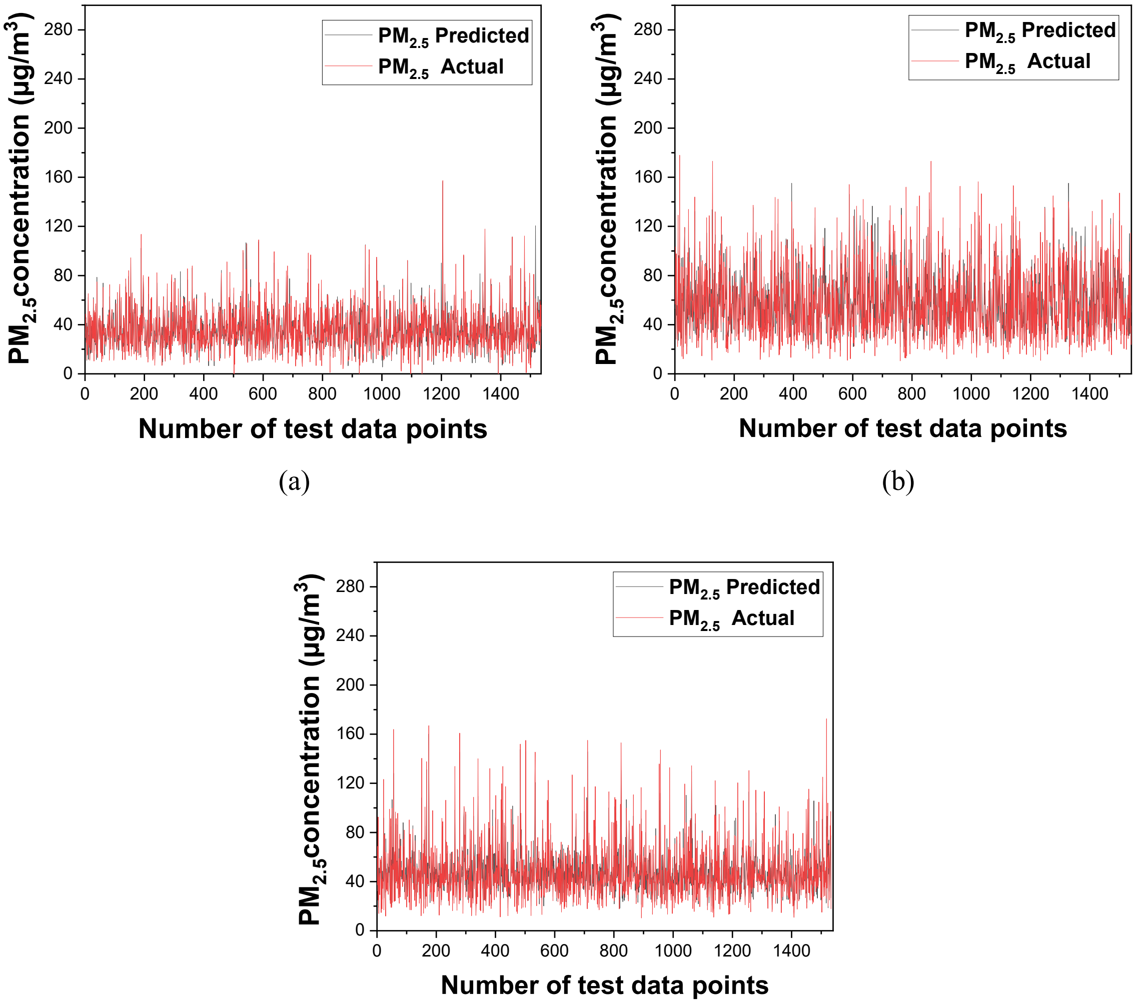
Figure 5: Improvement in model performance (R^2) after logarithmic transformation combined with bootstrapping at the three monitoring sites.

3.4. Selection of the best model

When the MLR model was applied, the initial results of R^2 varied from 0.27 to 0.71 for the three sites of Jaipur for $PM_{2.5}$ prediction. Hence, various data refinement techniques were explored to improve model performance. Firstly, the outlier's removal technique was analyzed, and the results displayed R^2 values from 0.51 to 0.71 at the three locations. Secondly, the extreme values were included, and data was three times logarithmically normalized, and the results exhibited significant improvement with R^2 values from 0.62 to 0.74 for the three sites. Furthermore, when the log normalized data was bootstrapped, R^2 values again improved, ranging from 0.77 to 0.80 for the three sites, reflecting the effectiveness of the model's prediction for $PM_{2.5}$ concentrations for Jaipur City. Therefore, it can be said that the best MLR models can be given by Equations 17, 18, and 19 for site-1, site-2, and site-3, respectively.

3.5. Predicted vs observed $PM_{2.5}$ concentrations

After selecting the best model, curves have been plotted between predicted and measured values to validate the accuracy of the developed MLR models, as shown in Figure 6. Predicted and observed $PM_{2.5}$ concentrations showed strong agreement for all the three monitoring sites, confirming the robustness and reliability of the optimized MLR models.



(a)

(b)

(c)

Figure 6: Comparison between observed and predicted $PM_{2.5}$ concentrations using optimized MLR models at (a) site-1, (b) site-2, and (c) site-3.

4. Conclusions

This study successfully developed a machine learning–based multiple linear regression framework for predicting $PM_{2.5}$ concentrations in Jaipur using hourly air quality data. PM_{10}, NO_x, and benzene were identified as significant predictors through correlation analysis. Initial model performance was substantially improved by applying outlier removal, logarithmic transformation, and bootstrapping techniques. The optimized models achieved strong predictive accuracy, with R^2 values ranging from 0.77 to 0.80 across three urban monitoring locations. The close agreement between predicted and observed values validates the effectiveness of the proposed approach. The findings highlight the importance of data refinement strategies in enhancing air quality prediction models. The developed methodology is computationally efficient, interpretable, and easily transferable to other urban environments with similar data availability. The results can support air quality management, policymaking, and exposure mitigation strategies in Indian cities. However, the model is developed using single-year air quality data and limited input parameters, which may restrict its ability to capture long-term and meteorological influences on $PM_{2.5}$ concentrations. Thus, future research can improve prediction accuracy by integrating multi-year datasets, meteorological variables, and advanced machine learning techniques for real-time $PM_{2.5}$ forecasting

Acknowledgments

The authors gratefully acknowledge the Rajasthan State Pollution Control Board (RSPCB) for providing access to the continuous ambient air quality monitoring data used in this study.

Conflict-of-interest Statement

The authors declare that there are no known financial or personal relationships that could have influenced the work reported in this paper.

References

[1] Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, AlMazroa MA, Amann M, Anderson HR, Andrews KG, Aryee M. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet*. 2012; 380:2224-60.
https://doi.org/10.1016/S0140-6736(12)61766-8

[2] Kumar P, Jain S, Gurjar BR, Sharma P, Khare M, Morawska L, Britter R. New directions: can a "blue sky" return to Indian megacities?. *Atmospheric Environment*. 2013;71:198-201.
https://doi.org/10.1016/j.atmosenv.2013.01.055

[3] Mittal AK, Van Grieken R. Health risk assessment of urban suspended particulate matter with special reference to polycyclic aromatic hydrocarbons: a review. *Reviews on environmental health*. 2001;16:169-90.
https://doi.org/10.1515/REVEH.2001.16.3.169

[4] Sharma R, Shilimkar G and Pisal S. Air Quality Prediction by Machine Learning. *Int. J. Sci. Res. Sci. Technol.* 2021;486–92.
https://doi.org/10.32628/IJSRST218396

[5] Langrish JP, Li X, Wang S, Lee MM, Barnes GD, Miller MR, Cassee FR, Boon NA, Donaldson K, Li J, Li L. Reducing personal exposure to particulate air pollution improves cardiovascular health in patients with coronary heart disease. *Environmental health perspectives*. 2012;120:367-72.
https://doi.org/10.1289/ehp.1103898

[6] Manojkumar N, Srimuruganandam B. Age-specific and seasonal deposition of outdoor and indoor particulate matter in human respiratory tract. *Atmospheric Pollution Research*. 2022;13:101298.
https://doi.org/10.1016/j.apr.2021.101298

[7] Shende P, Qureshi A. Burden of diseases in fifty-three urban agglomerations of India due to particulate matter (PM 2.5) exposure. *Environmental Engineering Research*. 2022;27.
https://doi.org/10.4491/eer.2021.042

[8] Davidson CI, Phalen RF, Solomon PA. Airborne particulate matter and human health: a review. *Aerosol*

Science and Technology. 2005;39:737-49.
https://doi.org/10.1080/02786820500191348

[9] Guo, B., Guo, Y., Nima, Q., Feng, Y., Wang, Z., Lu, R., Ma, Y., Zhou, J., Xu, H., Chen, L. and Chen, G. Exposure to air pollution is associated with an increased risk of metabolic dysfunction-associated fatty liver disease. *Journal of hepatology*. 2022;*76*:518-25.
https://doi.org/10.1016/j.jhep.2021.10.016

[10] Feng S, Gao D, Liao F, Zhou F, Wang X. The health effects of ambient PM2.5 and potential mechanisms. *Ecotoxicology and environmental safety*. 2016;128:67-74.
https://doi.org/10.1016/j.ecoenv.2016.01.030

[11] Xing YF, Xu YH, Shi MH, Lian YX. The impact of PM2. 5 on the human respiratory system. *Journal of thoracic disease.* 2016;8:E69.
https://doi.org/10.3978/j.issn.2072-1439.2016.01.19

[12] Lewis TC, Robins TG, Dvonch JT, Keeler GJ, Yip FY, Mentz GB, Lin X, Parker EA, Israel BA, Gonzalez L, Hill Y. Air pollution–associated changes in lung function among asthmatic children in Detroit. *Environmental Health Perspectives.* 2005;113:1068-75.
https://doi.org/10.1289/ehp.7533

[13] Ain NU, Qamar SU. Particulate matter-induced cardiovascular dysfunction: a mechanistic insight. *Cardiovascular Toxicology.* 2021;21:505-16.
https://doi.org/10.1007/s12012-021-09652-3

[14] Bhatnagar A. Environmental cardiology: studying mechanistic links between pollution and heart disease. *Circulation research*. 2006;99:692-705.
https://doi.org/10.1161/01.RES.0000243586.99701.cf

[15] Shahrbaf MA, Akbarzadeh MA, Tabary M, Khaheshi I. Air pollution and cardiac arrhythmias: A comprehensive review. *Current Problems in Cardiology*. 2021;46:100649.
https://doi.org/10.1016/j.cpcardiol.2020.100649

[16] Kumar S, Mishra S, Singh SK. Deep Transfer Learning-based COVID-19 prediction using Chest X-rays. *Journal of Health Management.* 2021;23:730-46.
https://doi.org/10.1177/09720634211050425

[17] Harishkumar KS, Yogesh KM, Gad I. Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. *Procedia Computer Science*. 2020;171:2057-66.
https://doi.org/10.1016/j.procs.2020.04.221

[18] Kumar S, Mishra S, Singh SK. A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere. *Heliyon*. 2020;6:e05618.
https://doi.org/10.1016/j.heliyon.2020.e05618

[19] Lv L, Wei P, Li J, Hu J. Application of machine learning algorithms to improve numerical simulation prediction of PM2. 5 and chemical components. *Atmospheric Pollution Research.* 2021;12:101211.
https://doi.org/10.1016/j.apr.2021.101211

[20] Yin S, Liu H, Duan Z. Hourly PM2. 5 concentration multi-step forecasting method based on extreme learning machine, boosting algorithm and error correction model. *Digital Signal Processing.* 2021;118:103221.
https://doi.org/10.1016/j.dsp.2021.103221

[21] Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electronic Markets.* 2021;31:685-95.
https://doi.org/10.1007/s12525-021-00475-2

[22] Nichols JA, Herbert Chan HW, Baker MA. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical reviews*. 2019;11:111-8.
https://doi.org/10.1007/s12551-018-0449-9

[23] Castelli M, Clemente FM, Popovič A, Silva S, Vanneschi L. A machine learning approach to predict air quality in California. *Complexity*. 2020.

https://doi.org/10.1155/2020/8049504
[24] Kang GK, Gao JZ, Chiao S, Lu S, Xie G. Air quality prediction: Big data and machine learning approaches. *Int. J. Environ. Sci. Dev*. 2018;9:8-16.
https://doi.org/10.18178/ijesd.2018.9.1.1066
[25] Gryech I, Ghogho M, Elhammouti H, Sbihi N, Kobbane A. Machine learning for air quality prediction using meteorological and traffic related features. *Journal of Ambient Intelligence and Smart Environments*. 2020;12:379-91.
https://doi.org/10.3233/AIS-200572
[26] Dangayach, R., Pandey, M., Gusain, D., Srivastav, A.L., Jain, R., Bairwa, B.M. and Pandey, A.K. Assessment of Air Quality Before and During COVID-19-Induced Lockdown in Jaipur, India. *MAPAN*. 2023;1-11.
https://doi.org/10.1007/s12647-022-00615-9
[27] Mao W, Wang W, Jiao L, Zhao S, Liu A. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustainable Cities and Society*. 2021;65:102567.
https://doi.org/10.1016/j.scs.2020.102567
[28] Liu Y, Mu Y, Chen K, Li Y, Guo J. Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Processing Letters*. 2020;51:1771-87.
https://doi.org/10.1007/s11063-019-10185-8
[29] Li M, Wang WL, Wang ZY, Xue Y. Prediction of PM2. 5 concentration based on the similarity in air quality monitoring network. *Building and Environment*. 2018;137:11-7.
https://doi.org/10.1016/j.buildenv.2018.03.058
[30] Singh, H. and Bawa, S. Predicting COVID-19 statistics using machine learning regression model: Li-MuLi-Poly. *Multimedia Systems*. 2022;28:113-20.
https://doi.org/10.1007/s00530-021-00798-2
[31] Lv L, Wei P, Li J, Hu J. Application of machine learning algorithms to improve numerical simulation prediction of PM2. 5 and chemical components. *Atmospheric Pollution Research.* 2021;12:101211.
https://doi.org/10.1016/j.apr.2021.101211
[32] Ciulla G, D'Amico A. Building energy performance forecasting: A multiple linear regression approach. *Applied Energy*. 2019;253:113500.
https://doi.org/10.1016/j.apenergy.2019.113500
[33] Li M, Wang WL, Wang ZY, Xue Y. Prediction of PM2. 5 concentration based on the similarity in air quality monitoring network. *Building and Environment*. 2018; 137:11-7. https://doi.org/10.1016/j.buildenv.2018.03.058
[34] Kleine Deters J, Zalakeviciute R, Gonzalez M, Rybarczyk Y. Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters. *Journal of Electrical and Computer Engineering*. 2017.
https://doi.org/10.1155/2017/5106045
[35] Wang H, Ma C, Zhou L. A brief review of machine learning and its application. In2009 international conference on information engineering and computer science. *IEEE*. 2009;1-4.
https://doi.org/10.1109/ICIECS.2009.5362936
[36] Bai W, Li F. PM2. 5 concentration prediction using deep learning in internet of things air monitoring system. *Environmental Engineering Research*. 2023;28.
https://doi.org/10.4491/eer.2021.456
[37] Maniruzzaman M, Rahman M, Al-MehediHasan M, Suri HS, Abedin M, El-Baz A, Suri JS. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *Journal of medical systems*. 2018;42:1-7.
https://doi.org/10.1007/s10916-018-0940-7
[38] Zhang Z, Liu J, Wang L, Guo G, Zheng X, Gong X, Yang S, Huang G. An enhanced smartphone indoor positioning scheme with outlier removal using machine learning. Remote Sensing. 2021;13:1106.

https://doi.org/10.3390/rs13061106

[39] Yang J, Rahardja S, Fränti P. Outlier detection: how to threshold outlier scores?. In Proceedings of the international conference on artificial intelligence, information processing and cloud computing; 2019;1-6.

[40] Curran-Everett D. Explorations in statistics: the log transformation. *Advances in physiology education*. 2018;42:343-7.
https://doi.org/10.1152/advan.00018.2018

[41] Fyodorov YV, Doussal PL. Moments of the position of the maximum for GUE characteristic polynomials and for log-correlated Gaussian processes. *Journal of Statistical Physics*. 2016;164:190-240.
https://doi.org/10.1007/s10955-016-1536-6

[42] Al-Najjar HA, Pradhan B, Kalantar B, Sameen MI, Santosh M, Alamri A. Landslide susceptibility modeling: An integrated novel method based on machine learning feature transformation. *Remote Sensing*. 2021;13:3281.
https://doi.org/10.3390/rs13163281

[43] Saraiva SV, de Oliveira Carvalho F, Santos CA, Barreto LC, Freire PK. Daily streamflow forecasting in Sobradinho Reservoir using machine learning models coupled with wavelet transform and bootstrapping. *Applied Soft Computing*. 2021;102:107081.
https://doi.org/10.1016/j.asoc.2021.107081

[44] Wu Y, Liu Y, Li J, Liu H, Hu X. Traffic sign detection based on convolutional neural networks. In The 2013 international joint conference on neural networks (IJCNN). *IEEE*. 2013;1-7.
https://doi.org/10.1109/IJCNN.2013.6706811.

[45] Tiwari MK, Adamowski JF. Medium-term urban water demand forecasting with limited data using an ensemble wavelet–bootstrap machine-learning approach. *Journal of Water Resources Planning and Management*. 2015;141:04014053.
https://doi.org/10.1061/(ASCE)WR.1943-5452.0000454

[46] Kang GK, Gao JZ, Chiao S, Lu S, Xie G. Air quality prediction: Big data and machine learning approaches. *Int. J. Environ. Sci. Dev*. 2018;8-16.
https://doi.org/10.18178/ijesd.2018.9.1.1066

[47] Central Pollution Control Board, (Ministry of Environment, Forests & Climate Change), Govt. of India 2019 *National Ambient Air Quality Status & Trends.* 2019;104.
https://cpcb.nic.in/upload/NAAQS_2019.pdf