Applied Ecology and Environmental Sciences, 2022, Vol. 10, No. 12, 717-722
Available online at http://pubs.sciepub.com/aees/10/12/3
Published by Science and Education Publishing
DOI:10.12691/aees-10-12-3

Estimation of Air Quality Index Using Multiple Linear Regression

A. Loganathan*, P. Sumithra, V. Deneshkumar

Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India
*Corresponding author: loganathan@msuniv.ac.in

Received October 22, 2022; Revised November 27, 2022; Accepted December 12, 2022

Abstract Air quality index is a numerical measure, which is computed to determine the air quality for various geographical locations. Human activities, industry functioning and climate conditions are some of the significant factors causing variations in the air quality index. Many methods are proposed in the literature and are applied to develop models for investigating the changing behaviour of the air quality index. Among them, regression model is a statistical tool, possessing established properties, recommended frequently for static data. This paper considers estimation of a multiple linear regression model based on the information pertaining to air quality index recorded in a monitoring station located in Chennai, India. Significance of the model to the data is detailed and residual analysis is carried out for testing validity of the fitted model.

Keywords: *regression analysis, air quality index*

Cite This Article: A. Loganathan, P. Sumithra, and V. Deneshkumar, "Estimation of Air Quality Index Using Multiple Linear Regression." *Applied Ecology and Environmental Sciences*, vol. 10, no. 12 (2022): 717-723. doi: 10.12691/aees-10-12-3.

1. Introduction

Analysis of air quality is essential for all locations in the globe, as it varies over time due to presence of chemical pollutants in the atmosphere and changing meteorological factors. National Ambient Air Quality Standards (NAAQS) have reported that, in general, twelve chemical pollutants *viz., sulphur dioxide, nitrogen dioxide, particulate matter* (size<2.5μm and size <10 μm), *ozone, lead, carbon monoxide, ammonia, benzene, benzo, arsenic* and *nickel* affect the air quality. Similarly, the meteorological factors like wind direction, wind speed, relative humidity and solar radiation make-changes in the air quality. Humans are susceptible to many serious diseases caused by air pollution, including respiratory infections, heart disease and lung cancer. Air quality can be assessed numerically by air quality index (AQI).

The Central Pollution Control Board (CPCB) of India categorised the air quality of a geographical location, based on AQI, into six classes *viz.*, *good* (0 - 50), *satisfactory* (51 - 100), *moderate* (101-200), *poor* (201-300), *very poor* (301-400) and *severe* (401-500). Estimation of models for determining the AQI corresponding to various levels of the chemical pollutants or for future time period is an integral part of AQI analysis.

Kumar and Goyal [1] proposed prediction of AQI for four seasons *viz.*, summer, monsoon, post-monsoon and winter of Delhi applying principal component regression (PCR) method. They found that PCR performs better for winter compared to other seasons. Kumar and Goyal [2] introduced an approach for forecasting the air quality index for Delhi, which integrates PCR and ARIMA models. They recommended the integrated PCR and ARIMA models for forecast accuracy. Ganesh *et al.* [3] have conducted a case study to predict AQI for Delhi and Houston, applying support vector regression model along with training algorithms for batch gradient descent, stochastic gradient descent, and mini-batch gradient descent. They have observed that multiple linear regression model with mini-batch gradient descent algorithm outperforms other linear regression models in predicting AQI. Recently, Abdullah *et al.* [4] have discussed finding stepwise multiple linear regression model to predict the *PM*10 concentration in the air. They have created three distinct prediction hours to formulate deterministic predictions for *PM*10.

Madan *et al.* [5] evaluated the performance of twenty different models of machine learning (ML) algorithms. They found that reinforcement and neural network models perform relatively better than the other ML algorithms. Kumar and Pande [6] have pointed out that AQI monitoring and forecasting have become essential and challenging especially for urban areas. Several ML models have been used to predict AQI.

Apart from forecasting accuracy, the fitted models are expected to possess some desirable properties. It is well known that statistical model building methods have been developed with a basis of verifiable theoretical reasoning. Appropriateness of the models can be investigated through the properties of the estimates concerned and model adequacy tests. Main objective of this paper is to construct a multiple regression model for AQI applying the

statistical procedures, which is adequate to study the influence of the chemical pollutants upon the changes in AQI. In this respect, a multiple regression model is constructed for AQI based on the information collected from the CPCB monitoring station located in Velachery, Chennai.

Section 2 describes the data and assesses the presence and strength of relationship of AQI with chemical pollutants and meteorological factors. The procedure to be followed for estimation of the model is also discussed. Results pertaining to estimation of the model and testing adequacy of the model are analysed in Section 3. Findings are summarized in Section 4.

2. Materials and Methods

2.1. Data and Its Description

The CPCB of India has an air quality monitoring station at Velachery, which is one of the largest commercial and residential areas in southern part of the Chennai, India.

Figure 1. Velachery Geographical location (Source: Google earth)

The station monitors wider area of the region including approximately 6.17 *sq. kms* of moderately dense forest, 1.301634 *sq.kms* of green cover of Indian Institute of Technology, Guindy and 55 *sq. kms* of marshland at Pallikaranai [7]. Figure 1 displays the satellite picture of Velachery.

The CPCB monitoring station collects and maintains the daily record of nine predominant air pollutants of the region *viz.*, Particulate Matter (PM2.5$\mu g/m^3$), Nitrogen Monoxide (NO $\mu g/m^3$), Nitrogen Dioxide (NO_2 $\mu g/m^3$), Nitrogen oxide (NOx $\mu g/m^3$), Sulphur Dioxide (SO_2 $\mu g/m^3$), Carbon Monoxide (CO mg/m^3), Ozone (O_3 $\mu g/m^3$), Benzene ($\mu g/m^3$) and Toluene ($\mu g/m^3$). In addition, meteorological readings such as wind direction (WD), wind speed (WS), relative humidity (RH) and solar ration (SR) are also measured.

Information about the above chemical pollutants and the meteorological factors for a period of 813 days from 01-01-2018 to 23-03-2020 are considered in this study. Information about some of these characteristics are found missing for 117 days. Also, information observed for 39 days are found to be extreme in magnitude. According to Fox [8], missing values and outliers in the data fail to capture the essential characteristics of the data. In environment related data, missing values can occur due to several reasons, including manual data input methods, equipment failure and inaccurate measurements. Though many algorithms have been introduced to replace missing values, none of them is proved preferable to the others [9]. When missing data are not handled appropriately, they can cause bias and can lead to invalid conclusions. Here, the *listwise deletion* method is applied to remove the missing values [10].

Outliers in the data increase the error variance, reduce power of statistical tests and lead to biased estimates. Therefore, the outlier detection process is an essential aspect of data analysis. As pointed by Ghorbani [11], Mahalanobis distance is used as a tool to detect the outliers, and these detected outlier observations have been dropped from the data. After removal of the missing values and outliers, information about the air pollutants and metrological factors are available for 657 days.

2.2. Characteristics of the Study Variables

The descriptive statistical measures calculated for the above-mentioned data are presented in Table 1. The AQI ranges in various levels from *good* (21.5) to *moderate* (184.8), with an average level of *satisfactory* (60.6). It is important to note that the air quality of Velachery has never been *poor* on any day. The average and minimum level concentrations of (*PM2.5*, *NO*, NO_2, *NOx*, SO_2, *CO* and O_3) are respectively (34.1, 8.0, 14.0, 18.8, 4.8, 0.8 and 29.4) and (4.6, 0.3, 1.2, 3.4, 0.9, 0.0 and 4.3). These are smaller than the concentration levels prescribed by National Ambient Air Quality Standard (NAAQS, 2009) guideline daily thresholds [12]. The maximum concentration levels of PM2.5 (123.3 $\mu g/m^3$) and O_3 (100.4 $\mu g/m^3$) are larger than the NAAQS 2009 guideline daily thresholds of 60 $\mu g/m^3$ and 100 $\mu g/m^3$ respectively.

Table 1. Descriptive Statistical Measures of AQI, Chemical Pollutants and Meteorological Characteristics

	PM2.5	NO	NO2	NOx	SO2	CO	O_3	Benzene	Toluene	RH	WD	SR	AQI
Mean	34.1	8.0	14.0	18.8	4.8	0.8	29.4	0.5	2.9	59.8	161.7	204.5	60.6
SD	18.6	4.4	6.4	7.2	1.4	0.2	14.6	1.4	3.7	13.8	55.0	67.5	27.0
Minimum	4.6	0.3	1.2	3.4	0.9	0.0	4.3	0.0	0.0	17.0	15.0	15.0	21.5
Quartile 1	20.0	4.3	10.0	13.6	4.0	0.6	18.4	0.0	0.0	50.0	125.0	166.0	38.0
Quartile 2	30.5	7.5	13.6	18.2	4.6	0.8	26.6	0.0	1.6	59.0	169.0	208.0	54.5
Quartile 3	45.4	10.4	17.5	23.2	5.3	0.9	37.3	0.4	4.4	69.0	211.0	250.0	75.3
Maximum	123.3	27.2	41.4	51.7	11.8	1.6	100.4	11.3	25.4	97.0	295.0	579.0	184.8

Since the AQI is determined from the records of the chemical pollutants, it is obvious that AQI depends upon the levels of chemical pollutants. Also, strength of relationship may vary. Koutsoyiannis [13] mentioned that correlation analysis determines the presence and strength of association among the study variables.

The chemical pollutants *NO*, NO_2, *NOx*, *Benzene*, *Toluene* and the meteorological factors are found to have very poor, in other words no, correlation with AQI. Level of correlation between AQI with *PM2.5* is high (0.85); and is moderate with *CO*, O_3 and SO_2 (0.28, 0.34, and 0.21). Figure 2 exhibits the level of correlation between each chemical pollutant with AQI. Since this study attempts to determine a regression model for AQI with appropriate regressors, information about *PM2.5*, *CO*, O_3 and SO_2 only are considered to fit the model.

A linear regression model can be fitted, only when the response variable has linear or approximately linear relationship with the regressors [14]. Also, the response variable is required to be distributed according to a normal distribution. If required, some suitable transformation may be applied to obtain a normal distribution to the response variable. When data have a large standard deviation compared to its respective mean, a logarithmic transformation affects dampening variability, reducing asymmetry and removing heteroscedasticity [15]. In this study, logarithmic transformation (*log*(x)) is applied to AQI and the regressors *PM2.5*, *CO*, O_3 and SO_2.

The diagonal cells in the pair plot of *log* (AQI) and the four pollutants in Figure 3, displays the histogram of the respective variable. The off-diagonal cells display the scatter diagram of each pair of study variables. It can be noted from the histogram displayed in the first diagonal cell that the distribution of *log* (AQI) may be approximately normal. Also, *log* (AQI) has approximate linear relationship with *CO*, *PM2.5* and O_3. More scatteredness of the values (log (AQI), SO_2) can be observed in Figure 3. However, it is assumed that *log* (AQI) has approximate, may be poor, linear relationship with SO_2. Also, the scatter diagrams corresponding to each pair of the four chemical pollutants indicate that the regressors may be uncorrelated.

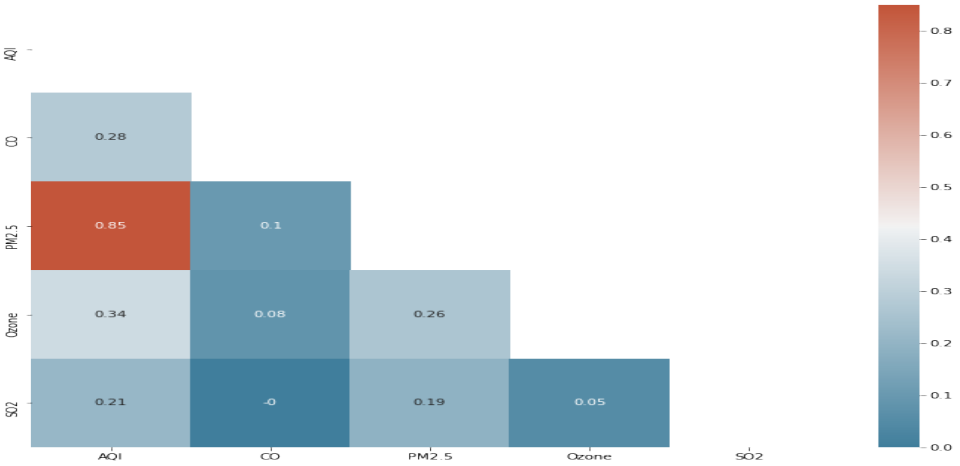

Figure 2. Correlation of AQI with *PM2.5*, *CO*, O_3 and SO_2

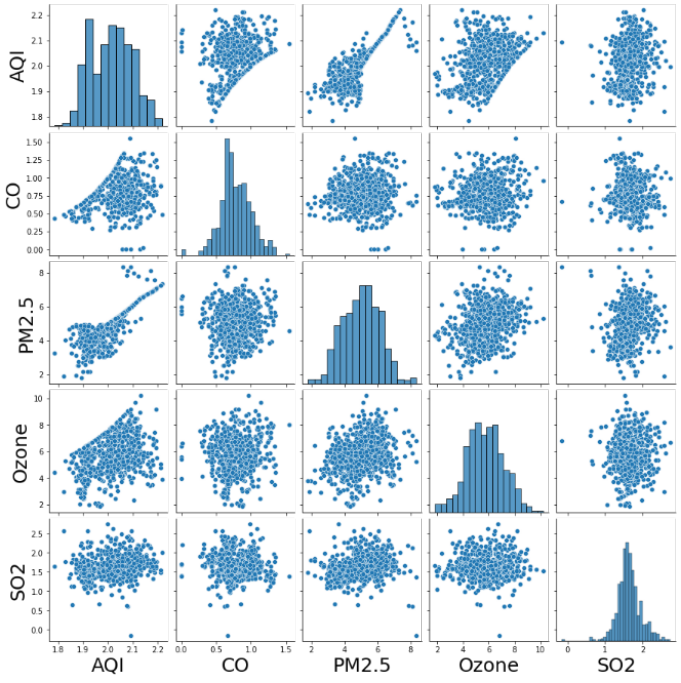

Figure 3. Pair plot

2.3. Model Estimation

The general form of the multiple linear regression model is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon$$

where
Y represents the response variable,
$X_1, X_2, \ldots, X_k$ are the regressors,
$\beta_1, \beta_2, \ldots, \beta_k$ are the regression co − efficients,
β_0 is the intercept term, and
ε denotes the error component.

Estimation of the model parameters $\beta_1, \beta_2, \ldots, \beta_k$, for given information about Y and X's is termed as fitting or estimation of the model. When ε is distributed according to a normal distribution with zero mean and constant variance, estimators of the model parameters can be obtained applying the maximum likelihood (ML) method or using the ordinary least squares (OLS) method. It is further assumed that the values of the residuals can be computed from the estimated model using

$$\varepsilon_i = Y_i - \hat{Y}_i, i = 1, 2, ..., n$$

where $\hat{Y}_i$ is the estimate of Y for the i^{th} sample.

Here, it is proposed to obtain estimators for the model parameters, based on the information on *log* (AQI), *log* (*PM*2.5), *log* (*CO*), *log* (O_3) and *log* (SO_2) for n = 657 days. Then, on successful investigation of model appropriateness, residual analysis will be performed towards analysing the validity of the model.

3. Results and Discussion

The OLS estimates of the model parameters are obtained using SPSS *ver*20. The ANOVA table displaying the results of testing the overall significance of the fitted model is exhibited in Table 2. The measures calculated for studying the overall fitness of the model are presented in Table 3. The OLS estimates of the model parameters and significance test results for each of the four chemical pollutants are displayed in Table 4.

The estimated multiple linear regression model for AQI of Velachery is

$$log(AQI) = 1.602 + 0.074 log(CO) + 0.058 log(PM2.5) + 0.007 log(O_3) + 0.016 log(SO_2)$$

The overall significance of this fitted model can be observed from Table 2 through the F-statistic value of 594.4, and the corresponding p-value of 0.0. These values indicate that the estimated model is significant for studying Velachery's AQI data. Also, the *Adjusted* R^2 value shows that 78.3 % of variation in the values of AQI are due to variations in the levels of *PM*2.5, *CO*, O_3 and SO_2. Standard errors of the OLS estimates of the co-efficients of *log* (*CO*), *log* (*PM*2.5), *log*(O_3) and *log* (SO_2) can be noted from Table 3 as 0.007, 0.001, 0.001 and 0.005 respectively. These measures point out that the corresponding OLS estimates, which can also be observed from the corresponding 95% confidence limits. The p-values calculated for testing the significance of each of these chemical pollutants are 0.00 except SO_2 (0.002), which also ensures the relevance of these components in estimating *log* (AQI) from this fitted model. It may also be recalled that the OLS estimates of the model parameters possess the desirable properties explained in Koutsoyiannis [13].

Residuals are calculated using the estimated model for five days and are presented in Table 5.

Table 2. ANOVA Table

	Sum of Squares	df	Mean Square	F	*p*
Regression	3.580	4	.895	594.414	.000
Residual	.982	652	.002		
Total	4.562	656			

Table 3. Overall Significance of the Fitted Model

R^2: 0.994	Adj.R^2: 0.783

Table 4. OLS Estimates, SE, 95% Confidence Limits and Significance of Model Components

Components	Coefficients	SE	t	*p*	95% Confidence Limits	
					Lower Bound	Upper Bound
(Constant)	1.602	.012	136.396	.000	1.579	1.625
log (*CO*)	.074	.007	10.617	.000	.060	.088
log (*PM*2.5)	.058	.001	41.390	.000	.055	.061
log (O_3)	.007	.001	6.117	.000	.005	.009
log (SO_2)	.016	.005	3.119	.002	.006	.026

Table 5. Residuals

S. No.	Observed *log* (AQI)	*log* (*CO*)	*log* (*PM2.5*)	*log*(O_3)	*log* (SO_2)	Estimated *log* (AQI)	Residual
1.	2.14	0.02	6.61	6.65	2.197	2.069	0.071
2.	2.16	0.78	6.803	7.127	2.186	2.13	0.03
3.	2.11	0.84	6.272	7.398	2.065	2.11	0.00
4.	2.14	0.76	6.655	7.352	2.317	2.13	0.01
5.	2.09	0.75	5.946	7.614	2.113	2.08	0.01

The above computations show that the magnitudes of the residual are very small.

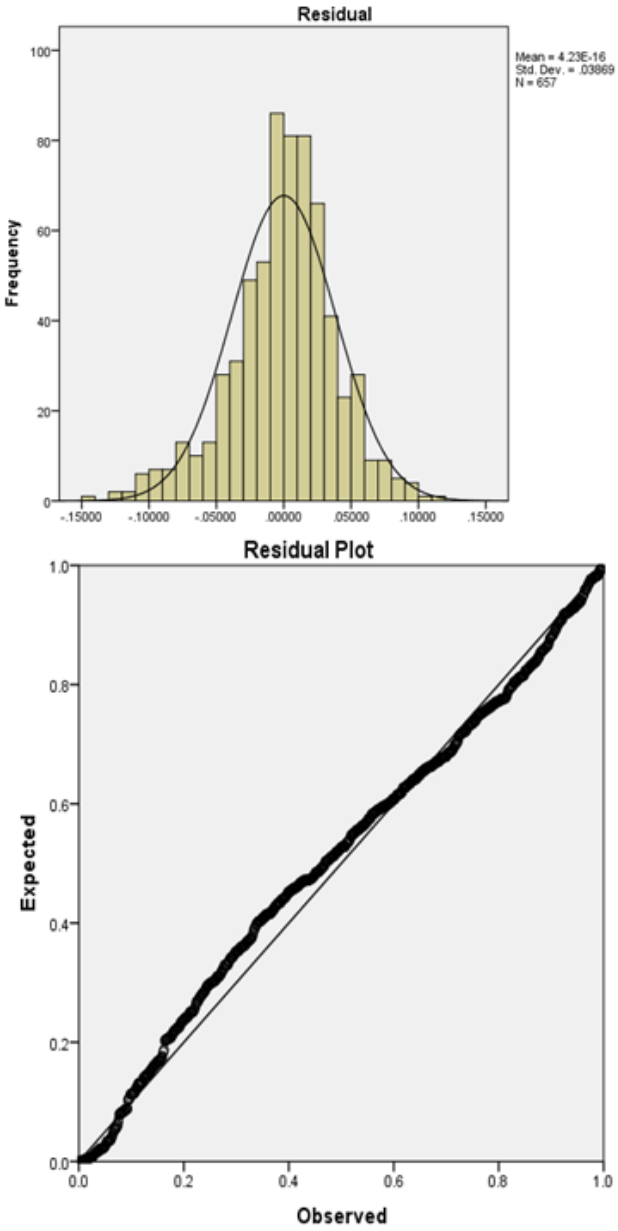

Figure 4. Residual Plot

The descriptive statistical measures for the residuals are computed for ~~of~~ all the 657 days as

$$Residual = Observed\log(AQI) - Estimated\log(AQI)$$

and the values are presented in Table 6. The residuals vary from -0.1405 to 0.1191 with shorter range of 0.2596. The quartiles are equidistant. The mean is, approximately, zero with standard deviation of 0.0387. Differences among mean, median and mode are marginal. The co-efficients of skewness and kurtosis are respectively -0.445 and 0.809. These numerical observations lead to diagnose that the distribution of the residuals can be a normal distribution with zero mean. The diagnosis can be justified from the histogram and the *P-P* plot drawn for residuals and presented in Figure 4. Moreover, the Anderson-Darling test statistic value is 0.4918 with significant *p*-value of 0.218. This ensures that the residual values fit to a normal distribution with zero mean. Thus, the estimated model can be regarded ~~considered~~ satisfying the assumptions considered for model construction.

It can be noted further from the estimated model that the estimate of AQI corresponding to the absence of the four chemical pollutants can be determined as

$$\log(AQI) = 1.602$$

$$AQI = 4.9629.$$

Also, contribution of the chemical pollutants in determining the AQI are not equal. The role of *CO* is relatively more in estimating AQI. The 7.4% of every unit change in the value of *log* (*CO*), 5.8% of *log* (*PM*2.5), 1.6% of *log* (SO_2) and 0.7% of *log* (O_3) influence one unit change in *log* (AQI).

Table 6. Descriptive Statistics of Residuals

Mean	0E-7
Mode	-.1405
Std. Deviation	.0387
Co-efficient of Skewness	-.445
Co-efficient of Kurtosis	.809
Range	.2596
Minimum	-.1405
Maximum	.1191
Quartile 1	-.0205
Quartile 2	.0026
Quartile 3	.0242

4. Remarks and Conclusions

AQI is a numerical indicator of the air quality of the region concerned. In general, air quality is affected by some chemical pollutants and meteorological factors, which may vary with respect to environment and geographical location of the region. Identifying such factors can be useful for planning to initiate preventive steps in that region. Statistical model constructed for AQI can be used for studying the influence of such factors. A statistical model is constructed in this work for AQI applying the regression modelling procedure to the information obtained from CPCB monitoring station in Velachery, Chennai city. Influence of individual chemical pollutants in the atmosphere and meteorological factors, which determine Velachery's air quality, are examined. It is found from the data that meteorological factors do not have a considerable impact on air quality of Velachery. Analysis of information about the chemical pollutants showed that *CO*, *PM2.5*, O_3 and SO_2 have strong association with AQI of the region compared to the other five predominant chemical pollutants *viz.*, *NO*, NO_2, *NOx*, *Benzene* and *Toluene*. A multiple linear regression model is fitted to *log* (AQI) applying the OLS method. The statistical hypotheses tests ensure the model's fitness and significance of *CO*, *PM2.5*, O_3, and SO_2 in determining AQI. Residual values computed for some sets of observations made on these four chemical pollutants are very small, rather near zero. The fitted model is found adequate to analyze Velachery's AQI based on the chemical pollutants *CO*, *PM2.5*, O_3, and SO_2. These four

pollutants are dominant for increasing tendency of AQI in Velachery.

References

[1] Abdullah, S., Napi, N. N. L. M., Ahmed, A. N., Mansor, W. N. W., Mansor, A. A., Ismail, M., Ramly, Z. T. A. (2020). Development of Multiple Linear Regression for Particulate Matter (PM10) Forecasting during Episodic Transboundary Haze Event in Malaysia. Atmosphere, 11(3), 289.

[2] Chatterjee, S., Hadi, A. S. (2006). Regression Analysis by Example. John Wiley & Sons.

[3] Fox, J. (2011). Regression Diagnostics: An Introduction. Saga University Paper Series on Quantitative Applications in the Social Sciences, 07-079. Newbury Park, CA:Sage.

[4] Ganesh, S. S., Modali, S. H., Palreddy, S. R., Arulmozhivarman, P. (2017). Forecasting Air Quality Index using Regression Models: A case study on Delhi and Houston. International Conference on Trends in Electronics and Informatics (ICEI). IEEE. 248-254.

[5] Gargava, P., Shukla, V. K., Darbari, T. (2021). National Ambient Air Quality Status and Trends 2019. Central Pollution Control Board, Ministry of Environment, Forest and Climate Change, Government of India. https://cpcb. nic. in/upload/NAAQS_2019. pdf. Accessed 25th June.

[6] Ghorbani, H. (2019). Mahalanobis Distance and its Application for Detecting Multivariate Outliers. Facta Univ Ser Math Inform, 34(3), 583-95.

[7] Kang, H. (2013). The Prevention and Handling of the Missing Data. Korean journal of anesthesiology, 64(5), 402-406.

[8] Koutsoyiannis, A. (1977). Theory of Econometrics. 2nd edition, Palgrave MacMillan.

[9] Koushik, Janardhan. "From 50 Sq Km to Just Three in 30 Years: Chennai's Pallikaranai Marsh Is Just about to Vanish." The Indian Express, (20 Aug. 2019), indianexpress.com/article/cities/chennai/chennai-pallikaranai-marshland-report-madras-high-court-5919329.

[10] Kumar, A., Goyal, P. (2011). Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research, 2(4), 436-444.

[11] Kumar, A., Goyal, P. (2011a). Forecasting of daily air quality index in Delhi. Science of the Total Environment, 409(24), 5517-5523.

[12] Kumar, K., Pande, B. P. (2022). Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology, 1-16.

[13] Madan, T., Sagar, S., Virmani, D. (2020). Air Quality Prediction using Machine Learning Algorithms–A Review. Second International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). 140-145. IEEE.

[14] Mohamed Noor, N., Zainudin, M. L. (2009). A Review: Missing Value in Enviromental Data Sets. Second International Conference and Workshops on Basic and Applied Sciences & Regional Annual Fundamental Science Seminar.

[15] Osborne, J. W., Waters, E. (2003). Four Assumptions of Multiple Regression that Researchers Should Always Test. Practical Assessment, Research and Evaluation, 8(2). 1-5.