Original Article
Bulletin of Pure and Applied Sciences.
Mathematics & Statistics Vol.44 E, No.1.
January-June 2025: P.1-6

Print version ISSN 0970 6577
Online version ISSN 2320 3226
Available online at www.bpasjournals.com

Multiple Regression Analysis of the Air Quality Index Using Time Series

[1]Lokesh Kumar*, [2]Gaurav Kumar

Author Affiliation:
[1]Research Scholar, Department of Mathematics, NAS College, Meerut, Uttar Pradesh 250003, India.
[2]Professor, Department of Mathematics, NAS College, Meerut, Uttar Pradesh 250003, India.

***Corresponding Author: Lokesh Kumar,** Research Scholar, Department of Mathematics, NAS College, Meerut, Uttar Pradesh 250003, India.
E-mail: lokesh181889@gmail.com

ABSTRACT

One of the biggest problems in Meerut, Uttar Pradesh, has been air pollution. The components that make up air pollution include PM10, NO2, and SO2. Forecasts of these pollutants can be used to design a plan to reduce air pollution. Using data gathered by the U.P. Pollution Control Board in prior years, this article examines the air quality index of Ghaziabad's Khora Colony in Uttar Pradesh, considering air pollution. The analysis makes use of multiple linear regression (MLR). This approach uses time series analysis to provide us with approximate findings. Four data points are used at each stage, and the first one is ignored in favor of the following four in the series at each subsequent step. The AQI's future values can be somewhat predicted by examining its historical values since we obtain 47.9% of the variability in the independent factors. We discovered that this approach outperformed previous prediction techniques.

Keywords: Air Quality Index (AQI), Air Pollution, MLR.

How to cite this article: Kumar L., Kumar G. (2025). Multiple Regression Analysis of the Air Quality Index Using Time Series. *Bulletin of Pure and Applied Sciences- Math & Stat.*, 44E (1), 1-6.

Received on 21.02.2025, Revised on 11.04.2025, Accepted on 22.02.2025

1. INTRODUCTION

Among the most important environmental problems is poor air quality that urban areas face globally, particularly in developing nations (Boznar et al., 1993 & 2002). The topic of air pollution has drawn a lot of attention lately because of a significant rise in awareness of the potential dangers of chemical contaminants and how they affect human health and the environment. Air pollution is defined as prolonged exposure to one or more pollutants in the atmosphere in quantities that are harmful to people's, animals', or plants' health or wellbeing (Bhavsar, 2019). The discharge of hazardous substances into the atmosphere may also be referred to by this term.

Air pollution comes in a variety of forms. These compounds are known as "air pollutants". Freeman created methods for valuing a variety of environmental factors, such as air pollution (1974). The Air Pollution Control Board of India is monitoring NO2, SO2, PM10, and AQI. Using time series analysis, the authorities can develop a strategy to keep an eye on the AQI. We have looked at the AQI in this work. Time series are used to describe the research's temporal ordering. This is frequently utilized in scientific domains such as signal processing and statistics, but it may also be applied to environmental economics and financial forecasts. Numerous techniques, including regression and the ANN approach, may be used to evaluate time series. The time series model for AQI prediction in this work is developed using regression analysis.

2. METHODS

The following is one definition for the MLR model:

$$Z = \sum_{k=1}^{p} \gamma_k X_k + \gamma_0 + \varepsilon \qquad (1)$$

Here Z is a dependent factor and X_k are defined as independent factors. Also ε is taken as an error term. Let us take $X = (X_1, X_2, X_3, \dots, X_p)$. Here, a few presumptions are made:

1. The distribution of errors is normal
2. There is no mean for the error term
3. No correlation between $X_j\ for\ j = 1,2,3,\dots,p$ and ε
4. The variance is finite for non-random variable X
5. There is no exact straight relationship between the free variables $X_j\ where\ j\ is\ taken\ as\ 1, 2, 3,\ \dots, p$

The least square method is applied to assess the parameters $\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_p$.

2.1 Regression Analysis

Let q perceptions are obtainable for dependent factor Z and permit p indicators $X_j\ where\ j = 1,2,3,\dots,p$. Allow Z_i being the i^{th} reaction stage of dependent factor Z and X_{ij} be the i^{th} phase of indicator X_j, Now, we have data made up of the q perceptions given in Table 1:

Table 1.

Perceptions i	Reaction Stage Z	(p) Indicators			
		X_1	X_2	…	X_p
1	Z_1	X_{11}	X_{12}	…	X_{1p}
2	Z_2	X_{21}	X_{22}	…	X_{2p}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
q	Z_q	X_{q1}	X_{q2}	…	X_{qp}

From equation (1), we get

$$Z_i = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3} + \cdots + \gamma_p X_{ip} + \varepsilon_i\,, i = 1, 2, 3, \dots, q$$

The parameters can be found using the least squares method.

Here Z as the dependent factor has two different kinds of variability:

1. Unexplained variability
2. Explained variability

Because of regression, the explained variability may alternatively be described as the sum of squares:

With effect of regression, the sum of squares (SSR)=$\sum_{i=1}^{q}(\hat{z}_i - \bar{z})^2$

Due to error, the sum of squares may also be used to describe the unexplained variability.

Due to error, the sum of squares (SSE) =$\sum_{i=1}^{q}(z_i - \hat{z})^2$

Thus, total changeability in the reliable variable Z may be given as follows:

Total variability (SST)=SSE+ SSR

Now the determination coefficient $R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1\text{-}\frac{SSE}{SST}$

It means 0≤R^2≤1

When the value of SSR is close to SST, then R^2 is close to 1, implying that The regression explains a significant amount of Z's changeability and characterizes that model as the best; when the value of SSE is close to SST, so R^2 is close to 0, implying that the regression study does not describe much Z fluctuation and characterizing that model as not great. The addition of an individual variable to the model raises the regression's value.

This increase is independent of the current independent variable's commitment. Therefore, modified R^2 is stated, as R^2 may be misleading.

Modified R^2 is specified by:

R^2_{mod} =1-$\frac{SSE/(q-p-1)}{SST/(q-1)}$

where p denotes how many independent variables there are.

One way to specify a standard error is $S_{ZX} = \sqrt{SSE/(q-p-1)}$

2.2 Time Series Interpretation

Multiple regression is applied to finish the analysis of time series, and it is defined by:

$$H_{q+1} = f(H_q, H_{q-1}, \dots, H_1)$$

Where H_1, H_2, H_3, …, H_q are inputs and H_{q+1} is output.

Multiple linear regression is used to form f.

The output is defined as the fourth information point, following the use of the first three as input.

3. DATA ANALYSIS

By filtering data on the constituents of air pollution, the State Pollution Control Board of Uttar Pradesh creates data for the state's various urban areas on its website. The city's Khora Colony area is used to estimate the AQI for Ghaziabad. The stage has been critical over the last few years. The air quality of Ghaziabad city will also deteriorate with an increase in the AQI. A study has been conducted using AQI data gathered for Khora Colony, which is located in Ghaziabad, between Jan. 2019 and June 2023.

Figure 1 provides an overview of the data:

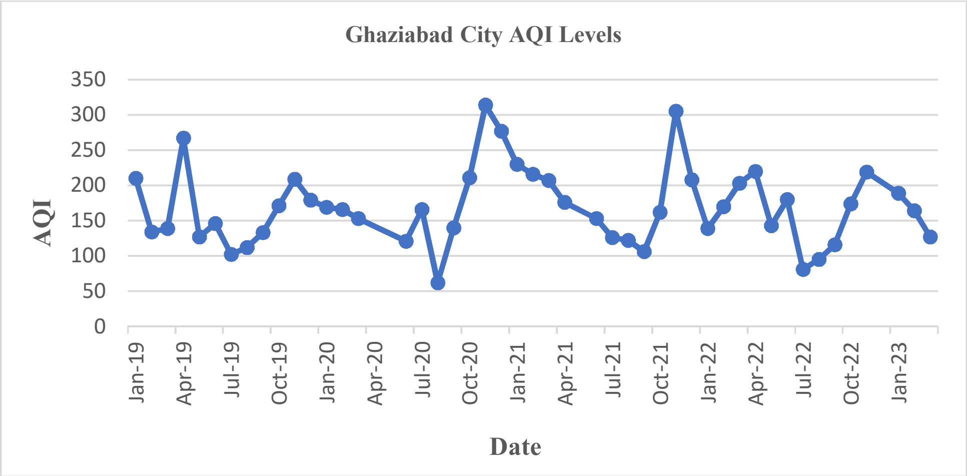

Figure 1: Actual values of AQI

4. REGRESSION ANALYSIS

For Ghaziabad City, we have developed a regression model using SPSS software. The total number of data points used in this investigation is 47. 44 groups were created with the help of these data points. For every category, we have four data points. The fourth AQI data point in the series was considered the output, and the initial three were considered the input. Now the starting information point is ignored, the following 3 information points remain as inputs, and the next in the series is an output. While the output information point is labeled as AQ_OUTPUT, the first three data points are designated as AQ_A, AQ_B, and AQ_C. The following model has been developed for the Time series:

Table 2: Model Specifications[b]

Model	R	R^2	Modified R^2	Standard Error (S.E.)	Durbin-Watson
1	0.479	0.230	0.172	50.78532	1.865

a. Forecasters: Constant, AQ_C, AQ_A, AQ_B
b. The dependent Factor is AQ_OUTPUT
c. According to Table 2, the independent factors have a 47.9% variability.

Table 3 explains the results of the ANOVA.

Table 3: ANOVA.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	30750.975	3	10250.325	3.974	.014[b]
	Residual	103165.934	40	2579.148		
	Total	133916.909	43			

The t-test values and coefficients of regression are shown in Table 4 below.

Table 4: The t-test values and coefficients of regression

Model		Non-Standardized Coefficients		Standardized Coefficients	t	Sig.
		B	S.E.	γ		
1	Constant	128.196	33.535		3.823	<0.001
	AQ_A	-0.165	0.154	-0.166	-1.071	0.291
	AQ_B	-0.070	0.172	-0.070	-0.408	0.686
	AQ_C	0.478	0.157	0.476	3.047	0.004

Table 4 is used to define the regression model as
AQ_OUTPUT=128.196-0.165*AQ_A-0.070*AQ_B+0.478*AQ_C.
The AQI graphs with actual and estimated values are shown below in Figure 2.

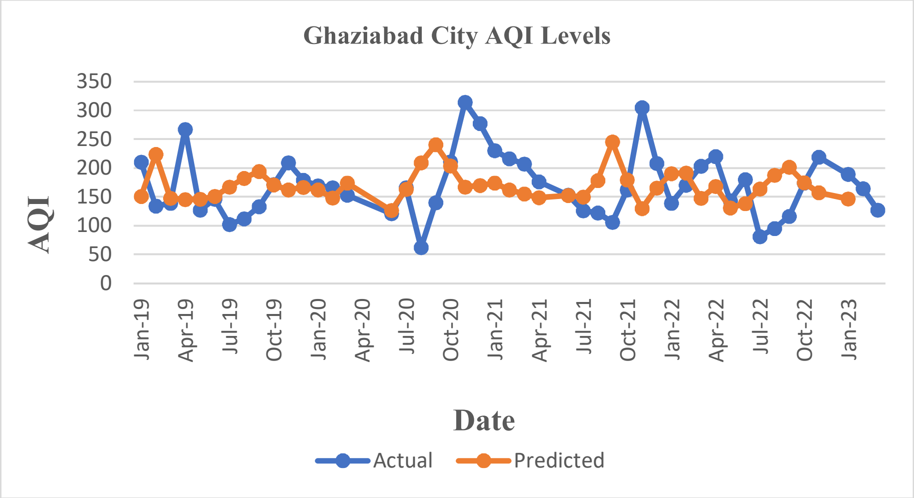

Figure 2: Actual and Estimated values of AQI

CONCLUSION

We forecasted the AQI data for Ghaziabad City in Uttar Pradesh using the study of time series. Multiple regression is utilized to generate the time series model. The AQI's future values can be somewhat predicted by examining its historical values, as evidenced by the independent factors accounting for 47.9% of the variability. There is nonlinearity in the model. Several regression methods, including the log-log and semi-log forms, are suitable to address this issue. Creating an ANN model is one method of dealing with non-linearity.

Funding details: No funding

Acknowledgement: No

REFERENCES

Asadollahfardi G., Zangooei H. and Aria S. H. (2016). Predicting PM2.5 Concentrations using Artificial Neural Networks and Markov Chain, a Case Study Keraj City. Asian Journal of Atmospheric Environment, 10(2), 67-79.

Bhavsar R. (2019). Air Pollution Monitoring Using Artificial Neural Network. International Journal of Scientific & Engineering Research, 10 (12), 515-519.

Boznar M., Lesjak M., and Mlakar P. (1993). A neural network-based method for short-term predictions of ambient So2 concentrations in highly polluted industrial areas of complex terrain. Atmospheric Environment, 27B, 221-230.

Boznar M.Z. and Mlakar P. (2002). Use of neural networks in the field of air pollution modelling. Air Pollution Modeling and Its Application XV, 375-383.

Cogliani E. (2001). Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. Atmospheric Environment, 35, 2871- 2877.

Comrie A.C. (1997). Comparing Neural Networks and Regression Models for Ozone Forecasting. Air & Waste Management Association, 47, 653- 663.

Freeman A. M. III (1974). Air pollution and property values, a further comment. Review of Economics and Statistics, 56, 554– 556.

Kumar G., Sharma R.K. (2017). Air Pollution Evaluation Methods. International Journal of Engineering Research and Development, 13 (9), 12-17.

Kumar G. (2018). Time series analysis of PM10 for Bulandhshahr Industrial Area in NCR using Multiple Linear Regression. International Journal of Engineering Research and Development, 14 (3), 56-62.

Kumar G. (2018). Time series analysis of PM10 for Noida Sector 1 Industrial Area in NCR using Multiple Linear Regression. Bulletin of Pure and Applied Sciences, Section E-Math. & Stat., 37 (2), 273-277.

****