**International Journal Of Engineering Research And Development**
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 14, Issue 3 (March Ver. I 2018), PP.56-62

# Time Series Analysis of $PM_{10}$ for Bulandhshahr Industrial Area in NCR using Multiple Linear Regression

Gaurav Kumar
*Associate Professor, Department of Mathematics, NAS College, Meerut*
*Corresponding Author: Gaurav Kumar*

**Abstract:** Air pollution has become a major problem around National Capital Region (NCR) of India. It has Particulate Matter ($PM_{10}$) as its one of the constituents. Prediction of air pollutant $PM_{10}$ can be of help in formulating a policy for air pollution abatement. This paper intends to introduce the reader about air pollution as a problem and also time series analysis of air pollutant PM10 for Bulandhshahr Industrial Area of Ghaziabad city in NCR India using past years data as published by the State Pollution Control Board is conducted using multiple linear regression.

**Keywords:** Regression, PM10, Air Pollution, Time Series

---

Date of Submission: 25-02-2018 Date of acceptance: 14-03 2018

---

## I. Introduction

Pollution is a pervasive phenomenon of economic growth and development causing harm to society. It refers to the residual flows that arise from anthropogenic sources and enter the environmental systems. The residuals concentration is directly proportional to economic activities and would increase with the rise in activity levels. An economic externality is said to be present when the activities of some economic agents affect the other agents, positively or negatively, which do not have control over them. Air and water pollution, deforestation and land degradation are some of the environmental externalities generated by the various development activities in the economy. The world health organization (WHO) has classified Delhi as one of the most polluted cities in the world along with Mexico City, Seoul and Beijing. The Central Pollution Control Board (CPCB) monitors air quality at various centers in Delhi. The State Pollution Control Board (SPCB) measures air quality at different major cities of the state. Ghaziabad and Noida are two big industrial cities of Uttar Pradesh state of India. These two are part of National Capital Region (NCR) of India and therefore have importance in the development and growth of NCR. These two cities have seen tremendous growth in last two decades. This growth has resulted in the growth of population, vehicles and industrial activities. The growth has resulted in the increase of air pollution beyond the safe level. The major sources of air pollution in these two cities are vehicles and industries. The reduction in the level of different pollutants of air pollution is expected to result in substantial benefits in various sectors of human life. Therefore forecasting of air pollutant becomes an important aspect of policy formulation to tackle problem of air pollution. There are three major air pollutants viz $NO_2$, $SO_2$ and $PM_{10}$. Time series analysis of $PM_{10}$ is conducted in this study. A time series is ordering of observation on time scale. This has been used heavily in scientific fields like statistics and signal processing but can also be used in financial forecasting and environment economics. Time series analysis can be done using various methods like Moving Average Method, Artificial Neural Network Method, Trend Reversal Pattern Method, Relative Strength Method, Elliot wave Theory and Regression among others. In this study, Regression method is used to develop time series model for $PM_{10}$.

## II. Regression Methodology

Multiple linear regression is used in this study. The methodology of the same is given below.
Consider the equation -

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ....\alpha_m X_m + \varepsilon \quad (1)$$

where $Y$ is dependent variable, $X_1, X_2, \ldots\ldots, X_m$ are explanatory (independent) variables, also called regressors or predictors and $\epsilon$ denotes the random error term. The above equation represents a linear regression model because the parameters $\alpha_0, \alpha_1, \alpha_2, ....\alpha_m$ occurring in this equation are linear in nature. Let $X = (X_1, X_2, ....., X_m)$. Here we make following assumptions:

1. Error is normally distributed

2. Error term has zero mean
3. All the predictors $X_j$'s, where $j = 1,2,....m$ and $\epsilon$ are uncorrelated i.e. we have Cov($X$, $\epsilon$) = 0
4. $X$ is nonrandom variable with finite variance
5. None of the predictor variable has perfect correlation with any other predictor variable or with linear combination of the other predictors i.e. there exists no exact linear relationship between the independent variables $X_j^{'}s, j = 1,2,...m$.

For fixed $X_j^{'}s, j = 1,2,...m$, the population regression hyperplane is given as conditional mean of $Y$ for the given values of $X_j^{'}s$ i.e.

$$E(Y \mid X_1,...X_m) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ....\alpha_m X_m \tag{2}$$

as $E(\epsilon) = 0$. This population hyperplane with parameters $\alpha_0, \alpha_1, \alpha_2, ....\alpha_m$ is estimated using sample data. From equation (2) –

$$\alpha_0 = E(Y \mid X_1 = 0,...X_m = 0) \tag{3}$$

And the coefficient $\alpha_j$ is given by

$$\alpha_j = \frac{\partial E(Y \mid X_1,...X_m)}{\partial X_j}, j = 1,2,...,m \tag{4}$$

It is called the $j^{th}$ partial regression coefficient. It represents the change in average value of $Y$ when there is increase of one unit in the value of $X_j$.

Once the values of $\alpha_j$, $j = 1,2,....,m,$ are estimated from sample information, the regression hyperplane is formulated as

$$\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 X_1 + ... + \hat{\alpha}_m X_m \tag{5}$$

where $\hat{Y}$ is the estimated value of $Y$ and the $\hat{\alpha}_j{}'s$ represent the estimates of the population parameters. Then the term $\epsilon_i = Y_i - \hat{Y}_i, i = 1,2,....n,$ is the $i^{th}$ residual or deviation from the sample regression hyperplane for $n$ observations of dependent variable. The population parameters $\alpha_0, \alpha_1, \alpha_2, ....\alpha_m$ are here estimated using least square method.

**2.1 Estimation of the Parameters in the Multiple Regression Model**

Suppose $n$ observations are available on the dependent variable $Y$ and let there be $m$ predictors $X_j$,where $j = 1,2,....,m$. Let $Y_i$ be the $i^{th}$ response level of dependent variable $Y$ and $X_{ij}$ be the $i^{th}$ level of predictor $X_j$, and then we can represent our sample information consisting of $n$ observations as in below table 1:

**Table 1:** Response Level and Regressors

| Observation No. | Response Level | (m) Predictors | | | |
|---|---|---|---|---|---|
| I | Y | $X_1$ | $X_2$ | … | $X_m$ |
| 1 | $Y_1$ | $X_{11}$ | $X_{12}$ | … | $X_{1m}$ |
| 2 | $Y_2$ | $X_{21}$ | $X_{22}$ | … | $X_{2m}$ |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| n | $Y_n$ | $X_{n1}$ | $X_{n2}$ | … | $X_{nm}$ |

Then from equation (1):

$$Y_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + ... + \alpha_m X_{im} + \varepsilon_i, i = 1,...n \tag{6}$$

The principle of least squares is being used to estimate the parameters. Let us choose the parameters $\alpha_j$'s so that the sum of the squared deviations from the sample regression hyperplane is minimized i.e.

$$\min\left\{\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \sum_{j=1}^{m}\hat{\alpha}_j X_{ij})^2 = F\left(\hat{\alpha}_0, \hat{\alpha}_1, ..., \hat{\alpha}_m\right)\right\} \quad (7)$$

For minima, we must have –

$$\frac{\partial F}{\partial \hat{\alpha}_0} = \frac{\partial F}{\partial \hat{\alpha}_1} = ... = \frac{\partial F}{\partial \hat{\alpha}_m} = 0 \quad (8)$$

Here we assume that the second order conditions for minima hold. Then the resulting simultaneous linear equation system can be expressed as -

$$\begin{aligned}
&\sum e_i = 0 \\
&\sum X_{i1} e_i = 0 \\
&\text{.....} \\
&\sum X_{im} e_i = 0 \\
&n\hat{\alpha}_0 + \hat{\alpha}_1 \sum X_{i1} + \hat{\alpha}_2 \sum X_{i2} + ... + \hat{\alpha}_m \sum X_{im} = \sum Y_i \\
&\hat{\alpha}_0 \sum X_{i1} + \hat{\alpha}_1 \sum X_{i1}^2 + \hat{\alpha}_2 \sum X_{i1} X_{i2} + ... + \hat{\alpha}_m \sum X_{i1} X_{im} = \sum X_{i1} Y_i \\
&\text{.....} \\
&\hat{\alpha}_0 \sum X_{im} + \hat{\alpha}_1 \sum X_{im} X_{i1} + \hat{\alpha}_2 \sum X_{im} X_{i2} + ... + \hat{\alpha}_m \sum X_{im}^2 = \sum X_{im} Y_i
\end{aligned} \quad (9)$$

This is a system of simultaneous linear equations. Solving this system of equations will give the set of parameter estimator $\alpha$'s. The system given by equation (9) can also be written as:

$$\hat{\alpha}_0 = \overline{Y} - \hat{\alpha}_1 \overline{X_1} - \hat{\alpha}_2 \overline{X_2} - ... - \hat{\alpha}_m \overline{X_m} \quad (10)$$

where

$$\overline{Y} = \sum \frac{Y_i}{n_i}, \overline{X_j} = \sum \frac{X_{ij}}{n}, j = 1, ..., m$$

Substituting equation (10) in remaining *m* least square normal equations (14), we get:

$$\begin{aligned}
&\hat{\alpha}_1 M_{11} + \hat{\alpha}_2 M_{12} + ... \hat{\alpha}_m M_{1m} = M_{1y} \\
&\hat{\alpha}_1 M_{21} + \hat{\alpha}_2 M_{22} + ... \hat{\alpha}_m M_{2m} = M_{2y} \\
&\text{.....} \\
&\hat{\alpha}_1 M_{m1} + \hat{\alpha}_2 M_{m2} + ... \hat{\alpha}_m M_{mm} = M_{my}
\end{aligned} \quad (11)$$

where

$$M_{jy} = \sum (X_{ij} - \overline{X_j})(Y_i - \overline{Y}) = \sum X_{ij} Y_i - n\overline{X_j}\,\overline{Y}$$

$$M_{jy} = \sum (X_{ij} - \overline{X_j})$$

$$(X_{il} - \overline{X_l}) = \sum X_{ij} X_{il} - n\overline{X_j}\,\overline{X_l}$$

for $j, l = 1,\ldots\ldots,m$.

This system of equations can be solved for the $\alpha$ 's and then value of $\alpha_0$ can be obtained from equation (10).

**2.2 Coefficient of Determination**

Let $Y = (y_1, y_2, \ldots\ldots, y_n)$ be the observations for output variable and $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots\ldots, \hat{y}_n)$ be the estimated value of output variable and $\overline{y}$ be the mean of actual observations of output variable.

The total variability in dependent variable *Y* can be divided into two parts viz explained variability and unexplained variability.

The explained variability is also called sum of squares due to regression (SSR) and is given by:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \tag{12}$$

The unexplained is also called sum of squares due to error (SSE) and is given by:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{13}$$

Therefore the total variability (SST) in *Y* is given by:

$$SST = SSR + SSE \tag{14}$$

The coefficient of determination is denoted by $R^2$. It evaluates the goodness of the fitted model and is given by:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} \tag{15}$$

$$\therefore R^2 = 1 - \frac{(SSE)}{(SST)} \tag{16}$$

It is evident that the value of $R^2$ lies between 0 and 1 i.e. $0 \le R^2 \le 1$.

When SSR is closed to SST then value of $R^2$ will be closed to 1. It means that the regression explains most of the variability in *Y* and the fitted model is good. When SSE is closed to SST then value of $R^2$ will be closed to 0. It means that regression does not explain much variability in *Y* and the fitted model is not good. The value of $R^2$ increases whenever an explanatory variable is added to the model. This increase is regardless of the contribution of newly added explanatory variable. Therefore value of $R^2$ may be misleading and so an adjusted value of $R^2$ is defined. It is called adjusted $R^2$ and is given by:

$$R_{adj}{}^2 = 1 - \frac{SSE/(n-m-1)}{SST/(n-1)} \tag{17}$$

where *m* is total number of explanatory variables.

Standard error of the estimate is given by:

$$S_{YX} = \sqrt{\frac{SSE}{n-m-1}} \tag{18}$$

**3. Time Series Analysis**

Time series analysis is done using multivariate regression method as described in section 2. The process of forecasting is described as:

$$A_{n+1} = f(A_n, A_{n-1}, \ldots, A_1) \tag{19}$$

where $A_1, A_2, \ldots, A_n$ are the inputs and $A_{n+1}$ is the output.

**3.1 Data Description**

State Pollution Control Board of U.P. monitors data of three components of Air Pollution viz Particulate Matter $PM_{10}$, SO2 and NO2 and publishes the same on their website for different cities of U.P. state. $PM_{10}$ for

Ghaziabad city being measured at two centers viz Sahibabad Industrial Area and Bulandshahr Industrial Area. This has been above critical level for past few years. Increase in level of $PM_{10}$ will further deteriorate the air quality of Ghaziabad city. $PM_{10}$ data from Jan'2014 to Nov'2015 for Bulandshahr Industrial Area has been considered for analysis. Total 69 data points are used to generate the model. Below in figure 1 is a snapshot of data:

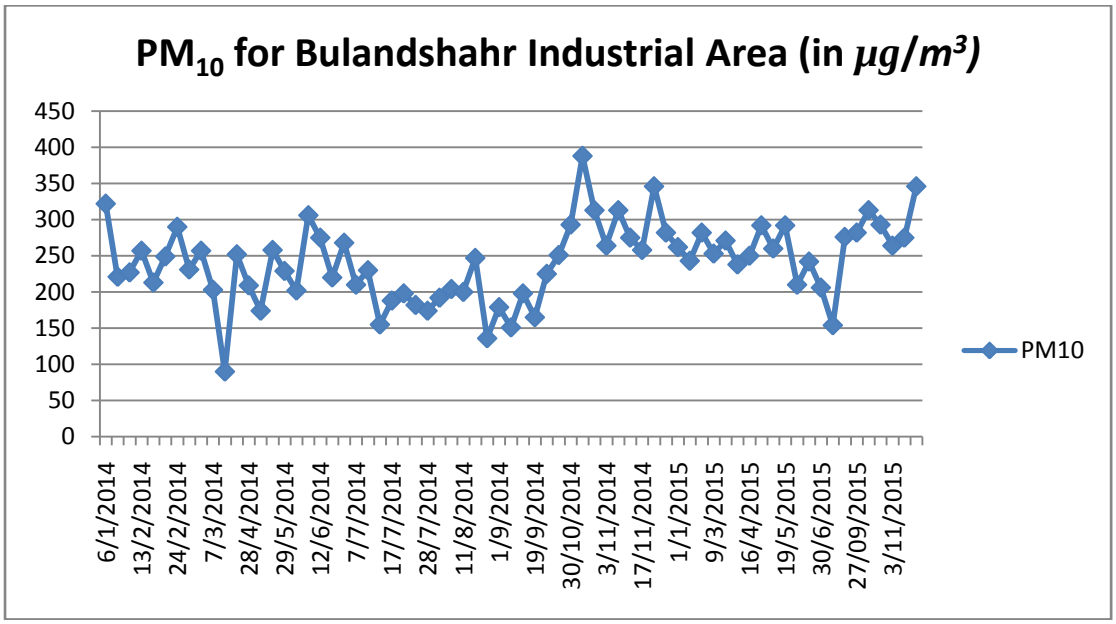


**Figure 1:** Value of $PM_{10}$ for Bulandshahr Industrial Area

### 3.2 Regression Model

Linear regression model for Bulandshahr Industrial Area in Ghaziabad city is formulated using IBM SPSS software. Total 69 data points are taken for this centre. These data points are grouped together into 66 groups. Each group contains 4 data points. First 3 data points for $PM_{10}$ have been considered as input data and $4^{th}$ in this series has been considered as output data. Again 3 data points, excluding the first data point, are considered as input and next data point as output. The first 3 data points are labeled as PM10_1, PM10_2 and PM10_3, while the output data point is labeled as Observed_PM10. Time series model has been generated as follows:

**Table 2: Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | | Durbin-Watson |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | R Square Change | F Change | df1 | df2 | Sig. F Change | |
| 1 | .548[a] | .300 | .267 | 46.45091 | .300 | 8.876 | 3 | 62 | .000 | 1.854 |

a. Predictors: (Constant), PM10_3, PM10_1, PM10_2

b. Dependent Variable: Observed_PM10

Table 2 shows that the independent variables explain 54.8% of the variability.
The result of ANOVA is shown in table 3:

**Table 3: ANOVA**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 57455.320 | 3 | 19151.773 | 8.876 | .000[b] |
| | Residual | 133776.619 | 62 | 2157.687 | | |
| | Total | 191231.939 | 65 | | | |

a. Dependent Variable: Observed_PM10

b. Predictors: (Constant), PM10_3, PM10_1, PM10_2

Table 4 shows the t-test results and gives the coefficient of regression equation:

**Table 4: Coefficients**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 72.721 | 33.451 | | 2.174 | .034 | 5.853 | 139.589 |
| | PM10_1 | .212 | .124 | .208 | 1.707 | .093 | -.036 | .459 |
| | PM10_2 | .156 | .132 | .151 | 1.182 | .242 | -.108 | .421 |
| | PM10_3 | .336 | .127 | .326 | 2.640 | .010 | .082 | .590 |

Based on table 4, the regression model is given as:

Predicted_PM10 = 72.721 + 0.212 * PM10_1 + 0.156 * PM10_2 + 0.336 * PM10_3 (13)

**Figure 2** shows chart of observed and predicted value of PM10.

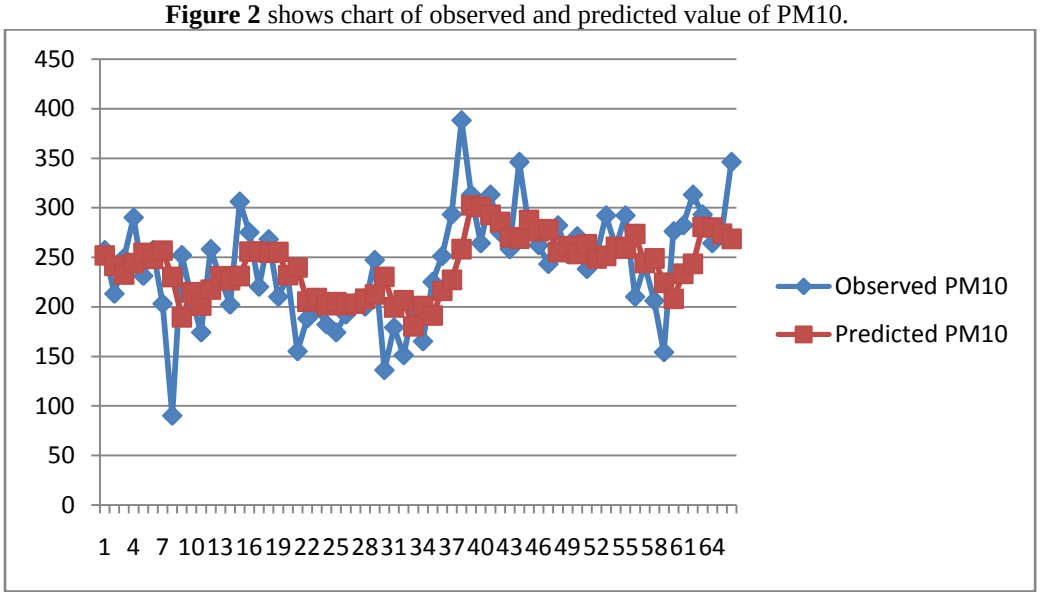


**Figure 2:** Observed and Predicted Value of PM10

## III. Conclusion

Time series analysis is conducted for prediction of value of PM10 for Bulandhshahr industrial area in NCR of India. Multiple linear regression is used to formulate time series model. The independent variables explained 54.8% of variability which shows that the future value of PM10 can be predicted upto some extent using its past values. There might be some non-linearity in the model. This can be addressed by considering different forms of regression viz log-linear, log-log forms among others. Artificial neural network model can also be built in case of non-linearity.

## References

[1] A. C. Goodman and T. G. Thibodeau, "Housing Market Segmentation", Journal of Housing Economics, vol. 7, no. 2, pp. 121–143, June, 1998

[2] Berry, J. A. and Lindoff, G., Data Mining Techniques, Wiley Computer Publishing, ISBN 0-471-17980-9, 1997

[3] Calhoum C.A., "Property Valuation Models and House Price Indexes for the Provinces of Thailand: 1992-2000", Housing Finance International, 17(3): 31-41, 2003

[4] Dasgupta, Purnamita , "Valuing Health Function Approach", Environment and Development Economics, 9(1): 83-106, 2004

[5] J. E. Zabel and K. A. Kiel, "Estimating the demand for air quality in four US cities," Land Economics, vol. 76, no. 2, 174–194, May 2000

[6] Kumar, Gaurav and Sharma R.K., "Air Pollution Evaluation Methods", International Journal of Engineering Research And Development, Vol 13, Issue 9, 12-17, Sept 2017

[7] Lutkepohl, H, "New Introduction to Multiple Time Series Analysis", Springer-Verlag, New York, 2005

[8] Murty, M.N. and A. Markandya, "The Cost Benefit Analysis of Cleaning Ganges: Some Emerging Environmental and Development Issues", Environmental and Economic Economics, Vol. 9, pp.61-81, 2004

[9] Montgomery, D.C., Peck, E.A. and Vining, G.G., "Introduction to Linear Regression Analysis", 5th Edition, John Wiley & Sons, Hoboken, NJ

[10] Wei, W.W.S., "Time Series Analysis: Univariate and Multivariate, Methods", Addison Wesley, New York, 2006

Gaurav Kumar "Time Series Analysis of PM10 for Bulandhshahr Industrial Area in NCR using Multiple Linear Regression" International Journal Of Engineering Research And Development , vol. 14, no. 03, 2018, pp. 56–62.