Original Article

Bulletin of Pure and Applied Sciences.
Vol. 37E (Math & Stat.), No.2, 2018.P.273-277
Print version ISSN 0970 6577
Online version ISSN 2320 3226
DOI: 10.5958/2320-3226.2018.00028.0

TIME SERIES ANALYSIS OF PM10 FOR NOIDA SECTOR 1 INDUSTRIAL AREA IN NCR USING MULTIPLE LINEAR REGRESSION

Gaurav Kumar*

Author Affiliation:
Associate Professor, Department of Mathematics, NAS College, Meerut, Uttar Pradesh 250003, India.

***Corresponding Author:**
Gaurav Kumar, Associate Professor, Department of Mathematics, NAS College, Meerut, Uttar Pradesh 250003, India.

E-mail: gauravkgv@gmail.com

Received on 10.01.2018, Accepted on 15.06.2018

Abstract

Time series analysis can be used to quantitatively explain and predict air pollutants. This technique offers the possibility of formulating policy to tackle problem of air pollution. This paper intends to develop time series model for air pollutant Particulate Matter (PM_{10}) for Sector-1 industrial area of Noida city in National Capital Region (NCR) of India using multiple linear regression.

Keywords: Multiple Linear Regression, PM10, Time Series

1. INTRODUCTION

Poor air quality is one of the most serious environmental problems in urban areas around the world especially in developing countries. The air pollution problem has received whose attention during the last decades whereby there has been a signification increases in public awareness of potential dangers caused by chemical pollutants and their effects on both human being and the environment. Air pollution is defined as the presence of one or more contaminants in the atmosphere in such quantity and for such duration as it is injurious or tends to be injurious to human health or welfare, animal or plant life. It can also be defined as the contamination of air by the discharge of harmful substances. There are various constituents of air pollution. These constituents are called air pollutants. Freeman [1] described methods for valuation of environment resources including air pollution. Bao and Wan [2] used hedonic regression method to analyze factors which determine the house prices. Diewert and Shimizu [4] used hedonic regression models for Tokyo condominium sales. Zabel and Kiel [6] valued air quality on four cities of US. A hedonic model is developed by Rogat Jorge [13] for the valuation of improved air quality in Santiago De Chile. Several authors like [8, 11 and 14] used hedonic method for the valuation of environment resources. In India, pollution control board measures NO2, SO2 and Particulate Matter (PM10) as air pollutants. Out of these three, PM10 is more dangerous as its size is very small. Time series analysis can help the authority to formulate policy for curbing of PM_{10}.

2. METHODOLOGY

In multiple linear regression, we have one dependent variable and more than one independent variables. The methodology of multiple linear regression is explained in this section.

Consider the equation -

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

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

1. Error is normally distributed.
2. Error term has zero mean
3. All the predictors X_j's, where $j = 1,2,\ldots.m$ and ϵ are uncorrelated i.e. we have $\mathrm{Cov}(X, \epsilon) = 0$
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, $j = 1,2, \ldots m$.

The values of parameters $\alpha_0, \alpha_1, \alpha_2, \ldots.\alpha_m$ are here estimated using ordinary least square method.

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:

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

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

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 \quad (3)$$

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

$$SST = SSR + SSE \quad (4)$$

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

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} \quad (5)$$

$$\therefore R^2 = 1 - \frac{(SSE)}{(SST)} \quad (6)$$

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:

$$R_{adj}{}^2 = 1 - \frac{SSE/(n-m-1)}{SST/(n-1)} \quad (7)$$

where m is total number of explanatory variables.

Standard error of the estimate is given by:

$$S_{YX} = \sqrt{\frac{SSE}{n-m-1}} \quad (8)$$

3. METHOD OF TIME SERIES ANALYSIS

Time series analysis is done using multiple linear regression defined as:

$$A_{n+1} = f\left(A_n, A_{n-1}, \ldots, A_1\right) \tag{9}$$

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

Here the function f is formulated using the multiple linear regression method.

Three consecutive data points are fetched as input and fourth data point is taken as output.

3.1 Data Description:

State Pollution Control Board of U.P. monitors data of three components of Air Pollution viz PM_{10}, SO_2 and NO_2 and publishes the same on their website for different cities of U.P. state. PM_{10} for Noida city is being measured at Industrial Areas of Sector 1 and Sector 6 of the city. This has been above critical level for past few years. Increase in level of PM10 will further deteriorate the air quality of Noida city. PM10 data from Jan'2014 to Nov'2015 for Noida Sector 1 Industrial Area has been considered for analysis. Below in figure 1 is a snapshot of data:

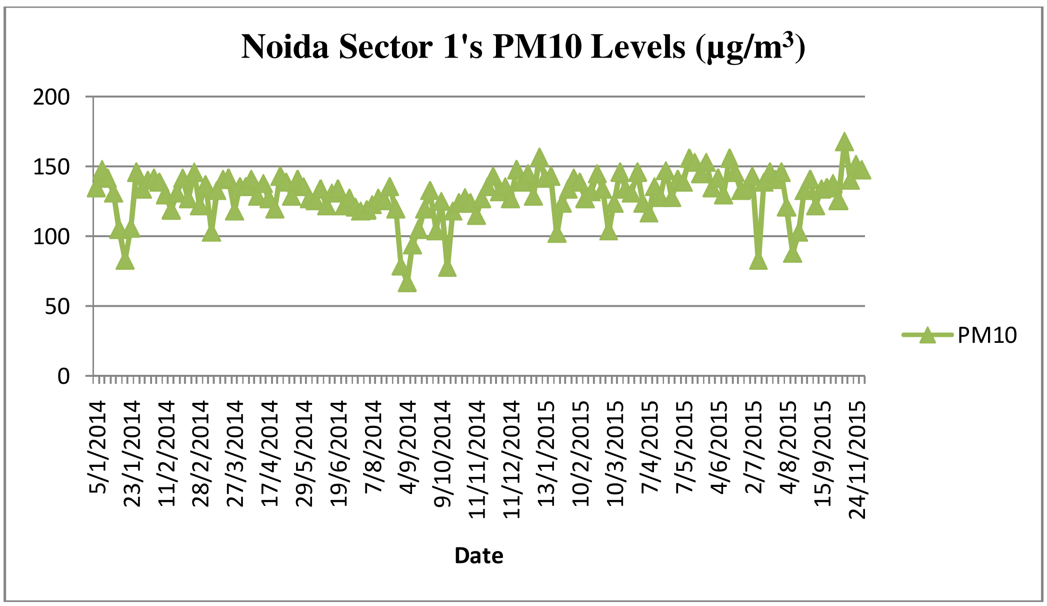

Figure 1: Value of PM10 for Noida Sector 1 Industrial Area

3.2 Regression Model:

Linear regression model for Noida Sector 1 Industrial Area is formulated using IBM SPSS software. Total 134 data points are taken for Sector 1. These data points are grouped together into 132 groups. Each group contains 4 data points. First 3 data points for PM10 have been considered as input data and 4th 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 PM_1, PM_2 and PM_3, while the output data point is labeled as Output_PM10. Time series model has been generated as follows:

Table 1: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.407[a]	.166	.146	15.463	1.984

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

b. Dependent Variable: Output_PM10

Table 1 shows that the independent variables explain 40.7% of the variability.
The result of ANOVA is shown in table 2.

Table 2: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6076.892	3	2025.631	8.472	.000[b]
	Residual	30603.494	128	239.090		
	Total	36680.386	131			

a. Dependent Variable: Output_PM10

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

The p value is 0.000 which shows that the model is significant.

Table 3 shows the t-test results and gives the coefficient of regression equation:

Table 3: Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	77.148	14.236		5.419	.000	48.979	105.317
	PM10_1	-.066	.088	-.066	-.751	.454	-.242	.109
	PM10_2	.101	.094	.100	1.072	.286	-.085	.286
	PM10_3	.372	.088	.371	4.217	.000	.198	.547

a. Dependent Variable: Output_PM10

Based on table 3, the regression model is given as:

$$Output\_PM10 = 77.148 - 0.066 * PM\_1 + 0.101 * PM\_2 + 0.372 * PM\_3 \quad (10)$$

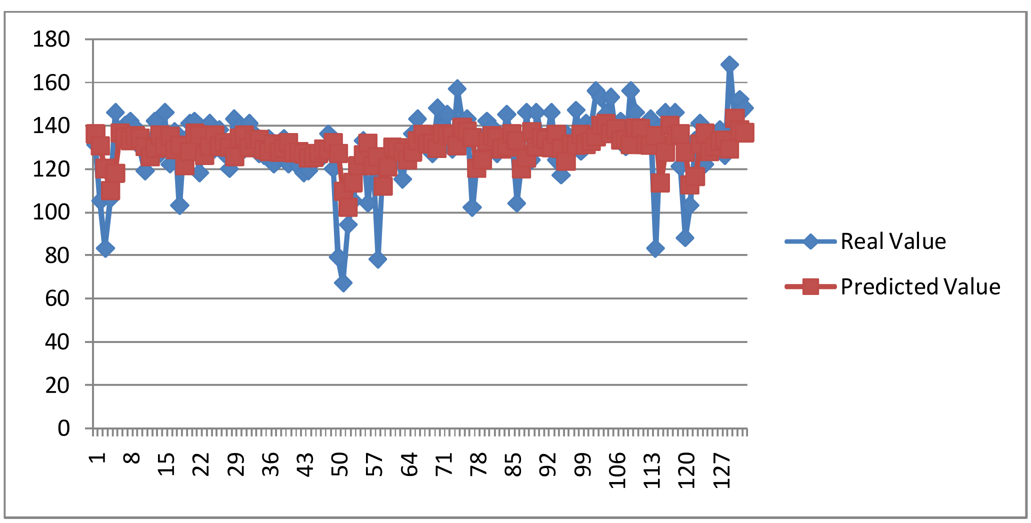

Figure 2: Real and Predicted Value of PM10

Figure 2 above shows the chart of real and predicted value of PM10.

4. CONCLUSION

Time series analysis is conducted for the value of Particulate Matter (PM_{10}) for Sector 1 industrial area of Noida city in NCR of India. Multiple linear regression is used for this purpose. The total data points taken are 134 which are grouped in 132 numbers of groups. The independent variables explained 40.7% of variability. In previous study of the author [7] on Bulandshahr Industrial Area of Ghaziabad city of NCR, multiple linear regression method is used for time series analysis of PM_{10}. The total data points taken were 69 which were grouped in 66 numbers of groups. There the independent variables explained 54.8% of variability. The non-linearity in data can be addressed if logarithmic values of independent variables are considered. Also, neural network method can be used to conduct time series analysis.

REFERENCES

[1] A. M. Freeman III, “Air pollution and property values, a further comment”, Review of Economics and Statistics, vol. 56, pp. 554–556, Nov. 1974

[2] Bao, H. X. H., Wan, A. T. K., “On the Use of Spline Smoothing in Estimating Hedonic Housing Prices Models: Empirical Evidence using Hong Kong Date”, Real Estate Economics, 32(3), 487-507, 2004

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

[4] Diewert, W. E. and C. Shimizu, “Hedonic Regression Models for Tokyo Condominium Sales,” Regional Science and Urban Economics 60, 300-315, 2016

[5] Hosmer, D. W., Lameshow, S., “Applied Logistic Regression”, New York: John Wiley & Sons, 1989

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

[7] Kumar, Gaurav, “Time Series Analysis of PM10 for Bulandshahr Industrial Area in NCR using Multiple Linear Regression”, International Journal of Engineering Research And Development, 56-62, Vol 14(3), March 2018

[8] Lee, T. H., Jung, S., “Forecasting Creditworthiness: Logistic vs Artificial Neural Network”, The Journal of Business Forecasting Methods and Systems, 18(4), 28-30, 2000

[9] Lutkepohl, H, “New Introduction to Multiple Time Series Analysis”, Springer-Verlag, New York, 2005

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

[11] Nguyen, N., Cripps, A., “Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks”, The Journal of Real Estate Research, 22(3), 313-336, 2001

[12] R. Valencia, G. Sanchez, I. Diaz, “A General Regression Neural Network for modeling the behavior of PM10 concentration level in Santa Marta, Columbia”, ARPN Journal of Engineering and Applied Sciences, 11(11), 2016

[13] Rogat Jorge, “The Value of Improved Air Quality in Santiago De Chile”, Printed in Sweden Kompendiet-Göteborg, ISBN 91-88514-36-6, 1998

[14] S. Chattopadhyay, “Estimating the demand for air quality: new evidence based on the Chicago housing market”, Land Economics, vol. 75, no.1, pp. 22-38, Feb 1999

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