ISSN 2722-2039

Regression Model to Analyse Air Pollutants Over a Coastal Industrial Station Visakhapatnam (India)

N.V. Krishna Prasad[a,1,*], M.S.S.R.K.N. Sarma[a], P.Sasikala[b], Naga Raju M[c], N. Madhavi[d]

[a] Department of Physics, G.S.S, GITAM University, Bengaluru, India.
[b] Department of Mathematics , GSS, GITAM University,Bengaluru, India.
[c] Department of Computer Science, G.S.S, GITAM University, Bengaluru, India
[d] Department of Statistics, Govt. Arts College, Rajhamundry, India
[1] drnvkprasad@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received May 10, 2020
Revised July 28, 2020
Accepted August 8, 2020

Keywords
multiple linear regression
correlation coefficient
$PM_{2.5}$, SO_2, NO_2, NO_x, CO
temperature
relative humidity

ABSTRACT

Particulate matter concentration and its study has gained tremendous significance in view of increase in air pollution. Since air pollution has many adverse effects on mankind, measures may be taken by observing the trends in $PM_{2.5}$ (particulate matter) and concentrations of pollutants like NO_2, SO_2, NO_2, NO, NO_x, CO, NH_3 and RH (Relative Humidity) as well as temperature. Even though continuous monitoring of air pollution in urban locations has been increasing in view of its huge impact on the sustainable development and ecological balance a regression model is essential always to analyse large sets of data. These regression models also play vital role in some cases where data was not observed due to unavoidable circumstances and during times when the measuring instruments do not work. In this context an attempt was made to develop a regression model exclusively for Visakhapatnam (India) a coastal, urban and industrial station and to analyse the trends in particulate matter concentration at this staion. A regression model was developed with $PM_{2.5}$ as dependent variable and SO_2, NO_x, NO_2, CO, NH_3, temperature(Temp) and relative humidity(RH) as independent variables. The efficiency of the model was tested with known independent variables and $PM_{2.5}$ was estimated. It is found that observed and estimated $PM_{2.5}$ values are highly correlated.

1. Introduction

India being one of the fast developing countries has been experiencing huge population growth as well as rapid urbanization and industrialization leading to increase in vehicular emissions. These changes significantly affect atmospheric environment which leads to impurement of particulate matter [1] – [4]. Correia et al., 2013 [5] reported the affect of particulate matter on global climate, health of mankind as well as visibility. At the same time the report of Padoan et al., 2016 [6] was of significance which stated that PM_{10} (particulate matter having diameter <10 μm) is responsible for injecting chemicals of harmful nature into the respiratory system of human beings. This increases diseases related to heart, blood vessels as well as human respiratory system. As per the report of Lim et al., 2012 [7] death count of people due to air pollution in 2010 was 3.2 millions. In view of the severity, Central Pollution Control Board (CPCB) has fixed a 24 hour limit of 100 $\mu g/m^3$ and yearly limit of 60 $\mu g/m^3$ for PM_{10} all over India. Even then, increase in unavoidable industrial and vehicular emissions increase the levels of PM_{10} concentrations beyond the limits specified by CPCB in majority of the cities [8] – [10]. Delhi is one of the examples which recently experienced

huge air pollution during November 2019. This alarming situation has raised concern to study the concentration of particulate matter in polluted cities of India. Even though particulate matter and its concentration is not only the parameter in terms of respiratory issues, still an attempt was made to study the particulate matter concentration of Visakhapatnam which is a fast developing city of Andhra Pradesh (India).

Visakhapatnam is one of the fast growing cities in the state of Andhra Pradesh (India). It is a coastal, urban and industrial station with continuously increasing population and traffic on day to day basis. At present it has an approximate population of 24 lakhs in 2019 which is increased by 20% when compared to 2011 (Population Research Centre, Andhra University). As per the statistics of Visakhapatnam City Police 8 lakhs vehicles of all types move in the city every day. Every year nearly 20,000 new vehicles are being added. The city is encountering a huge construction activity on daily basis. In view of these activities and tremendous anthropogenic activities air pollution is of prime concern which was due to uncontrolled industrial emissions, municipal waste incineration, vehicular traffic etc. Obviously there will be a drastic degradation of air quality as well as instability of various environmental parameters. These parameters need to be analyzed in view of their affect on health and they need to be stabilized by taking proper measures..

Review of CPCB and Comprehensive Environmental Pollution Index (CEPI) by Ministry of Environment and Forests identified Visakhapatnam as one of the critically polluted stations. Regular measurements of the atmospheric parameters like solar UV-B radiation were started in Visakhapatnam in 1985 as part of Indian Middle Atmosphere Programme (IMAP). However emphasis on measurement of air quality index and its related parameters gained significance only after initiation of NAMP (National Air Quality Monitoring Program) in 1984.

As per the standards of National Ambient Air Quality set by CPCB (Central Pollution Control Board) the concentration of RSPM in ambient air (24 hours) at an industrial site area is 150 $\mu g/m^3$ and that of other areas is 100 $\mu g/m^3$ while in particular sensitive areas is 75 $\mu g/m^3$ while the annual mean values are found to be 120 $\mu g/m^3$, 60 $\mu g/m^3$, and 50 $\mu g/m^3$ respectively. It is reported that the SPM level in the residential areas of Visakhapatnam is high while the NO_2 level in the residential areas is moderate [1]. As per the CPCB the RSPM levels are reported to exceed by a factor of 1.0 - 1.5 at almost all urban locations of Visakhapatnam while it is moderate with exceedence factor of 0.5 to 1 in industrial locations of Visakhapatnam. This situation needs to be addressed immediately. Here it is note worthy that very few research papers have reported about the monitoring and analysis of PM for Visakhapatnam station till date. Hence the present study may help us to know the status of existing air quality at this industrial station and to instigate some strategic plans for ensuring better and healthy environment to the people of Visakhapatnam. In this context an attempt was made to develop a regression model with $PM_{2.5}$ as dependent variable and PM_{10}, NO_2, SO_2, NO, NOx, CO, NH_3 ,RH as well as temperature as independent variables. The idea is to check the dependence of $PM_{2.5}$ concentration on other variables as well as to use the developed model for forecasting the concentration of $PM_{2.5}$ for this urban station Visakhapatnam.

2. Data & Modelling

Hourly values of $PM_{2.5}$, NO_2, SO_2, NO, NOx, CO, NH_3 ,RH as well as temperature for the period October 2016 – November 2019 were obtained from the official website of Central Pollution Control Board (Source:https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data) [11],[12]. From these values monthly means and standard deviations were calculated. To assess the dependence of $PM_{2.5}$ on other parameters relevant analysis was done.

3. Relevant Feature Analysis between $PM_{2.5}$ and other parameters

Relevant feature analysis was done in order to develop a regression model to know about the concentration of $PM_{2.5}$ and its correlation with other parameters. Hence correlation analysis between $PM_{2.5}$ concentration and other parameters was done with double variable correlation indexes calculate formula given by

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2]\,[\, n\Sigma y^2 - (\Sigma y)^2]}} \quad (1)$$

where x and y are two variables, n is the sample size. Then the bivariate correlation analysis between PM 2.5 and relevant parameters is obtained as follows.

Table 1. Significance level between PM2.5 and other parameters

Variables	Correlation coefficients	Significance level	Sample size
PM 2.5 & SO2	0.228894392	0.00	16582
PM 2.5 & NO	0.279472053	0.00	16582
PM2.5 & NO2	0.475810546	0.00	16582
PM2.5 & NOX	0.42346651	0.00	16582
PM2.5 & NH3	0.24655445	0.00	16582
PM2.5 & CO	0.309112155	0.00	16582
PM 2.5 & RH	-0.055293291	0.00	16582
PM 2.5 & TEMP	-0.025037937	0.001262257	16582

Table 1 indicates that the significance levels between $PM_{2.5}$ and PM_{10}, SO_2, NO, NO_2,NO_X, NH_3, CO, RH, Temperature. If the significance level less than 0.05, it shows that the respective parameter has passed the significance test and can be used for building $PM_{2.5}$ concentration calculation model. From the table it is clear that there is a significant +ve correlation existing between $PM_{2.5}$ and SO_2, NO, NO_2,NO_X,NH_3 and CO while a significant negative correlation exists between $PM_{2.5}$ & RH and $PM_{2.5}$& Temperature.

4. Regression Model

Multivariate Linear Regression is a statistical technique that estimates a single regression model with more than one variable as the outcome. In order to predict the dependent variable (Y) the independent variables are fitted in a linear equation given by

$$\mathbf{Y = C_0 + C_1X_1 + C_2X_2 + C_3X_3 + \ldots.. + C_nX_n + \varepsilon} \quad (2)$$

where X_1,X_2 ....., X_n represent independent variables, C_1, C_2,.....C_n represent regression coefficients, C_0 a constant while 'ε' stands for estimated error term obtained from random sampling of normal distribution with mean zero and constant variance. We constructed a Multivariate Linear Regression with $PM_{2.5}$ as dependent variable(Y) and SO_2(X_1), NO (X_2), NO_2 (X_3), NO_X (X_4) , NH_3 (X_5) , CO (X_6) , RH (X_7) , Temp (X_8) as independent variables.

If m represent the number of independent variables, n represent the number of samples, we can set up multiple linear regression forecast model as follows:

$$\left.\begin{aligned} Y_1 &= C_0 + C_1X_{11} + C_2X_{12} + \ldots..+ C_m X_{1m} + \varepsilon_1 \\ Y_2 &= C_0 + C_1X_{21} + C_2X_{22} + \ldots..+ C_m X_{2m} + \varepsilon_2 \\ &\ldots\ldots \\ Y_n &= C_0 + C_1X_{n1} + C_2X_{n2} + \ldots..+ C_mX_{nm} + \varepsilon_n \end{aligned}\right\} \quad (3)$$

where C_0 is constant, ε represents the observation error: $\varepsilon_i \sim N(0,\sigma^2)$ (i= 1,2,… n), and $\varepsilon_1, \varepsilon_2,\ldots \varepsilon_n$ are independent to each other.

Let

$$Y_{nx1} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \qquad X_{nx(m+1)} = \begin{pmatrix} 1 & X_{11} & \ldots. & X_{1m} \\ 1 & X_{21} & \ldots. & X_{2m} \\ . & . & & . \\ . & . & & . \\ 1 & X_{n1} & \ldots. & X_{nm} \end{pmatrix}$$

$$C_{(m+1)x1} = \begin{pmatrix} C_0 \\ C_1 \\ . \\ . \\ C_m \end{pmatrix} \qquad \varepsilon_{nx1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

We show (3) formula with a matrix:

$$Y = CX + \varepsilon$$

$$E(\varepsilon) = 0, \mathrm{Cov}(\varepsilon,\varepsilon) = \sigma^2 I_n \quad (4)$$

We can use the least square method to get the unbiased estimate, effective and consistent estimate about the multiple linear regression model. The estimate formula is as follows:

$$\hat{C} = \begin{pmatrix} C_0 \\ C_1 \\ . \\ . \\ C_m \end{pmatrix} = (X^T X)^{-1} X^T Y \quad (5)$$

Then, we can get the estimates of $C_0, C_1, \ldots C_m$ are $\hat{C}_0, \hat{C}_1, \ldots, \hat{C}_m$ where

$$\sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y_i - C_0 - C_1X_i - \ldots. - C_mX_i)^2 \quad (6)$$

We can get the regression equation:

$$\hat{Y} = \hat{C}_0 + \hat{C}_1X_1 + \hat{C}_2X_2 + \ldots + \hat{C}_mX_m \quad (7)$$

In this model m = 8 and n= 16582, the output of the model is given by Table 2.

Table 2. Regression Statistics for PM2.5 and other independent parameters

Regression Statistics				
	Multiple R value	0.802239725		
	R^2 value	0.643588577		
	Adjusted R^2 value	0.643395015		
	Standard Error	21.53639017		
	No. of observations	16582		
Parameter	**Coefficients**	**Standard Error**	**t Stat**	**P-value**
$PM_{2.5}$	4.437978522	2.772956527	1.600450089	0.11
RH	-0.007484115	0.019606153	-0.381722787	0.70
Temp	-0.184544889	0.079358555	-2.325456777	0.02
SO_2	0.026204307	0.014391022	1.820878738	0.07
NO	0.140197814	0.032433116	4.322674843	0.00
NO_2	0.255607455	0.020793711	12.29253648	0.00
NO_x	-0.329747188	0.03937006	-8.375582638	0.00
NH_3	0.323146197	0.02110109	15.31419471	0.00
CO	3.338825723	0.24473705	13.64250213	0.00

Table 2 indicates the probability value of regression coefficient in the significance test. Here it is observed that all the significance levels are less than 0.05, hence the model is feasible. The final regression equation can be expressed as

$$\mathbf{Y = 4.4380 + 0.0262\ X_1 + 0.1402\ X_2 + 0.2556\ X_3 - 0.3297\ X_4}$$
$$\mathbf{+ 0.3231\ X_5 + 3.3388\ X_6 - 0.0075\ X_7 - 0.1845\ X_8}$$

5. Results and Discussion

5.1. Modelled $PM_{2.5}$ versus Observed $PM_{2.5}$

A graph (Fig.1) was plotted with number of day on x-axis and the corresponding observed and estimated concentrations of $PM_{2.5}$ on y-axis. .

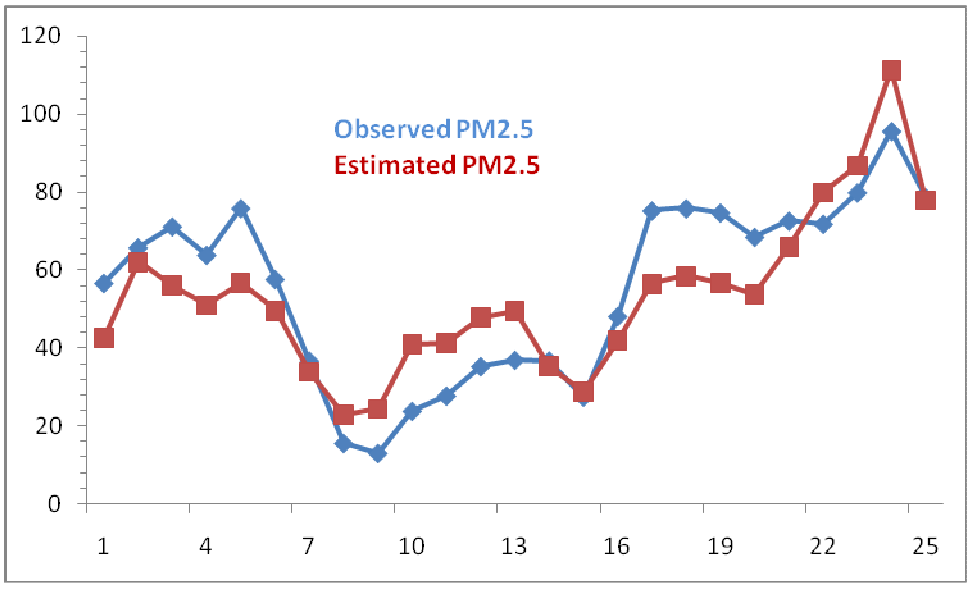

Fig 1. Plot showing Modelled $PM_{2.5}$ versus Estimated $PM_{2.5}$ concentrations

The regression equation defined above is used to estimate the $PM_{2.5}$ concentration for the month of December 2019 by substituting the known values of SO_2, NO, NO_2 , NO_X , NH_3 , CO , RH and Temperature. This exercise was done to check the efficiency of the developed regression model. From the plot it is very clear that both observed and modelled $PM_{2.5}$ concentrations are highly correlated. This allows us to forecast the concentration of $PM_{2.5}$ for Visakhapatnam station in future.

5.2. $PM_{2.5}$ concentration and its dependence on other variables at Visakhapatnam

The above regression analysis indicate the dependence of PM2.5 concentration on other concentrations like SO_2 , NO , NO_2, NO_X , NH_3 , CO & RH and Temperature. Since all the parameters pass the significant test the dependence of PM2.5 concentration on other parameters to be analysed and hence the equation is developed. From Table 2 it is clear that P-value for relative humidity (RH) is 0.7 which is much greater than 0.05. This value signifies that RH fails the significance test. At the same time it is observed that the correlation coefficient of RH is -0.007484115 indicting negative correlation between PM2.5 and Relative Humidity(RH). Since Visakhapatnam is a location with high relative humidity its effect on $PM_{2.5}$ need to be analysed in more detailed manner.

6. Conclusion

In this paper, we used the latest three years monitoring data from 2016-2019 of coastal, urban, industrial station Visakhapatnam as experimental data. A double variant correlation analyses between $PM_{2.5}$ concentrations and other monitoring elements was made. The result shows that the concentrations of $PM_{2.5}$ with SO_2, NO, NO_2,NO_X, NH_3, CO, RH and temperature have linear relationship during this period. However correlation exists between different pollutants due to production, formation and involvement of changing processes. Hence we developed a multivariate linear regression model between the concentrations of $PM_{2.5}$ and SO_2, NO, NO_2,NO_X, NH_3, CO, RH, Temperature. The developed model was tested with reserved 25 days monitoring data of December 2019 which was not used in developing the model. This exercise was done to verify the efficiency of the developed model.

The main conclusions drawn are significant relevance exist between the concentration of $PM_{2.5}$ and SO_2, NO, NO_2,NO_X, NH_3, CO, RH, Temperature. Hence all these elements can be used for building $PM_{2.5}$ regression model. The values of correlation coefficients indicate the correlation with $PM_{2.5}$

from higher to lower value in the order of CO(3.33), NH_3(0.32), NO_2(0.255), NO(0.14), SO_2(0.026), RH(-0.0074), Temperature (-0.18) and Nox (-0.32). Effieciency of the developed model was tested for 25 days data of December 2019 which was not used for developing the model. This indicated excellent correlation as shown in Figure 1.and hence the model can be used to forecast $PM_{2.5}$ for this station. This model can be used at times when the $PM_{2.5}$ data was not available and other data was available. In the same lines models can be developed by changing the dependent and independent variables so that non availability of any instrument or data may not be a hurdle for regular analysis of particulate matter or pollutants on daily basis.

Acknowledgment

The authors deeply acknowledge Central Pollution Control Board for providing valuable data through their official website.

References

[1] D.S. Satish Kumar, 2013. Air Pollution in Visakhapatnam – An Overview. International Journal of Civil Engineering (IJCE) Vol. 2, Issue 4, Pg. 11-14.

[2] Sandeep Police, Sanjay Kumar Sahu, Gauri Girish Pandit (2016). Chemical characterization of atmospheric particulate matter and their source apportionment at an emerging industrial coastal city, Visakhapatnam, India. Atmopsheric Pollution Research. Vol(7), Issue 4, Pg 725-733

[3] Pradyumn Singh, Renuka Saini ,AjayTaneja(2014). , Physicochemical characteristics of $PM_{2.5:}$ Low, middle, and high–income group homes in Agra, India–a case study. Atmopsheric Pollution Research. Vol 5, Issue 3, Pg 352-360.

[4] Reshmi Das , Bahareh Khezri, Bijayen Srivastava , Subhajit Datta , Pradip K.(2015). Trace element composition of $PM_{2.5}$ and PM_{10} from Kolkata – a heavily polluted Indian metropolis. Atmos. Pollut. Res., 6, pg. 742-750

[5] Correia AW, Pope CA , Dockery DW, Wang Y, Ezzati M Dominici F.Correia(2013). Effect of air pollution control on life expectancy in the United States an analysis of 545 US counties for the period from 2000 to 2007. Epidemiology, 24 (2013), pp. 23-31.

[6] Padoan E, Malandrino M,Giacomino A,Grosa M M,Lollobrigida F(2016). Spatial distribution and potential sources of trace elements in PM10 monitored in urban and rural sites of Piedmont Region. Chemosphere, vol.145 , pg. 495-507

[7] Lim S.S, Vos T,Flaxman AD, Danaei G,Shibuya K et.al.,(2012)..A comparative risk assessment of burden of disease and injury attributable to67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010.Lancet, 380 pg. 2224-2260.

[8] Danting Zhao , Hong Chen, Erze Yu and Ting Luo (2019). PM2.5/PM10 Ratios in Eight Economic Regions and their relationship with Meteorology in China. Advances in Meteorology,Volume 2019, Article ID 5295726, 15 pages

[9] Cairong Lou, Hongyu Liu,Yufeng Li, Yan Peng, Juan Wang, and Lingjun Dai(,2017). Relatioships of relative humidity with PM2.5 and PM10 in the Yangtze River Delata, China. Environmental Monitoring and Assessment. Vol(189).Article No.582.

[10] Wang Hua, Jiang Nan, Yang Naiwang (2014). The study of atmospheric environment quality and its change trend in Xi 'an during 1991-2012. Journal of environmental engineering, pg 526-529.

[11] Central Pollution Control Board (http://cpcb.nic.in/)

[12] Environmental Protection Agency (http://www.epa.gov.in/)