Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 132 (2018) 1077–1085

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

# Forecasting air pollution load in Delhi using data analysis tools

Nidhi Sharma[a], Shweta Taneja[b*], Vaishali Sagar[c], Arshita Bhatt[d]

[a]*Associate Professor, Deptt. of Applied Chemistry, Bhagwan Parshuram Instt. Of Technology, Rohini, New Delhi-110085*
[b]*Assistant Professor, Deptt. of Computer Science, Bhagwan Parshuram Instt. Of Technology, Rohini, New Delhi-110085*
[c,d] *Student, Deptt. of Computer Science, Bhagwan Parshuram Instt. Of Technology, Rohini, New Delhi-110085*

**Abstract**

The enormity of air pollution has always been a matter of concern due to rapid development and urbanization over a long period. The increasing level of pollutants in ambient air in 2016-2017 has deteriorated the air quality of Delhi at an alarming rate. This brought us to focus our study on air quality in Delhi region. The prediction of future air quality has been carried out by analyzing the pollutants using data analysis techniques.

In our previous study, we had analyzed the data from 2011-2015. Detailed analysis from 2009-2017 of air pollutants has been proposed in this extended paper along with the critical observation of 2016-2017 air pollutants trend in Delhi. Descriptive analysis and predictive analysis have been used to study the trends of various air pollutants like sulphur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), suspended particulate matter (PM), ozone ($O_3$) carbon monoxide (CO), benzene, and forecast the future trend. We have observed through data analytics techniques that $SO_2$ is likely to increase by 1.24ug/$m^3$, $NO_2$ is likely to increase by 16.77ug/$m^3$, O3 is likely to increase by 6.11 mg/$m^3$, benzene is likely to reduce by 1.33 mg/$m^3$ and $NO_2$ is predicted to reduce by 0.169 mg/$m^3$ in the coming years.


Peer-review under responsibility of the scientific committee of the International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

* Corresponding author. Tel.: 9212388121.
*E-mail address:shweta.madhur21@gmail.com*


Peer-review under responsibility of the scientific committee of the International Conference on Computational Intelligence and Data Science (ICCIDS 2018).
10.1016/j.procs.2018.05.023

*Keywords:* Air pollution; Ambient air; Descriptive analysis; Predictive analysis; Time series regression forecasting.

## 1. Introduction

Air is an invisible substance surrounding the Earth and providing us all with the breathable oxygen and performs a vital role in supporting life on Earth. But with the passage of time the fresh and pure air is gradually getting contaminated due to increase in air pollution. Air pollution is the presence of one or more substance at a concentration above their natural levels, with the potential to produce an adverse effect.
According to World Population Review, Delhi, the National Capital Territory (NCT) of India, is the densely populated metropolitan city with a large influx of population from other states of India. As per the last Census carried out in 2011, population of Delhi was 16.7 million [2] and estimated 2016 population of 18.6 million. In recent years, rapid industrialisation and urbanisation posed detrimental effect on environment. Problem of air pollution is increasingly getting more serious. Increasing levels of pollutants in air is causing extreme health disorder. It directly affects a population of millions who are suffering from shortness of breath, eye irritation to chronic respiratory disorders, pneumonia, acute asthma etc [3] [4].

### *1.1. Causes of Air Pollution*

Some major causes of air pollution are discussed below.

- Industrial exhaust
  Emission of harmful gases such as $SO_2$ and $NO_x$ from thermal power plants of Rajghat, Badarpur, Indraprastha and other industrial regions adds to the major pollutants of the Delhi air pollution.
- Vehicular emission
  Traffic congestion and vehicular emission contributes majorly to degrading the Delhi air quality. The data accessed from the transport department of Delhi government up to 31-Dec 2016 puts the total number of registered vehicles to 1, 01,06,791. The largest number of the registered vehicles in the city are motor cycles and scooters , numbering 63,40,136 . These are considered to be major contributors towards air pollution [2].
- Agricultural stubble burning in Punjab and Haryana .Farmers of Punjab and Haryana burning their rice crop stubble to quickly prepare their field for rabi crop wheat [5].
- Construction and demolition
  Continuous construction and demolition contributing to increased level of dust-borne particulate matters in the air and are, therefore, considered hazardous [6].
- Other factors
  Some factors that may indirectly involve in worsening air quality are over -population, road dust, diwali cracker smoke etc.

### *1.2. Type of pollutants in ambient air*

The major concentration of pollutants in the Delhi air is:-
1. Particulate Matter, RSPM and SPM (PM2.5 and PM10): The principle source of particulate matter in Delhi is vehicular emissions, particularly from heavy motor diesel vehicle, kerb-side dust, thermal power plants, industrial and residential combustion processes. Respirable suspended particulate matter (PM2.5) is considered to be more hazardous to human health than PM10. The average limit of PM2.5 pollution is 60 microgram per cubic meter but all the areas of Delhi have the level of PM 2.5 exceeding 300 microgram per cubic meter [7].
2. Nitrogen Oxide ($NO_x$): Oxides of Nitrogen are produced during industrial combustion processes and primarily as vehicular exhaust. $NO_x$ levels are highest in urban areas as it is related to traffic. It is an important ingredient in generation of photochemical smog which envelops the urban air with haze like blanket. It has harmful effects such as wide-range of respiratory problems in adults and children.
3. Sulphur Dioxide ($SO_2$): It is formed mostly by burning of fossil fuels particularly from thermal power plant. This

pollutant is the reason for acid rain and has adverse effects on lung functions.
4. Benzene: The main sources of benzene are from vehicle exhaust and other industrial processes since it is an industrial solvent. Benzene is a component of crude oil and petrol. Apart from vehicle exhaust, evaporation from petrol filling stations can raise benzene levels [7].
5. Ozone ($O_3$): Formed by chemical reaction of volatile organic compounds and nitrogen dioxide in the presence of sunlight, so level of ozone is generally higher in the summer. Ground level ozone also contributes in formation of photochemical smog.
6. Toluene**:** Toluene is another industrial volatile solvent whose short term exposure causes irritation of eyes and the respiratory tract. The substance is a known carcinogen and affects the central nervous system also.
7. Carbon Monoxide (CO): CO is a toxic air pollutant which is produced by incomplete combustion of carbon-containing fuels. Vehicle deceleration and idling vehicle engines are one of its main causes.

### *1.3. Study Area*

Delhi has been considered for the study whose geographical regions are shown in figure 1. Delhi is one of the most polluted cities in the world according to World Health Organization (WHO) 2016 [8]. The level of the airborne particulate matter- PM2.5 is very high in Delhi. Also, PM 10 levels is the highest among the 11 mega cities of the world having more than 14 million habitants (2011 to 2015). PM is considered to be the most harmful pollutants to health [8].
There are 20 air monitoring stations at present in Delhi located at different locations. Different areas of Delhi have been selected for the study: Anand Vihar, Shadipur, DTU, Dwarka. As a representative sample, Anand Vihar(AV) and Shadipur, due to their high population density and heavy vehicular movement are represented and shown in the results.

Fig.1. Map of Delhi [9]

## 2. Related Work

Considerable research has been carried out in the subject area of air pollution data analysis along with its interpretation. An inventory of emission data for Delhi region along with study of climate and atmospheric chemistry from 1900 t0 2000 have been prepared by B.R. Gurjar [10]. Other researchers working in the field of air quality studies are Rati Sindhwani and Pramila Goyal , have studied the trend of atmospheric pollution in Delhi from 2000 to 2010 are Rati [11]. Dr Aaron J Cohen et al have carried out ambient air study from 1990 to 2015 (25 years) on global scale to ascertain the burden of disease related to pollution of ambient air. Their research identified air pollution as the main cause of global disease burden [3]. Shaddick, G et al. have used model of data

integration to study air quality .Thereafter the author have taken an hierarchical approach to estimate the exposures of ambient air pollution on global scale [12].
Authors have studied the long term air pollution characteristics in Wuhan, China and found the correlation between PM 2.5 and PM 10 [13]. Ozgur Kisi, Kulwinder Singh have made a model to ascertain pollutants concentration by adopting statistical techniques like multivariate adaptive regression spline, least square support vector regression and M5 model tree models [14]. Pramila Goyal has used Artificial Neural Networks for predicting the air pollution at Taj Mahal in Agra in [15]. Ping-Wei Soh, Kai-Hsiang Chen et al have studied spatial-temporal pattern analysis for prediction of air quality in Taiwan [16].Authors have carried out assessment of economic cost pertaining to adverse health effects and disability adjusted years of life due to $PM_{2.5}$ and $PM_{10}$ pollution in Mumbai and Delhi, in India from 1991 to 2015 in [17].

S.Taneja and N. Sharma have studied the air quality trend from 2011-2015 using multi perceptron technique and proposed detailed trends in air pollution in Delhi for 2011to 2015 in [18].However there is a scope for further studies due to susceptibility of population towards increasing air pollution and to find ways to monitor and control it in effective manner.

In the presented work, time series regression forecasting for analyzing the pollution trends of 2016 to 2017 in Delhi has been used. The critical analysis is followed by predicting the trend till 2022 by considering historic data from 2009-2017. Techniques that have been used are: descriptive, predictive analysis and time series regression forecasting.

## 3. Proposed work

### *3.1 Data set used*

The data has been collected from CPCB [19]. A snapshot of the dataset used is shown in figure 2. The dataset contains twelve attributes: Date, Benzene, NO, $NO_2$, Toluene, $NO_x$, O3, PM10, PM2.5, $SO_2$ and CO. The 'Date' attribute describes the sampling date and other parameters give their individual concentration in air. The data has been collected from 2009 to 2017 for critical analysis and accurate prediction for 'Shadipur station'.

| | Date | benzene(u | NO | NO2 | tolune | Nox | O3 | pm2.5 | pm10 | PXY | SO2 | CO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 07/01/20 | 3.04 | 250.71 | 112.15 | 7.75 | 442.3 | 22.44 | 469.61 | 742.25 | 0.56 | 32.61 | 1.14 |
| 3 | 08/01/20 | 2.28 | 209.9 | 95.16 | 4.03 | 371.36 | 12.09 | 519.68 | 727.35 | 0.51 | 13.9 | NA |
| 4 | 09/01/20 | 0.75 | 151.82 | 85.5 | 1.01 | 283.39 | 14.22 | 169.14 | 476.08 | 0.39 | 16.06 | 1.21 |
| 5 | 10/01/20 | 1.3 | 267.57 | 108.85 | 1.04 | 462.4 | 15.1 | 280.7 | 519.8 | 14.65 | 18.22 | 2.5 |
| 6 | 11/01/20 | 1.87 | 400.27 | 125.3 | 6.28 | 653.32 | 39.62 | 408.91 | 681.16 | 18.11 | 22.41 | 2.64 |
| 7 | 12/01/20 | 1.35 | 168 | 93.15 | 6.15 | 312.42 | 25.69 | 289.21 | 560.1 | 9.16 | 26.45 | 2.61 |
| 8 | 13/01/20 | 0.81 | 63.31 | 81.95 | 1.44 | 159.49 | 15.06 | 312.28 | 510.49 | 10.62 | 16.66 | 2.71 |
| 9 | 14/01/20 | 0.62 | 98.16 | 69.33 | 0.65 | 195.81 | 10.74 | 258.55 | 475.71 | 5.61 | 13.92 | 1.75 |
| 10 | 15/01/20 | 0.43 | 98.76 | 59.14 | 0.71 | 187.52 | 10.88 | 183.25 | 344.5 | 19.14 | 13.34 | 2.53 |
| 11 | 16/01/20 | 0.34 | 75.9 | 60.6 | 0.48 | 157.64 | 13.65 | 143.24 | 299.33 | 30.5 | 15.44 | 3.52 |
| 12 | 17/01/20 | 0.61 | 81.51 | 69.2 | 0.67 | 172.97 | 15.16 | 200.8 | 386.68 | 41.13 | 20.15 | 2.04 |
| 13 | 18/01/20 | 1.33 | 301.02 | 96.74 | 1.79 | 497.27 | 14.01 | 339.6 | 672.07 | 33.38 | 16.95 | 2.53 |
| 14 | 19/01/20 | 0.75 | 105.51 | 67.62 | 0.97 | 204.08 | 9.79 | 323.85 | 566.72 | 15.61 | 13.01 | 2.36 |
| 15 | 20/01/20 | 0.43 | 89.06 | 73.91 | 0.63 | 187.5 | 10.61 | 307.65 | 531.77 | 13.92 | 12.44 | 3.81 |
| 16 | 21/01/20 | 0.53 | 84.65 | 69.85 | 0.66 | 177.81 | 11.7 | 235.53 | 422.96 | 25.84 | 13.29 | 1.65 |
| 17 | 22/01/20 | 10.93 | 66.98 | 73.52 | 19.58 | 156.97 | 13.54 | 300.89 | 466.73 | 56.18 | 15.53 | 1.59 |
| 18 | 23/01/20 | 37.35 | 123.87 | 68.47 | 105.7 | 230.17 | 15.71 | 360.79 | 544.08 | 3.81 | 17.07 | 4.9 |
| 19 | 24/01/20 | 39.61 | 87.29 | 111.41 | 59.48 | 218.47 | 10.74 | 388.06 | 586.09 | 9.3 | 24.01 | 2.77 |
| 20 | 25/01/20 | 31 | 58.86 | 79.39 | 54.3 | 151.12 | 10.81 | 275.88 | 510.62 | 14.29 | 14.92 | 1.61 |
| 21 | 26/01/20 | 31.33 | 168.42 | | 47.21 | 317.6 | 12.58 | 374.62 | 569.33 | 22.81 | 21.06 | 1.76 |
| 22 | 27/01/20 | 35.19 | 275.8 | 120.94 | 109.93 | 484.38 | 15.16 | 338.67 | 532.39 | 17.62 | 18.31 | 3.82 |
| 23 | 28/01/20 | 41.98 | 330.54 | 124.23 | 97.87 | 562.08 | 10.51 | 354.83 | 652.46 | 7.58 | 16.24 | 4.29 |

Fig.2. Snapshot of the dataset

### *3.2. Proposed approach*

A systematic approach has been followed in this analysis which is depicted in figure 3. The approach starts with the collection of dataset from CPCB [19]. Collected data has been pre processed to remove the redundancy. Pre processing of data includes steps like parsing of dates, noise removal, cleaning, training and scaling. Further, descriptive analysis has been carried out on two different platforms- Rstudio [20] and Tableau [21] for different stations. For observing the forecasted results, predictive analysis has been done.

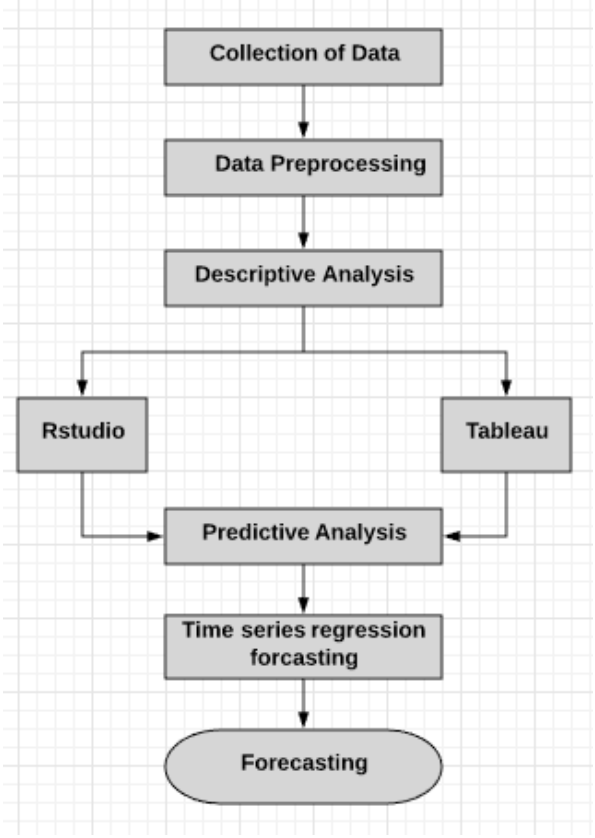


Fig.3. Flowchart of proposed approach

### *3.3 Methods and Techniques involved*

1. Descriptive analysis

Descriptive analysis has been used in the study to analyze the basic characteristics of data. Summaries about the sample and the measures have been observed through descriptive analysis.

2. Predictive analysis

Predictive analytics is the use for statistics and machine learning techniques to predict about the future (unknown data). The goal is to predict the probable future through past experience. Similarly in our analysis, predictive analysis has been done by implementing time series regression forecasting.

3. Time series regression forecasting

Time series regression forecasting is used for analyzing time series data in order to study the behaviour of data with respect to time. This method involves the use of a model to forecast future values on the basis of previously observed values. Time series method is used in various applications like weather forecasting, earthquake prediction and trajectory forecasting and pattern recognition etc.

## 4. Results

Results for the stations, Anand Vihar (AV) and Shadipur are presented. Descriptive analysis carried for Anand Vihar Station, 2016-2017 shows drastic increase in PM10 level. $NO_2$ and PM2.5 have evidently increased contributing to the increased pollution in Delhi in figure 4.The comparative study shows four quarters where Q1 is Jan'16- July '16, Q2 is Aug'16-Dec'16, Q3 is Jan'17-July'17 and Q4 is Aug'17-Dec'17 in fig.4. The trend followed by carbon monoxide is predicted to reduce present year by 0.169 mg/m$^3$. Also, the reducing trend is seemed to be followed for consecutive years. This can be seen in the fig 6.
The trend of nitrogen dioxide ($NO_2$) in air is predicted to drastically increase for the coming years by 16.77ug/m$^3$. A gradual rise in $NO_2$, often breaching the safe standard, is shown in figure 7.Ozone ($O_3$) is estimated to increase by 6.11 mg/m$^3$ in the upcoming years .Benzene trend and the predicted trend seems to reduce by 1.33 mg/m$^3$.

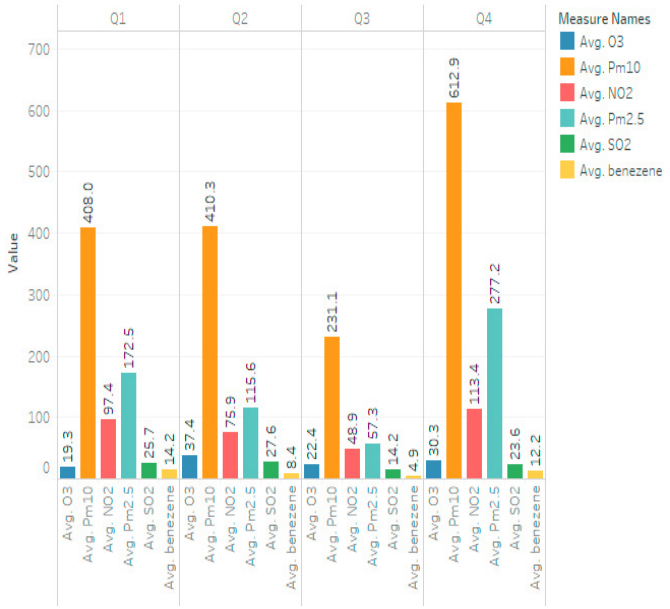

Fig.4. Descriptive analysis of AV station

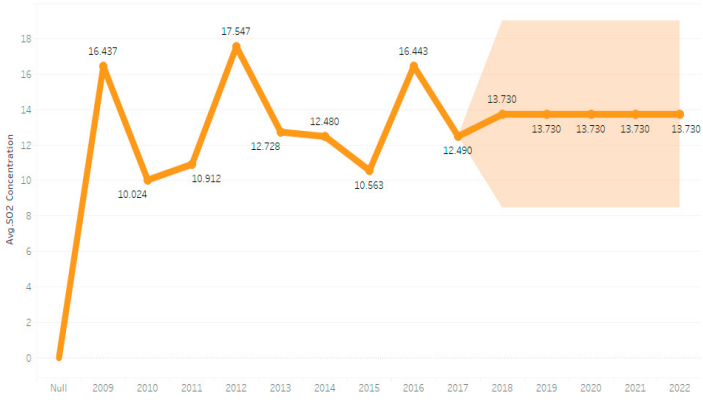


Fig.5. Trend of $SO_2$ in ug/$m^3$

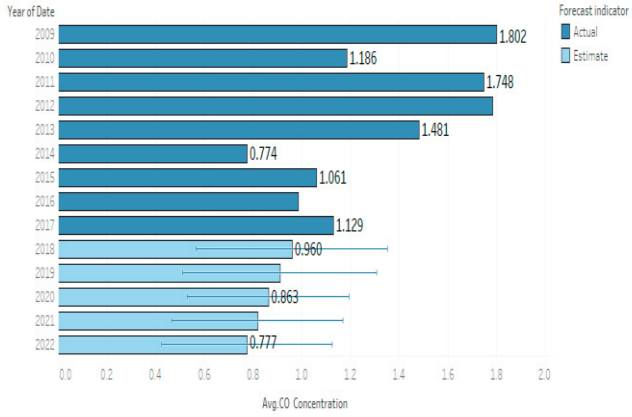


Fig.6. Trend of CO in mg/$m^3$


Fig.7. Trend of $NO_2$ in ug/$m^3$

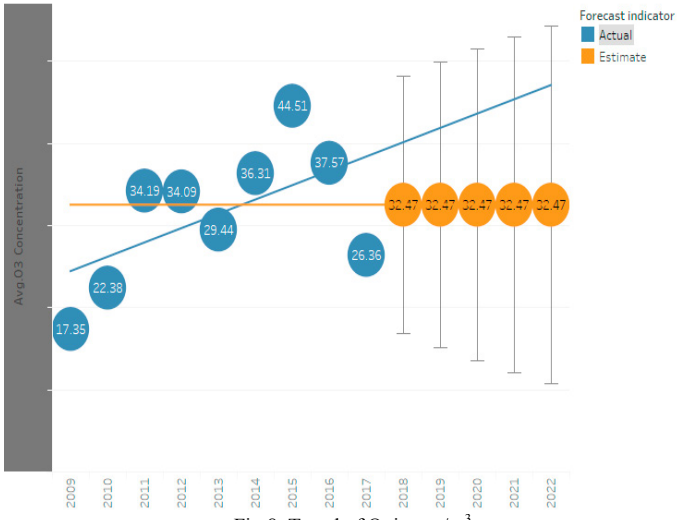


Fig.8. Trend of $O_3$ in mg/m$^3$

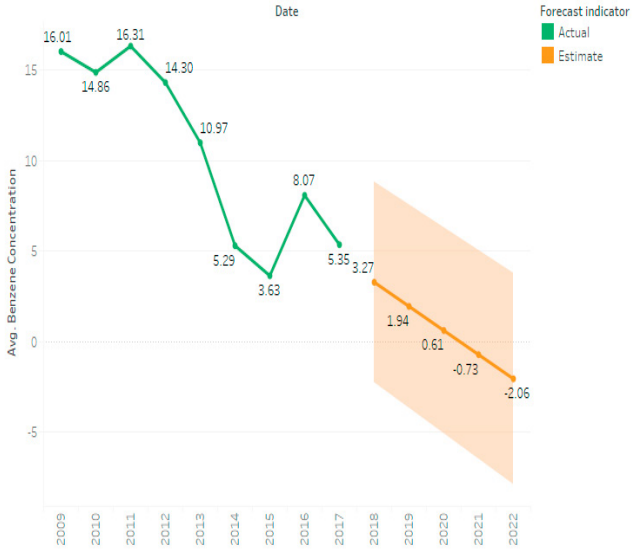


Fig.9. Trend of benzene in mg/m$^3$

The observed result indicates that pollutants like $NO_x$, PM10, PM 2.5 are likely to drastically increase in future, while $SO_2$ level may increase marginally in future. $O_3$ levels will increase in initial years. Though the amount of CO and Benzene are showing reducing trend.

## 5. Conclusion and future scope

The agenda of this study is to use technology for creating awareness to decrease pollution by adopting proper measures. New Delhi, which has already been ranked among the world's most polluted city, is considered in the study. Anand Vihar and Shadipur areas having high population load were selected. This study focused on the implementation of time series regression forecasting, data mining techniques to fathom the different patterns in various types of pollutants. R studio platform and tableau have been used for predicting future pollution levels in Delhi using R language.

As the result indicates, the increasing trend of NOX in coming years can be attributed to increasing number of

vehicles on road, emissions from nearby industrial zone and thermal power plant operations. Increase in PM10 and PM 2.5 can be attributed to kerb-side dust, construction work and crop stubble burning in adjacent states . The aggravated NOX, particulate matter (PM10 and PM2.5) and ground level O3 also leads to development of smog pollution.

Pollution prevention requires public-policymakers participation which sometimes becomes a major limitation. Atmospheric variables like wind, precipitation, humidity do not remain static which also pose a problem in long range forecasting.

The results proposed are based on the historic data, hence cannot take account the future events that can manipulate the proposed results. The predicted results can be verified further from different techniques.

The future scope of the work is to explore and work upon various data analytical techniques to build a forecasting model which is adaptable to dynamic atmospheric variables. Appropriate forecasting model of ambient air pollution is prerequisite in developing stringent pollution control technology and measures thereof. Emphasis to explore clean energy fuels and gradually phasing out fossil fuels, employing zero waste technology with integrated waste management will curb air pollution menace in due course.

## References


[1] Seinfeld J. H, and Pandis S. (2006). *Atmospheric chemistry and physics*. 2nd ed. Hoboken (NJ): John Wiley.

[2] Statistical Abstract (2016) Delhi Govt Portal, www.delhi.gov.in.

[3] Cohen, Aaron J, Brauer, Michael et al. (2017) "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study", *The Lancet*389 (10082): 1907-1908.

[4] Rizwan SA, Nongkynrih B, Gupta SK. (2013) "Air pollution in Delhi: Its Magnitude and effects on health." *Indian J Community Med*38 (1):4-8.

[5] Kumar, P, Kumar, S and Joshi, Laxmi. (2014) "Socioeconomic and Environmental Implications of Agricultural Residue Burning, A Case Study of Punjab, India." *Springer Briefs in Environmental Science*, DOI 10.1007/978-81-322-5_1.

[6] Technical report .(2017) "CPCB Guidelines on Environmental management on C & D Wastes (Prepared in compliance of Rule 10 sub-rule 1(a) of C & D Waste Management Rules, 2016).

[7]Realtime ambient air quality data of Delhi, DPCClink:http://www.dpccairdata.com/dpccairdata/display .

[8]WHO's Urban Ambient Air Pollutiondatabase Update 2016 http://www.who.int/phe/health_topics/outdoorair/databases/AAP_database_summary_results_2016_v02.pdf.

[9]Sindhwani, R. (2012) "Assessment of gaseous and respirable suspended Particulate matter (PM10) emission estimates over megacity Delhi: past trends and future scenario (2000–2020)." *13th Annual CMAS Conference,* Chapel Hill, NC, USA.

[10] Gurjar, B.R., Aardenne, J.A. van, Lelieveld, J., Mohan, M. (2004) "Emission estimates and trends (1990–2000) for megacity Delhi and implications." *Atmospheric Environment* :5663–5681.

[11] Sindhwani, R., Goyal, P. (2014) "Assessment of traffic–generated gaseous and particulate matter emissions and trends over Delhi (2000–2010)." *Atmospheric Pollution Reseach*5 (3): 438-446.

[12] Shaddick, G, Thomas, ML, Jobling, A et al. (2016 ) "Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution."

[13] Ma, Xin, Gong, Wei, Zhu, Zhongmin. (2016) "The study of long-term air pollution characteristic in Wuhan, China." *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International* DOI: 10.1109/IGARSS.2016.7730073

[14] Kisi, Ozgur Singh, Kulwinder Soni, Kirti. (2017) "Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models." *Air Quality, Atmosphere & Health*10 (7): 873–883

[15] Mishra, D., Goyal, P. (2015) "Development of artificial intelligence based NO2 forecasting models at Taj Mahal, Agra centre for atmospheric sciences", *Atmospheric Pollution Research*: 99-106

[16] Soh, Ping-Wei, Chen, Kai-Hsiang, Huang, Jen-Wei. (2017) "Spatial-temporal Pattern Analysis and Prediction of Air Quality in Taiwan." *10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)*, DOI: 10.1109/UMEDIA.2017.8074094

[17] Kumar, Anil, Kamal, Dikshit, Maji, Jyoti, Deshpande, Ashok. (2017) "Disability-adjusted life years and economic cost assessment of the health effects related to $PM_{2.5}$ and $PM_{10}$ pollution in Mumbai and Delhi, in India from 1991 to 2015." *Environmental Science and Pollution Research*24 (5): 4709–4730.

[18] Taneja, Shweta, Sharma, Nidhi, Oberoi, Kettun, Navoria, Yash. (2016) "Predicting Trends in Air Pollution in Delhi using Data Mining." Information Processing (IICIP), 2016 1st India International Conference, DTU, New Delhi.

[19] CPCB report on Air pollution in Delhi ; An analysis (2016) .

[20] RStudio. (2018) [Online] Available at: https://www.rstudio.com/products/rstudio2/ [Accessed 15 Jan. 2018].

[21]Tableau Software. (2018). Tableau Desktop. [Online] Available at: https://www.tableau.com/products/desktop [Accessed 16 Jan. 2018].