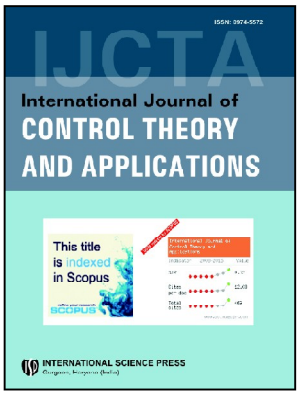

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 6 • 2017

Forecasting of Air Pollutants Particulate Matter and Carbon Mono-oxide concentration in Delhi City using ARIMA Model and Artificial Neural Network

Monika Singh[1], S.P. Mahapatra[2], Sudhir Nigam[3], Pradeep Singh[4] and Kuhu Gupta[5]

[1,2] *Department of Chemistry, National Institute of Technology, Raipur, Chhattisgarh, India*
[3] *Department of Civil, Laxmi Narain College of Technology and Science, Bhopal, Madhya Pradesh,, India*
[4] *Department of Computer Science, National Institute of Technology, Raipur, Chhattisgarh, India*
[5] *Department of Information Technology, National Institute of Technology, Raipur, Chhattisgarh, India*
E-mails: *moniica.singh21@gmail.com[1], spmahapatra.chy@nitrr.ac.in[2], nigam.sudhir@hotmail.com[3], psingh.cs@nitrr.ac.in[4], kuhu.gupta08@gmail.com[5]*

Abstract: The study generally deals with the significance of stochastic modeling technique in forecasting of possible concentration of various pollutants present in the air. The ITO square in New Delhi, the capital of India is the site of analysis. For the study, the dataset has been collected from CPCB (Central Pollution Control Board) for the period of nine years i.e from 2007 to 2015. In this study, a Feed-forward neural network with a single hidden layer and a seasonal ARIMA (Autoregressive Integrated Moving Average) model is applied for forecasting of particulate matter (PM10 and PM2.5) and Carbon monoxide, using the R Package. R-square, R-square, Root Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error are studied for pollutants Carbon Mono-oxide and Particulate Matter PM 10 and PM 2.5. The selected model has been further used to forecast concentration of pollutants for the year 2016 from the previous data of 2007 to 2015. Suitable results have been obtained using the selected ARIMA models and Neural Network and then compared to find which generally reveals its application and signifying its use as a modeling technique for determining the future values of concentration of pollutants and its use in different areas.

Keywords: Air Pollutant, ARIMA Model, carbon mono oxide, particulate matter, forecasting, time series analysis, Neural Network

I. INTRODUCTION

India is developing country in this era of industrialization and modern technology where we are facing major problem of Air pollution. Now a day's air pollution has become an important apprehension because of its adverse effect on the biosphere[1]. Air pollution is caused by both natural and by many anthropogenic activities which results in the release of many unwanted substances in the air. The natural source which causes pollution are fires of dense and dry forest, volcanic eruption, wind erosion, pollen scattering, vaporization of organic compounds

and radioactivity [2]. Human activities such as deforestation, urbanization, lack of knowledge, high birthrate over death rate, vehicle exhaust, lack of technology and unawareness toward air policy management are some of the frequent causes of air pollution and are treated as major causes of air pollution [3]. Due to these activities air pollutants concentration are increasing at hurried rate .Massive quantity of air pollutants which can be carbon mono oxide , oxides of Nitrogen, oxides of Sulphur, particulate matter may causes health issues such as breathing problem , heart disease, lung infections and even cancer and death of the person which is in continuous contact with the abundant quantity of these pollutants .Pollutants also cause adverse consequence on the terrestrial ecosystem also and they damage vegetation, flora and fauna [4]. The World Health Organization reports states that 2.4 million folks pass away per annum from the effect of air pollution, 1.5 million of this number pass away from indoor air pollution. All over world, the poor quality of air has caused more deaths than from automobile accidents [5]. Therefore with the increase in sources of pollutants it has become important to keep an eye on the variation of concentration of pollutants in ambient air in urban as well as rural areas [6]. For the successful remedy of these environmental pollutants we needs analysis at the base and identification of pollution sources, conveying ability of pollutants in the atmosphere, anticipation and its controlling methodologies, and economic impact on the particular area. After the consequences obtained from such an analysis, the government and pollution control board and authorities should take necessary steps to overcome the harmful hazardous caused air pollution of that particular area [7]. Different analysis and computational methods are developed to know and predict the concentration of pollutants respectively. Analysis of regional air pollution problems and their solutions is effectively performed through the application of computer modeling techniques. Therefore, it becomes essential to compare these techniques and find out the best technique for forecasting. In this study, we use Box e Jenkins autoregressive integrated moving average (ARIMA) model based on time series modeling and Neural Network. A vital postulation in time series analysis is forecasting of standards in future period which depends on the chronological sequence of observations of the variable under study [8]. Method is used to generate synthetic series with the same persistence structure as an observed series, and also to predict behavior of a time series from past values [9]. In this study, ARIMA models and Artificial Neural Network (ANN) are used in air pollution modeling with respect to future prediction of concentrations of particulate matter (PM10 and PM2.5), and carbon mono-oxide .The current work suggests models based on a stochastic process (ARIMA Model) and Neural Network that uses the characteristics of the actual data and forecast its future values [10].

II. MATERIALS AND TECHNIQUES

2.1. Site description and data collection

The study area is capital of India "Delhi" ITO SQUARE. Being the busiest place of India and is the combination of commercial and industrial hub due to which it show variation in air quality. The area Selected is ITO SQUARE because it has been observed an area of interaction of residential, commercial as well as industrial area. Being an area of interaction area is facing the problem of pollution and amalgam of pollutants. New Delhi, the urban capital of India, is situated between the latitudes of 28_0.210 to 28_0.530 North and longitudes of 76_0.200 to77_0.370 East and 213.3 to 305.4 m above the mean sea level. The megacity (Fig. 1) is among the highly polluted cities in the world, with a rapidly increasing number of vehicles and industries as well as educational and job hub in the central India [11].The major anthropogenic sources of pollution in Delhi are increasing day by day by increase in migration of population, transportation ,incomplete combustion of fuel and biomass, energy consumption which result in many health problem such as breathing disorder, lungs dysfunction, vision impairment, lose in commercial properties and effects are seen on flora and fauna [12].

CPCB (central pollution control board) has an automatic monitoring station in ITO interaction in New Delhi. At this station respirable suspended particles, carbon monoxide, carbon mono oxide, Sulphur di-oxide, nitrogen dioxide and suspended particulate matter are being monitored and information is weekly updated. The data used for this experiment is collected from CPCB's websitehttp://www.cpcb.gov.in[13] .Data set consist of

Figure 1: The Projected view of Delhi from the map of India

monthly average data pollutants from2007 to 2015of pollutants Particulate matter 2.5 and 10, and Sulphur oxides. The missing values are calculated by taking simple mean and linear interpolation method.

2.2. Box e Jenkins ARIMA model

The Autoregressive Integrated Moving Average Model is a vital method of time series forecasting proposed by Box and Jenkins in 1970s [14]. When this ARIMA model is recognized, the future values can be foreseen by the time series data of past and future observations [15].The purpose of the time series modeling is to evaluate and forecast the prospect variation of concentration structured on previous concentration measured. Extensively used model for fulfill the purpose is the Box e Jenkins Autoregressive Integrated Moving Average commonly known ARIMA model. The autoregressive (AR) section of ARIMA demonstrates that the target variable is regressed on its own former values. The regression error is linear conjunction of error terms whose values appeared concurrently and occasionally in the past which is specified by the Moving average (MA) section. The Integration (I) section replaces the data values with the difference between their past and present observations. Building model to fit the data as much as possible is the basic aim of these sections [16].

2.2.1. Seasonal ARIMA (SARIMA) model

Seasonality in a time series is a habitual trend of changes that repeats after S time periods until the pattern repeats again. The monthly dataset presents a seasonal period of 12 months. To determine seasonality in dataset, the Auto correlation function (ACF) of the time series is examined for it being significantly different from zero. In order to obtain an inactive seasonal time series data, seasonal differencing is performed by taking difference between the present and corresponding observation from the previous data. In this variant of ARIMA model, data values and errors from former times with lags that are multiples of S are used by seasonal AR and MA terms for forecasting [17]. The seasonal ARIMA model is denoted as

$$\text{ARIMA}\ (p, d, q) \times (P, D, Q)\ S \tag{1}$$

p = non-seasonal AR (AUTO REGRESSION) order,

d = non-seasonal differencing,

q = non-seasonal MA (MOVING AVERAGE) order,

P = seasonal AR order,

D = seasonal differencing,

Q = seasonal MA order, and

S = time span of repeating seasonal pattern. [18]

The seasonal ARIMA model includes both non-seasonal and seasonal elements in an augmented model. The used ARIMA model is a variation of the Hyndman and Khandakar algorithm in combination with unit root tests, minimization of the AICc and MLE [19]. The model having the minimum Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) values is selected as the best fit model. We have used default setting of auto ARIMA of the forecast package in R.

2.3. Artificial Neural Network (ANN)

An artificial neural network (ANN) is a computational system made up by the arrangement and interconnection of simple processing biological neurons which processes information. There are variations in artificial neural network according to the nature of the task to be processed by the network, therefore, it possesses the capability to give answer to few problems which have been unsolved by ordinary computers. Neural Network aims to give human-like performance when working in fields of speech processing, image recognition, machine vision, and robotic control. Multiple inputs can be given to one neuron, while there can be only one output. The inputs could be external stimuli or could be output from other neurons. A copy of the neuron's output could be input to itself as a feedback. Certain weight is associated with the each connection of neuron and on exceeding a certain threshold, the neuron is fired and output signal is produced. In the proposed work, a neural network is used for forecasting.

2.3.1. Implementation of neural network model

In the proposed work, Feed-forward neural networks is used with a single hidden layer and lagged inputs for forecasting univariate time series.

1. The input layer with lagged values of y (a numeric vector) and single hidden layer with size nodes is given to model a feed forward network.
2. The inputs are for lags 1 to p [number of non-seasonal lags] and lags m to mP where m=frequency(y).
3. Then many networks are fitted with random starting weights which are averaged to produce forecasts.

The fitted model is depicted as NNAR (p, k) model, where k is the number of hidden nodes for non-seasonal data while for seasonal data, the fitted model is called an NNAR(p, P, k)[m] model.

III. RESULTS AND DISCUSSION

3.1. Accuracy Tests

The prediction results are tested by error testing indexes to inspect the validity and precision of the results. Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Mean Absolute Scaled Error (MASE), Autocorrelation Function (ACF) and Box Pierce Test are the error testing indexes [20]. The difference between the values forecasted by a model and the observed values is an often used measure known root-mean-square error (RMSE) while mean absolute error (MAE) is a measure used to quantify the difference between predictions and eventual outcomes. This index has the same units as the data, and magnitude but is slightly smaller than the root mean squared error. Since it does not involve squaring the errors in calculation,

it is less sensitive to large error. In statistics, it is necessary to measure prediction accuracy for several forecasting method. Therefore, this measure is known as the mean absolute percentage error (MAPE).The prediction accuracy is expressed as a percentage. Kohler and Hyndman introduced mean absolute scaled error (MASE) as a measure of forecast accuracy. It is the ratio of mean absolute error produced by the actual forecast to MAE in-sample, naive is the mean absolute error generated by a naive forecast, calculated on the in-sample data. While autocorrelation function is the association of a time series with its own past and future values. It is also known as "lagged correlation" or "serial correlation". It represents the relation between numbers that are arranged in time [21]. After finding the error testing indexes; perform the Box-pierce test to suggest a better approximation to the null hypothesis distribution. In the proposed work, Box Pierce test is performed on both ARIMA model and artificial neural network model. These indexes are calculated and the results are shown in Table 1.

3.2. ARIMA Results

By ARIMA Modeling technique, following plots are obtained for PM 10, PM 2.5 and CO. Figure 2 shows the best fitted ARIMA (0,1,0)(1,0,1)[12] model for forecasting of pollutant PM 10 by using data of 2007 to 2015 as known data and data of 2016 is forecasted. And Figure 3 shows the variations of standard residual, ACF of residuals and p value of L-Jung statics of PM 10 pollutant. Figure 4 shows the best auto fit ARIMA(1,0,2)(2,0,1)[12] model for forecasting of pollutant PM 2.5 by using data of 2007 to 2015 as known data and data of 2016 is forecasted. Figure 5 shows the variations of standard residual , ACF of residuals and p value of L Jung statics of PM 2.5 pollutant while Figure 6 shows the best auto fit ARIMA(1,1,1)(1,0,0)[12] model for forecasting of pollutant Carbon Monoxide (CO) by using data of 2007 to 2015 as known data and data of 2016 is forecasted. Figure 7 shows the variations of standard residual, ACF of residuals and p value of L-Jung statics of Carbon Monoxide (CO) pollutant. Different error values are discussed in Table 1 for ARIMA Model.

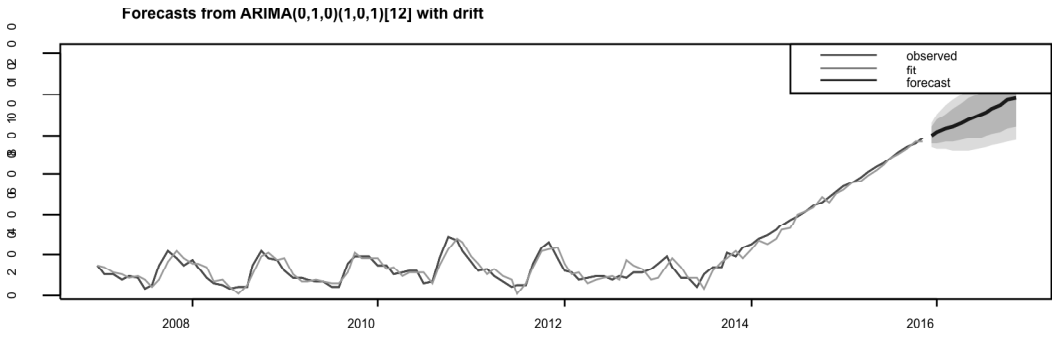

Figure 2: ARIMA(0,1,0)(1,0,1)[12] model of PM 10

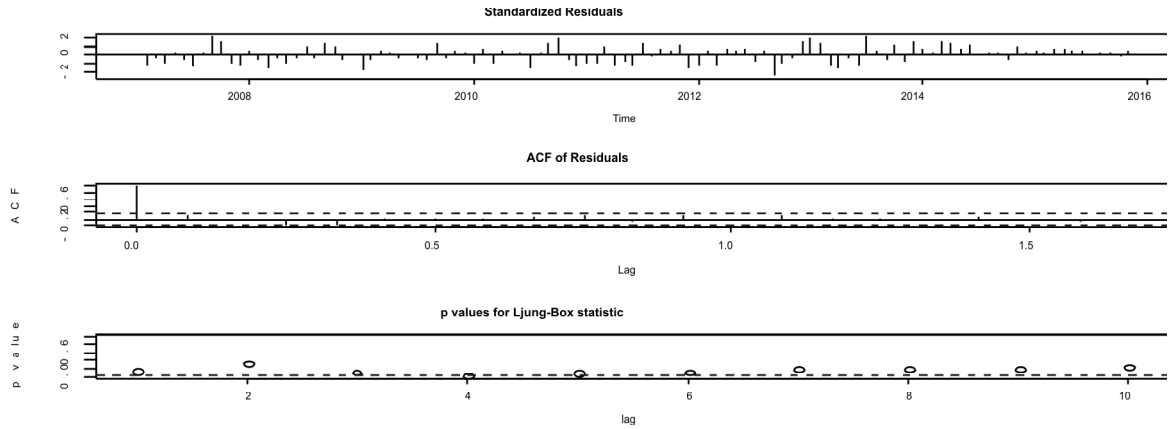

Figure 3: Variations of standard residual , ACF of residual and p value of L-Jung statics of PM 10

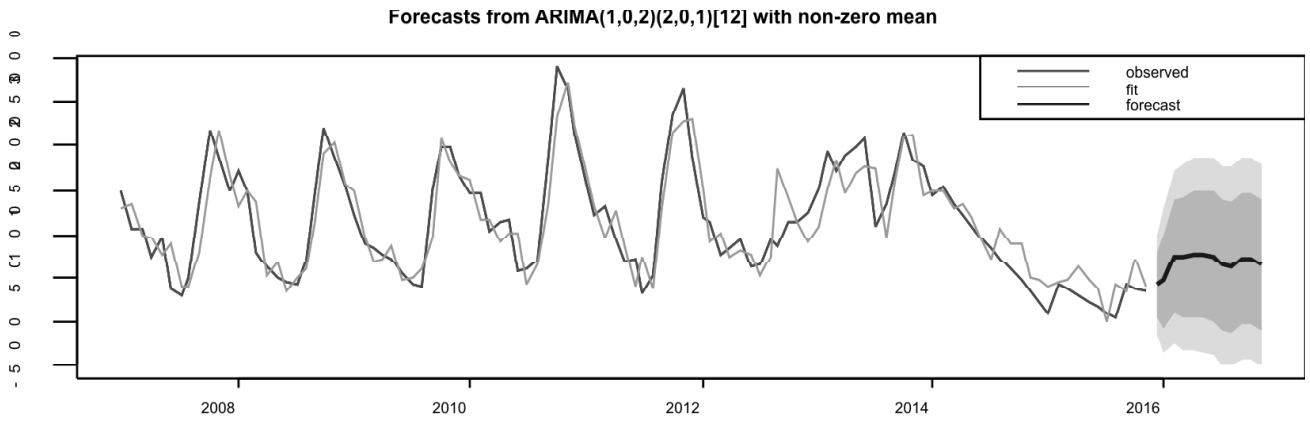

Figure 4: ARIMA(1,0,2)(2,0,1)[12] model of PM 2.5

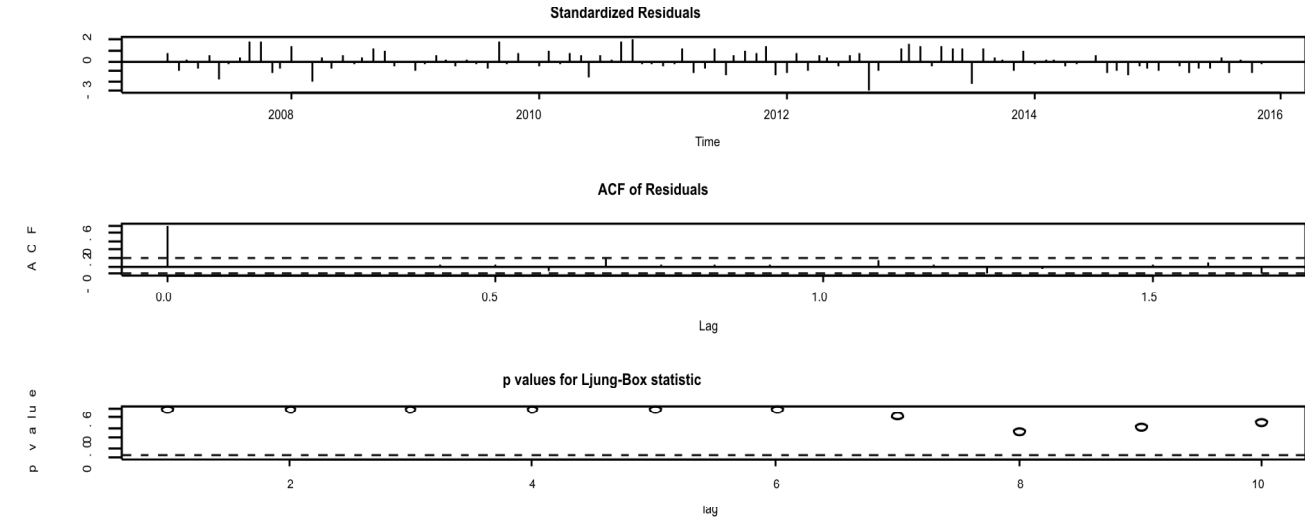

Figure 5: Variations of standard residual, ACF of residual and p value of L-Jung statics of PM 2.5

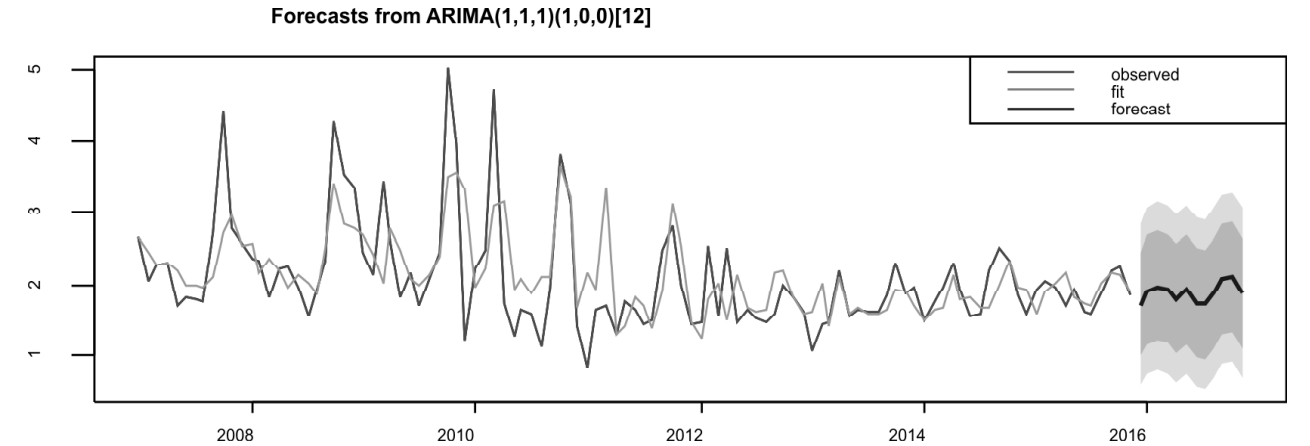

Figure 6: ARIMA(0,1,0)(1,0,1)[12] model of Carbon Monoixde (CO)

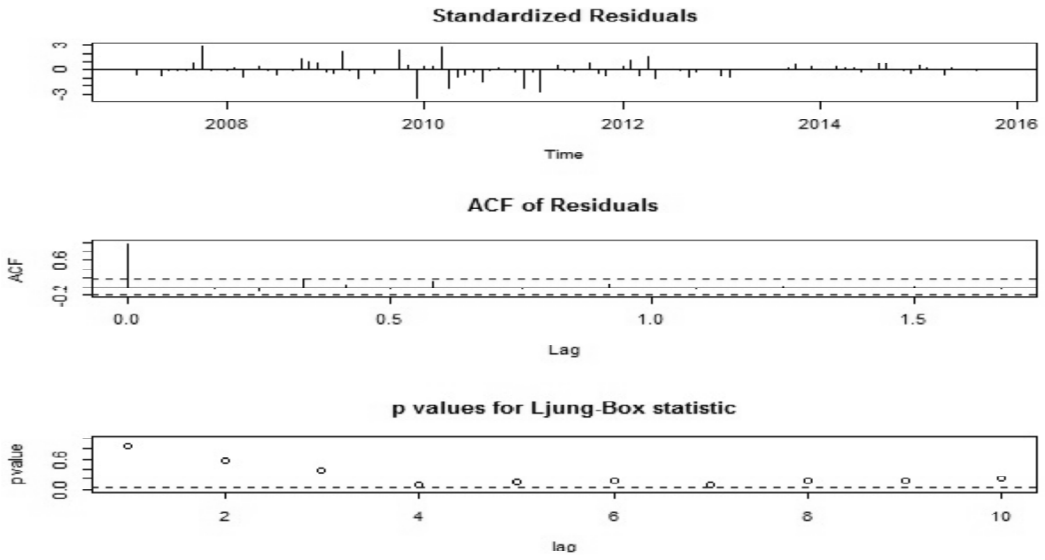

Figure 7: Variations of standard residual , ACF of residual and p value of L-Jung statics of CO

3.3. Neural Network Results

On application of Neural Network Modeling technique, following plots are obtained for PM 10, PM 2.5 and CO. For forecasting Neural network model of (1, 1, 2)[12] is found best fitted as shown in Figure 8 for PM 10 while for PM 2.5 forecasting Neural network model of (1,1,2)[12] is found best fitted as shown in Figure 9. For CO forecasting Neural network model of (1, 1, 2)[12] is found best fitted as shown in Figure 10.

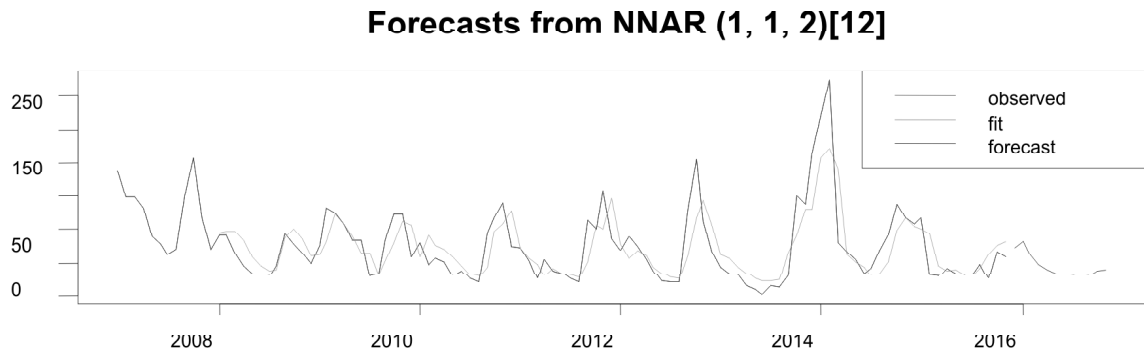

Figure 8: Forecast by Feed Forward Neural Network for PM 10

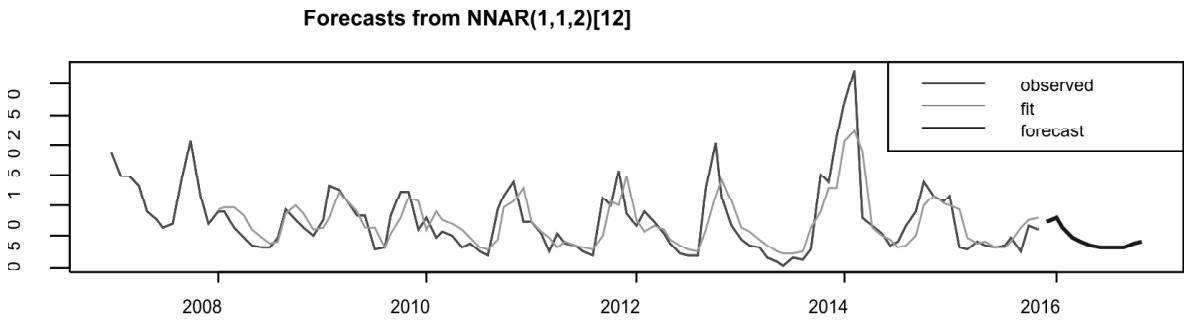

Figure 9: Forecast by Feed Forward Neural Network for PM 2.5

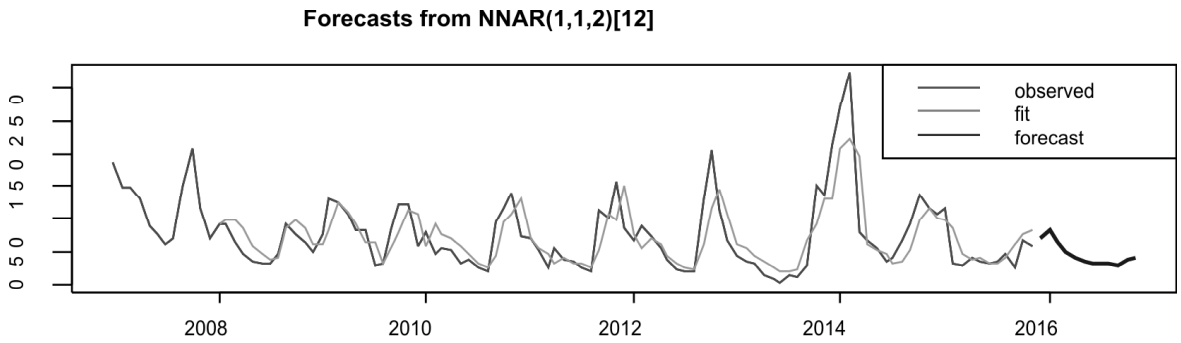

Figure 10: Forecast by Feed Forward Neural Network for CO

In the box-pierce test, the p-value calculated helps to determine the importance of data. It weighs the strength of the evidence, i.e, what is the data telling about the population. While df stands for the degrees of freedom and X-squared is the "goodness of fit" between observed and expected data. In Table 2 and 3 box pierce test for ARIMA as well as for neural network are discussed.

Table 1
Different accuracy data of ARIMA model

Pollutants	*sigma^2*	*ME*	*RMSE*	*MAE*	*MPE*	*MAPE*	*MASE*	*ACF1*
PM 2.5	839.6	-0.85	28.0	22.7	-18.1	36.4	0.47	-0.00
PM 10	839.6	-0.85	28.0	22.7	-18.1	36.4	0.47	-0.00
CO	**1520**	**1.81**	**38.2**	**25.8**	**-36.7**	**53.4**	**0.57**	**0.05**

Table 2
Results of Box pierce test of Feed Forward Neural Network

Pollutants	*Box-Pierce test*		
	X-squared	*df*	*p-value*
CO	0.30022	1	0.5837
PM 2.5	0.6705	1	0.4129
PM 10	**0.4069**	**1**	**0.5235**

Table 3
Results of Box pierce test of ARIMA Model

Pollutants	*Box-Pierce test*		
	X-squared	*df*	*p-value*
CO	0.045913	1	0.8303
PM 2.5	0.00013	1	0.9907
PM 10	**0.00013**	**1**	**0.9907**

IV. CONCLUSION

In this study, the methods used is Neural Network and ARIMA (Autoregressive Integrated Moving Average) model for simulating the long term forecasting, i.e. yearly average of pollutants particulate matter and carbon mono oxide present in ITO square in New Delhi, the capital of India. The evaluation of the selected model, the

ARIMA (0,1,0)x(1,0,1)12 is best suitable model for PM10 , ARIMA(1,0,2)(2,0,1)[12] is best suitable model for PM 2.5 and ARIMA(1,1,1)(1,0,0)[12] is best suitable model for CO. Feed forward neural network with one hidden layers Neural network model of (1,1,2)[12] is identified as the best fit model for PM10, PM2.5,CO . R-square, Root Mean Squared Error, Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) is studied for different pollutants as discussed in Table 1 for ARIMA model To conclude, the comprehensive forecasting ARIMA model and neural Network is studied for the data of particulate matter and carbon mono oxide. The scope of this research lied in the fact that pollution levels can be forecasted by several methods, but enhancing techniques like ARIMA model and Neural Network will make it more effective and it can be used for generation of best fitting curves in diverse research problems.

REFERENCES

[1] Homer-Dixon, T. F. (2010). Environment, scarcity, and violence. Princeton University Press.

[2] Smith, W. H. (2012). Air pollution and forests: interactions between air contaminants and forest ecosystems. Springer Science & Business Media

[3] Kampa, M., & Castanas, E. (2008). Human health effects of air pollution.Environmental pollution, 151(2), 362-367.

[4] World Health Organization. (2006). Air quality guidelines: global update 2005: particulate matter, carbon mono oxide, nitrogen dioxide, and sulfur dioxide. World Health Organization.

[5] Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. Plos med, 3(11), e442

[6] Murena, F. (2004). Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples. Atmospheric Environment, 38(36), 6195-6202

[7] Bilgen, S. (2014). Structure and environmental impact of global energy consumption. Renewable and Sustainable Energy Reviews, 38, 890-902.

[8] Manfren, M., Caputo, P., & Costa, G. (2011). Paradigm shift in urban energy systems through distributed generation: Methods and models. Applied Energy, 88(4), 1032-1048.

[9] Leland, W. E., Taqqu, M. S., Willinger, W., & Wilson, D. V. (1994). On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Transactions on networking, 2(1), 1-15.

[10] Sen, A., Ahammed, Y. N., Arya, B. C., Banerjee, T., Begam, G. R., Baruah, B. P., ... & Dhyani, P. P. (2014). Atmospheric fine and coarse mode aerosols at different environments of India and the Bay of Bengal during Winter-2014: implications of a coordinated campaign. Mapan, 29(4), 273-284.

[11] Goyal, P. (2003). Present scenario of air quality in Delhi: a case study of CNG implementation. *Atmospheric environment*, *37*(38), 5423-5431

[12] Thomas, M. D. (1961). Effects of air pollution on plants. Air pollution, 239

[13] PCB (Central Pollution Control Board). 2009. Indian National Ambient Air Quality Standards, New Delhi. Available on http://www.cpcb.gov.in

[14] Box, G.E.P.; Jenkins, G.M. Time Series Analysis: Forecasting and Control; Holden-Day: San Francisco, CA, USA, 1970.

[15] Zhang, G.P. Time series forecasting using a hybrid ARIMA and neural network model.Neurocomputing **2003**, 50, 159–175.

[16] Harvey, Andrew C. Forecasting, structural time series models and the Kalman filter. Cambridge university press, 1990.

[17] Ghosh, Bidisha, Biswajit Basu, and Margaret O'Mahony. "Bayesian time-series model for short-term traffic flow forecasting." *Journal of transportation engineering* 133.3 (2007): 180-189.

[18] Borak, Szymon, Wolfgang Karl Härdle, and Brenda López-Cabrera. "ARIMA Time Series Models." *Statistics of Financial Markets*. Springer Berlin Heidelberg, 2013. 143-161.

[19] Wei, William Wu-Shyong. *Time series analysis*. Reading: Addison-Wesley publ, 1994.

[20] Sadat, Noshin Nawar. Implementation of Time Series Approaches to Financial Data. Diss. BRAC University, 2016.