Jñānābha, Special Issue, 2018, 5-12
Proceedings : MSSCID-2017, 20th Annual Conference of VPI, Jaipur, India

STATISTICAL TIME SERIES AND PREDICTABILITY ANALYSIS OF NITROGEN DIOXIDE

By
Rashmi Bhardwaj*, Dimple Pruthi
Non Linear Dynamics Lab, University School of Basic and Applied Sciences
Guru Gobind Singh Indraprastha University, Delhi, India
Email: *rashmib22@gmail.com
(*Received : November* 24, 2017 ; *Revised: September* 06, 2018)

Abstract

Main component of deadly smog is ground-level ozone. Ozone pollution is triggering pulmonary disorders and respiratory infections at high rate. The chemical reaction of oxides of nitrogen with volatile organic compound in sunlight leads to the formation of ozone. Ozone is a complex pollutant to be controlled due to its formation process. The primary source in the formation of ozone is nitrogen dioxide. According to CPCB, the NO_2 level has increased by 1.8 times from $36\mu g/m^3$ in 2000 to $65\mu g/m^3$ in 2016. Alarming ozone concentration has been found in the residential area RK Puram, Delhi. To control ozone level, nitrogen dioxide has to be considered. The predictability index of nitrogen dioxide not close to zero indicates NO_2 concentration is predictable. In order to raise alarm for increasing ozone level, nitrogen dioxide can be predicted. The present study attempts to predict the daily future concentration of nitrogen dioxide using developed autoregressive integrated moving average model.This will assist regulatory bodies to warn about poor air quality and carry out preventive measures.

Keywords and phrases: ARIMA, Predictability Index, Ground level ozone, Nitrogen Dioxide, Air Pollutant.
2010 Mathematics Subject Classification: 62P12

1 Introduction

The sources of NO_2 are classified into natural and anthropogenic. Natural sources comprise of lightning, forest fires etc. and anthropogenic sources as biomass, fossil fuels burning and combustion processes. Increasing levels of nitrogen dioxide are harmful for human health and environment. It leads to pulmonary disorders, increased susceptibility to respiratory infections. Major environmental effects are aerosol format and ozone formation in troposphere. Here nitrogen dioxide is considered because of its environmental effects. As the ozone level in some regions of Delhi are alarming, so level of nitrogen dioxide is taken into consideration. Ozone effects on human are respiratory problems, asthma, bronchitis etc. Long exposure leads to premature death. Ozone assists in the formation of peroxyacetylnitate which results in death of plant tissues, thus contributing in environmental deterioration. Hence O_3 has severe effects on health and climate.

In figure.1 nitrogen oxides role is very much clear in formation of ozone and particulates. Particulate matter and ozone effects the respiratory system. NO_2 and other NO_x causes acid rain which harm sensitive ecosystems such as lakes and forests. Nitrogen oxide leads to

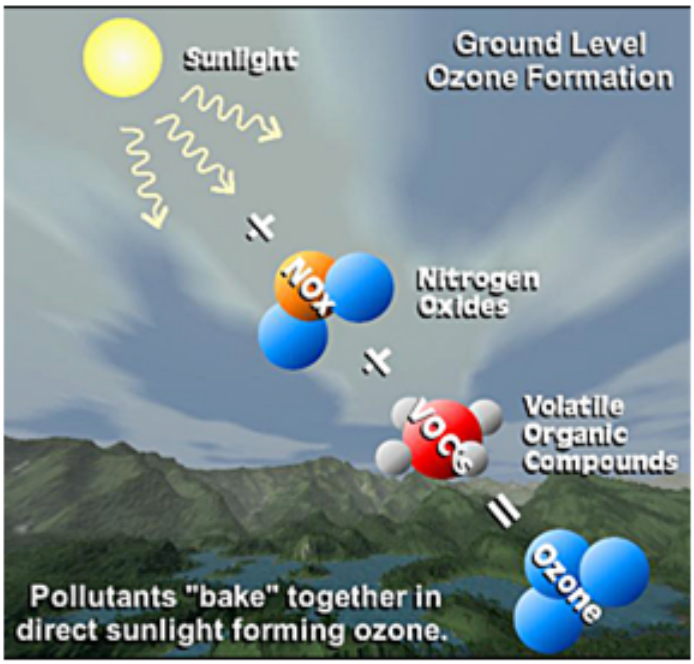

Figure 1: Formation of Ozone

nitrate particles which make the air hazy. National Capital Territory of India has faced the worst haze condition marked in the history as Great Smog of Delhi. The levels of nitrogen dioxide should be controlled otherwise same phase can be repeated.

Bhardwaj and Pruthi [1] studied the Predictability and Wavelet Analysis of Air Pollutants for Commercial and Industrial Regions in Delhi. Bhardwaj [2-3] estimates carbon mono-oxide using wavelet and fractal methods. Onate [4] analyzed precipitation records in Spain using fractals. Jorquera et al [5] developed ozone forecasting model for Santiago. Kumar and Anand[6] applied ARIMA for predicting sugarcane production in India. Rangarajan and Sant[7-8] studied Fractal dimensional analysis of Indian climatic dynamics. Siew et al [9] discussed varios models based on auto regressive and moving average for predicting pollution index in Shah Alam.

None of the authors have studied the time series, predictability analysis of NO_2. In this paper, we studied the impact of time on amount of nitrogen oxides with ozone particles. Time series, predictability index and behavior of NO_2 is observed.

2 Methodology

2.1 Predictability Analysis

The measure of long term memory is given by Hurst Exponent. Dependence of rescaled range on span of n observations is estimated. For sample series of m elements (original series of M elements is breakdown into shorter series of length m) $U_1, U_2, U_3, \cdot \cdot \cdot \cdot, U_m$, rescaled range $(R(m)/S(m))$ is calculated as:

$$R(m) = -(\min(V_1, V_2, ..., V_m - \max(V_1, V_2, ..., V_m))$$

$$\text{where, } V_t = \Sigma_{i=1}^{t}(U_i - (\frac{1}{m}\Sigma_{i=1}^{m}U_i)), t = 1, 2, ..., m$$

$$S(m) = \sqrt{\frac{1}{m}\Sigma_{i=1}^{m}(U_i - (\frac{1}{m}(\Sigma_{i=1}^{m}U_i))^2}$$

Average rescaled range over partial time series. H satisfies the following law:

$$E(R(m)/S(m)) = Cm^H$$

Further, fractal dimension is 2 minus Hurst exponent. The index of predictability gives time series behavior. Predictability index is two times modulus of fractal dimension minus 1.5. If the index of prediction approaches zero, no amplitude trend can be anticipated and series is unpredictable. Decrease in amplitude most likely inflate in future as index approaches one. Hence, predictability again increases.

2.2 Time Series Modeling

A Time Series $\{x_t : t \in T\}$ is collection of random variables usually parameterized by $T = (-\infty, \infty), [0, \infty), \{..., -2, -1, 0, 1, 2, ...\}$ or $\{0, 1, 2, ...\}$. The probability measure of a time series is defined by specifying the joint distribution (in a consistent manner) of all finite subsets of $\{x_t : t \in T\}$. Time series analysis comprises of extracting characteristics, interpretation, forecasting, control, hypothesis testing and simulation. ARIMA linear models have governed field of time series forecasting. $AR(p)$ is represented as :

$$U_t = \Theta_0 + \phi_1 U_{t-1} + \phi_2 U_{t-2} + ... + \phi_p U_{t-p} + E_t \tag{2.1}$$

where p is cardinality of AR terms, U_t forecasted output U_{t-p} observation at $t - p$. ϕ are determined by linear regression. Θ_0 is intercept and E_t is error associated with regression. This series depends on p past values of itself and random term E_t. $MA(q)$ is represented as

$$U_t = \mu - \Theta_1 E_{t-1} - \Theta_2 E_{t-2} - ... - \Theta_q E_{t-q} + E_t \tag{2.2}$$

where q is cardinality of MA terms,Θ_1, Θ_2,...,Θ_q are finite weights and μ is mean of series. U_t is build upon E_t present and q past random terms. ARMA is build upon series p past values and q past random terms E_t. ARIMA(p, d, q) with $d = 0$ is

$$U_t = \Theta_0 + \phi_1 U_{t-1} + \phi_2 U_{t-2} + ... + \phi_p U_{t-p} + \mu - \Theta_1 E_{t-1} - ...\Theta_q E_{t-q} + E_t \tag{2.3}$$

In ARIMA$(p, d, q), d = 0$ for stationary time series and for non-stationary depends on number of times series is differenced. The time series $\{u_t : t \in T\}$ is stationary if joint distribution of $u_{t_1}, u_{t_2}, ..., u_{t_l}$ is similar as joint distribution of $u_{t_{1+h}}, u_{t_{2+h}}, ..., u_{t_{l+h}}$ for all finite subsets $t_1, t_2, ..., t_h$ of T and all choices of h. Augmented Dickey-Fuller(ADF) test series is stationary or non-stationary. The null hypothesis is series is non-stationary and alternative hypothesis is stationary. If alternative hypothesis is true that means series is stationary in its mean and variance thus there is no need of further differencing and value of d can be obtained. Mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE) measure model accuracy and calculated as:

$$MAE = \frac{\Sigma_{i=1}^{n}|x_i - \hat{x}|}{n} \tag{2.4}$$

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \tag{2.5}$$

$$MAPE = \frac{\Sigma_{i=1}^{n} |\frac{x_i - \hat{x}_i}{x_i}|}{n} \times 100\% \tag{2.6}$$

Another criteria used is Bayesian information criteria(BIC), criterion for model selection. Based on likelihood function, lowest BIC gives preferred model. BIC measures efficacy of parameterized model given by :

$$BIC = ln(m)l - 2ln(\hat{K}) \tag{2.7}$$

where, x is observed vector m data points, $\hat{K}$ maximized value of likelihood function and l free parameters to be estimated.

2.3 Case Study: R.K. Puram, Delhi

R.K. Puram situated in South West Delhi, is a Central Government Employees residential colony. It is roughly rectangular, surrounded by Ring Road, Outer Ring Road facing Vasant Vihar, Rao Tula Ram Marg and Africa Avenue to the north, south, west and east respectively. It was observed that in the daytime ozone levels in R.K. Puram increased from $90\mu g/m^3$ to $211\mu g/m^3$ in April 2017 which was around 134 increase in percent. The same trend was followed in May. Ozone is formed in the presence of sunlight, it is dependent on temperature. As the rise in temperature was observed from $32°C$ as on May3 to $36°C$ on May 8, the ozone concentration spiked to $230\mu g/m^3$ from $100\mu g/m^3$ which was 130 approximately in percent. These spikes are major in the region. The ozone level if not controlled will lead to severe atmospheric condition. As a result, there is decrease in National Capitals economy and health.

Figure 2: Study Area: R.K. Puram, Delhi

3 Result and Discussion

The data for daily concentration of nitrogen dioxide is collected for R.K. Puram from 1st January, 2016 to 30th June, 2017. In this study predictability index and time series modeling is carried out for daily concentration of nitrogen dioxide. The time series in Figure 3 depicts the daily concentration of nitrogen dioxide, the maximum value of NO_2 is much above the prescribed limit. Most of the values cross the National Ambient Air Quality Standard which is $80\mu g/m^3$ for nitrogen dioxide (24 hours).

The fractal dimension and predictability index of series is 1.471 and 0.058 respectively depicting predictable behavior. Further, ARIMA model is used for forecasting daily

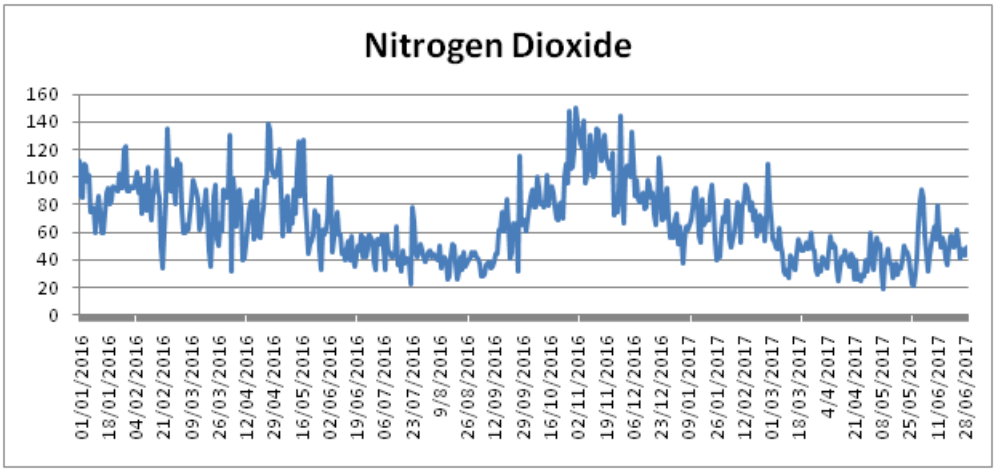

Figure 3: Time series of Nitrogen Dioxide at R.K.Puram from 1stJanuary, 2016 to 30thJune, 2017.

concentration of nitrogen dioxide using SPSS. Firstly ADF test is carried out to test whether series is stationary or not. The ADF test result, as obtained is Dickey-Fuller= -8.2351 and p-value= 0.001. The result indicates the p-value of ADF statistic is less than 5percent level of significance. Therefore, we accept the null hypothesis. Hence, with 95 % confidence level presence of unit root in series can be rejected. Now, it can be concluded that series is stationary and fit for applying ARIMA. Thus, $d = 0$ for our ARIMA(p, d, q) model.

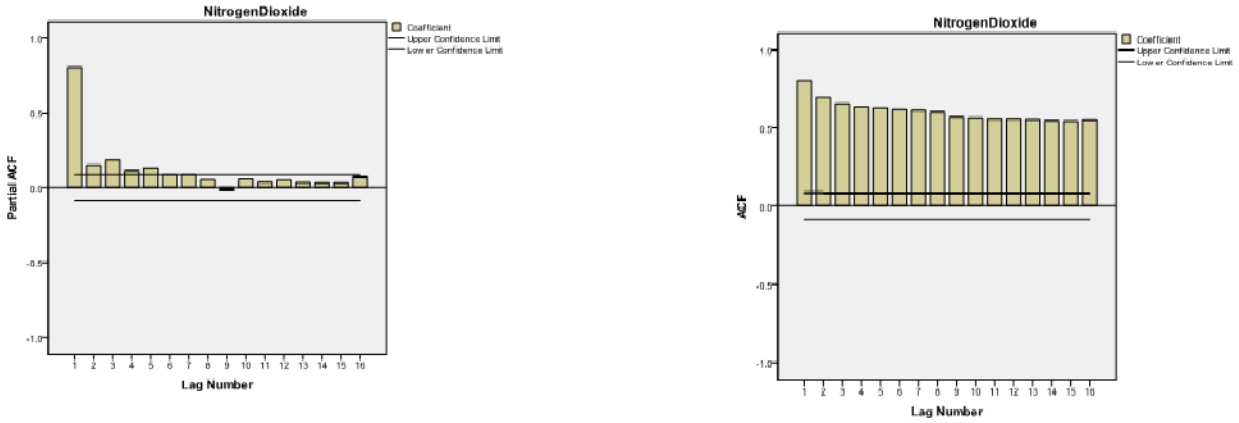

Figure 4: PACF and ACF plots.

Based on correlogram the values of autoregressive and moving average is determined. Autocorrelation Function plot shows continuous peak till lag 16 with exponential tail. Partial Autocorrelation Function plot show significant peaks at 1^{th}, 2^{nd}, 3^{rd} and 5^{th} lags. ARMA model seems to be applicable. Now different combinations of AR(p) and MA(q) depending on the spikes are applied. It is observed that ARIMA(3,0,1) model is preferred based on the

least value on normalized BIC. So based on ARIMA model we have model

$$U_t = 67.764 + 1.426U_t - 0.500U_{t-1} + 0.065U_{t-2} + 0.846U_{t-3} \tag{3.1}$$

Table 1: LCL,UCL and forecasted values using ARIMA model.

Month	Observed	Forecast	LCL	UCL	Absolute Error
Jan-16	88.466	86.3	55.752	116.85	0.0245
Feb-16	87.263	87.164	57.585	116.74	0.0011
Mar-16	75.802	77.032	47.453	106.61	0.0162
Apr-16	82.208	78.49	48.912	108.07	0.0452
May-16	76.746	80.641	51.064	110.22	0.0507
Jun-16	54.596	56.07	27.391	86.548	0.0435
Jul-16	45.78	47.238	17.659	76.816	0.0319
Aug-16	40.166	42.235	12.658	71.813	0.0515
Sep-16	51.297	49.295	19.717	78.873	0.0390
Oct-16	86.824	81.451	51.873	111.03	0.0619
Nov-16	113.52	111.12	81.544	140.7	0.0212
Dec-16	87.227	89.022	59.446	118.6	0.0206
Jan-17	65.394	67.022	37.444	96.6	0.0248
Feb-17	72.169	69.921	40.343	99.499	0.0311
Mar-17	46.043	51.642	22.063	81.219	0.1215
Apr-17	37.461	40.291	10.712	69.869	0.0755
May-17	39.317	39.693	10.114	69.27	0.0095
Jun-17	54.479	54.426	24.847	84.003	0.0009

Table 2: Fit Statistic for ARIMA(3,0,1)

Fit Statistic	Mean
R-squared	0.685
MAPE	18.302
RMSE	15.143
MaxAPE	239.463
MAE	10.884
MaxAE	75.143
Normalized BIC	5.493

Figure 5 depicts the forecasted values using ARIMA modeling where value of y at $1(x-\text{axis})$ depicts the concentration of nitrogen dioxide on 1st January, 2016 and so on. The lower and upper confidence limits are also depicted in the figure. It is calculated that the Normalized BIC value is 5.493 for ARIMA(3,0,1). The fit statistics for the linear model are summarized in Table1.

4 Conclusion

The prediction of daily concentration of nitrogen dioxide is carried out for one year approximately. The preferred linear model observed is ARIMA(3,0,1) with root mean square

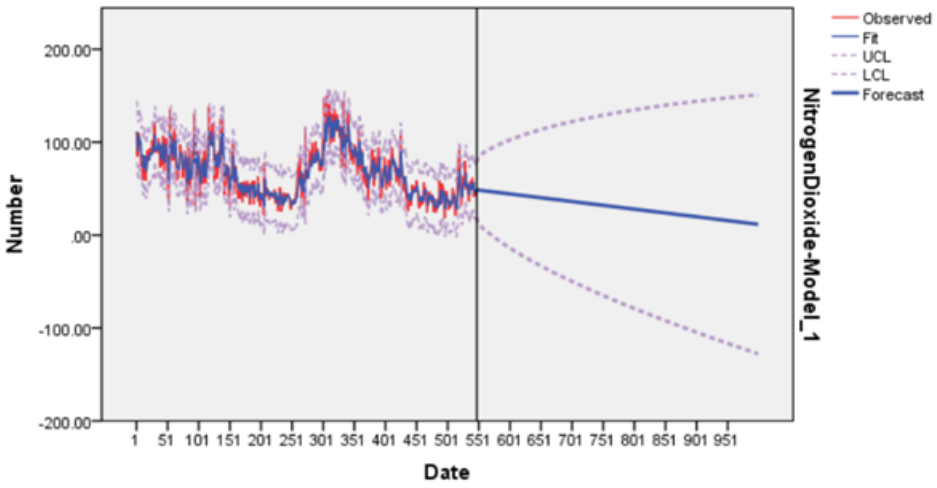

Figure 5: ARIMA modeling for nitrogen dioxide at R.K. Puram.

error 15.109. The future values of nitrogen dioxide can be used to calculate the ozone concentration as the formation of ozone depends on nitrogen dioxide. It would be helpful in taking precautions and raising alarm. As NO_2 is one of the primary components for Air Quality Index, the prediction will contribute in forecasting the air quality index. The predicted daily concentration of nitrogen dioxide demands for the necessary action. As nitrogen dioxide plays major role in formation of ozone and aerosol as a result contributing in spiking the respiratory disorders and hence decrease in nations economy.

Acknowlegement
The authors are thankful to Guru Gobind Singh Indraprastha University, Delhi (India) for providing research facilities and financial support.

References

[1] R. Bhardwaj and D. Pruthi, Predictability and Wavelet Analysis of Air Pollutants for Commercial and Industrial Regions in Delhi,*Indian J. of Ind. & App. Math.,* **7**(2) (2016), 165-174.

[2] R. Bhardwaj, Wavelets and Fractal Methods with environmental applications, *Mathematical Models, Methods and Applications,* Eds: Siddiqi, A.H., Manchanda, P., Bhardwaj, R.; 173-195. Springer.

[3] R. Bhardwaj R., A.H. Siddiqi and A. Mittal , Predictability Index, Fractal Dimension and Hurst Exponent Estimation of Carbon Mono-Oxide at different locations of Delhi,*Indian J. of Ind. & App. Math.,* **3**(2) (2012), 91-97.

[4] J.J.Onate Rubalcaba, Fractal analysis of climatic data: annual precipitation records in Spain, *Theor. Appl. Climatol.,* **56** (1997), 83-87.

[5] H.Jorquera ,W. Palma and J. Tapia , Ground level Ozone Forecasting Model for Santiago, *J. Forecast.,* **21** (2002), 451-472.

[6] M. Kumar and M. Anand , An Application of Time Series Arima Forecasting Model for Predicting Sugarcane Production In India, *Studies in Business and Economics,* **9**(1) (2014), 81-94.

[7] G. Rangarajan and D.A. Sant , Fractal dimensional analysis of Indian climatic dynamics, *Chaos Solitons Fractals.,***19** (2004), 285-291.

[8] G. Rangarajan and D.A. Sant , A climate predictability index and its applications, *Geophys. Res. Lett.,* **24** (1997), 1239-1242.

[9] L.Y.Siew,L.Y.Chin and P.M.J.Wee, ARIMA And Integrated ARFIMA Models For Forecasting Air Pollution Index In Shah Alam, Selangor, *Malaysian Journal of Analytical Sciences,* **12**(1) (2008), 257-263.