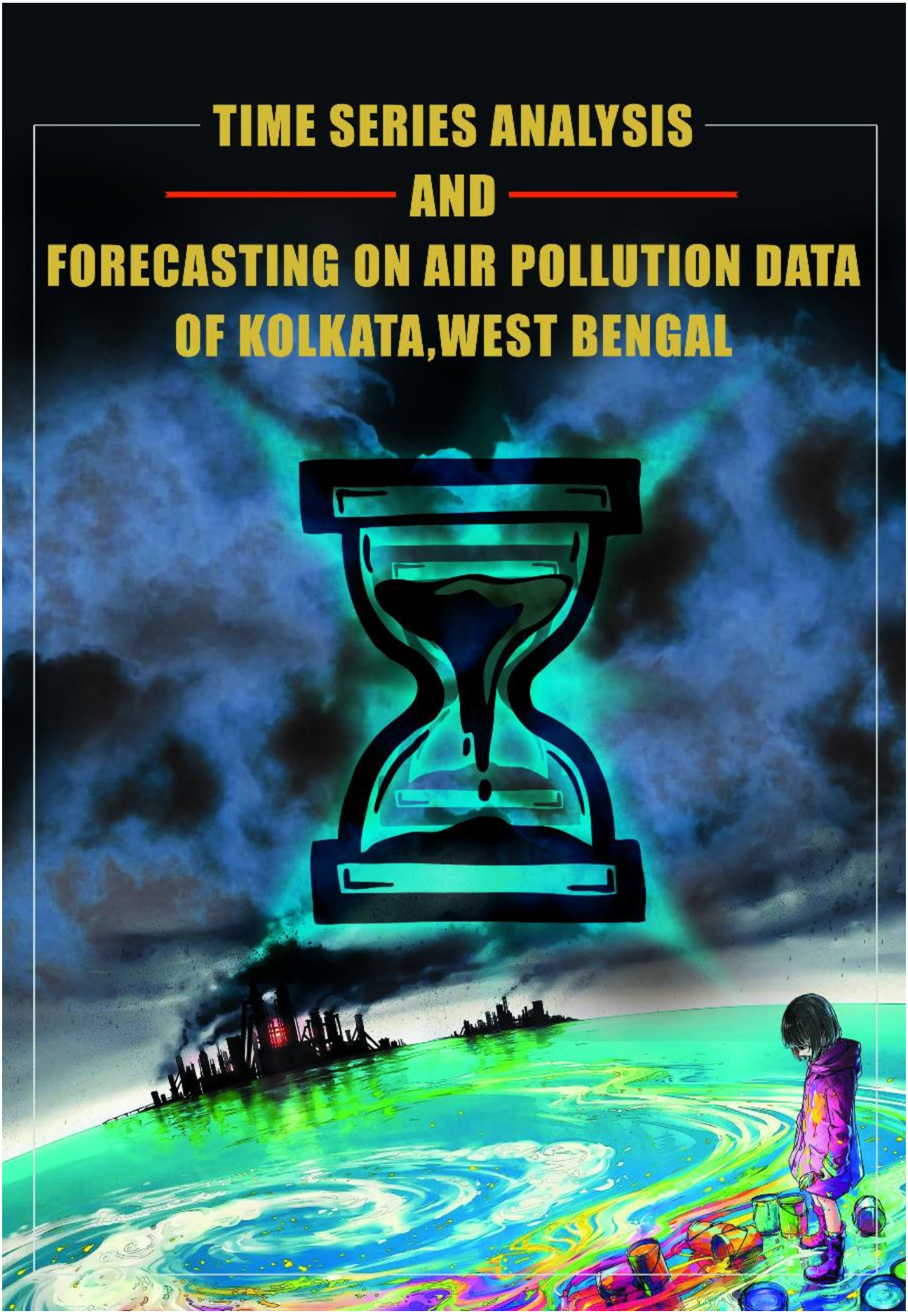
TIME SERIES ANALYSIS AND FORECASTING ON AIR POLLUTION DATA OF KOLKATA,WEST BENGAL



TIME SERIES ANALYSIS AND FORECASTING OF AIR POLLUTION DATA OF KOLKATA,WEST BENGAL

By Soumallya Mitra , Sayak Kolay

and Titas Chatterjee

DEPARTMENT OF STATISTICS, KU

AIR POLLUTION

DEFINITION: The presence in or introduction into the air of a substance which has harmful or poisonous effects.

Air pollution is the single biggest environmental health risk, causing roughly 7 million deaths annually. Short-lived pollutants – which include black carbon, methane, ozone, and airborne particles produced by industrial operations and the burning of diesel, coal, kerosene or biomass – are responsible for about one third of deaths from stroke, chronic respiratory disease and lung cancer and one quarter of deaths from heart attack. These pollutants are also contributing to global warming, lowering labour productivity, and increasing food insecurity around the world.

Among some of the major air pollutants, we discuss here:

1. NO2: NO2 exposures outdoors is complicated by the fact that in most urban locations, the nitrogen oxides that yield NO_2 are emitted primarily by motor vehicles, making it a

strong indicator of vehicle emissions (including other unmeasured pollutants emitted by these sources). NO_2 (and other nitrogen oxides) is also a precursor for a number of harmful secondary air pollutants, including nitric acid, the nitrate part of secondary inorganic aerosols and photo oxidants (including ozone). Nitrogen dioxide is an important air pollutant because it contributes to the formation of photochemical smog, which can have significant impacts on human health. The main effect of breathing in raised levels of nitrogen dioxide is the increased likelihood of respiratory problems. Nitrogen dioxide inflames the lining of the lungs, and it can reduce immunity to lung infections. This can cause problems such as wheezing, coughing, colds, flu and bronchitis. Increased levels of nitrogen dioxide can have significant impacts on people with asthma because it can cause more frequent and more intense attacks.

2. SO2: About 99% of the sulfur dioxide in air comes from human sources. The main source of sulfur dioxide in the air is industrial activity that processes materials that contain sulfur, eg the generation of electricity from coal, oil or gas that contains sulphur. Some mineral ores also contain sulfur, and sulfur dioxide is released when they are processed. In addition, industrial activities that burn fossil fuels containing sulfur can be important sources of sulfur dioxide. It causes acid rain, haze and many health-related problems.

3. PM2.5: PM2.5 refers to atmospheric particulate matters (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair. Commonly written as $PM_{2.5}$, particles in this category are so small that they can only be detected with an electron microscope. Fine particles can come from various sources. They include power plants, motor vehicles, airplanes, residential wood burning, forest fires, agricultural burning, volcanic eruptions and dust storms. Some are emitted directly into the air, while others are formed when gases and particles interact with one another in the atmosphere. For instance, gaseous sulfur dioxide emitted from power plants reacts with oxygen and water droplets in the air to form sulfuric acid as a secondary particle. Since they are so small and light, fine particles tend to stay longer in the air than heavier particles. This increases the chances of humans and animals inhaling them into the bodies. Owing to their minute size, particles smaller than 2.5 micrometers are able to bypass the nose and throat and penetrate deep into the lungs and some may even enter the circulatory system. Studies have found a close link between exposure to fine particles and premature death from heart and lung disease. Fine particles are also known to trigger or worsen chronic disease such as asthma, heart attack, bronchitis and other respiratory problems.

Polluted air is creating a health emergency

There is no doubt today that air pollution is a global public health emergency. It threatens everyone from unborn babies to children walking to school, to women cooking over open fires. On the street and inside the house, the sources of air pollution can be very different, yet their effects are equally deadly: asthma, other respiratory illnesses and heart disease are among the adverse health effects known to be caused by polluted air. According to the World Health Organization, every year around 7 million premature deaths are attributable to air pollution—a staggering 800 people every hour or 13 every minute. Overall, air pollution is responsible for more deaths than many other risk factors, including malnutrition, alcohol use and physical inactivity.

Children are most at risk

Globally, 93 per cent of all children breathe air that contains higher concentrations of pollutants than the World Health Organization (WHO) considers safe to human health. As a result, 600,000 children die prematurely each year because of air pollution. As if that were not enough, exposure to dirty air also harms brain development, leading to cognitive and motor impairments, while at the same time putting children at greater risk for chronic disease later in life

The right to clean air is a human right

The right to a healthy environment enjoys constitutional status—the strongest form of legal protection available—in more than 100 countries. At least 155 states are legally obligated, through treaties, constitutions and legislation, to respect, protect and fulfil the right to a

healthy environment. The right to clean air is also embedded in the Universal Declaration of Human Rights and the International Covenant on Economic, Social and Cultural Rights, and fully enshrined in the Sustainable Development Goals—the global blueprint for peace and prosperity.

INTRODUCTION

The time series forecasting approach is of useful for predicting future air quality status from various aspects of development in each country. The forecasting method analyzes the sequence of historical data in a period of time to establish the forecasting model. The ARIMA method has been extensively studied and used in previous research proven to be effective in the forecasting field. Forecasting methods applying the ARIMA time series method for pollution field have been expounded upon in many previous publications.

Air pollution data are obtained from the West Bengal Pollution Control Board Air Quality Information System. Currently, there are 91 monitoring stations for air quality monitoring operations, managed and maintained by WBPCB throughout West Bengal,India.

Two critical pollutants, sulfer dioxide and nitrogen dioxide and atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers(PM2.5) , which is about 3% the diameter of a human hair are considered because each of the data sets covers at least 3 years with no missing data in between and shows a fairly apparent either trend or seasonality, or both. Scientific research has proven that these two gases and PM2.5 have many negative health effects, including some deadly diseases.

MATERIALS AND METHODS

By using time series, this study aims to analyse the API performance. The time series approach used in this study is based on Box-Jenkins model. Box-Jenkins is referred as Autoregressive Integrated Moving Average (ARIMA) method. Until nowadays, a lot of researchers still use this model in many area of research because the result effectiveness in forecasting field.

Sampling sites: The Capital of West Bengal, Kolkata has been chosen as the study site. There are eight air quality monitoring stations located in Kolkata. However only three monitoring stations, namely as Behala Chowrasta,Moulali,Shyambazar has been chosen in this study. This is because there was not enough complete data to study.

Box-Jenkins modeling: The Box-Jenkins is taken in honour of its discoverers, Box and Jenkins (1976). This method is classified as linear models that capable in presenting both stationary and non-stationary time series. Most of researchers use this model to forecast univariate time series data. Box-Jenkins methods is a practical importance in forecasting which inclusive Autoregressive (AR) models, the Integrated (I) models and the Moving Average (MA) models.

To obtain the model by the Box-Jenkins methodology, there are four steps that must be considered which are tentative identification, parameter estimation, diagnostic checking and finally model is used in prediction purposes. This step is the important procedure in order to determine the best ARIMA model for time series data.

Autoregressive (AR), Moving Average (MA) and Autoregressive Integrated Moving Average (ARIMA) Models: Autoregressive (AR) model is suitable for stationary time series data patterns. A p-th order of autoregressive or AR (p) model can be written in the form

Eq. 1:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t \quad (1)$$

The current value of the series y_t is a linear combination of the p most recent values of itself. The coefficient f_0 is related to the constant level of series.

For AR models, forecast depend on observed values in previous time periods. Meanwhile, the dependent variable y_t of Moving Average (MA) depends on previous values of the errors rather than on the variable itself. MA models provide forecasts of y_t based on linear combination of a finite

number of past errors. The errors involved in this linear combination move forward as well.A moving average with qth-order or MA (q) model takes the form Eq. 2:

$$Y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$
(2)

A mixed between autoregressive and moving average terms develop Autoregressive moving Average Model (ARMA). The notation is ARMA (p, q) where, p is the order of the autoregressive part and q is the order of the moving average part which represent this models.

The ARMA (p,q) is in the form below Eq. 3:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$
(3)

A wide variety of behaviors for stationary time series can be described by ARMA models. Since ARMA is a mixture of AR and MA, the forecasting is depend on both current and past values of response Y as well as current and past values of the residuals.

Set of data could be a non-stationary time series data patterns since the data did not fluctuate around a constant level or mean. One way to make the data stationary is by taking the difference. Therefore, the series of data generally donated as y_t after difference is said to follow an integrated autoregressive moving average model, ARIMA (p, d, q). Normally for practical purpose, the difference would be one or at most two (d≤2).

By considering d = 1, we can obtain ARIMA (p,1,q) process with $\Delta_d y_t = W_t$ or may written as $W_t = y_t - y_{t-1}$. Then, (3) becomes Eq. 4:

$$W_t = \varphi_1 W_{t-1} + \varphi_2 W_{t-2} + \cdots + \varphi_p W_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$
(4)

Rewrite (3) as Eq. 5:

$$Y_t = (1+\varphi_1)Y_{t-1} + (\varphi_2 - \varphi_1)Y_{t-2} + (\varphi_3 - \varphi_2)Y_{t-3} + \cdots + (\varphi_p - \varphi_{p-1})Y_{t-p} - \varphi_p Y_{t-p-1} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}$$

(5)

Seasonal ARIMA model (SARIMA): The Box- Jenkins approach for modeling and forecasting has the advantage in analyze the seasonal time series data. In this case where the seasonal components are included, the model is called as seasonal ARIMA model or SARIMA model.

The model can be abbreviated as SARIMA (p, d, q) (P, D, Q)^S where the lowercase for non-seasonal part meanwhile the uppercase for seasonal part. The generalized form of SARIMA model can be written as Eq. 6:

$$\varphi_p(B)\Phi_P(B^s)(1-B)^d(1-B^S)^D Y_t = \theta_q(B)\Theta_Q(B)^S\varepsilon_t \quad \textbf{(6)}$$

where:

$$\varphi_p(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$$

$$\Phi_P(B) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS}$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

$$\Theta_Q(B) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \cdots - \Theta_Q B^{QS}$$

Measure of accuracy: Primary criterion in selecting the order of ARIMA model is by using Akaike Information Criteria (AIC) . AIC is a measure of the relative quality of various statistical models and AICc is Akaike Information Criteria with a correction for finite sample sizes. Among the numerous values of AICc , the smallest value of AICc , are selected as the finest model to be used in forecasting purpose . Model accuracy and performance was identified using the Akaike Information Criteria (AICc)For identification of the models performance, the criteria chosen are Akaike Information Criteria (AICc) with a correction, Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), the Mean Square Error (MSE) and the Root Mean Square Error (RMSE). Given as:

$$AICc = -2L + 2k + 2k(k+1)/(n-k-1)$$

$$MAE = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{n}$$

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|}{n} \times 100$$

$$MSE = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}}$$

Where:

y_t **= The actual value at time t**

$\hat{y}_t$**= The fitted value at time t**

n = The number of observations

The smallest values of MAE, MAPE, MSE and RMSE are chosen as the best model to be used in forecasting.

RESULTS

All the computations involved in this research have been performed by using RStudio. Before applying the ARIMA model, the out sample of API data will be kept out that will be used as to check the forecasting performances based on models built.

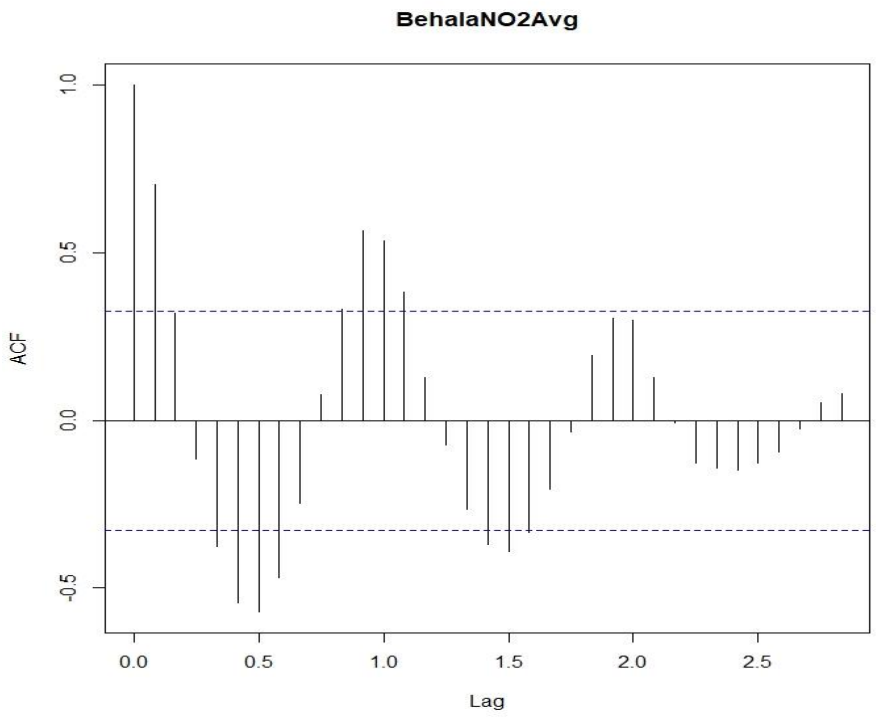
API raw data plots: The data used in this study obtained from Kolkata air quality monitoring located at Behala Chowrasta , Moulali and Shyambazar from January 2016-May 2019.

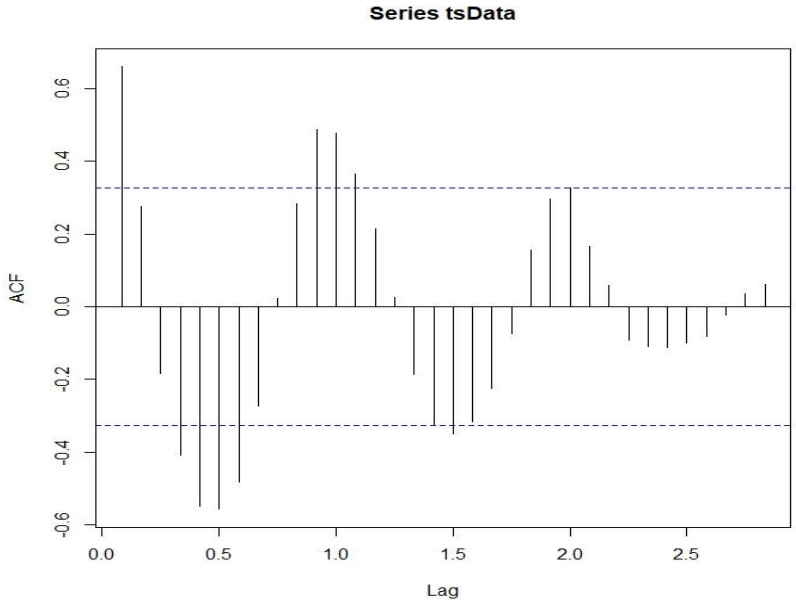
The data divided into two samples data set, in sample and out sample. The in sample data set contains 36 data (January 2016 until December 2018) meanwhile the last 5 data (January until May 2019) as out sample used to test the model performance.

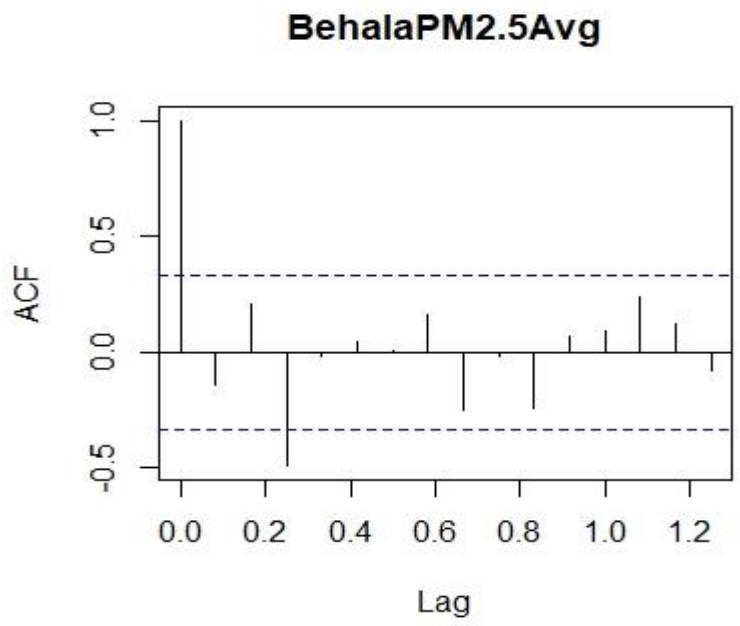
The time series plot illustrate that the data have seasonal pattern which indicates that the data non- stationary. In that case, the differencing process is necessary to obtain the stationary data set before model development can be made. By taking difference d = 1/2/3 for non-seasonal and D = 1 with S = 12 for seasonal then the data become stationary series. As stated before, there are three main stages in building ARIMA model before used in forecasting, tentative identification, parameter estimation and diagnostic checking.

Model identification: Based on stationary series, the Autocorrelation (ACF) and Partial Autocorrelation (PACF) were examined to determine the best combination order of ARIMA model for each data set.

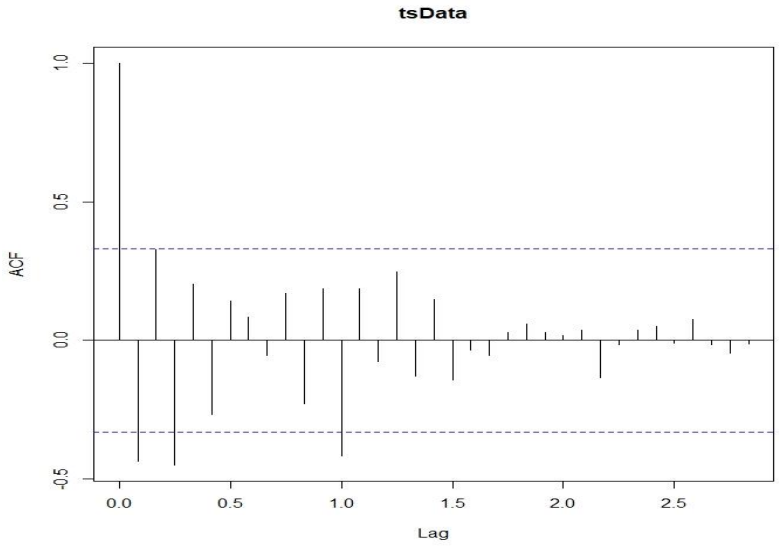
For Behala Chowrasta the ACF plot for NO2,SO2 and PM2.5 are give below :

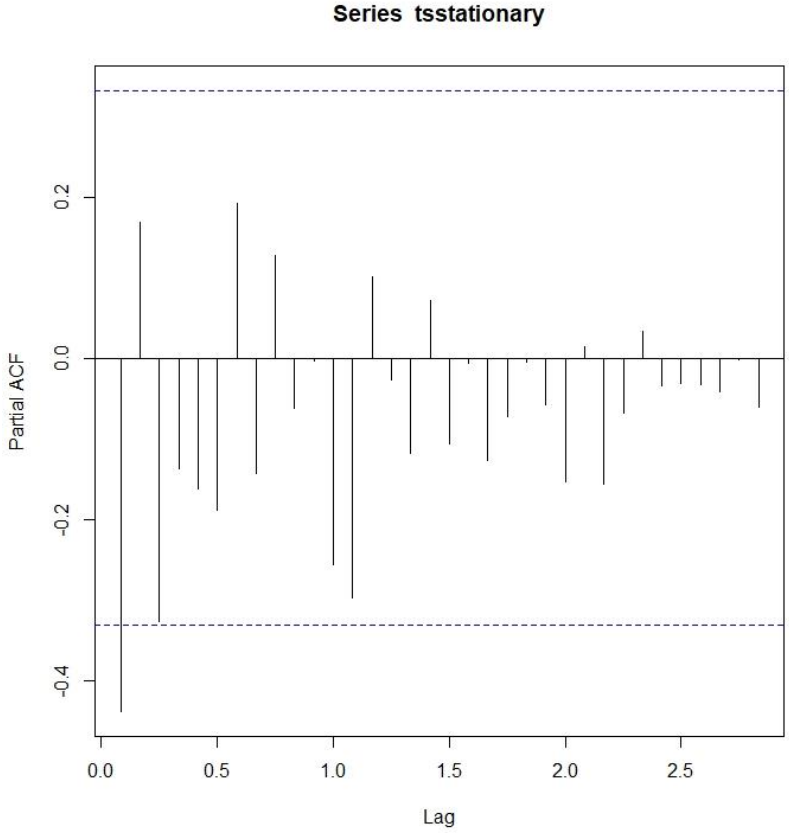

BehalaNO2Avg
ACF
1.0
0.5
0.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Lag

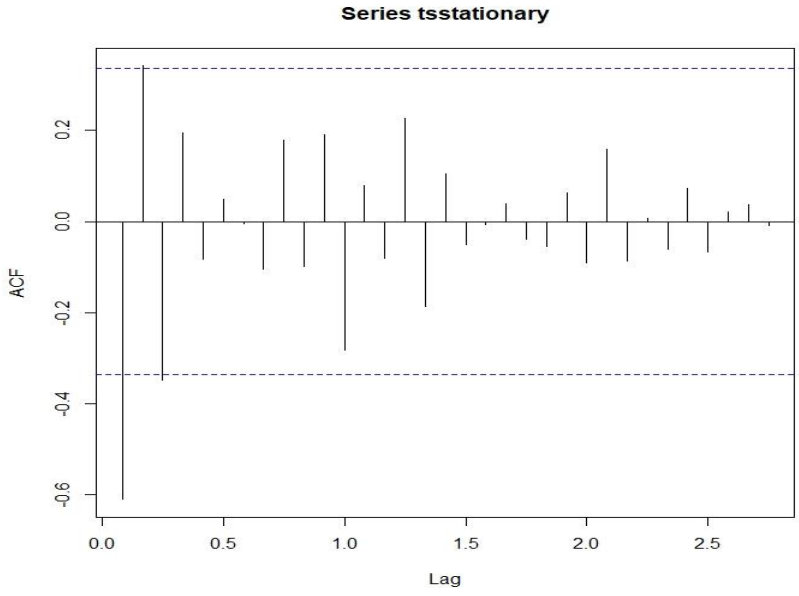

Series tsData
ACF
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
0.0
0.5
1.0
1.5
2.0
2.5
Lag

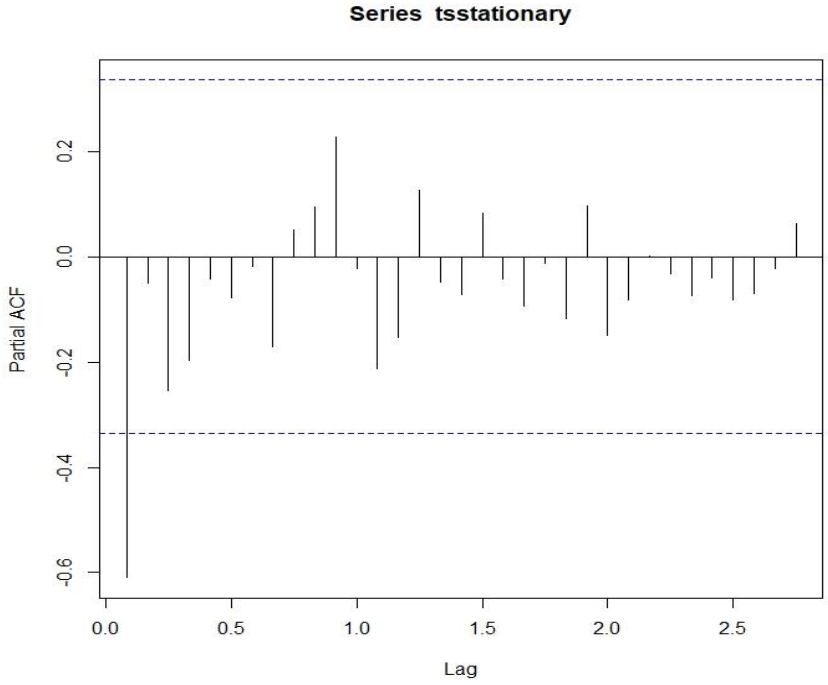


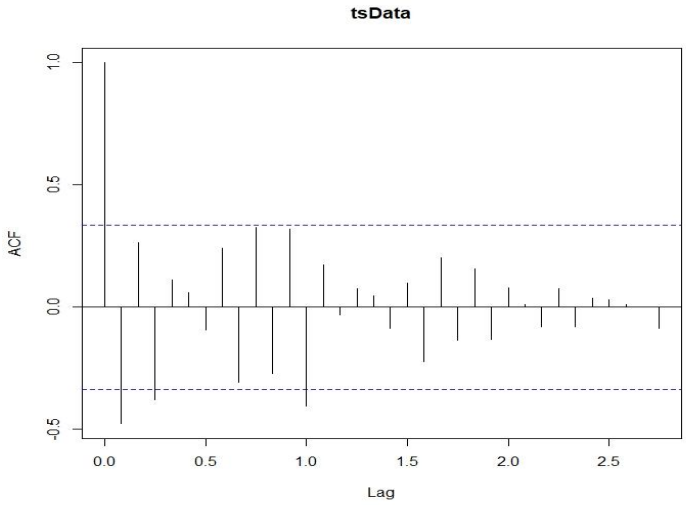
And after Seasonality removes and making the data stationary with corresponding differences needed the ACF and PACF plots for NO2,SO2,PM2.5 are given below :

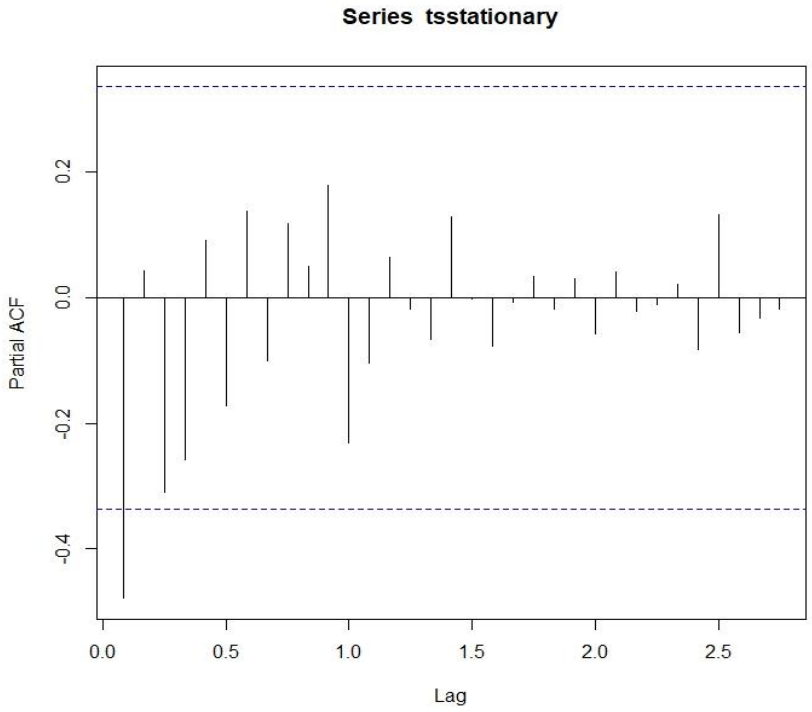



Series tsstationary
Partial ACF
0.2
0.0
-0.2
-0.4
0.0
0.5
1.0
1.5
2.0
2.5
Lag


Series tsstationary
ACF
0.2
0.0
-0.2
-0.4
-0.6
0.0
0.5
1.0
1.5
2.0
2.5
Lag


Series tsstationary
Partial ACF
0.2
0.0
-0.2
-0.4
-0.6
0.0
0.5
1.0
1.5
2.0
2.5
Lag


tsData
ACF
1.0
0.5
0.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Lag



The possible models are:

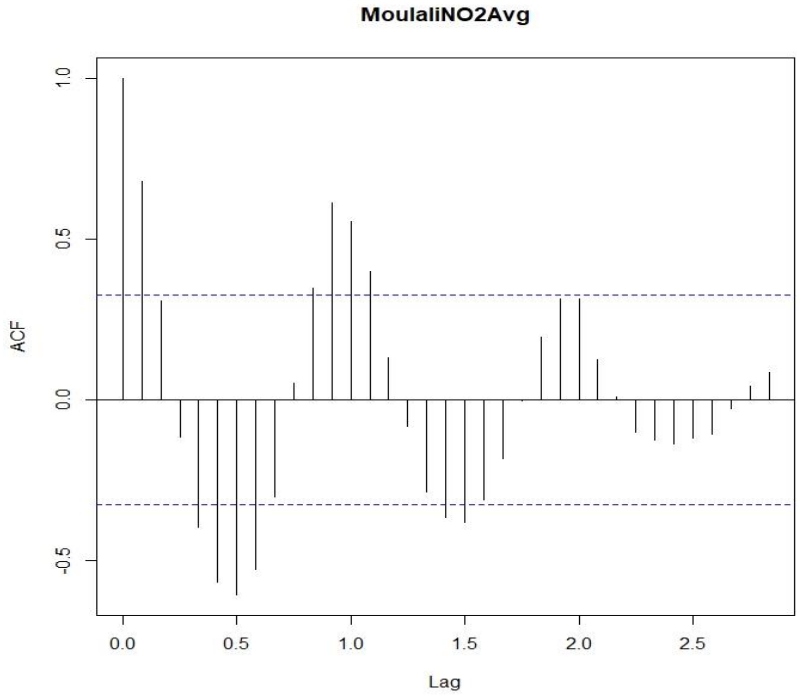
Behala Chowrasta :

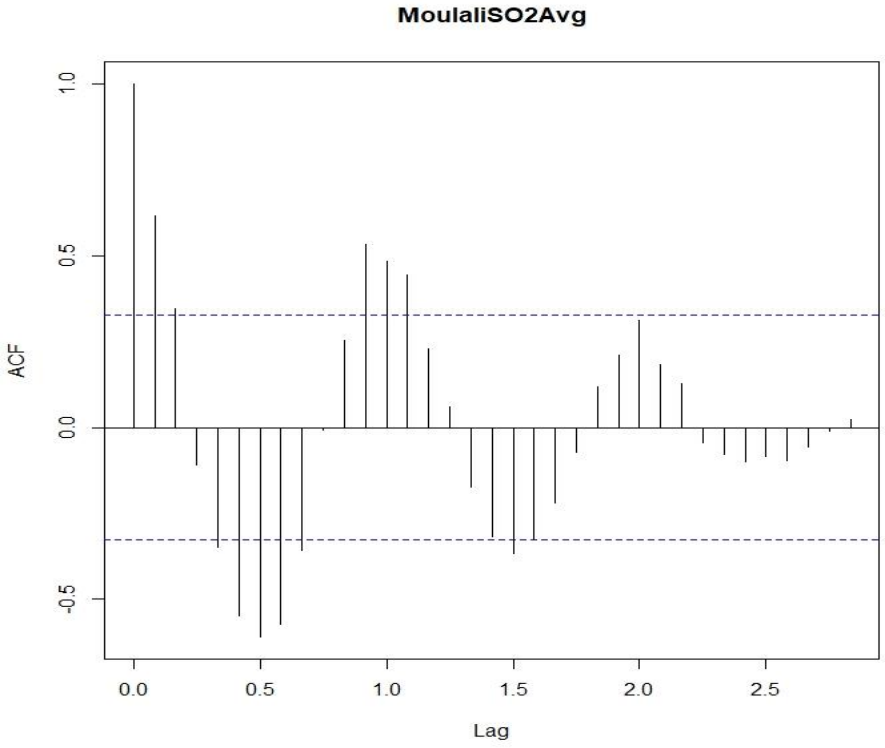
Best Model	AICc
NO2 :	
ARIMA(0,1,0)(0,1,0)[12]	: 163.2526
ARIMA(1,1,0)(1,1,0)[12]	: 154.0777
ARIMA(1,1,0)(0,1,0)[12]	: 159.2384
ARIMA(0,1,0)(1,1,0)[12]	: 156.9554
ARIMA(2,1,0)(1,1,0)[12]	: 155.6976
ARIMA(1,1,1)(1,1,0)[12]	: 154.8128
ARIMA(0,1,1)(1,1,0)[12]	: 155.4064
ARIMA(2,1,1)(1,1,0)[12]	: 158.0341
SO2 :	
ARIMA(0,0,0)(0,1,0)[12] with drift	:100.1501
ARIMA(1,0,0)(1,1,0)[12] with drift	: 97.2726
ARIMA(0,0,0)(0,1,0)[12]	: 97.9383
ARIMA(1,0,0)(0,1,0)[12] with drift	: 100.1385
ARIMA(0,0,0)(1,1,0)[12] with drift	: 95.39114
ARIMA(0,0,0)(1,1,1)[12] with drift	: 98.29641
ARIMA(0,0,1)(1,1,0)[12] with drift	: 97.35681
ARIMA(1,0,1)(1,1,0)[12] with drift	: 100.4844
ARIMA(0,0,0)(1,1,0)[12]	: 93.30443
ARIMA(0,0,0)(1,1,1)[12]	: 95.933

ARIMA(1,0,0)(1,1,0)[12] : 94.77956
ARIMA(0,0,1)(1,1,0)[12] : 94.86856
ARIMA(1,0,1)(1,1,0)[12] : 97.66106

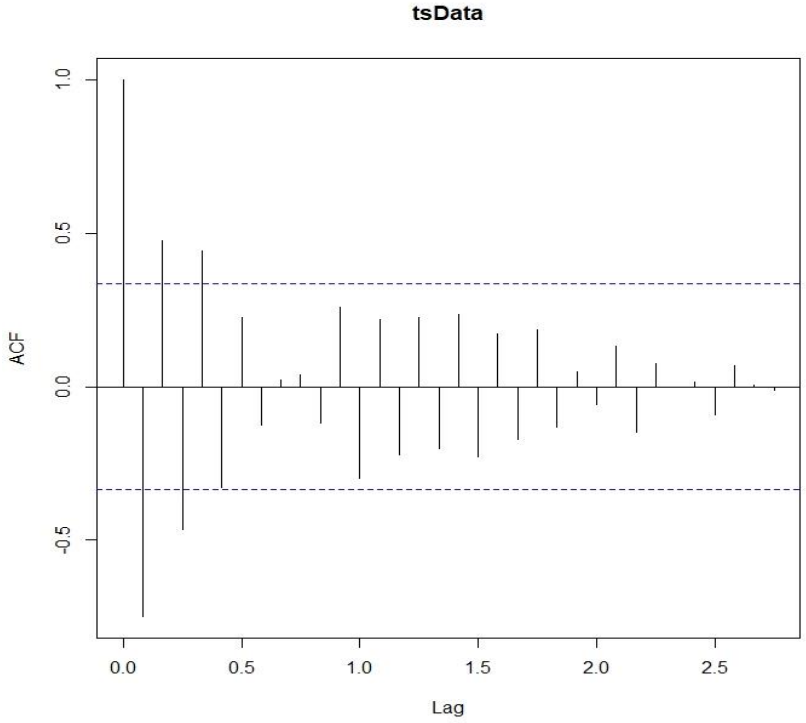
PM2.5 :
ARIMA(0,0,0)(0,1,0)[12] with drift : 215.6926
ARIMA(0,0,0)(0,1,0)[12] : 213.324
ARIMA(0,0,0)(1,1,1)[12] with drift : 197.9722
ARIMA(0,0,0)(1,1,1)[12] : 195.3457

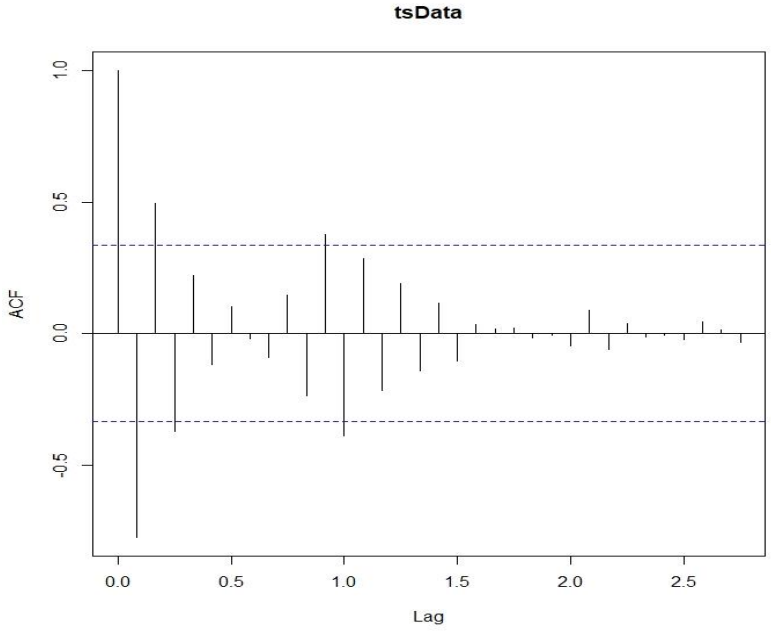
For Moulali the ACF plot for NO2,SO2 and PM2.5 are give below :

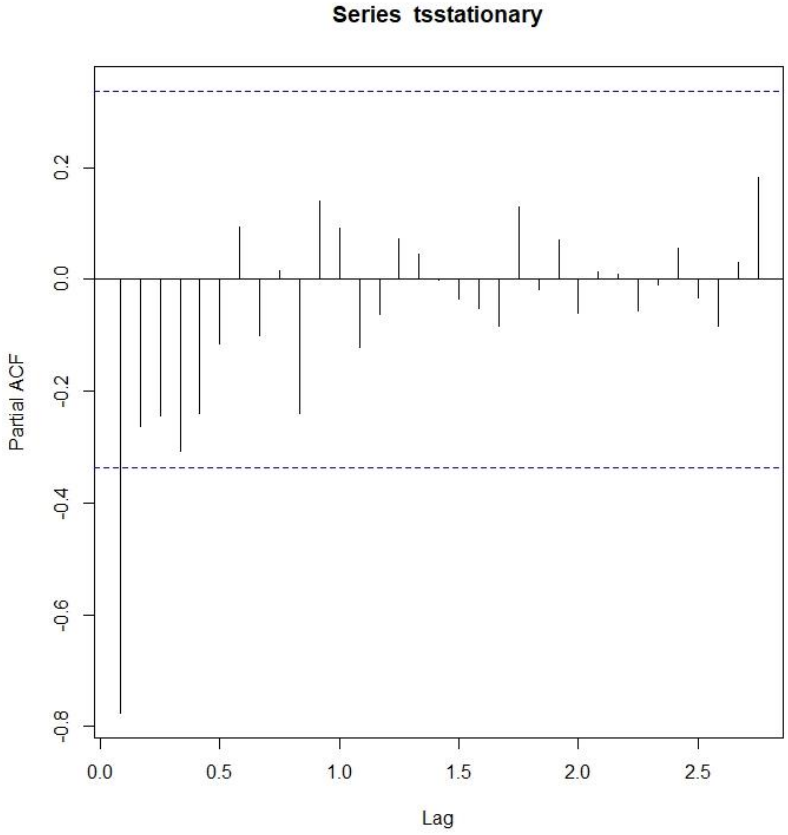


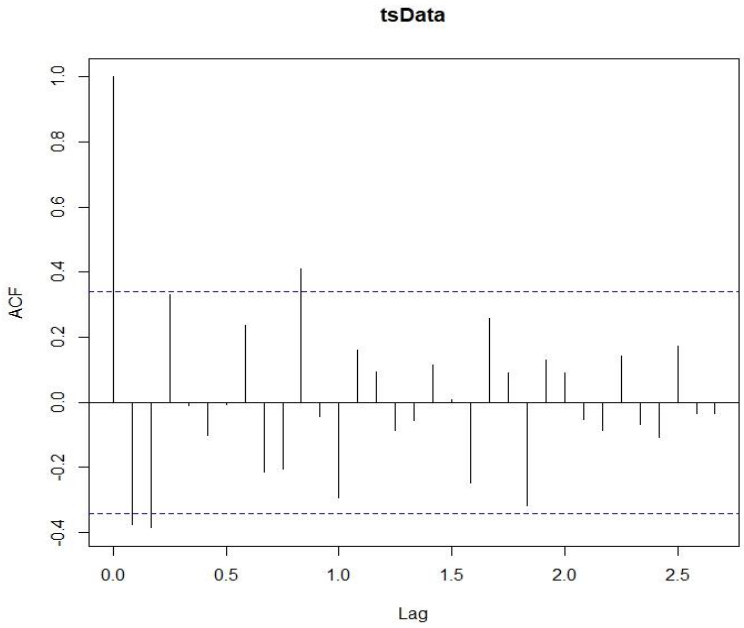


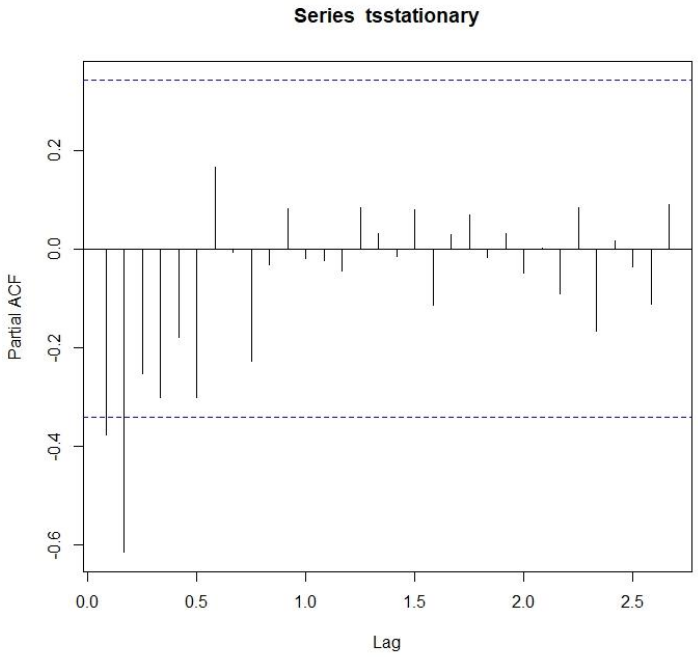
And after Seasonality removes and making the data stationary with corresponding differences needed the ACF and PACF plots for NO2,SO2,PM2.5 are given below :


tsData
ACF
1.0
0.5
0.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Lag


tsData
ACF
1.0
0.5
0.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Lag


Series tsstationary
Partial ACF
0.2
0.0
-0.2
-0.4
-0.6
-0.8
0.0
0.5
1.0
1.5
2.0
2.5
Lag


tsData
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0.0
0.5
1.0
1.5
2.0
2.5
Lag



The possible models are:

Moulali :

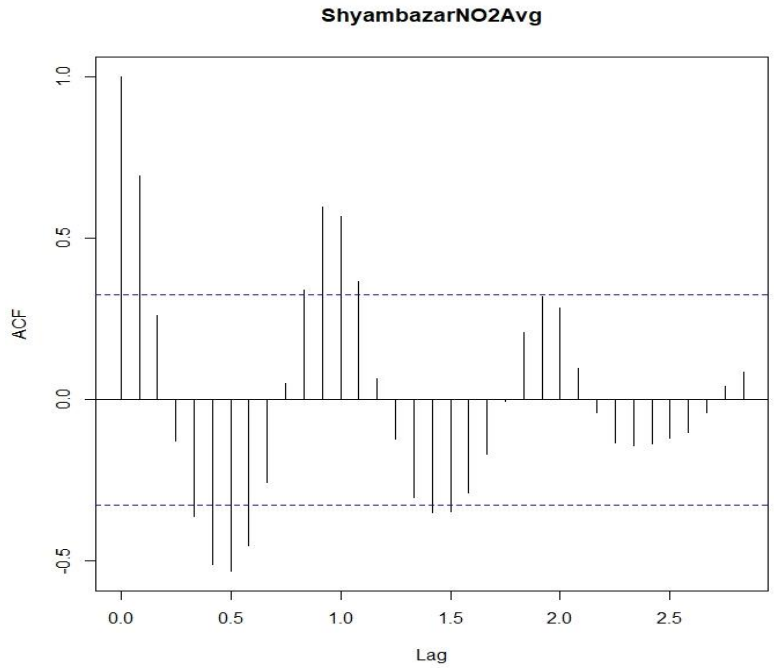
Best Model	AICc
NO2 :	
ARIMA(1,1,0)(1,1,0)[12]	: 160.3292
ARIMA(0,1,1)(0,1,1)[12]	: 160.6037
ARIMA(1,1,0)(0,1,0)[12]	: 160.316
ARIMA(1,1,0)(0,1,1)[12]	: 160.3696
ARIMA(2,1,0)(0,1,0)[12]	: 162.9672
ARIMA(1,1,1)(0,1,0)[12]	: 162.3477
ARIMA(0,1,1)(0,1,0)[12]	: 159.8668
ARIMA(0,1,1)(1,1,0)[12]	: 160.6037
ARIMA(0,1,1)(1,1,1)[12]	: 163.5628
ARIMA(0,1,2)(0,1,0)[12]	: 162.428
ARIMA(1,1,2)(0,1,0)[12]	: 163.8292
SO2 :	
ARIMA(0,0,0)(0,1,0)[12] with drift	: 101.833
ARIMA(1,0,0)(1,1,0)[12] with drift	: 103.751
ARIMA(0,0,1)(0,1,1)[12] with drift	: 103.7946
ARIMA(0,0,0)(0,1,0)[12]	: 100.4035
ARIMA(0,0,0)(1,1,0)[12] with drift	: 101.2888
ARIMA(0,0,0)(0,1,1)[12] with drift	: 101.2888
ARIMA(0,0,0)(1,1,1)[12] with drift	: 104.1941

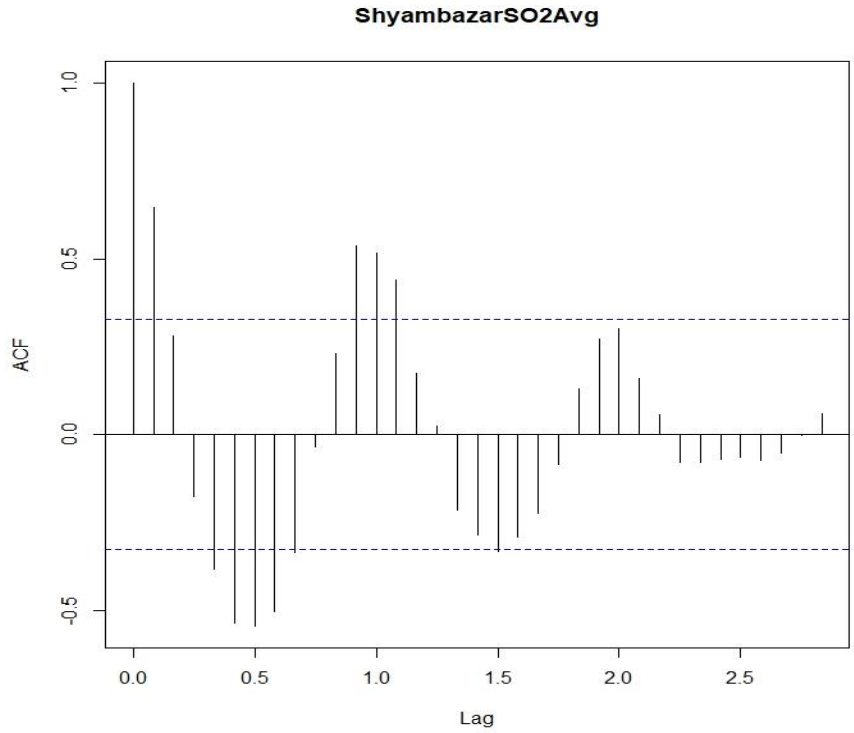
ARIMA(1,0,0)(0,1,0)[12] with drift : 103.2447
ARIMA(0,0,1)(0,1,0)[12] with drift : 103.441
ARIMA(1,0,1)(0,1,0)[12] with drift : 105.2843

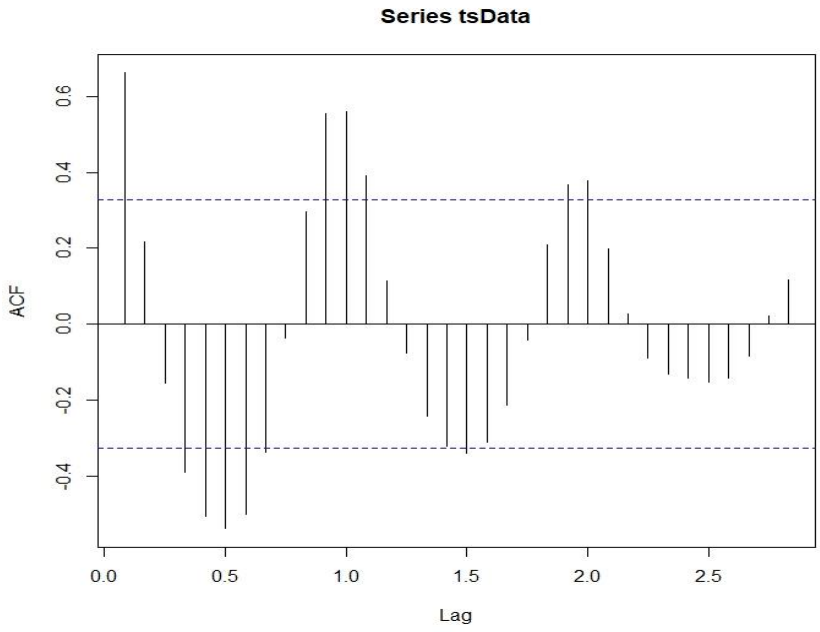
PM2.5 :

ARIMA(1,0,0)(0,1,0)[12] with drift : 211.0778
ARIMA(2,0,0)(0,1,0)[12] with drift : 211.3359
ARIMA(1,0,1)(0,1,0)[12] with drift : 205.7578
ARIMA(1,0,1)(1,1,0)[12] with drift : 208.2859
ARIMA(1,0,1)(0,1,1)[12] with drift : 208.4449
ARIMA(2,0,1)(0,1,0)[12] with drift : 208.9754
ARIMA(1,0,2)(0,1,0)[12] with drift : 208.9144
ARIMA(0,0,2)(0,1,0)[12] with drift : 205.7152
ARIMA(0,0,2)(1,1,0)[12] with drift : 208.4011
ARIMA(0,0,2)(0,1,1)[12] with drift : 208.5152
ARIMA(0,0,3)(0,1,0)[12] with drift : 208.9406
ARIMA(0,0,2)(0,1,0)[12] : 202.8395
ARIMA(0,0,2)(1,1,0)[12] : 205.2267
ARIMA(0,0,2)(0,1,1)[12] : 205.3263
ARIMA(1,0,2)(0,1,0)[12] : 205.7444
ARIMA(0,0,3)(0,1,0)[12] : 205.7446
ARIMA(1,0,1)(0,1,0)[12] : 202.8649
ARIMA(1,0,3)(0,1,0)[12] : 208.9455

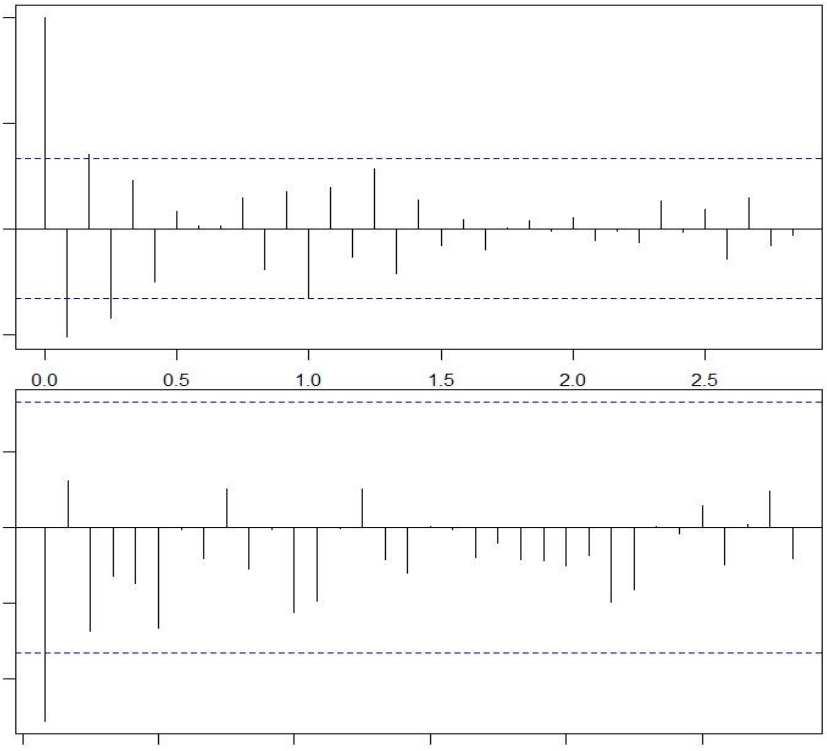
For Shyambazar the ACF plot for NO2,SO2 and PM2.5 are give below :

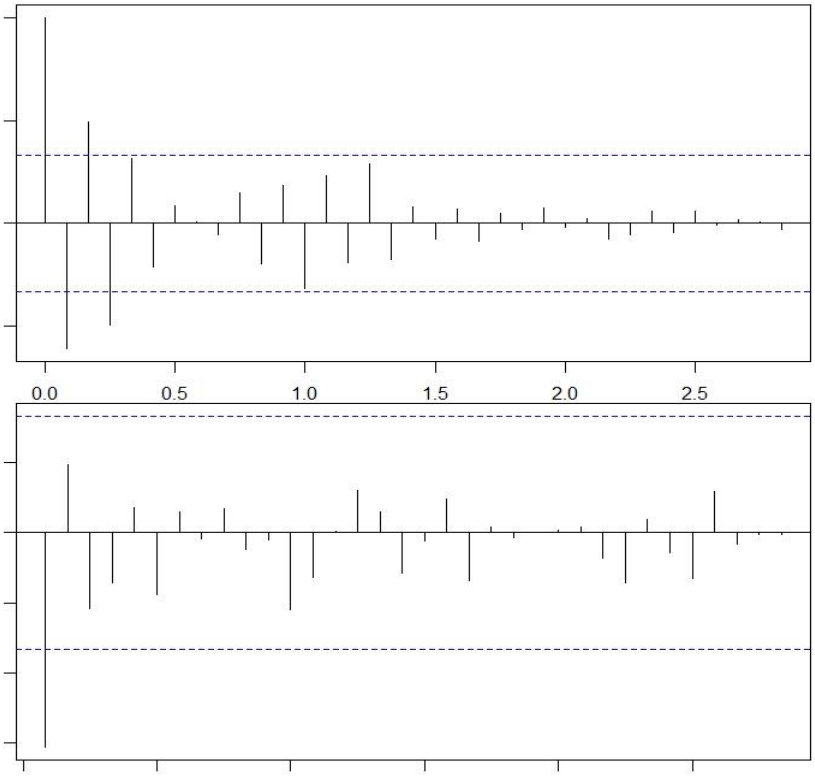


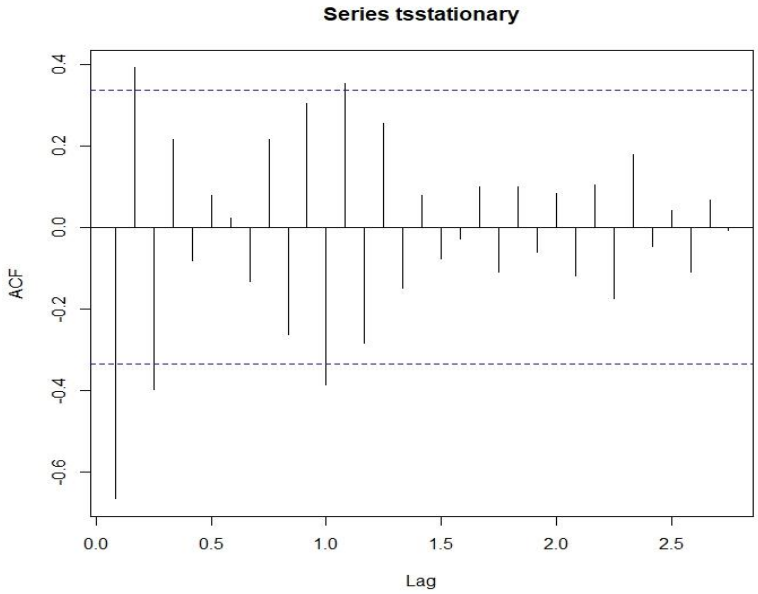


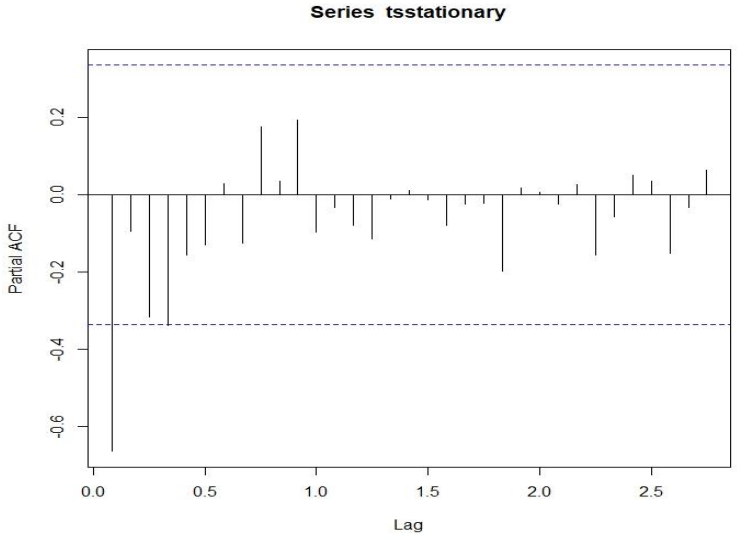


And after Seasonality removes and making the data stationary with corresponding differences needed the ACF and PACF plots for NO2,SO2,PM2.5 are given below :


0.0
0.5
1.0
1.5
2.0
2.5


0.0
0.5
1.0
1.5
2.0
2.5





The possible models are:

ShyamBazar :

Best Model	AICc
NO2 :	
ARIMA(0,1,0)(0,1,0)[12]	: 160.8736
ARIMA(1,1,0)(1,1,0)[12]	: 152.2833
ARIMA(1,1,0)(0,1,0)[12]	: 154.1877
ARIMA(0,1,0)(1,1,0)[12]	: 156.8657
ARIMA(2,1,0)(1,1,0)[12]	: 154.3874

ARIMA(1,1,1)(1,1,0)[12] : 153.5242
ARIMA(0,1,1)(1,1,0)[12] : 153.0072
ARIMA(2,1,1)(1,1,0)[12] : 157.1484

SO2 :

ARIMA(0,0,0)(0,1,0)[12] with drift : 92.25638
ARIMA(1,0,0)(1,1,0)[12] with drift : 95.0082
ARIMA(0,0,1)(0,1,1)[12] with drift : 95.07876
ARIMA(0,0,0)(0,1,0)[12] : 89.91997
ARIMA(0,0,0)(1,1,0)[12] with drift : 92.33395
ARIMA(0,0,0)(0,1,1)[12] with drift : 92.33396
ARIMA(0,0,0)(1,1,1)[12] with drift : 95.23921
ARIMA(1,0,0)(0,1,0)[12] with drift : 94.44924
ARIMA(0,0,1)(0,1,0)[12] with drift : 94.61112
ARIMA(1,0,1)(0,1,0)[12] with drift : 94.80931

PM2.5 :

ARIMA(0,0,0)(0,1,0)[12] with drift : 208.0394
ARIMA(1,0,0)(1,1,0)[12] with drift : 201.9418
ARIMA(0,0,0)(0,1,0)[12] : 205.6566
ARIMA(1,0,0)(0,1,0)[12] with drift : 207.8714
ARIMA(0,0,0)(1,1,0)[12] with drift : 201.3211
ARIMA(0,0,0)(1,1,1)[12] with drift : 204.2264
ARIMA(0,0,1)(1,1,0)[12] with drift : 201.9922
ARIMA(1,0,1)(1,1,0)[12] with drift : 205.0305
ARIMA(0,0,0)(1,1,0)[12] : 198.7184
ARIMA(0,0,0)(1,1,1)[12] : 201.347
ARIMA(1,0,0)(1,1,0)[12] : 199.0376
ARIMA(0,0,1)(1,1,0)[12] : 199.091
ARIMA(1,0,1)(1,1,0)[12] : 201.8024

From here we take the models with lower AICc values.

For Behala :

NO2 : Best model: ARIMA(1,1,0)(1,1,0)[12]

Coefficients:

	ar1	sar1
	-0.4831	-0.7372
s.e.	0.1895	0.1403

sigma^2 estimated as 24.92: log likelihood=-73.41

AIC=152.81 AICc=154.08 BIC=156.22

SO2 : Best model: ARIMA(0,0,0)(1,1,0)[12]

Coefficients:

	sar1
	-0.6656
s.e.	0.1609

sigma^2 estimated as 1.839: log likelihood=-44.37

AIC=92.73 AICc=93.3 BIC=95.09

PM2.5 :

Best model: ARIMA(0,0,0)(1,1,1)[12]

Coefficients:

	sar1	sma1
	-0.8859	-0.272
s.e.	NaN	NaN

sigma^2 estimated as 58.48: log likelihood=-94.07

AIC=194.15 AICc=195.35 BIC=197.68

For Moulali :

NO2 :

Best model: ARIMA(0,1,1)(0,1,0)[12]

Coefficients:

ma1
-0.7120
s.e. 0.1533

sigma^2 estimated as 50.73: log likelihood=-77.63

AIC=159.27 AICc=159.87 BIC=161.54

SO2 :

Best model: ARIMA(0,0,0)(0,1,0)[12]

sigma^2 estimated as 3.507: log likelihood=-49.11

AIC=100.22 AICc=100.4 BIC=101.4

PM2.5 :

```
Best model: ARIMA(0,0,2)(0,1,0)[12]

Coefficients:
         ma1     ma2
      1.1723  0.2587
s.e.  0.2166  0.2273

sigma^2 estimated as 202.5:  log likelihood=-97.82
AIC=201.64   AICc=202.84   BIC=205.17
```

Shyambazar :

NO2 :

Best model: ARIMA(1,1,0)(1,1,0)[12]

Coefficients:

ar1	sar1
-0.5457	-0.5708
s.e. 0.1800	0.1982

sigma^2 estimated as 28.15: log likelihood=-72.51

AIC=151.02 AICc=152.28 BIC=154.43

SO2 :

Best model: ARIMA(0,0,0)(0,1,0)[12]

sigma^2 estimated as 2.266: log likelihood=-43.87

AIC=89.74 AICc=89.92 BIC=90.92

PM2.5 :

```
Best model: ARIMA(0,0,0)(1,1,0)[12]

Coefficients:
         sar1
      -0.7351
s.e.   0.1328

sigma^2 estimated as 135:  log likelihood=-97.07
AIC=198.15   AICc=198.72   BIC=200.5
```

When the tentative model is identified, the next step is to achieve the most efficient estimates of the parameters.

Model estimation: Then, based on ARIMA model, the parameter of each coefficient in AR and MA for both seasonal and non-seasonal can be identified. However, before choosing the coefficient values, the estimation output must satisfy the significant criteria in the next step, model checking.

Also for PM2.5 Data(API for the time series analysis) we check the accuracy of models chosen for each places and find that these are minimum values. The minimum values of MAPE,MAE,MSE,RMSE are given as :

Stations	Model	MAPE	MAE	MSE	RMSE
Behala	ARIMA(0,0,0)(1,1,1)[12]	4.893	3.163	0.146	4.789
Moulali	ARIMA(0,0,2)(0,1,0)[12]	10.683	7.087	0.331	11.125
ShyamBazar	ARIMA(0,0,0)(1,1,0)[12]	8.977	6.209	0.284	9.288

Model checking and forecasting: By using Ljung-Box test and checking the p-value of the

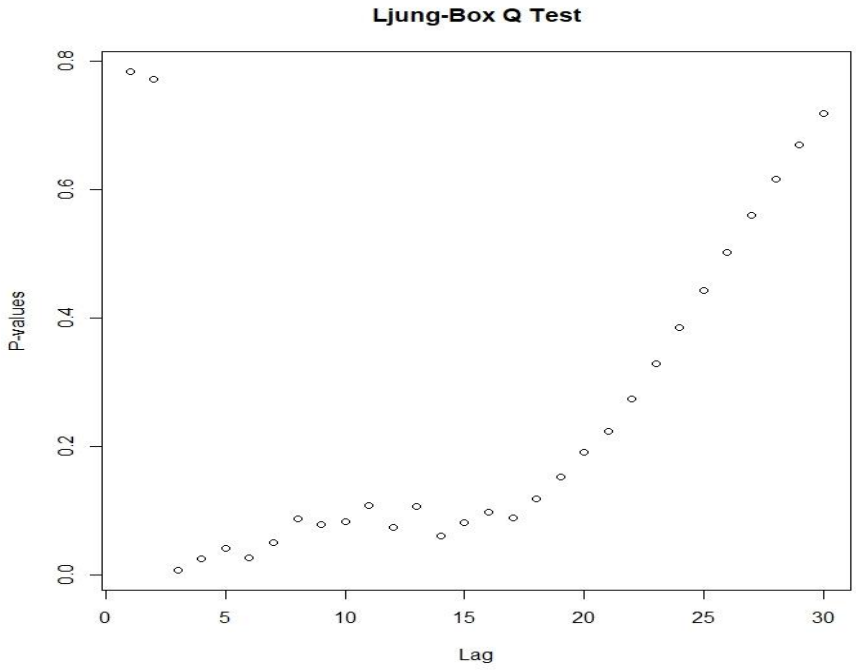
coefficient, then the significant model can be determined. Without taking the constant value into the model, the possible model stated for each station satisfied both of these tests.

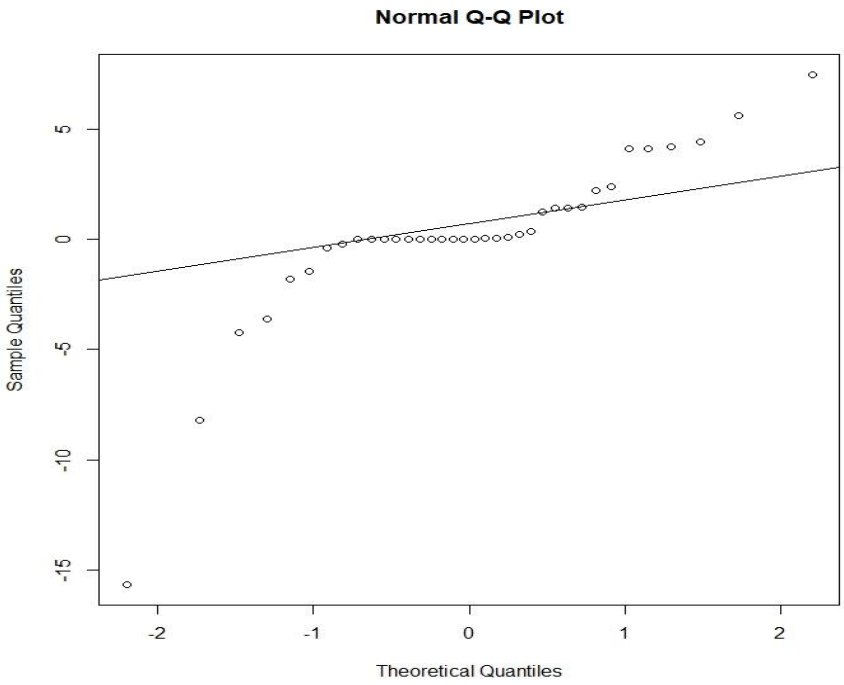
Most parametric tests require that residuals be normally distributed and that the residuals be homoscedastic. One approach when residuals fail to meet these conditions is to transform one or more variables to better follow a normal distribution. Often, just the dependent variable in a model will need to be transformed. However, in complex models and multiple regression, it is sometimes helpful to transform both dependent and independent variables that deviate greatly from a normal distribution.
Here we checked for Log transformation , Square root transformation and finally the Tukey's ladder of power transformation and the best result was obtained from Tukey transformation.

The Ljung-Box Q tests and Normal Q-Q plots of Behala,Moulali and Shyambazar for NO2,SO2 and PM2.5 respectively are given below :

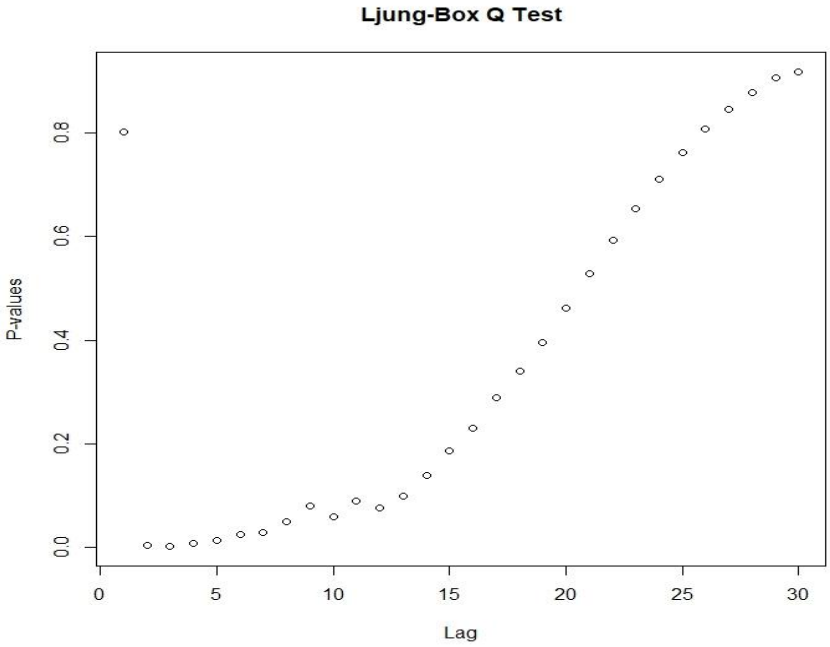
For Behala :

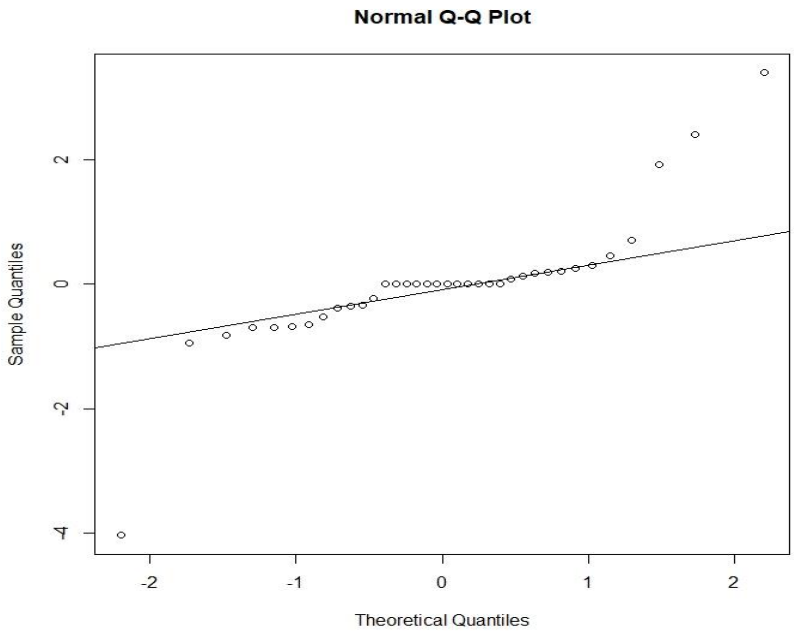
Fot pollutant NO2 the test corresponding to the best fit :



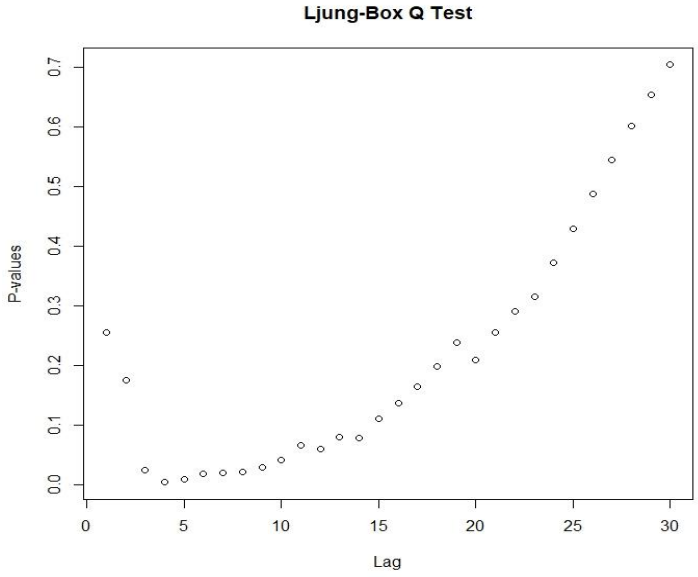


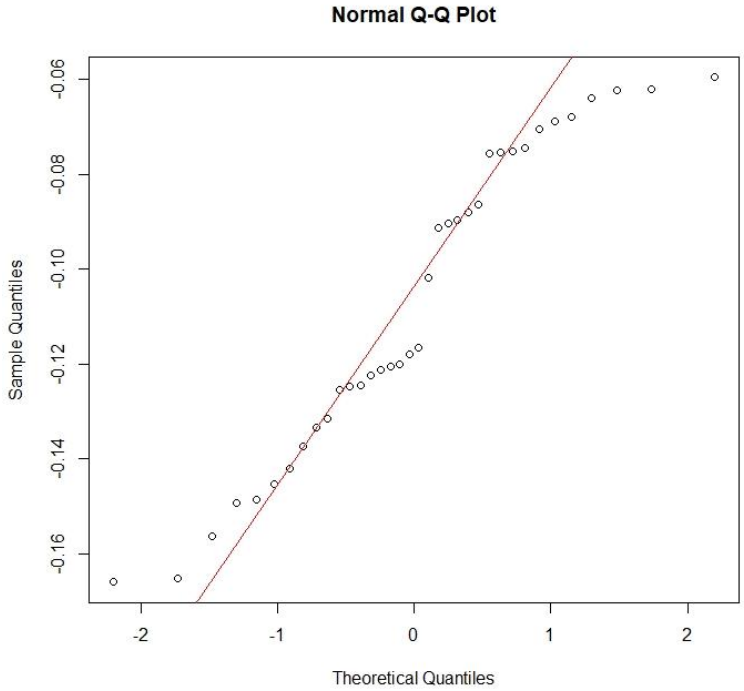
Fot pollutant SO2 the test corresponding to the best fit :





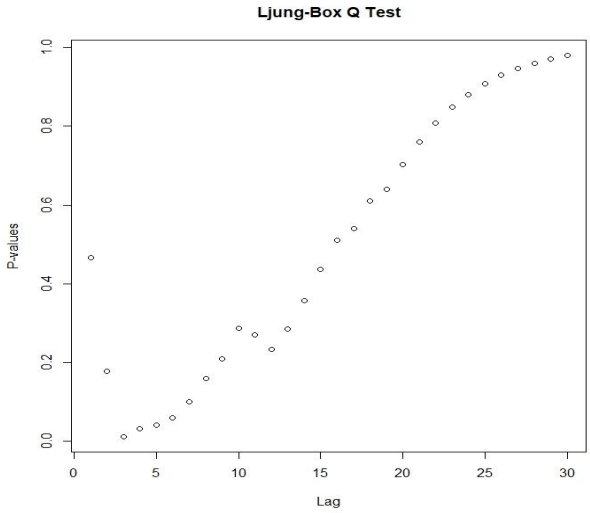
Fot pollutant PM2.5 the test corresponding to the best fit :

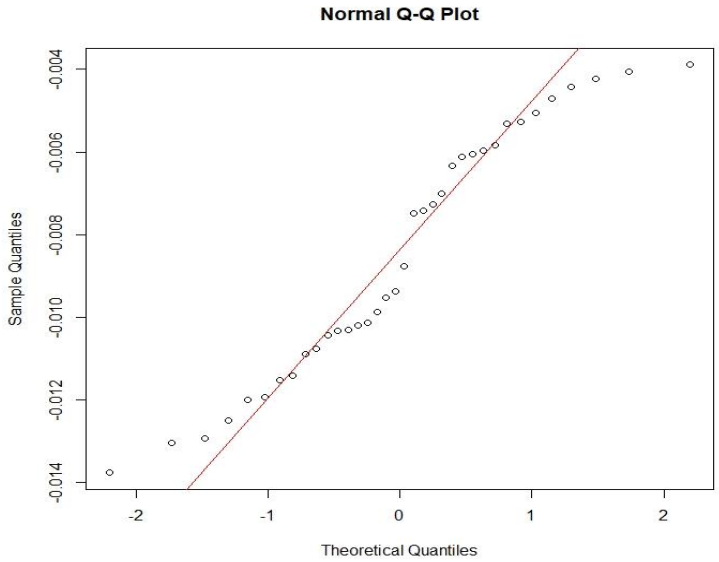




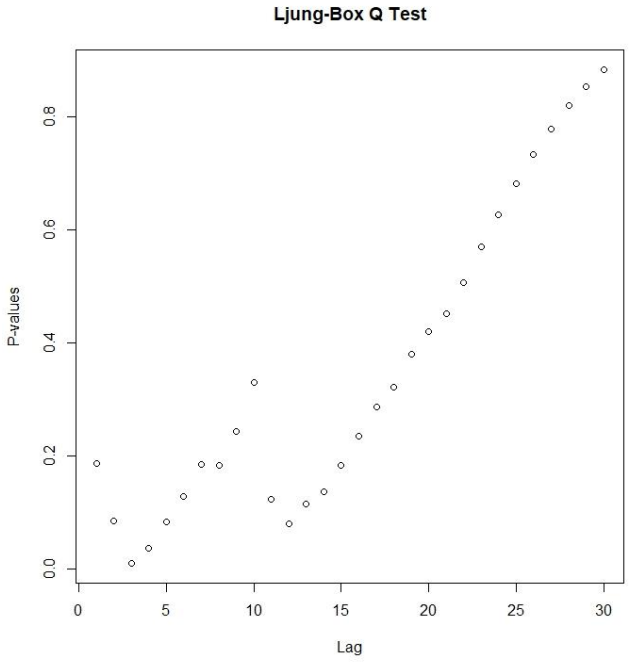
For Moulali :

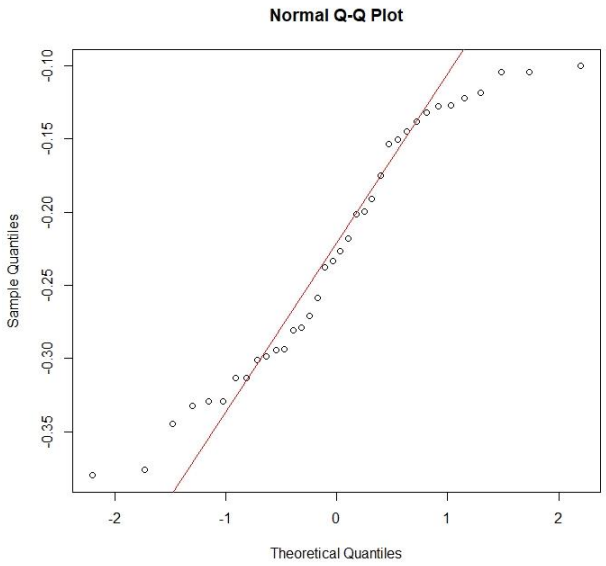
Fot pollutant NO2 the test corresponding to the best fit :



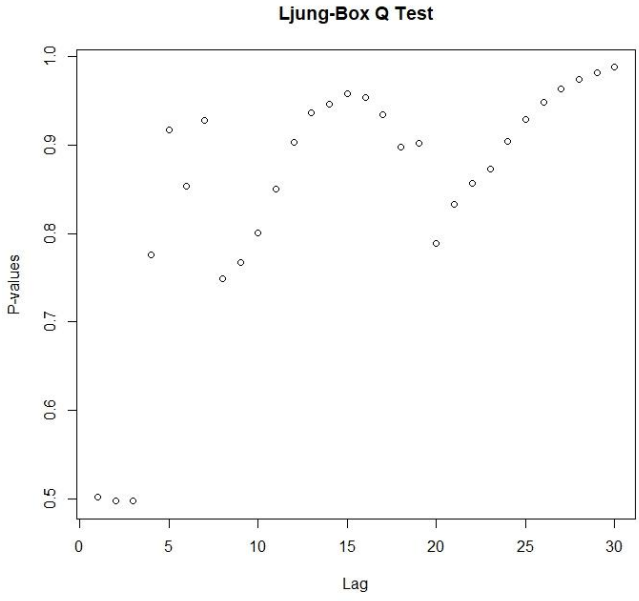


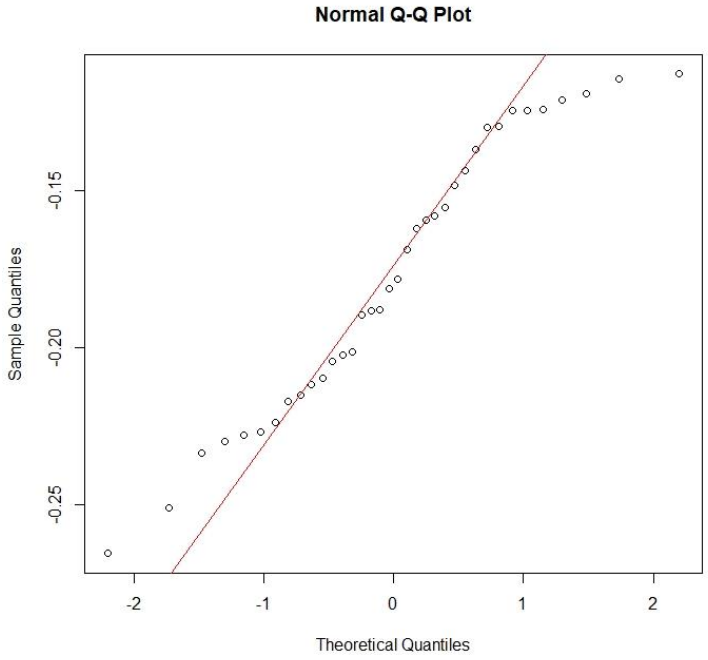
Fot pollutant SO2 the test corresponding to the best fit :





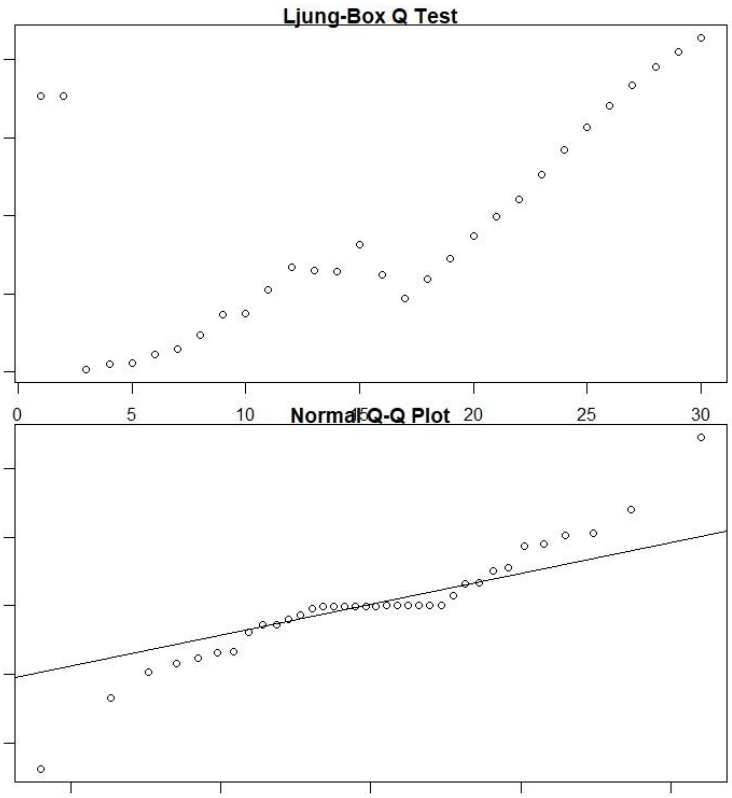
Fot pollutant PM2.5 the test corresponding to the best fit :



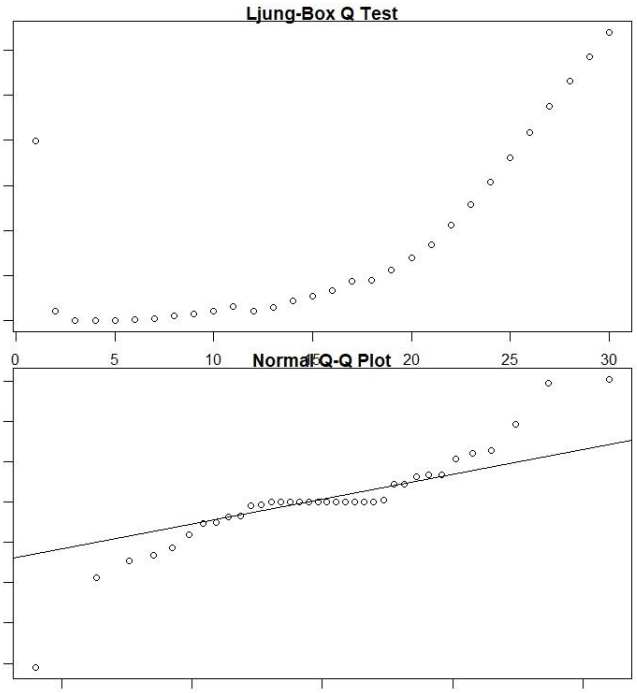


For Shyambazar :

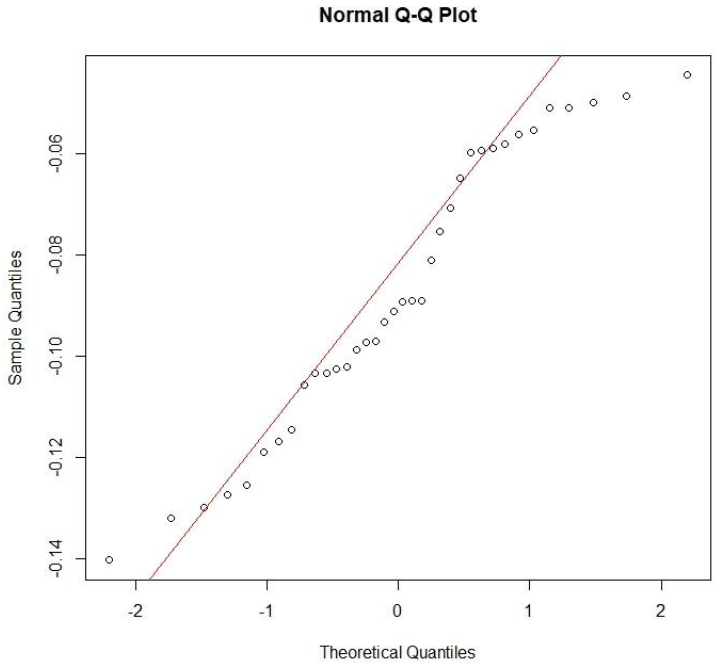
Fot pollutant NO2 the test corresponding to the best fit :



Fot pollutant SO2 the test corresponding to the best fit :



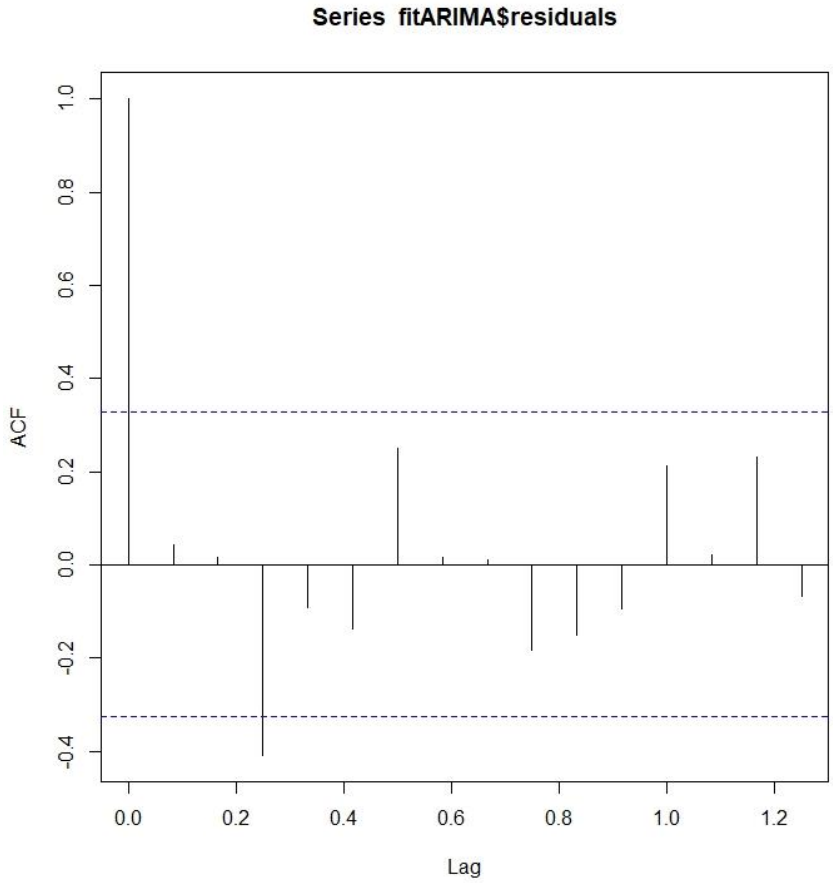
Fot pollutant PM2.5 the test corresponding to the best fit :



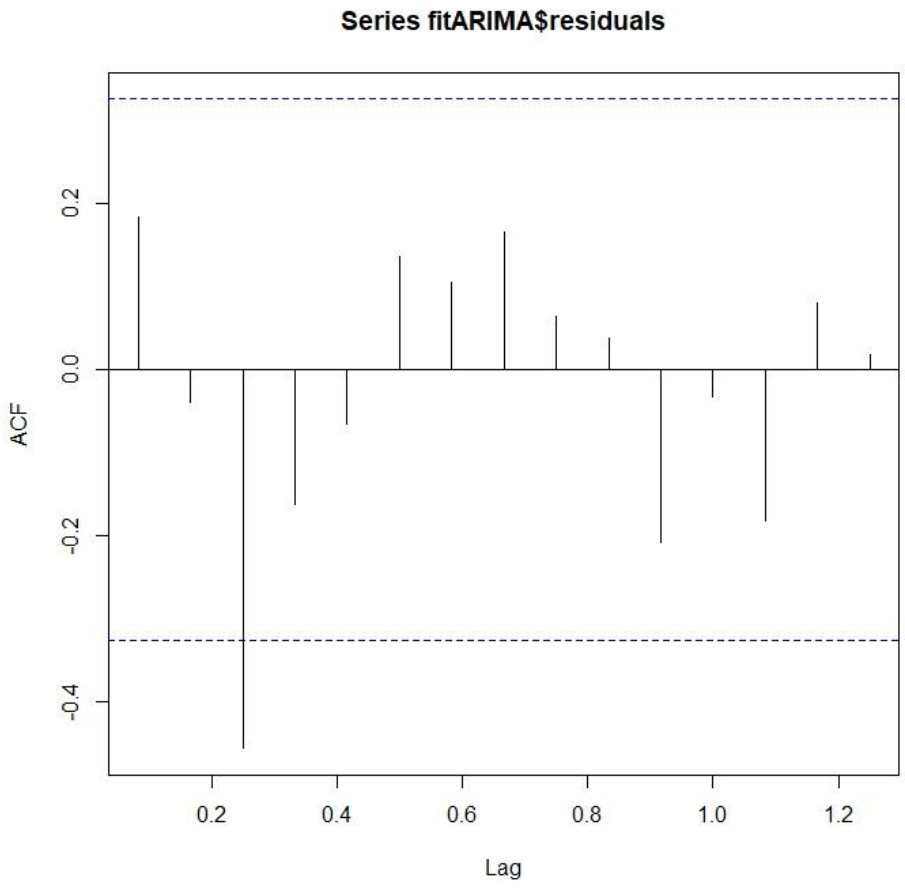
Then, the process continues to calculate the forecasting values based on satisfied model. Forecasting ARIMA model: The best model is chosen based on the smallest values of accuracy measurement.

ACF of ARIMA models with Residuals :

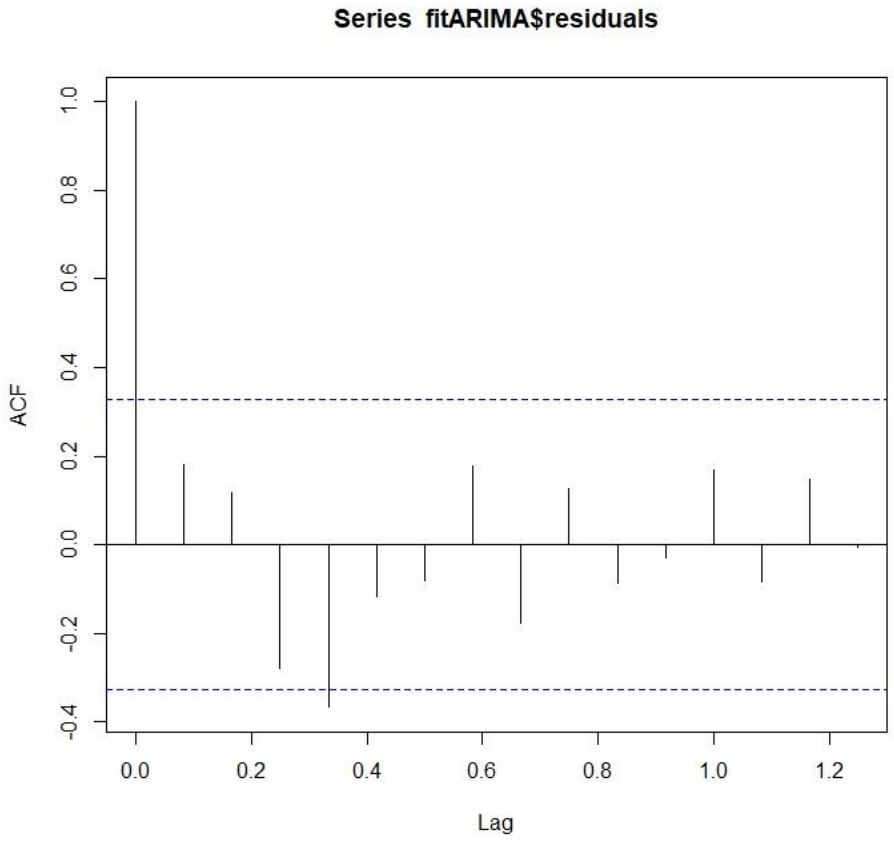
Behala NO2 :


Series fitARIMA$residuals
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Lag

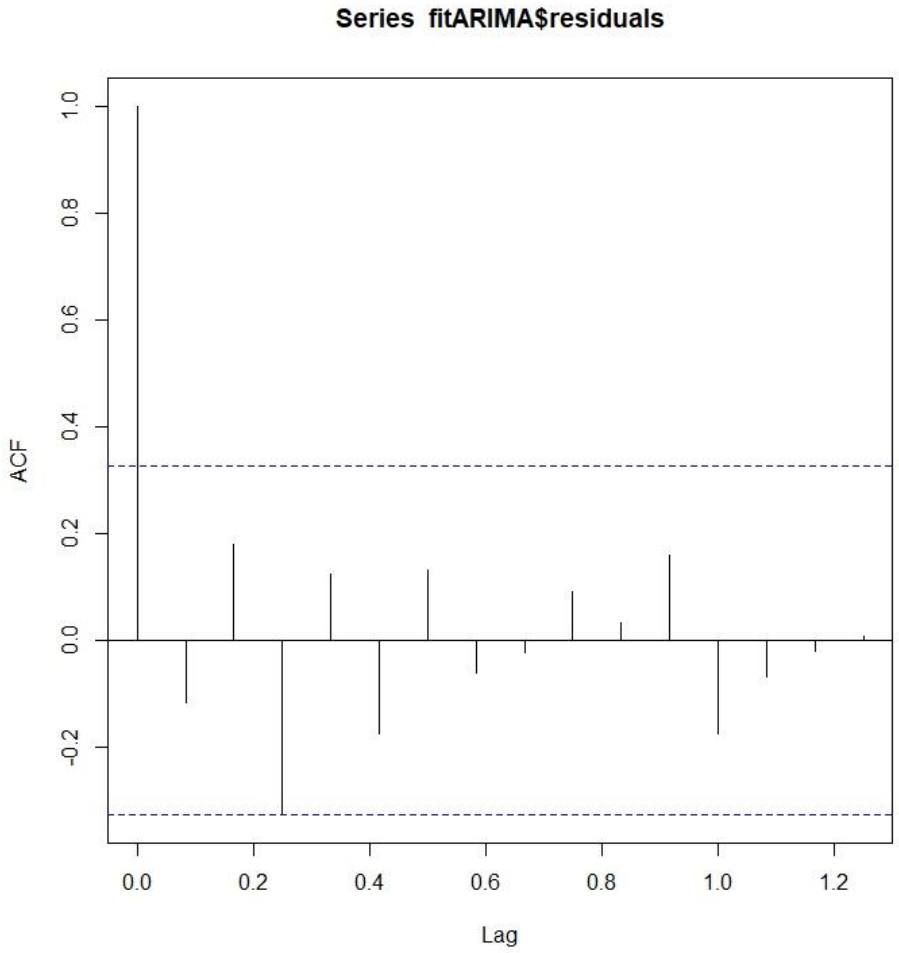
Behala SO2 :



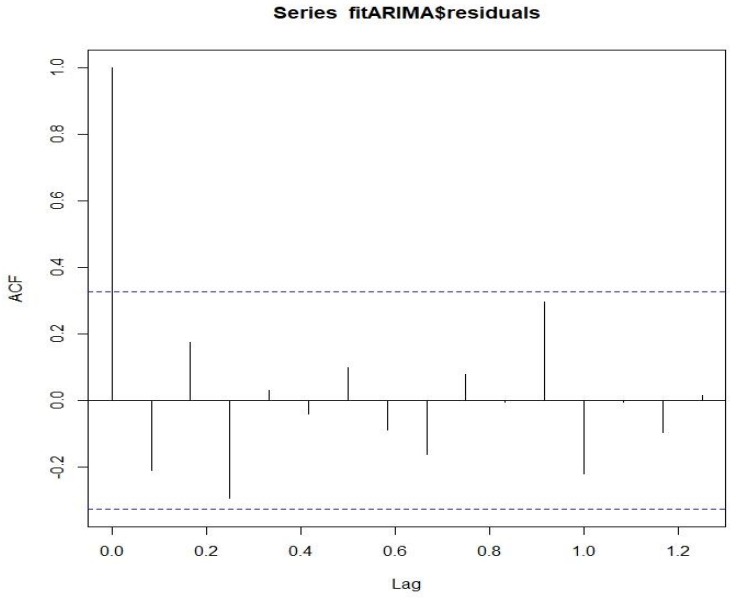
Behala PM2.5 :


Series fitARIMA$residuals
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Lag

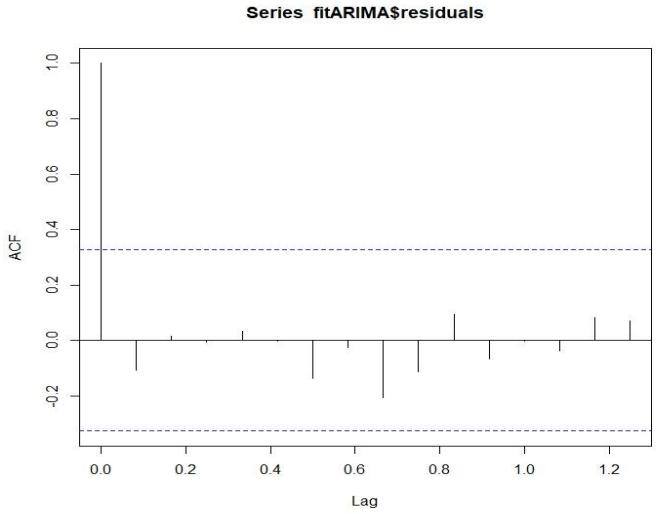
Moulali NO2 :


Series fitARIMA$residuals
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Lag

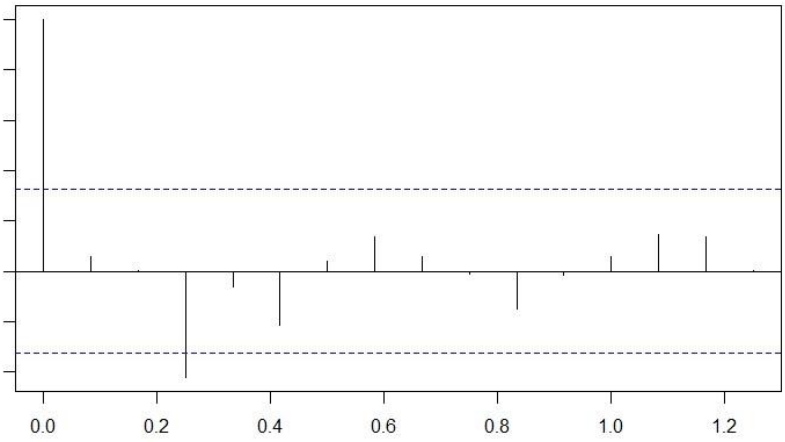
Moulali SO2 :


Series fitARIMA$residuals
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Lag

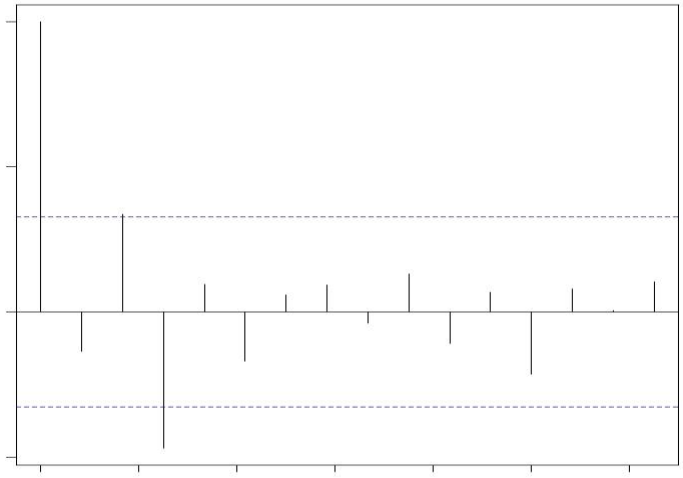
Moulali PM2.5 :



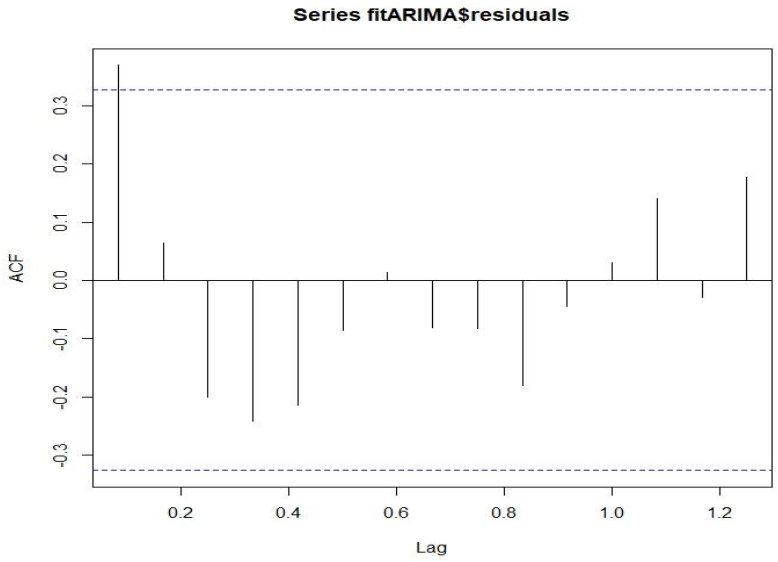
Shyambazar NO2 :



Shyambazar SO2 :



Shyambazar PM2.5 :



From the ACF plots of residuals we can see that they are White Noise.

So finally we create the Model equations from the coefficients obtained in RStudio.

Equations for Forecasting:

Behala NO2 :

$$Y_t = 0.5169Y_{t-1} + 0.4831Y_{t-2} + 0.2628Y_{t-12} - 0.1358Y_{t-13} - 0.127Y_{t-14} + 0.7372Y_{t-24} - 0.3811Y_{t-25} - 0.3561Y_{t-26} + \varepsilon_t$$

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.48310	0.18951	-2.5491	0.0108
sar1	-0.73723	0.14033	-5.2536	1.492e-07

Behala SO2 :

$$Y_t = 0.3344Y_{t-12} + 0.6656Y_{t-24} + \varepsilon_t$$

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
sar1	-0.66562	0.16088	-4.1373	3.514e-05

Behala PM2.5 :

$$Y_t = 0.1141Y_{t-12} - 0.8859Y_{t-24} + 0.272\varepsilon_{t-12} + \varepsilon_t$$

Moulali NO2 :

$$Y_t = Y_{t-1} + Y_{t-12} - Y_{t-13} + \varepsilon_t + 0.712\varepsilon_{t-1}$$

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
ma1	-0.71225	0.15313	-4.6514	3.298e-06

Moulali SO2 :

$$Y_t = Y_{t-12} + \varepsilon_t$$

Moulali PM2.5 :

$$Y_t = Y_{t-12} + \varepsilon_t - 1.1723\varepsilon_{t-1} - 0.2587\varepsilon_{t-2}$$

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ma1  1.17234    0.21659  5.4128 6.204e-08
ma2  0.25866    0.22733  1.1378    0.2552
```

Shyambazar NO2 :

$$Y_t = 0.4543Y_{t-1} + 0.5457Y_{t-2} + 0.4292Y_{t-12} - 0.195Y_{t-13} - 0.2342Y_{t-14} + 0.5708Y_{t-24} - 0.2593Y_{t-25} - 0.3115Y_{t-26} + \varepsilon_t$$

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.54567	0.18004	-3.0309	0.002438
sar1	-0.57076	0.19824	-2.8792	0.003987

Shyambazar SO2 :

$$Y_t = Y_{t-12} + \varepsilon_t$$

Shyambazar PM2.5 :

$$Y_t = 0.2649Y_{t-12} + 0.7351Y_{t-24} + \varepsilon_t$$

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
sar1	-0.73509	0.13278	-5.536	3.094e-08

The actual and predicted values of NO2,SO2 and PM2.5 for next 5 months are given below for different monitoring stations :

Table 1. Actual and predicted values for the year 2019 in Behala

Month-Year	NO2		SO2		PM2.5	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
Jan-19	65.03	61.17	14.06	9.52	162	138.83
Feb-19	52.93	50.45	8.2	7.51	108.6	105.4
Mar-19	44.6	34.52	5.8	4.64	76.4	64.87
Apr-19	36.95	28.65	6.88	3.78	39.51	46.27
May-19	36.47	28.21	5.78	3.77	35.38	47.55

Table 2. Actual and predicted values for the year 2019 in Moulali

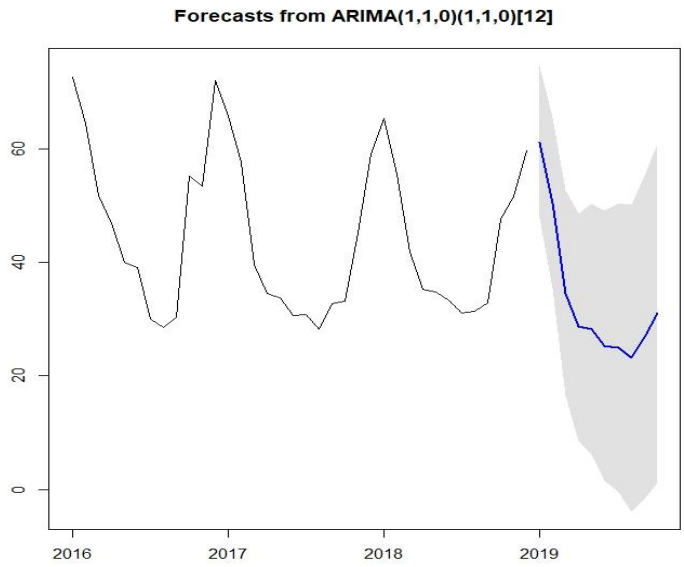
Month-Year	NO2		SO2		PM2.5	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
Jan-19	71.27	77.66	13.9	12.87	109.2	108.14
Feb-19	56.57	59.82	8.43	9.9	71.8	70.17
Mar-19	45.3	45.89	5.97	6.3	49.8	51.2
Apr-19	43.74	40.99	7.6	4.5	35.53	33.8
May-19	37.86	40.62	5.2	3.43	34.23	38.6

Table 3. Actual and predicted values for the year 2019 in Shyambazar

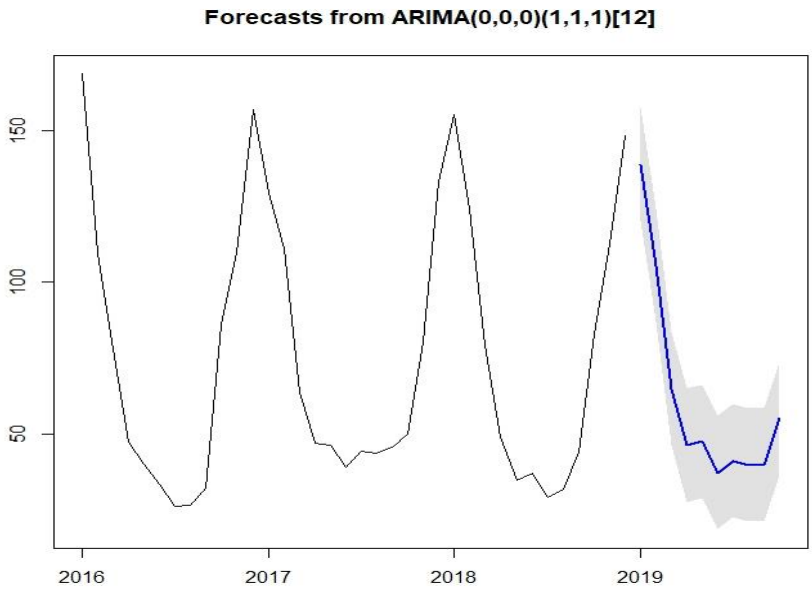
Month-Year	NO2		SO2		PM2.5	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
Jan-19	65.03	65.49	14.06	11.83	162	146.62
Feb-19	52.93	51.43	8.2	7.56	108.6	113.52
Mar-19	44.6	36.21	5.83	5.47	76.4	61.17
Apr-19	40.89	31.86	5.55	3.73	41	48.65
May-19	37.9	30.29	6.04	3.37	35.12	42.15

Finally we give the forecast of future values graphically :

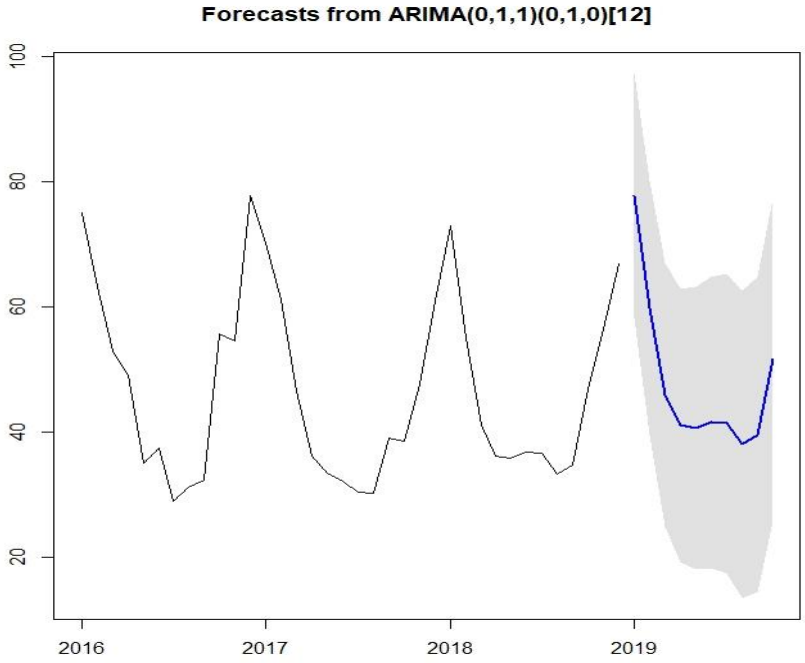
Behala NO2 :



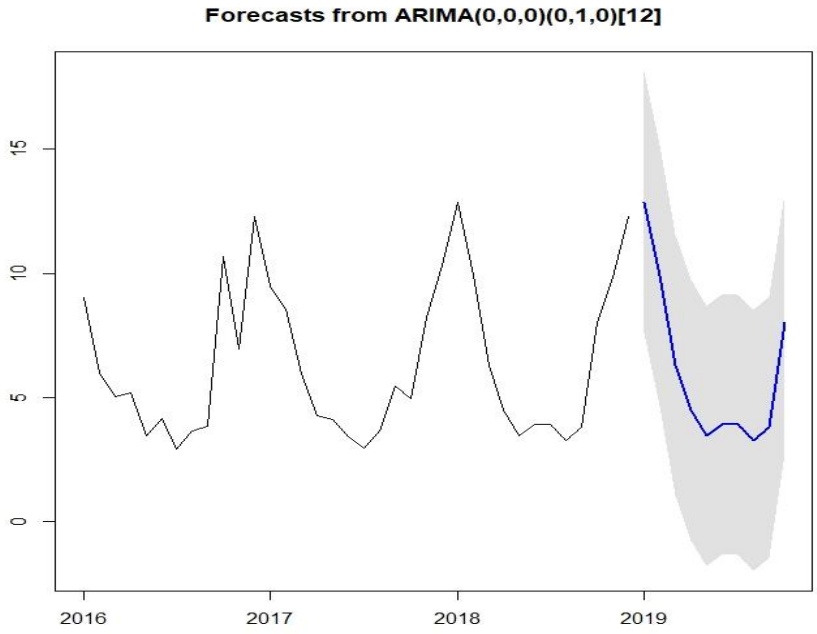
Behala PM2.5 :



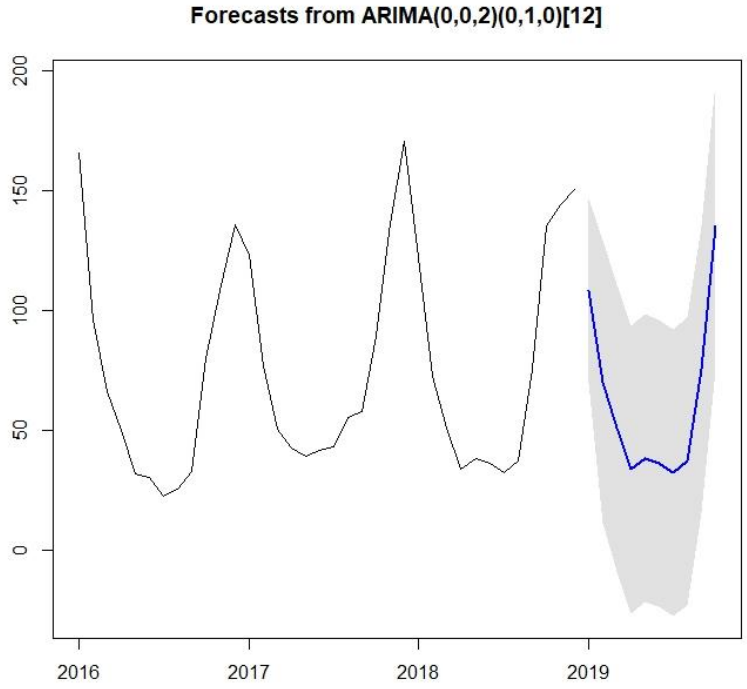
Moulali NO2 :



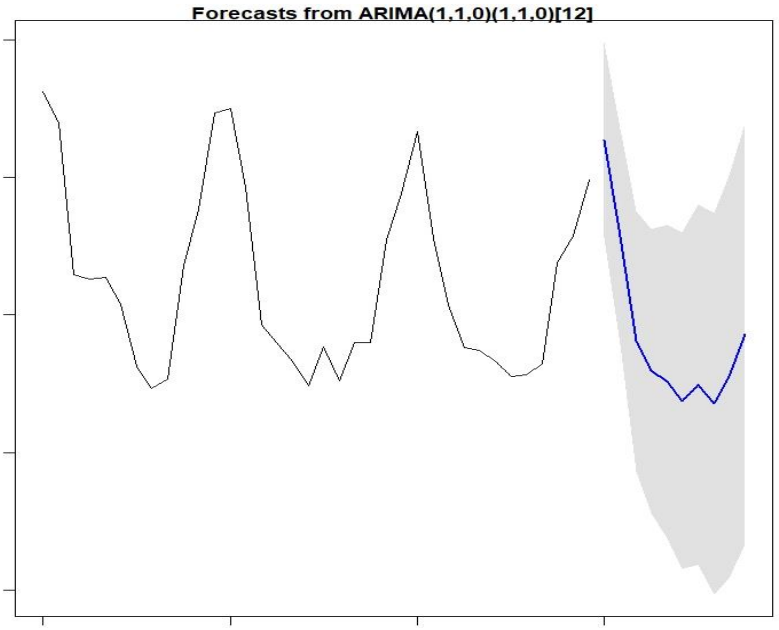
Moulali SO2 :



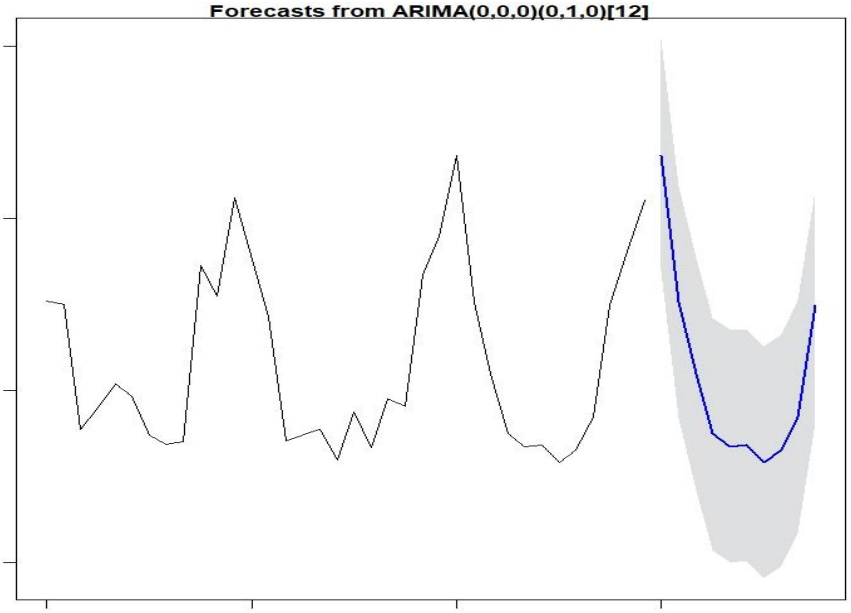
Moulali PM2.5 :



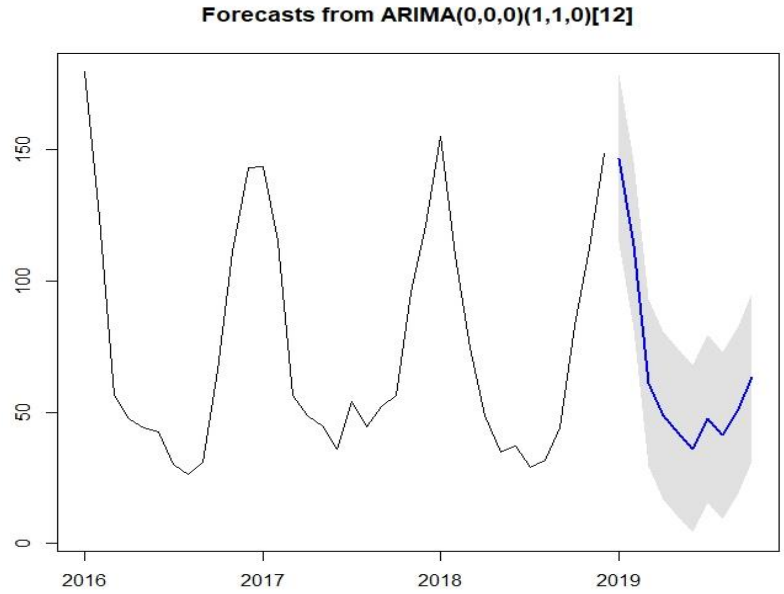
Shyambazar NO2 :



Shyambazar SO2 :



Shyambazar PM2.5 :



Model Updation:

Now from the Model equation we can see that the Models for Moulali SO2 and Shyambazar SO2 are seasonal random walk models. So these two models need to be updated after every 12 months. Also as the dataset used was small(only 3 years) the other models fail to predict more than next 26 months. So we have to update the models every 2 years with relevant data obtained.

Conclusion :

In this study, the time series analysis and forecasting be used to analyze the Air Pollution Data in Kolkata. It is caused that this method has been proven as an effective way in most of research area. Moreover, the ARIMA model has become the most popular methods in forecasting.

By looking at the forecast plots of the model chosen, high values recorded at Moulali,Kolkata. This situation indicates that the most polluted area in Kolkata located in Moulali.

In summary, the time series model used in forecasting is an important tool in monitoring and controlling the air quality condition. It is useful to take quick action before the situations worsen in the long run. In that case, better model performance is crucial to achieve good air quality forecasting. Moreover, the pollutants must in consideration in analysis air pollution data.

Source Code : Here we give the libraries and codes used in RStudio to obtain all the graphs,models and forecasting values which helped us see if Time Series Analysis is really fruitful for Air Pollution data or not.

```
#Load Libraries

library(fUnitRoots)

library(lmtest)

library(forecast)

library(FitAR)

library(readxl)

library(TSA)


#import data
```

```
RawData<-read_excel("data.xlsx",header=TRUE)

#convert to time series

tsData<- ts(RawData,start=c(2016,1),frequency=12)

plot(tsData)


#decompose into time series components

timeseriescomponents <- decompose(tsData)

plot(timeseriescomponents)


#detemine stationarity of data

tsstationary<-diff(tsData, differences=d)

plot(tsstationary)

acf(tsData,lag.max=n)


#remove seasonality

timeseriesseasonallyadjusted <- tsData- timeseriescomponents$seasonal

plot(timeseriesseasonallyadjusted)

tsstationary <- diff(timeseriesseasonallyadjusted, differences=d)

plot(tsstationary)

acf(tsstationary, lag.max=n)

pacf(tsstationary, lag.max=n)


#fit the model

fitARIMA<-arima(tsData, order=c(p,d,q),seasonal = list(order = c(P,D,Q), period = 12),method="ML")

fitARIMA
```

```
#significance of coefficients

coeftest(fitARIMA)

acf(fitARIMA$residuals)


#checking accuracy

accuracy(fitARIMA)


#residual diagnostics

boxresult<-LjungBoxTest (fitARIMA$residuals,k=2,StartLag=1) # residual?? or the original series?

plot(boxresult[,3],main="Ljung-Box Q Test", ylab="P-values", xlab="Lag")

qqnorm(fitARIMA$residuals)

qqline(fitARIMA$residuals)


#checking if the model taken is correct according to R

auto.arima(tsData, trace=TRUE)


#forcast future values

par(mfrow=c(1,1))

predict(fitARIMA,n.ahead = 5)

futurVal <- forecast (fitARIMA,h=10, level=c(99.5))

plot (futurVal)
```