EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Development of time series models for various pollutants in Bangalore city using the Akaike information criterion

Vivekanand Venkataraman*, Shashank Prasad, Balakrishna Aswathanarayana, Susmith Barigidad, Vinayak Nayak and Sai Tarun Kumar N

Industrial Engineering and Management, Ramaiah Institute of Technology, MSRIT post, Bangalore -54, India

Received 11 October 2019
Revised 1 February 2020
Accepted 5 February 2020

Abstract

Pollution levels in developing countries, such as India, have become a major source of health problems. They need to be monitored and controlled. Bangalore, one of the major cities in India, faces a huge amount of pollution. Due to the dire need to control these pollutants, a sound mathematical modeling approach needs to be created for forecasting, controlling and monitoring. One such approach is time series modeling. The current work addresses a time series model that has been developed for the major pollutants in Bangalore city. These pollutants include PM_{10}, $PM_{2.5}$, NO_x and SO_2. The models used vary from AR (autoregressive), ARMA (autoregressive moving average) and ARIMA (autoregressive integrated moving average) for modeling air pollution in Bangalore city. Additionally, the selection of the best models was based on the Akaike Information Criterion, p-value and Box-Pierce test. Various steps were followed to build the model, which included identification of missing and extreme values followed by creating an appropriate imputing method and then identification of time series models using autocorrelation and partial autocorrelation plots to obtain various time series models. The best time series models were chosen based on the Akaike Information criterion (AIC) and various other statistical tests.

Keywords: Time series, Air pollution, Akaike information criterion, ARIMA, Statistics

1. Introduction

During the early 1900s, time series analysis was conducted on wheat price fluctuations. The analysis was termed harmonic analysis and was done using periodograms [1-2]. Further advancement of time series analysis was achieved wherein the concept of differencing was used to show the periodic behavior of time series, through which insights can be obtained on the behaviour of a dataset [3-4]. Later Wold [5] had highlighted the usage of moving average model for time series by making use of correlogram/autocorrelation plot, the models developed led to the development of auto regressive(AR) models, MA (moving average), ARMA (autoregressive moving average) models. As time progressed, further development of models for analysis took place, advancing the state-of-the-art in time series analysis. A more detailed description is given by Box et al. [6-7]. A good narration of historical development of time series and various other models has been provided [8].

Concurrently, application of time series modelling began to appear in various areas, especially in economics and environmental science. Time series modelling began to appear in air pollution studies during the 1960's. During this time period, Time series analysis using models such as MA, ARMA was applied to study pollutants such as SO_2, NO_2, Los Angeles smog and predict their behavior [9-10]. Further development was achieved by building univariate and multivariate time series regression models for prediction of pollutant concentration based on previous pollutant values and weather parameters [11]. Applying these models improved the understanding of the effects of various pollutants and weather parameters on mortality rates on given days [12].

Some authors started to investigate multivariate techniques such as principal component analysis, factor analysis, vector autoregressive models and had applied to various pollutants such as particulate matter, O_3, SO_2, and CO, for a obtaining a sound & effective model [13-19]. Applications of time series models also started to be developed in the Middle East and Asia. The ARIMA model was used to predict the behavior of various pollutants such as particulate matter, SO_2, NO_x in Delhi City and the models build had sound prediction power [20]. Similarly, other time series models for pollutants such as NO, NO_2, SO_2, CO, O_3 were used for Middle Eastern cities including Beirut and Isfahan [21-22]. Such modeling approaches were also applied in developing countries such as Nigeria as evidenced by various modeling work [23].

In India, Bangalore is a major city where there has been a huge investment in infrastructure development and industrial growth. Due to rapid growth and urbanization, the city now experiences significant pollution. There are various

*Corresponding author. Tel.: +9199 5271 6390
Email address: vivek999hyderabad@gmail.com; vivekanand@msrit.edu
doi: 10.14456/easr.2020.28

pollutants which are harmful to the environment and health, these include fine particulate matter ($PM_{2.5}$, PM_{10}), NO_x, and SO_2. There is a need to improve its air quality, which can be achieved through monitoring, prediction and control. To do so, mathematical models need to be developed. The amount of modeling achieved for air pollution is quite less in India. The number of models developed and literature addressing air pollution in India and with respect to Bangalore city is minimal. Hence, understanding the behavior of air pollutants becomes crucial as this can form a further basis of research growth. This particular work addresses the use of time series models. Its overall goal is to provide a complete approach in conducting time series analysis for various pollutants in Bangalore city. The objective of the work is to develop a sound methodology using time series techniques for air pollution data, to provide the importance of time series analysis in understanding the behavior of pollutants and also to highlight areas where the analysis is successful and where further research is needed for improvement.

2. Mathematical models for time series analysis

Typical time series data can be stationary, weakly stationary or non-stationary. In order to create sound mathematical approach stationarity of data becomes crucial, . The conditions for stationarity need to be met to create a sound mathematical approach [6-7, 24]. To do so, a transformation process is necessary. Various models that arise out of these transformations are autoregressive (AR), moving average (MA), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) Models [3, 7]. A data dependent system of time series can be transformed to a summation of independent and uncorrelated past random shocks ε_t which termed as white noise and the transformation of the data is given by equation 1 [3]. The mathematical formulation are discussed below

$$x_t = \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \cdots \quad (1)$$

$$x_t = \mu + \psi(B)\varepsilon_t \quad (2)$$

where B is called the backward shift operator and is defined as B $x_t = x_{t-1}$ and μ the process average. $\psi(B)$ is called the linear operator which transforms the input white noise to data values. It is given by:

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots \quad (3)$$

The above model forms the basis for development of the AR, MA, ARMA and ARIMA models.

2.1 Autoregressive (AR) & Moving Average (MA) models

Based on Equations 1-3, x_t can be expressed in terms of their previous values as given below and is termed as an autoregressive process or AR of order p, termed AR(p).

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots \phi_p x_{t-p} + \varepsilon_t \quad (4)$$

x_t can also be expressed in terms of the previous random shocks as and ϕ is the weight

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} \ldots . - \theta_q \varepsilon_{t-q} \quad (5)$$

Equation 5 is a process and is termed a moving average process with order q, expressed as MA(q). The weights are given by $-\theta_1, -\theta_2, \ldots - \theta_q$.

2.2 Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models

ARMA models encompass the properties of AR and MA and are given by:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} \ldots . - \theta_q \varepsilon_{t-q} \quad (6)$$

The ARIMA model represented by Equation 7 is called ARIMA(p,d,q) where w_t represents difference term $x_t - x_{t-1}$.

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots \phi_p w_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} \ldots . - \theta_q \varepsilon_{t-q} \quad (7)$$

3. Methodology

Figure 1 provides the methodology of model development.

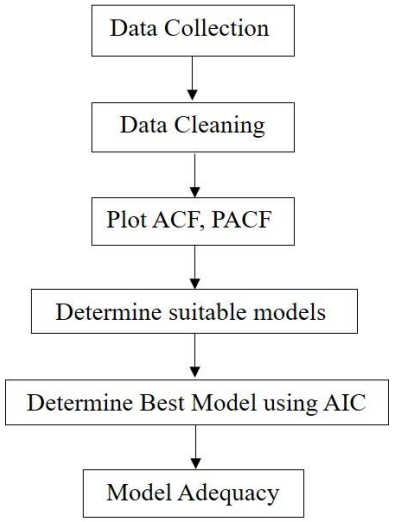

Figure 1 Methodology for time series model building

3.1 Data collection

The data required for the study was extracted from Central Pollution Control Board databases [25]. The Board monitors pollutant levels at various places in Bangalore. Data collection is done every half an hour and is displayed on their website on a real-time basis. For this study, the daily average values were considered. The data was collected for SGHalli and BTM Stations.

The BTM Layout Station was chosen because of its proximity to Belandur and Koramangala, the IT hubs in Bangalore. Here, traffic movement is high compared to other areas. SGHalli was chosen because of its proximity to a highly industrialized area where pollution levels are high.

Pollution data for SGHalli Station was extracted over the time period 18.11.2015 to 01.03.2017. Since continuous data is required for ARIMA modeling, data for the time frame, 01.08.2016 to 28.02.2017, was taken because it had few outliers and missing data. As such, it was continuous throughout the time frame. The same BTM layout can be used where data was extracted from 30.03.2015 to 01.03.2017, but the time frame of 07.04.2015 to 08.05.2016 was considered.

3.2 Data cleaning

Several methods for data cleaning are available, it involves identifying missing data, outliers, extreme values and accordingly making a decision on whether imputation is necessary [26]. When dealing with air pollution data set, caution needs to be exercised on removing extreme values. The importance and usefulness of extreme values for analysis have been highlighted and researched [27-29]. Thus, while removing of extreme values, it would be better to understand the correlation structure with other pollutants and accordingly make decisions on removal. In case of missing values, the time series data as a whole needs to be examined based on the number of missing data points and the averaging method used for imputation.

3.3 Autocorrelation and partial autocorrelation function with plots

The first step in analysis is to examine the autocorrelation plots. The autocorrelation function is given by the following equation [30]:

$$\rho_k = \frac{Covariance(x_t, x_{t-k})}{\sigma_{x_t}\sigma_{x_{t-k}}} \tag{8}$$

$$Covariance(x_t, x_{t-k}) = E[(x_t - \mu)(x_{t-k} - \mu)]$$

The autocorrelation estimate r_k is given as:

$$r_k = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{i=1}^{n}((x_t - \bar{x})^2)} \tag{9}$$

The plot of autocorrelation values for different lags is called the autocorrelation function (ACF). In time series analysis, the ACF plots find their usefulness in identifying whether a model can be categorized as AR, MA, ARMA or ARIMA. Various models can be suggested using ACF plots and previous authors have provided guidance [7, 24]. The ACF is more useful for order identification of MA models. Its structure can be identified based on a cut off after a certain lag k (ACF value cuts after a certain lag), In AR models, the plots can show an exponential decay, or sinusoidal damping. In such cases, identifying the order for the model becomes difficult and therefore partial autocorrelation functions and plots are used. The partial autocorrelation function for a particular lag k is defined as:

$$PACF = Correlation(x_t - \hat{x_t}, x_{t-k} - \widehat{x_{t-k}}) \tag{10}$$

Based on the PACF values, the order of an AR process can be determined. Typically, the way to identify a suitable model is to first plot the ACF and PACF and then identify the pattern. For example, if the ACF values cuts off after lag k, then a MA(k) process can be chosen, provided that PACF plots have an exponential decay, sinusoid dampening or a combination of the two. Suppose ACF shows an exponential decay, sinusoid dampening or a combination of both, and the PACF cuts off at lag k, then it is an AR(k) process. In case PACF shows exponential decay or sinusoidal damping or a combination of both then ARMA model needs to be looked. When the data is of non-stationary behavior, differencing data should be obtained and the ACF and PACF of differencing data needs to be examined to choose the p and q parameters.

3.4 Determination of suitable models

One of the complexities of a time series modeling is determination of the weights for the various models. Furthermore, the weight estimate can be obtained using one of several approaches [7, 24]. Based on ACF and PACF plots, various types of models can be employed. If a moving average model is selected based on the plots, then the choice of q becomes important, which is determined based on the cut off value of ACF at a particular lag. However there would be variety of models which can be looked when there is ambiguity in the ACF and PACF plot, in such cases based on plots the different combination of p,q can be chosen and the same is applicable for a non stationary time series wherein ACF and PACF plot for differencing is looked into and when ambiguity in these plots arises various combinations of ARIMA model are obtained. The weights of the model are listed based on the ACF and PACF plots. They are typically calculated considering the type of model.

For a moving average model of order q, the autocorrelation is function is expressed in terms of the weights as shown below [24]:

$$\rho_x(k) = \begin{cases} \frac{-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + .. \theta_q^2} & k = 1, \ldots q \\ 0 & k > q \end{cases} \tag{11}$$

Since an estimate for an autocorrelation function can be obtained using Equation 9, there would be q nonlinear equations with q unknowns to be solved.

For an autoregressive process of order p, the relation between autocorrelation and weights are given by the Yule – Walker Equation [3, 31] shown below [24]:

$$\rho(k) = \sum_{i=1}^{p} \phi_i \rho(k-i) \quad k = 1,2, \ldots \ldots \tag{12}$$

Using Equation 9, an estimate of autocorrelation can be obtained. For each autocorrelation value, there will be one set of linear combination of weights, likewise for a lag of p, there would be a p set of equations with p unknown weights that can be solved to obtain the weights. Once the weights are obtained for the various models, significance tests can be performed for the coefficients to determine the suitability of the model. Typically, a p-value less than .05 indicates that the coefficient is significant. Therefore, for various models, the p-values are accordingly checked and the models chosen. The Box-Pierce test is also conducted to check whether white noise is uncorrelated and independent. A larger p-value indicates that the there is no autocorrelation among the residues or white noise. After a set of suitable models is obtained, Minitab and R software are used to do various plots, tests and building the model

3.5 Identification of the best model

Once the models are identified the next step is to find the best suitable model, there are different criteria which can be used to determine the best model. One such criterion is the Akaike Information Criterion (AIC) developed by [32]. The AIC is given as:

$$AIC = ln\left(\frac{\sum_{i=1}^{N} e_i^2}{N}\right) + \frac{2p}{N} \tag{13}$$

where N is the number of periods or number of data points, p is an independently attributed parameter. Various other criteria are also available and can be used for evaluation [24].

Table 1 Basic statistics of SGHalli, BTM air pollutants before imputation

	Location	Data Before Imputation	Missing data	Extreme Values	Mean	Stdev	Median	Q1 (25th Percentile)	Q3 (75th Percentile)	IQR (Interquartile Range)
PM_{10}(µg/m^3)	SGHalli	185	17	0	41.3	15.91	40.34	30.09	49.97	19.88
SO_2 (µg/m^3)	SGHalli	185	17	0	4.83	7.34	3.58	2.76	4.61	1.85
NO_x (µg/m^3)	SGHalli	185	17	0	26.68	13.047	28.58	14.58	37.42	22.84
$PM_{2.5}$(µg/m^3)	BTM	396	2	31	224.4	8.72	52.2	17.2	76	52.8
SO_2 (µg/m^3)	BTM	397	1	5	3.61	15.39	1.77	1.61	2.04	.43
NO_x (µg/m^3)	BTM	397	1	0	9.6	8.71	7.89	4.5	11.03	6.53

Table 2 Mean values for imputation after removal of extreme values

Location	Pollutants	Imputed	Mean (µg/m^3)	Data after Imputation	Data imputed due to error (Extreme values)	Missing Data Imputed
SGHalli	PM_{10}	Missing Data	41.3	202	0	17
SGHalli	SO_2	Extreme Values & Missing Data	3.81	202	0	17
SGHalli	NOx	Missing Data	26.66	202	0	17
BTM	$PM_{2.5}$	Extreme Values & Missing Data	45.77	398	31	2
BTM	SO_2	Extreme Values & Missing Data	2.19	398	5	1
BTM	NOx	Missing Data	9.6	398	0	1

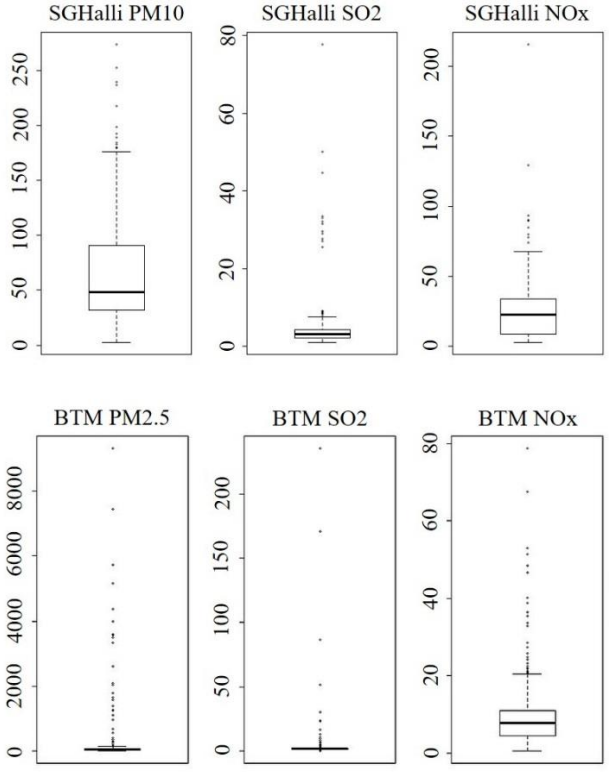

Figure 2 Box Plot for SGHalli & BTM pollutants with Y axis units of µg/m^3

The greater the number of parameters included in the model, the higher the penalty incurred and a larger value of AIC is incurred, thus a smaller AIC is preferred among set of models to be identified. Therefore, based on p-values and AIC values, the best model is identified.

3.6 Model adequacy

Model adequacy can be validated by understanding the behavior of the residues or white noise. The basic assumption for building a time series lies in understanding the white noise or residues. White noise should be independent and uncorrelated. Thus, to identify the independence of the residue, the behavior of residue over time is examined to determine if there is pattern. ACF plots can be used to understand correlation. If the values are small then it can be said that residues are uncorrelated. Also, the performance of the model is also quite important in terms of root mean square and mean absolute error. The smaller these values, the better the performance. These equations are given below.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(e_t)^2}{n}} \text{ where } e_t = x_t - \hat{x_t} \quad (14)$$

$$MAE = \frac{\sum_{t=1}^{n}|e_t|}{n} \quad (15)$$

4. Results and discussion

4.1 Data cleaning

For SGHalli Station, the dataset for PM_{10}, NO_x, SO_2 was considered. It is also important to note that particulate matter (PM_{10} or $PM_{2.5}$) is formed due to complex interaction of various pollutants such as NO_x, SO_2, CO, hydrocarbons and environmental factors. To conduct time series analysis, missing data, outliers and any other abnormality must be identified. Based on the dataset considered, it was found that 17 data points were missing for PM_{10} out of a dataset containing 212 points, i.e., the percentage of missing points was less than 10%. Typically in such cases, if instances of missing data are sufficiently low, then any kind of approach for imputation can be applied [23]. Several other researchers have done study on imputation of missing data and found that for low percentage of missing data most of the imputation methods work well [26, 33-34]. In the current study, the mean value of the data is substituted for the missing data points. The mean value and other statistics are shown in Table 1.

The Box plot of the dataset before imputation is shown in Figure 2. It can be seen there are few extreme values and they were not large. The outliers were not removed since they provide useful information and could be part of normal processes. The missing values were imputed using the mean value. Imputed values are shown in Table 2. The time series data after imputation of the mean value is shown in Figure 3. Time series data provides useful information on the behavior of the dataset. It can be seen that the mean of the data set for SGhalli Station PM_{10} seems to be stable. However somewhere around 90 days, there seem to be changes in the variance and the data seems to be autocorrelated (as noticed by downward and upward values). This needs to be verified by considering an autocorrelation plot.

In the Box plot of the dataset for SGHalli Station, NO_x before imputation is shown in Figure 2. Based on this Box plot, there are no extreme values or outliers. From the time series data in Figure 3, a change in mean around 50 days with an upward trend and there seems upward and downward trend as evidenced in the plot (data from 150 days onwards). This shows that data may be non-stationary, and hence a more detailed study of the ACF, PACF and differencing plot is essential.

The Box plot for SGHalli Station shows extreme values for SO_2. Therefore, the relationships between the extreme values and other variables were examined. It was found that extreme values did not exhibit correlation with other variables, so these values were removed and the mean was calculated (Table 2) and accordingly imputed using the mean value approach. The time series data is plotted after imputation in Figure 3. For the SGHalli Station SO_2 level, the mean seems to be stable except at the beginning stages in the dataset. This could be indicative of a process that is weakly stationary. However to further examine this assertion, ACF, PACF plots are needed.

There are numerous large extreme values in the BTM station $PM_{2.5}$ data in Figure 2 Box plots. Typically such very large extreme values do not occur. However, since they are many and sufficiently larger values and large amount, a scatter plot of the extreme values of $PM_{2.5}$ with other variables were looked into, to identify whether other pollutants are the cause for such extreme value and the scatter plot is shown in Figure 4. The relation between extreme values of the variables can be judged based on the correlation value or by using the red lines in this figure. The correlation values indicate that there is no relationship between the extreme values for any of the variables. The red lines in the graphs indicate the fit of data between two variables. Regression lines (red line) show the fit between $PM_{2.5}$ and SO_2, $PM_{2.5}$ and NO_X, $PM_{2.5}$ and CO. It can be seen that this line is almost parallel to the y-axis for $PM_{2.5}$ vs. SO_2 and parallel to x-axis for $PM_{2.5}$ vs. NO_X. In case of $PM_{2.5}$ vs. CO, it can be seen that most of the values are not correlated, as indicated by the straight line parallel to the x-axis (for CO values between 0 to 2). Thus, these values may have been caused due to instrumentation or human error, hence a decision was made to remove them and accordingly impute with the recalculated mean values shown in Table 2. Likewise, it is noted after imputation that there was a sharp decrease in mean value. This was mainly from large extreme values due to instrumentation or human error. After imputation, the time series dataset is plotted as shown in Figure 3. It can be seen from the plot that there is a change in the mean value and also there seems to be a trend in mean indicating that the dataset is non-stationary. Further details on modeling need to be obtained by understanding ACF, PACF and difference data.

A Box plot of BTM SO_2 shows extreme values. These values were removed and accordingly imputed based on the recalculated mean as shown in Table 2. The time series plot after imputation shows that the mean seems to be stable, indicating a weakly stationary process.

The Box plot for NO_x (Figure 2) shows extreme values and outliers, but these values are not large enough to be removed. Hence they were retained. The dataset contains one missing value that was imputed using recalculated mean values as shown in Table 2.

The NO_x time series data shows that at the beginning stages there seems to be trend, but after a certain period the mean seems to be stable, whereas there seems to be variance changes occurring after a certain period of time, an analysis of ACF and PACF plot would provide a clear picture on the behavior.

4.2 Model building using ACF, PACF plots for actual and differencing data

4.2.1 Model building for SGHalli Station

The autocorrelation and partial autocorrelation plots for PM_{10}, NO_X and SO_2 of SGHalli station are shown in Figure 5. The autocorrelation plot of PM_{10} indicates a gradual dip around the fourth lag and there seems to be a sinusoidal pattern of decay. This indicates that the model can be either an AR or ARMA type. For confirmation, the PACF plots were examined. The PACF cuts at lag 1, indicating an AR(1) model. However, there seems to be sinusoidal pattern and since the cut off is around lag 4, so it could also be ARMA (1,1), ARMA(1,2), ARMA(1,3) and ARMA(1,4) models. Table 3 gives these possibilities.

For NO_x, there is long exponential decay in the ACF plot, indicating a non-stationary time series. This can be verified by observation of the time series plot in Figure 3 (trend in data). Additionally, the PACF plot indicates large values at various lags, thereby creating a need to build an ARIMA model with various p-values. In order to model a non-stationary series, a differencing method needs to be used. Based on the stationarity of the order of differences, an ARIMA model can be created. The first order differencing plot was calculated and the ACF and PACF differencing

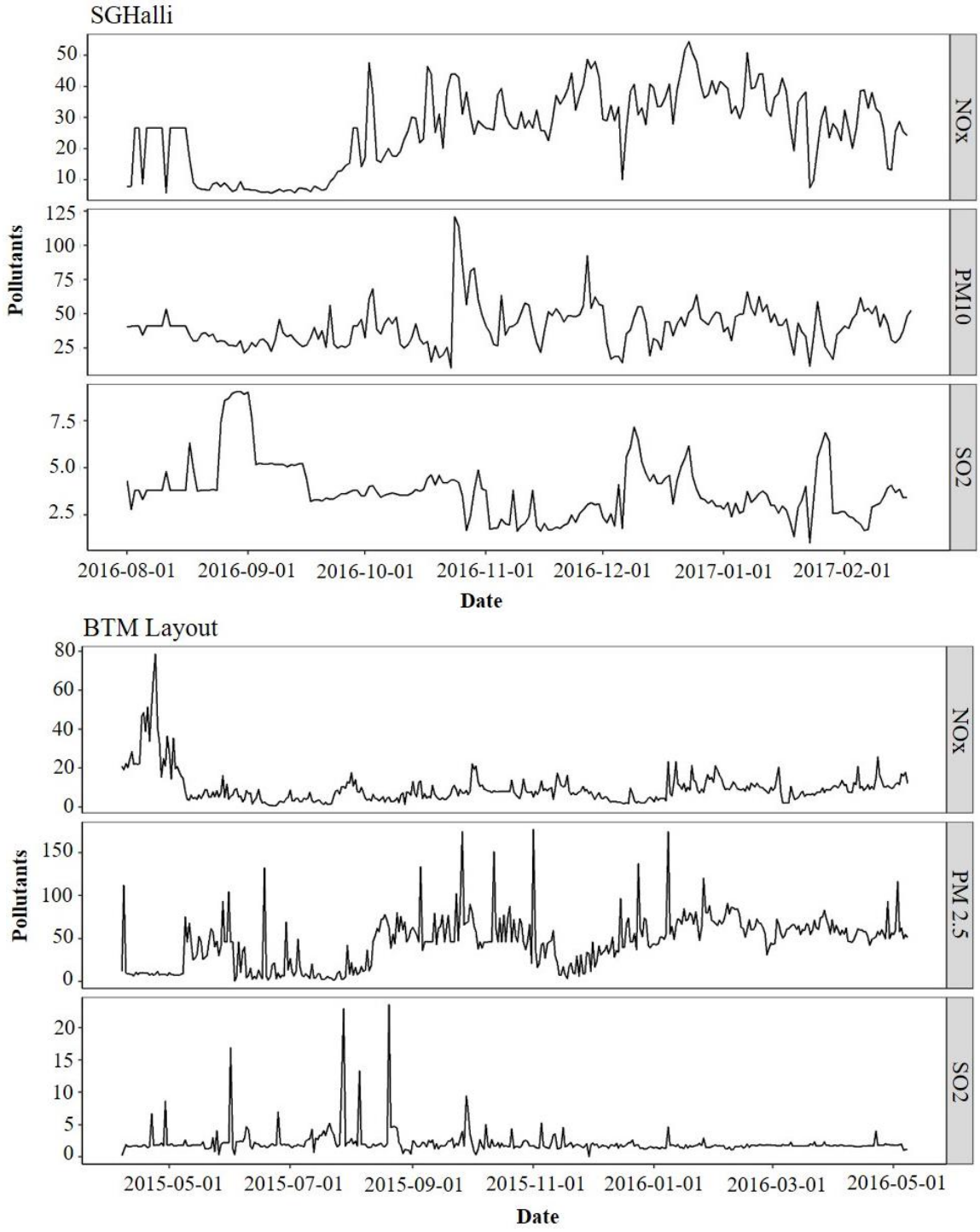

Figure 3 Time series plot (SGHalli and BTM) for NO_x, PM_{10}, SO_2, y-axis units in µg/m^3

Table 3 Time series model types for various pollutants

Pollutants	Time Series Model types
PM10 (SGHalli)	AR(1), ARMA(1,1), ARMA(1,2), ARMA(1,3), ARMA(1,4)
NOx (SGhalli)	ARIMA(1,1,1), ARIMA(1,1,2), ARIMA(1,1,3), ARIMA(2,1,1), ARIMA(2,1,2), ARIMA(2,1,3)
SO_2 (SGhalli)	ARIMA(1,1,1), ARIMA(1,1,2)
$PM_{2.5}$ (BTM)	ARIMA(0,1,1)
NOx (BTM)	ARIMA(0,1,1), ARIMA(0,1,2), ARIMA(0,1,3), ARIMA(0,1,4)
SO_2 (BTM)	AR(1), AR(2)

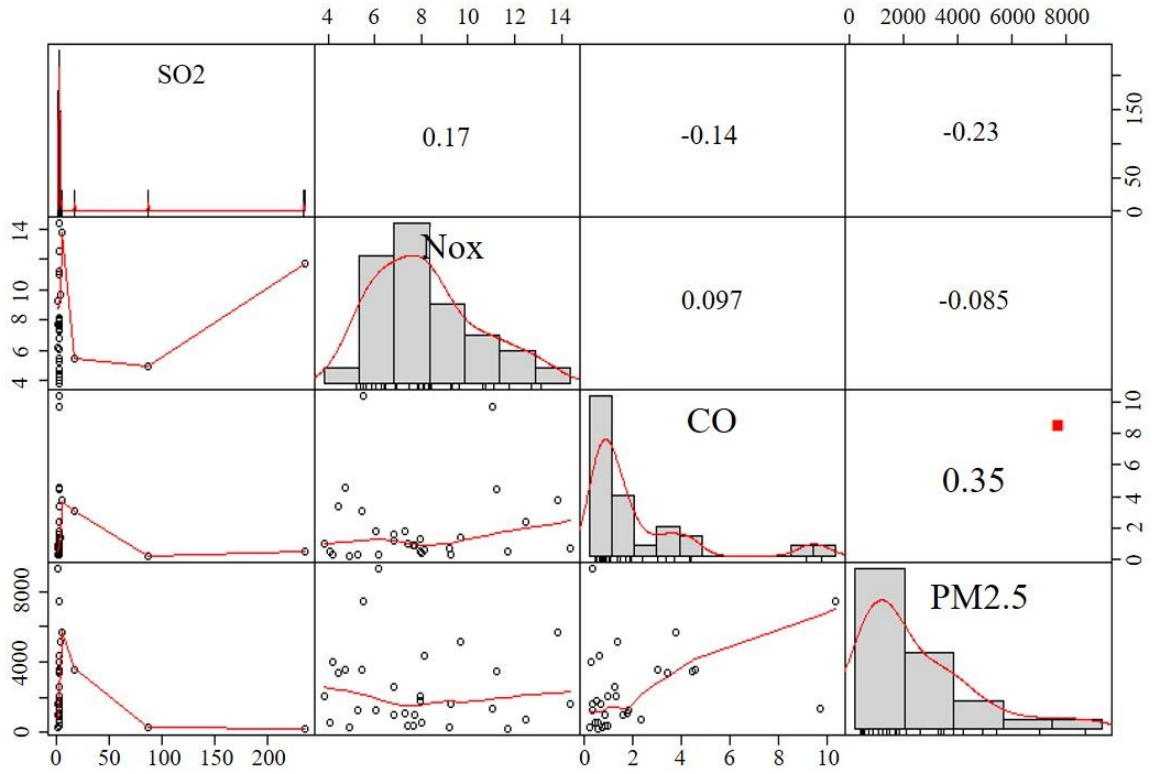

Figure 4 Correlation and scatter plots of extreme values of pollutants (x, y-axis units in µg/m^3)

plots for NO_x are shown in Figure 6. It is seen from the plot that the autocorrelation is low and PACF values are well within limits, indicating that first order differencing makes the dataset stationary, and hence suitable for creating an ARIMA model. There are various possible models that can be built using ARIMA. Table 3 gives these possibilities.

For SO_2, the ACF plot indicates a non-stationary model. This is evidenced by observing the time series plot in Figure 3, which is indicative of mean changes. The differences are taken between consecutive data. Accordingly, the ACF and PACF are examined. Based on the plots in Figure 6, it can be seen that first order differencing is stationary, indicating an ARIMA model with a first order difference. Based on the order difference and the PACF plot suitable ARIMA models can be built. Table 3 presents the various combination for a difference of order 1.

4.2.2 Model building for BTM Station

It is seen from the plots of ACF (long decay) and PACF (long decay) from Figure 5 for $PM_{2.5}$, that the dataset is non-stationary. Thus, differencing of data was conducted and the ACF and PACF plots for differencing were obtained as shown in Figure 6. Based on these plots, ACF value cuts off after lag 1 can be seen and the PACF has an exponential decay thereby leading to the conclusion that a differencing model with MA(1) is needed. In other words, ARIMA (0,1,1) would be a suitable model.

For NO_x, the ACF and PACF plots indicate that the dataset is non-stationary as seen by the long ACF decay process. The time series plot is indicative of a non-stationary process. Based on the ACF and PACF, it can be also be interpreted as a MA model. In such cases, it is worthwhile to look at differencing plots. The ACF and PACF plots for a first order difference are drawn. It seen from the plot in Figure 5 that the ACF cuts after lag 1 and the PACF seems to have to low values. However, there seems to be an exponential decay indicating that an MA model for differencing should be used. It is seen from the PACF plot that various ARIMA models with different p and q values can be used. The various models that can be used are shown in Table 3.

The ACF and PACF for SO_2 indicates that there is an exponential decay for ACF and the PACF values are small which cut after lag 1. This indicates either an AR(1) or an AR(2) model.

4.2.3 Determining suitable models using a t-test for parameters

Table 4 provides a list of models with significant p-values for the various parameters used in model building. For SGHalli Station PM_{10}, the p-values for model AR(1) were significant. For the other models from ARMA(1,1) to ARMA (1, 4), the p-values were not significant. Thus, AR(1) was a suitable model. It can be seen from Table 4 that AIC, BIC values of AR(1) are smaller than for ARMA(1, 4), thus confirming suitability.

For SGHalli Station NO_x, p-values for the coefficients of the ARIMA(1,1,2), ARIMA(1,1,3), ARIMA(2,1,2) and ARIMA(2,1,3) models are large and thus not significant. The ARIMA(1,1,1) and ARIMA(2,1,1) models have p-values less than .05, so in the final model, the AIC, BIC criteria are used. It is seen from Table 5 that the AIC, BIC values for ARIMA (2,1,1) are smaller than for ARIMA(1,1,1). At the same time, the p-values for the lags using a modified Box-Pierce test are large. In a Box-Pierce test, larger p-values indicate that residual autocorrelation is minimal, indicative of white noise [24] in comparison to ARIMA(1,1,1). Based on this, ARIMA (2,1,1) was chosen as the model.

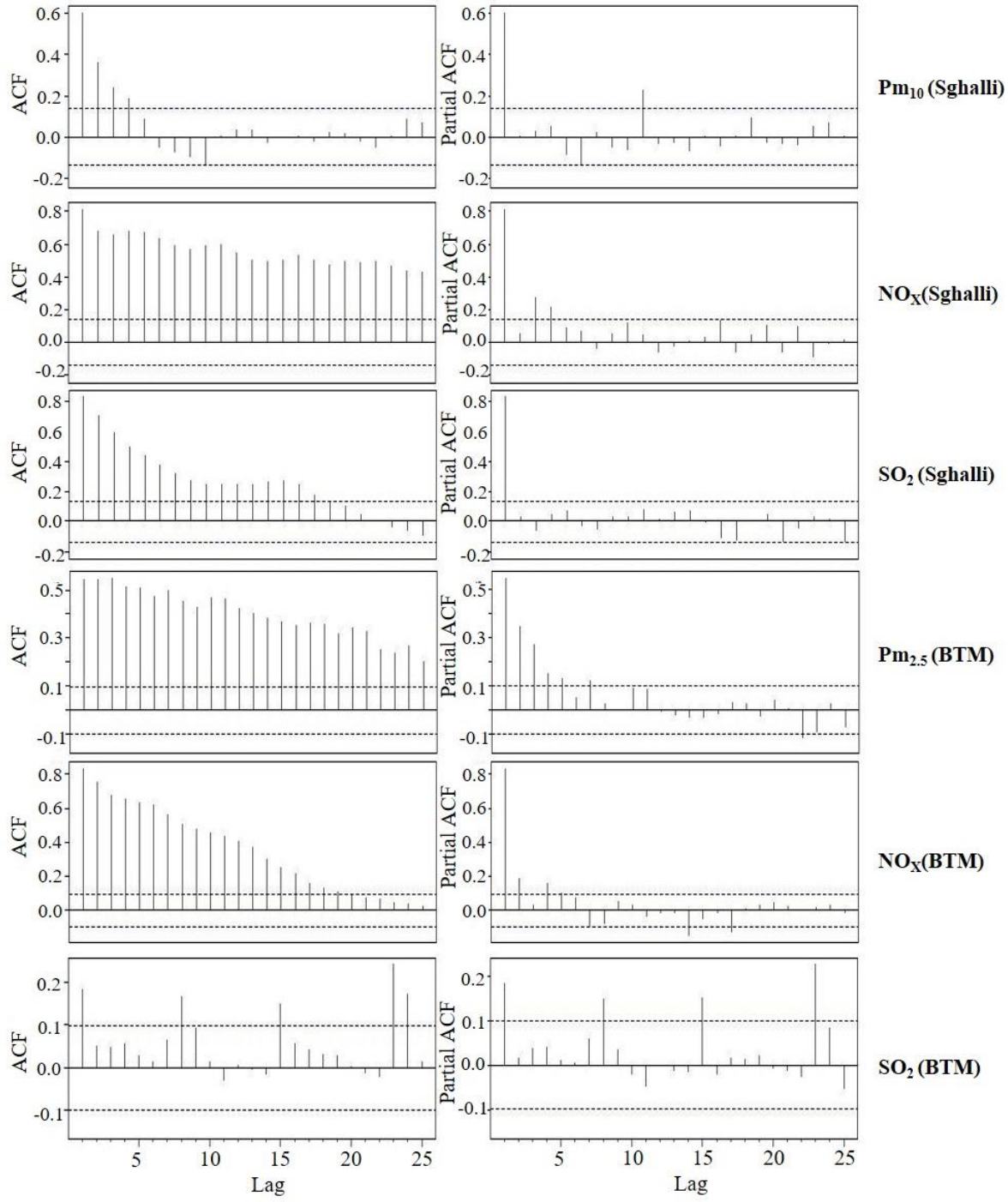

Figure 5 ACF, PACF plots for SGHalli (PM_{10}, NO_x, SO_2), BTM pollutants ($PM_{2.5}$, NO_x, SO_2)

Table 4 Coefficients values and p-values for various significant models

Pollutants	Model types	Coefficients							p-value					
		Const	AR1	AR2	MA1	MA2	MA3	MA4	AR1	AR2	MA1	MA2	MA3	MA4
PM_{10} SGHalli	AR(1)	16.29	.602	--	--	--	--	--	0.0	--	--	--	--	--
NO_x SGHalli	ARIMA(1,1,1)	0.025	.478	--	.866	--	--	--	0.0	--	0.0	--	--	--
	ARIMA(2,1,1)	.036	.386	-.24	-.70	--	--	--	0.0	.03	0.0	--	--	--
SO_2 SGHalli	ARIMA(1,1,1)	$8*10^{-4}$	.821	--	1.00	--	--	--	0.0	--	0.0	--	--	--
$PM_{2.5}$ BTM	ARIMA(0,1,1)	.086	--	--	.803	--	--	--	--	--	0.0	--	--	--
NO_x BTM	ARIMA(0,1,1)	-.017	--	--	.371	--	--	--	--	--	0.0	--	--	--
	ARIMA(0,1,2)	-.017	--	--	.363	.134	--	--	--	--	0.0	0.0	--	--
SO_2 BTM	AR(1)	1.79	.183	--	--	--	--	--	0.0	--	--	--	--	--

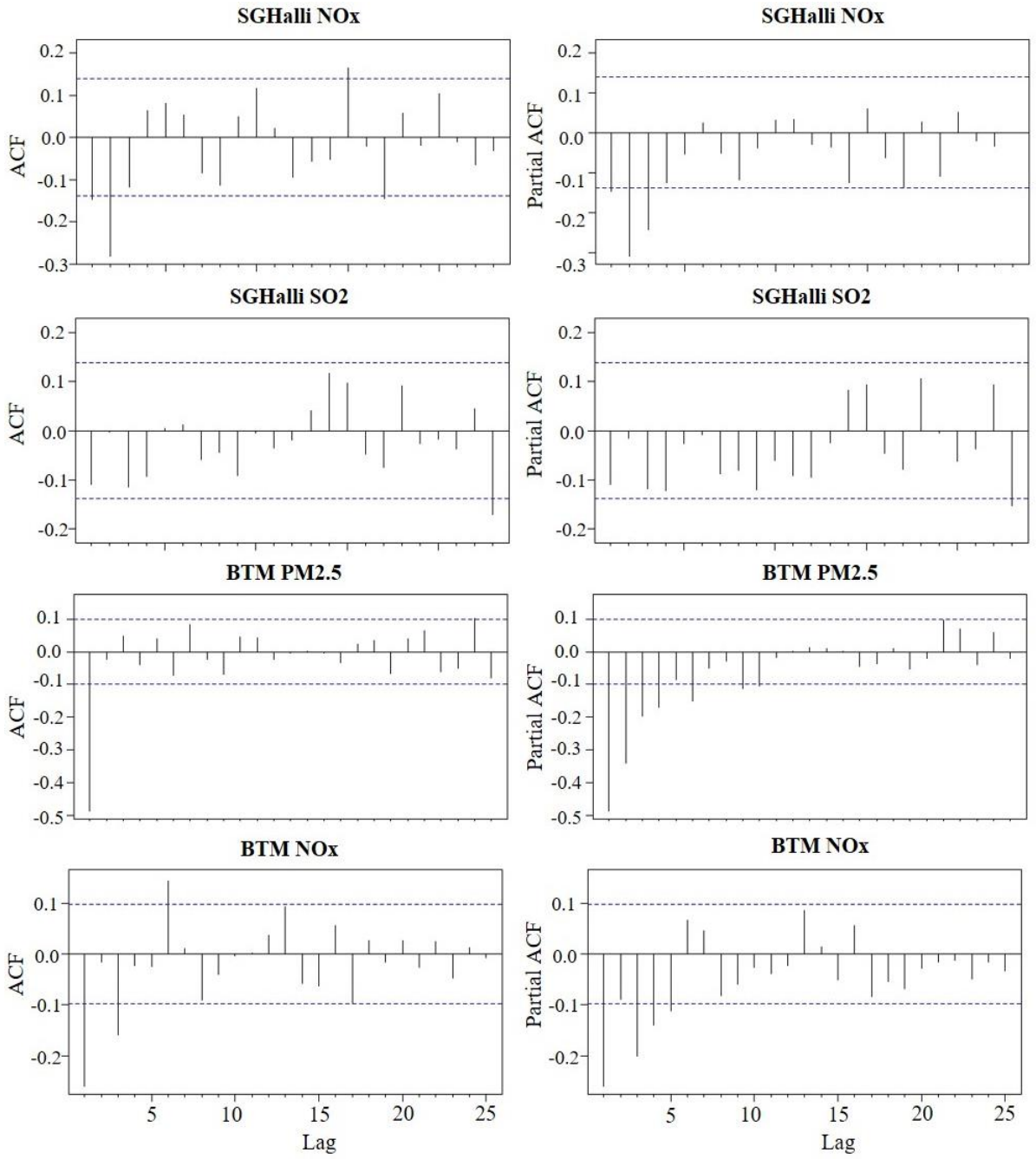

Figure 6 ACF, PACF plots for differences of SGHalli (NO_x, SO_2) and BTM pollutants ($PM_{2.5}$, NO_x)

Table 5 AIC, BIC and Box-Pierce p-value statistics for various model types

Pollutants	Model types	Box-Pierce Test p-Value					
		AIC	BIC	Lag12	Lag24	Lag36	Lag48
PM_{10} SGHalli	AR(1)	1595	1604	0.080	0.265	0.009	0.011
NO_x SGHalli	ARIMA(1,1,1)	1356	1366	0.035	0.081	0.096	0.108
	ARIMA(2,1,1)	1351	1364	0.648	0.574	0.667	0.726
SO_2 SGHalli	ARIMA(1,1,1)	507	517	0.925	0.466	0.875	0.822
$PM_{2.5}$ BTM	ARIMA(0,1,1)	3587	3594	0.677	0.778	0.820	0.828
NO_x BTM	ARIMA(0,1,1)	2370	2378	0.000	0.000	0.009	0.051
	ARIMA(0,1,2)	2364	2376	0.003	0.012	0.130	0.337
SO_2 BTM	AR(1)	1700	1712	0.227	0.000	0.001	0.002

Table 6 Best identified models for various pollutants and ADF test for stationarity

Pollutants	Model	ADF value	Significant
PM_{10} SGHalli	AR(1)	-7.16	Yes
NO_x SGHalli	ARIMA(2,1,1)	-4.99	Yes
SO_2 SGHalli	ARIMA(1,1,1)	-4.24	Yes
$PM_{2.5}$ BTM	ARIMA(0,1,1)	--8.3	Yes
NO_x BTM	ARIMA(0,1,2)	-4.75	Yes
SO_2 BTM	AR(1)	-16.89	Yes

For SGHalli Station SO_2, it is seen that the for ARIMA(1,1,1), p-values are significant for AR1, MA1 coefficients and for ARIMA(1,1,2), the p-value is not significant for MA2 coefficients (p-value =. 64), indicating ARIMA(1,1,1) as a suitable model.

For BTM Station $PM_{2.5}$, only one model was found suitable. The p-value shows significance for MA1, thus this model is suitable for BTM Station $PM_{2.5}$.

For BTM Station NO_x it is seen that the ARIMA(0,1,3) model has large p-values for MA2 (p value = .0167), ARIMA(0,1,4) has large p values for MA2 (p-value = .2) and MA4 (p-value = .67) coefficients, indicating that these two models are not suitable. The p-values for ARIMA(0,1,1) and ARIMA(0,1,2) models have significant p-values, as shown in Table 4. Since there is a tie, AIC values are used to make a decision, and the ARIMA(0,1,2) model has a lower value. Additionally, the Box-Pierce test statistics show a large p-value for ARIMA (0,1,2), indicating that errors are white noise. Thus, based on this, ARIMA(0,1,2) is chosen.

For BTM Station SO_2, AR(1) has significant p-values for AR1 coefficients and AR(2) has a large p-value for AR2 coefficients (p-value = .743). Therefore, the AR1 model is chosen. The final model selected is listed in Table 6. Additionally, the ADF (Augmented Dickey Fuller) test was conducted to check for stationarity of the selected models. It is seen from Table 6 that the ADF values have large negative values, suggesting rejection of the null hypothesis (non-stationary). Hence, it can be concluded that the models are stationary.

4.3 Model adequacy

The typical way of checking model adequacy is to examine whether the residuals are white noise. This can be done by checking the autocorrelation structure of the residues using the Box-Pierce test. Changes in variances can be identified from the plot by observing the general behavior of residues using histograms, QQ plots, and residue vs. fit data. Table 5 provides the p-values for the Box-Pierce test conducted for the various models. It is seen that for most of models selected, the p-values are larger than .05, indicating that residues are white noise, which is a primary requirement for time series modeling.

For SGHalli PM_{10}, the p-value for the Box-Pierce test for AR(1) shows large values indicating that the residuals are white noise and not correlated. This can be cross-checked from the ACF and PACF plot shown in Figure 7. It is seen from Figure 7 that the histogram is symmetrical and there is random order as seen in plot of residue vs order.

For SGHalli NO_x ARIMA (2,1,1), the p-value for the Box-Pierce test shows very large values indicating that the residuals are white noise and not correlated. The QQ plot shows normality of the data (however, normality is not a criteria for white noise) as evidenced by the histogram. Also there is randomness in data shown by residue vs. order plot.

For SGHalli SO_2 ARIMA (1,1,1), the Box-Pierce test shows large-p values indicating no correlation among residues, which is also evidenced by the residue vs. order plot, ACF, PACF plot, thus indicating a sound model.

For BTM $PM_{2.5}$ ARIMA(0,1,1), the Box-Pierce test, based on Table 5, shows large p-values as well as the ACF and PACF shown in Figure 7, with no correlation, thereby indicating a white noise structure. The QQ plot shows skewness towards the right and is shown by the histogram. The residue vs. order structure indicates that the dataset is random.

For BTM NO_x ARIMA(0,1,3), the Box-Pierce test shows large p-values. Furthermore, the ACF and PACF values of residues show that there is no autocorrelation, indicating that the residues are white noise. The residue vs. order plot shows that the data is random. Therefore the model is sound.

For BTM SO_2 AR(1), the Box-Pierce test result is small, indicating there could be an autocorrelation structure. However, the ACF plot for residue indicates that there is no autocorrelation.

It is essential to check how much of a difference exists between the actual and the fitted values. There are measures, such as root mean square error, that can indicate the performance of the model. The lower this value, the better the performance. Additionally, a plot of actual and fitted values would indicate how well the model can perform. Figure 8 shows the actual and predicted values for SGHalli and BTM pollutants. It is seen from the plot, the models obtained are able to depict similar behavior and there seems to be little variation between the actual and fitted values. However, this needs to be verified statistically. Equations 14 and 15 provide the RMSE and MAE, where smaller values indicate better fit.

The mathematical formulations for the selected models are given below, in equation form, for SGHalli PM_{10}, NO_x and SO_2 and BTM $PM_{2.5}$, NO_x and SO_2, in Equations 16 through 21.

$$x_t = 16.29 + .602x_{t-1} + \varepsilon_t \quad (16)$$

$$w_t = .036 + .386w_{t-1} - .24w_{t-2} + \varepsilon_t - .752\varepsilon_{t-1} + .06\varepsilon_{t-2} \quad (17)$$

$$w_t = .0008 + .821w_{t-1} + \varepsilon_t - \varepsilon_{t-1} \quad (18)$$

$$w_t = -.017 + \varepsilon_t - .371\varepsilon_{t-1} \quad (19)$$

$$w_t = -.022 + \varepsilon_t - .342\varepsilon_{t-1} + .073\varepsilon_{t-2} - .155\varepsilon_{t-3} \quad (20)$$

$$x_t = 1.74 + .183x_{t-1} + \varepsilon_t \quad (21)$$

4.4 Discussion

Since there is no complex time series model built for Bangalore city and Sharma et al. [20] has presented a detailed study on the performance based on Delhi pollution data, it

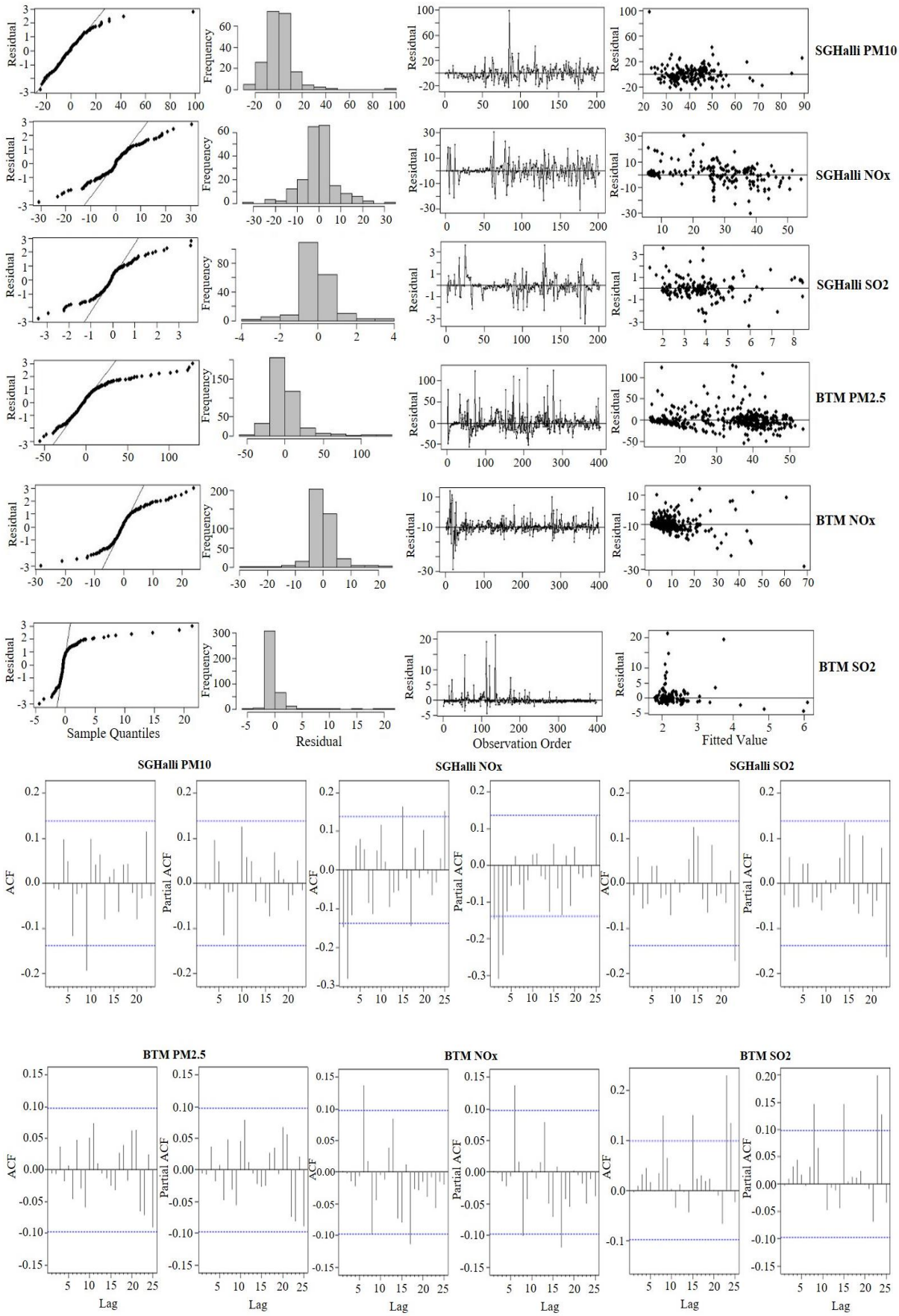

Figure 7 QQ plot, residue vs. fit and order, histogram, ACF&PACF for residues of SGhalli and BTM pollutants (units of residue and fitted values are $\mu g/m^3$)

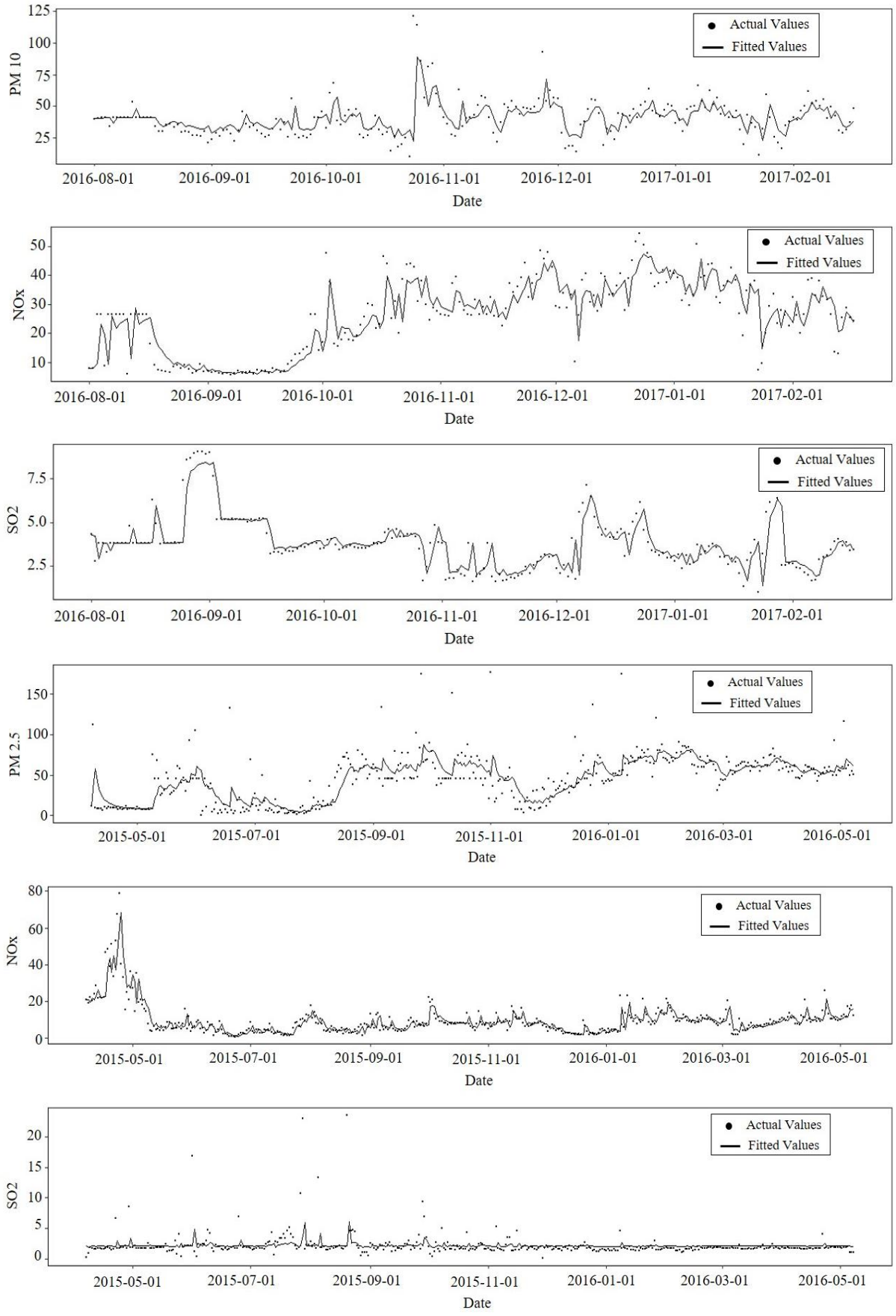

Figure 8 Actual and predicted values for SGHalli and BTM pollutants, y-axis units in µg/m^3

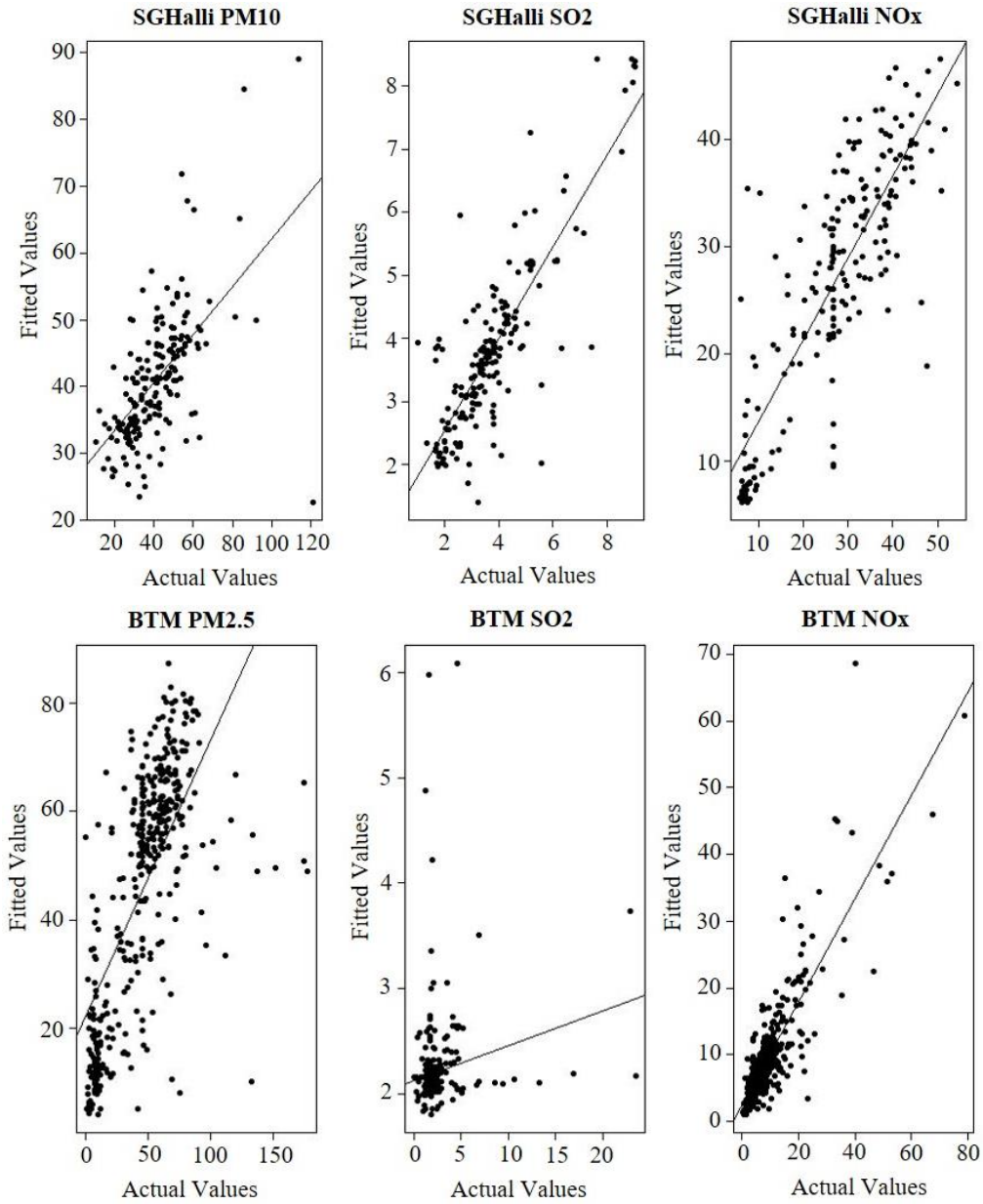

Figure 9 Scatter plot of fitted value vs. actual values of various pollutants (SGhalli, BTM) (µg/m^3)

would be useful to understand the differences and provide a comparison based on performance.

There are various measures to find the performance. Sharma et al. [20] used RMSE based on fitting a regression model for the predicted and observed values. The formula obtained by Sharma et al. [20] has been based on Willmott et al. [35] and are shown in Equations 23 and 24. In similar lines for comparison with Sharma et al. [20] and understanding the prediction power, the formulas are applied to the final applied model.

It is to be noted regression modeling is more suitable when the data set is stationary. However, that scenario would be more applicable when the regression is done on a set of independent variable, and in this case since the regression is based on the fitted value vs observed value, regression model can be applied. Additionally, a scatter plot for the fitted and observed values shows a linear relationship except for BTM SO_2. The formulas are shown in Equations 22, 23,24 and is applicable more often while conducting linear regression analysis where $RMSE_s$ is a systematic component and $RMSE_u$ an unsystematic component.

$$RMSE = \sqrt{RMSE_s^2 + RMSE_u^2} \quad (22)$$

$$RMSE_s = \left(\frac{\sum_{i=1}^{n} w_i|\widehat{p_i} - o_i|^2}{\sum_{i=1}^{n} w_i}\right)^2 = \left(\frac{\sum_{i=1}^{n}|\widehat{p_i} - o_i|^2}{n}\right)^2 \ (when\ w_i = 1) \quad (23)$$

$$RMSE_u = \left(\frac{\sum_{i=1}^{n} w_i|\widehat{p_i} - p_i|^2}{\sum_{i=1}^{n} w_i}\right)^2 = \left(\frac{\sum_{i=1}^{n}|\widehat{p_i} - p_i|^2}{n}\right)^2 \ (when\ w_i = 1) \quad (24)$$

where $\widehat{p_i}$ is the estimated predicted value, p_i is the predicted value obtained by modeling time series and $\widehat{p_i}$ is obtained by regressing fitted values p_i on observed values o_i. Thus, the regression equation is $\widehat{p_i} = \widehat{\beta_0} + \widehat{\beta_1} o_i$. Based on the regression model of predicted vs. observed values, R^2 values are obtained. The $RMSE_s$, $RMSE_u$ and lower proportionality

Table 7 RMSE, $RMSE_s$, $RMSE_u$ and $\left(\frac{RMSE_s}{RMSE}\right)^2$ values for predicted vs. observed values in a regression model

Station	Model	R^2	$RMSE_s$	$RMSE_u$	RMSE	$\left(\frac{RMSE_s}{RMSE}\right)^2$	$\left(\frac{RMSE_u}{RMSE}\right)^2$	Authors
PM_{10} SGHalli	AR(1)	.501	9.89	7.32	12.34	64.21%	35.79%	Current work
NO_x SGHalli	ARIMA(2,1,1)	.71	2.98	6.25	6.92	90.31%	9.69%	
SO_2 SGHalli	ARIMA(1,1,1)	.703	.178	.532	.5612	10.0%	90.00%	
$PM_{2.5}$ BTM	ARIMA(0,1,1)	.46	14.58	16.46	21.99	43.97%	56.03%	
NO_x BTM	ARIMA(0,1,3)	.718	7.97	3.62	8.75	82.88	17.12%	
SO_2 BTM	AR(1)	.0303	4.00	.1376	2.03	98.8%	1.2%	
SPM Ashokvihar Delhi	ARIMA(1,0,1)	.9681	.436	.2841	.515	70.22%	29.78%	[20]
NO_2 Ashokvihar Delhi	ARIMA(0,1,1)	.9065	.3238	.1921	.376	73.96%	26.04%	
SO_2 Ashokvihar Delhi	ARIMA(1,1,1)	.9060	.7614	.1178	.770	96.04%	3.96%	
SPM Shahzada bagh Delhi	ARIMA(1,0,1)	.9437	.4668	.1476	.499	87.42%	12.58%	[20]
NO_2 Shahzada bagh Delhi	ARIMA(0,1,1)	.9001	.4029	.2781	.489	67.89%	32.11%	
SO_2 Shahzada bagh Delhi	ARIMA(1,1,1)	.9431	.289	.0952	.304	90.21%	9.79%	

of $\left(\frac{RMSE_u}{RMSE}\right)^2$, indicate a more precise model [35].

It is to be noted that Sharma et al. [20] have not used the proportionality of $\left(\frac{RMSE_u}{RMSE}\right)^2$ but this research has used because it has been mentioned by Willmott et al. [35] on the usefulness of the proportionality to understand the precision of the model wherein lower the ratio of $\left(\frac{RMSE_u}{RMSE}\right)^2$ the more precise the model. For this reason, a comparison is made based on the above formula with Sharma et al. [20] and details are provided in Table 7.

A regression model was built using an ordinary least squares method determining the values of RMSE based on Equations 20, 21 and 22. A scatter plot was made and examined for linear relationships. From Figure 9, it is seen that fitted and the actual values are linearly related except for those of SO_2. The models are build, the predicted values, coefficient of determination (R^2), residue values and the coefficients β_0, β_1 are calculated. Accordingly, $\widehat{p_i}$ is calculated based on $\widehat{p_i} = \widehat{\beta_0} + \widehat{\beta_1} o_i$, in which RMSE, $RMSE_s$, $RMSE_u$ are calculated. The values are shown in Table 7.

Table 7 gives the RMSE and the ratio values for the current work based on the regression model and a comparison is shown with other work. To compare in similar ways, two locations were chosen, one of which is an industrialized area (Shahzada bagh, Delhi) and the other a residential area (Ashok vihar, Delhi) from Sharma et al. [20]. As far as model precision is concerned, most of the models in the current and earlier work seems to be sound, and have low $\left(\frac{RMSE_u}{RMSE}\right)^2$ values except for SO_2 in SGHalli. Also, it is notable that most of the models built for Bangalore and Delhi pollutants are based on 1st differencing. Thus as far as time series modeling is concerned, there seems to be similar behavior of the pollutants.

An important question arises. Why is it that there is high value of $\left(\frac{RMSE_u}{RMSE}\right)^2$ for SO_2 at SGHalli Station? When the time series plot from Figure 3 is examined, there seems to be a complex data structure for SO_2. At a casual glance of the plot, it looks like the data set has mixture of changes such as trend at the beginning, change in mean, sharp change in mean and after certain period of time, a change in mean, change in variance and downward trends, indicating that time series model might not be suitable. In the case of time series data for other pollutants, such marked changes are not seen. Hence, better precision was obtained. This can be only be ascertained when more variety of time series data at different places and for various pollutants are obtained and analyzed. This research thus leads to questions of how to tackle such complex scenarios and what would be suitable models when such complexity arises, a wider area of exploration.

5. Conclusions

It is shown that time series modeling can be developed for modeling air pollutants in Bangalore city. Suitable models can be obtained by selecting a sound methodology. It is also seen that time series seems to perform better for stationary processes as far prediction is concerned compared to processes that are highly non-stationary. Therefore to understand the complexity of the process and the effectiveness of time series model, more amount of modeling work needs to be looked at various cities in India to understand the reason for superior performance and moderate performance. This can be looked as a future research objective for developing air pollution models in INDIA.

6. Acknowledgment

We would like to thank the anonymous reviewers for providing valuable inputs, thereby enhancing the quality of the manuscript.

7. References

[1] Beveridge WH. Weather and harvest cycles. Econ J. 1921;31:429-52.
[2] Beveridge WH. Wheat prices and rainfall in western Europe. J Roy Stat Soc.1922;85:412-59.
[3] Yule GU. On the time-correlation problem, with special reference to the variate-difference correlation method. J Roy Stat Soc. 1921;84:497-526.
[4] Slutsky E. The summation of random causes as the source of cyclic processes. Econometrica. 1927;5:105-46.

[5] Wold HO. A study in the analysis of stationary time series. Stockholm: Almqvist and Wiksell; 1938.
[6] Box GEP, Jenkins GM., Bacon DW. Models for forecasting seasonal and nonseasonal time series. In Wiley HB, editor. Advanced seminar on spectral analysis of time series. New York: Wiley; 1967. p. 271-311.
[7] Box GEP, Jenkins GM, Reinsel G. Time series analysis, forecasting and control. 4th ed. Englewood Cliffs: Prentice-Hall; 2008.
[8] Diebold FX, Kilian L, Nerlove M. Time series analysis. In: Durlauf SN, Blume LE, editors. Macro econometrics and time series analysis. London: Palgrave Macmillan; 2010. p. 317-42.
[9] Merz PH, Painter LJ, Ryason PR. Aerometric data analysis time series analysis and forecast and atmospheric smog diagram. Atmos Environ. 1972;6: 319-42.
[10] Tiao GC, Box GEP, Hamming WJ. Analysis of Los Angeles photochemical smog data: a statistical overview. J Air Pollut Contr Assoc. 1975;25(3):260-8.
[11] Chock PD, Terrell TR, Levitt SB. Time series analysis of riverside, California air quality data. Atmos Environ. 1975;9(11):978-89.
[12] Schwartz J, Marcus A. Mortality and air pollution in London: a time series analysis. Am J Epidem. 1990;131:85-194.
[13] Liu PWG, Johnson R. Forecasting Peak daily ozone levels—I. a regression with time series errors model having a principal component trigger to fit 1991 ozone levels. J Air Waste Manag Assoc. 2002;52(9):1064-74.
[14] Liu PWG, Johnson R. Forecasting peak daily ozone levels: part 2—a regression with time series errors model having a principal component trigger to forecast 1999 and 2002 ozone levels. J Air Waste Manag Assoc. 2003;53(12):1472-89.
[15] Liu PWG. Establishment of a Box-Jenkins multivariate time series model to simulate ground-level peak daily one-hour ozone concentrations at Ta-Liao in Taiwan. J Air Waste Manag Assoc. 2007;57:1078-90.
[16] Liu PWG. Simulation of the daily average PM10 concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis. Atmos Environ. 2009;43:2104-13.
[17] Liu PWG, Tsai JH, Lai HC, Tsai DM, Li LW. Establishing multiple regression models for ozone sensitivity analysis to temperature variation in Taiwan. Atmos Environ. 2013;79:225-35.
[18] Gocheva-Ilieva SG, Ivanov AV, Voynikova DS, Boyadzhiev DT. Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. Stoch Environ Res Risk Assess. 2014;28:1045-60.
[19] Dhurafi NAA, Masseran N, Zamzuri ZH. Compositional time series analysis for air pollution index data. Stoch Environ Res Risk Assess. 2018;32:2903-11.
[20] Sharma P, Chandra A, Kaushik SC. Forecasts using Box Jenkins models for the ambient air quality data of Delhi City. Environ Monit Assess. 2009;157(1-4):105-12.
[21] Modarres R, Dehkordi AK. Daily air pollution time series analysis of Isfahan City. Int J Environ Sci Tech. 2005;2(3):259-67.
[22] Farah W, Nakhle MM, Abboud M, Maesano IA, Zaarour R, Saliba N, et al. Time series analysis of air pollutants in Beirut, Lebanon. Environ Monit Assess. 2014;186:8203-13.
[23] Abulude FO, Fagbayide SD, Akinnusotu A, Elisha JJ, Makinde OE. Particulate matter and source identification: a case study of Nigeria. Eng Appl Sci Res. 2019;46(2):151-69.
[24] Montgomery DC, Jenning CL, Kulahci M. Introduction to time series analysis and forecasting. 2nd ed. New York: John Wiley & Sons; 2015.
[25] CPCB. Central Pollution Control Board, Bangalore [Internet]. India: Ministry of Environment, Forest and Climate Change; 2017 [cited 2017 Mar 7]. Available from: http://cpcb.nic.in/
[26] Hair JFJr, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 7th ed. Noida: Pearson Education; 2013.
[27] Roberts EM. Review of statistics extreme values with applications to air quality data part I: review. J Air Pollut Contr Assoc. 1979;29(6):632-7.
[28] Roberts EM. Review of statistics of extreme values with applications to air quality data – part II: applications. J Air Pollut Contr Assoc. 1979;29(7):733-40.
[29] Achcar JA, Fernández Bremauntz AA, Rodrigues ER, Tzintzun G. Estimating the number of ozone peaks in Mexico City using a non-homogeneous Poisson model. Environmetrics. 2008;19(5):469-85.
[30] Kolarik WJ. Creating quality: process design for results. New York: Mcgraw Hill; 1999.
[31] Walker G. On periodicity in series of related terms. Proc Roy Soc A. 1931;131:518-32.
[32] Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19:716-23.
[33] Raymond MR, Roberts DM. A comparison of methods for treating incomplete data in selection research. Educ Psychol Meas. 1987;47(1):13-26.
[34] Yozgatligil C, Aslan S, Iyigun C, Batmaz I. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor Appl Climatol. 2013;112:143-67.
[35] Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, et al. Statistics for the evaluation and comparison of models. J Geophys Res. 1985;90:8995-9005.