Time Series Analysis of Air Pollution in Bengaluru Using ARIMA Model

M. S. K. Abhilash, Amrita Thakur, Deepa Gupta and B. Sreevidya

Abstract Air pollution control measures in India are still in its infancy, while the country is developing at a faster rate. Development is known to affect the air quality of a place adversely. The key to manage the air quality of a place is proper planning, and for that, robust forecasting system based on continuous monitoring is required. Bengaluru is a city which has grown in size and population in the past decades. This rapid growth has affected its environmental quality. The present work deals with development of air quality prediction model based on Autoregressive Integrated Moving Average (ARIMA). For this, pollution data of NO_2, PM_{10} and SO_2 from January 2013 to March 2016, 14 pollution monitoring stations has been used. The results show that data which satisfies the stationary condition can be used as an accurate prediction model. NO_2 residential and RSPM residential satisfy this condition.

Keywords Air pollution · Bengaluru · ARIMA

1 Introduction

Substances like CO, CO_2, NO_X, SO_X, Particulate matter, Lead, VOC, Benzene and their photochemical products, in concentration which are harmful for humans and environment cause air pollution. Several studies aimed at correlating human health with the quality of air have been conducted in various parts of the world [1–3].

M. S. K. Abhilash · B. Sreevidya
Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: b_sreevidya@blr.amrita.edu

A. Thakur (✉)
Department of Chemistry, Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: t_amrita@blr.amrita.edu

D. Gupta
Department of Mathematics, Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: g_deepa@blr.amrita.edu

G. M. Perez et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 696,
https://doi.org/10.1007/978-981-10-7386-1_36

Short-term exposure even in very small quantity of pollutants such as particulate matter and ozone is found to cause respiratory and cardiovascular disease, leading to increase in mortality and hospital admissions [4], while oxides of nitrogen and sulphur (NO_X, SO_X) are known to affect the respiratory organs and cause asthma symptoms. Increasing number of cities in developing countries shows severe air pollution due to rapid urbanization. An estimation for Kanpur city in India reports a gain of 165.45 INR per individual per year in the absence of polluted air [5]. The Central Pollution Control Board (CPCB) of India in its annual report has shown that 67 monitoring stations (for NO_2) and 295 stations (for PM_{10}) exceed national ambient air quality standards (NAAQSs). SO_2 concentration has been reported to be under limits. Observations in case of 24 hourly average data show that those 11 stations (SO_2), 57 stations (NO_2) and 241 stations (PM_{10}) exceeded NAAQS. Even for sensitive area the annual average for 4 stations (for NO_2) and 17 stations (for PM_{10}) was found to exceed NAAQS [6].

Air pollution in a city requires immediate attention as large number of people live in city, and hence, air pollution may affect more people. Strict laws, constant monitoring of air pollutants and their trend prediction are key aspects of air pollution management strategy [7, 8]. Prediction being key to air pollution, our paper focuses on developing a suitable model of prediction. The remaining part of the paper has been organized into six sections. Section 2 deals with the literature survey and Sect. 3 with the area of study and the description of data used for investigation in the study. ARIMA-based time series prediction model is described in Sect. 4, while the fifth has been dedicated for the result and analysis. The last section deals with inference and the future scope.

2 Literature Survey

Time series analysis is a proven tool for air pollution forecasting as presented in the following literature survey. Factor analysis and Box–Jenkins methodology were used to evaluate concentrations of air pollutants such as NO, NO_2, NO_x, PM_{10}, SO_2 and ground-level O_3 in the town of Blagoevgrad, Bulgaria, with one-year hourly data of the pollutants using factor analysis with PCA and promax rotation. Results indicated high multicollinearity between the six pollutants [9]. The application of an intelligent hybrid system consisting of an artificial neural network combined with a particle swarm optimization algorithm for time series forecasting of air pollutant's concentration levels indicated a fair prediction of the presented pollutant time series by using compact networks [10, 11]. Box–Jenkins ARIMA approach was investigated for modelling the time series of monthly maximum 1 h concentration of CO and NO_2 in the east coast states of Peninsular Malaysia. The results have shown consistency with the observed values [12]. PM_{10} and SO_2 air pollution and residential natural gas consumption (RNGC) in Turkey were modelled by various multiparameter time series modelling methods (TSMs). Short-term estimation of RNGC, PM_{10} and SO_2 for 2014–2015, temperature-dependent ARIMAX (1, 1, 2)

(R_2 = 0.944) and RNGC and meteorological factor-dependent SARIMAX (0, 1, 1) (1, 1, 0)12 (R_2 = 0.761) and ARIMAX (1, 1, 0) (R_2 = 0.698) models, respectively, yielded the best-fitting scores and accuracy measures [13]. Artificial neural networks (ANNs) and genetic programming (GP) were used to predict the AQI of SO_X, NO_x, RSPM for Pune city using daily average data of 7 years. The results of the models developed were compared with GP and forecasting, and performance of the models has been compared using r, RMSE and d. It was found that GP models were robust and better than ANN [14]. The study was conducted on air pollutants data from Bahadur Shah Zafar Marg near ITO intersection, Delhi, on the varying trends of air quality and the levels of related air pollutants using Seasonal Autoregressive Integrated Moving Average (SARIMA) approach, implemented by Box–Jenkins. The performance evaluations of the adopted forecast models when done on the basis of correlation coefficient (R_2) and root-mean-square error (RMSE) provided reliable and satisfactory air quality predictions [15]. Air pollution studies in Bengaluru have shown that the air pollution problem is mainly because of two wheelers, construction activity and diesel consumption [16]. Pollution trend analysis of criteria pollutant using time series analysis for representative monitoring stations in this city has shown results similar to the actual values in most of the cases [17]. Poisson regression models were developed to study short-term impacts of PM_{10} and temperature, a factor that controls the climate on mortality for Indian cities including Bengaluru. It showed that temperature and pollution interactions do not significantly impact mortality [18]. Time series studies in Bengaluru has indicated that the correlation between model and observed values varies from 0.4 to 0.7 for SO_2, 0.45 to 0.65 for NO_x and 0.4–0.6 for SPM. About 80% of data is observed to fall within the error range of ±50%. The deviation in results observed was attributed to change in fuel quality, increased traffic, LPG as transport fuel, poor infrastructure and meteorological conditions [19]. It is observed that statistical and time series models have been successfully employed for air pollution prediction. In case of Bengaluru, studies such as time series analysis, Poisson regression modelling, surveys related to the factors contributing to air pollution and health impact on air pollution have been carried out. We have also observed that recent data have not been used for investigations, and since Bengaluru is one of the fast growing cities in India, it is necessary to investigate the air pollution scenario and develop a prediction model for better pollution control. In this study, our approach is to develop class of ARIMA model for prediction based on the AQI. After a brief introduction about the city, the following sections describe the proposed model in detail.

3 Study Area and Data Description

Bengaluru, the capital of Karnataka, is a landlocked city, positioned at 12.97° N 77.56° E and covers an area of 2,190 km^2 (850 $miles^2$). Its average rainfall is 1286.6 mm, and temperature variation is between 7.8 and 38.9 °C [20]. Situated in the heart of Mysore Plateau (a larger part of Deccan Plateau) at an average elevation of 900 m, it has a pleasant climate round the year. Growth of IT industry has caused its rapid growth in population (51.39%) [21] and urban area (466%). Land area expansion is due to urbanization of nearby villages and illegal conversion of the green belt area [22, 23]. This transition from Garden City to Silicon Valley affected its ecological services adversely. It is witnessing shortage of water, polluted lakes, shrinking green area and decreasing air quality. Traffic movements and construction activity are greatly affecting its air quality.

CPCB under nationwide air quality monitoring program (NAMP) has identified four air pollutants such as SO_2, oxides of nitrogen as NO_2, suspended particulate matter (SPM) and respirable suspended particulate matter (RSPM/PM_{10}) for regular monitoring. In Bengaluru, the monitoring is done at 14 places (monitoring stations) which are divided into three categories—residential, industrial and sensitive [24] which are shown in Fig. 1. Monitoring stations under these categories are as follows:

Industrial	AMCO, Peenya, Graphite, KHB are industrial areas such as AMCO batters industry, textile industry at Peenya, KHB and electrode industry at Graphite India.
Residential	These are normal everyday streets, busy places, schools, offices, etc. Traffic congestion and construction activities are main sources of pollution. City railways, TERI, Yeshwantpur, DTDC, and Central Silk Board.
Sensitive Areas	Places such as hospitals and garbage dumps are termed as sensitive areas. Vicotia hospital, Indira Gandhi children care and Kajisonnehalli fall under this category.

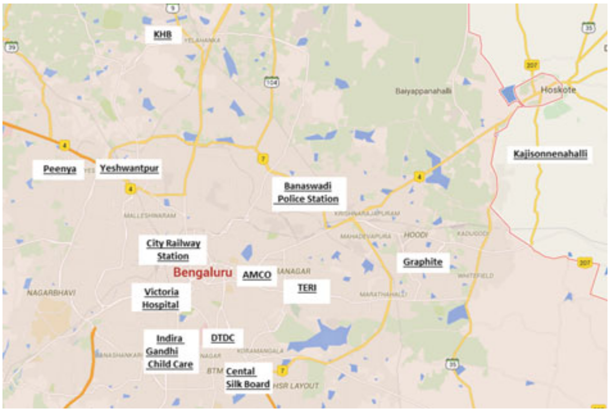

Fig. 1 Air pollution monitoring stations, Bengaluru, India

4 Proposed Approach

In this study monthly average concentration in (μg/m^3), of pollutants SO_2, NO_2 and RSPM received from the Karnataka State Pollution Control Board from January 2013 till March 2016, of three constituent pollutants SO_2, NO_2 and RSPM has been considered for modelling. Model development has been done using data from January 2013 to November 2015. Four-month data that is from December 2015 to March 2016 has been used for the testing of developed model. The raw data has been processed and used for forecasting purposes. Since the data has a time stamp of less than 5 years, ARIMA (Autoregressive Integrated Moving Average) method has been employed for the analysis. This model also helps in identification of parameters to be inputted for the plot [25–27] (Fig. 2).

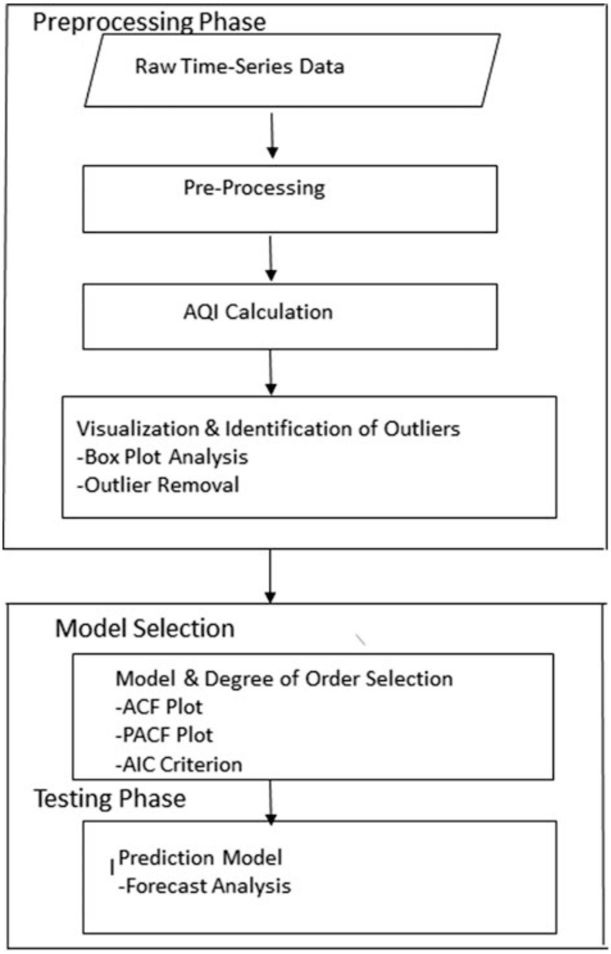

Fig. 2 Flow chart depicting stages of ARIMA model

The model selection can be divided into three phases:
i. preprocessing, ii. model selection and iii. testing.

4.1 Preprocessing Phase

Preprocessing involved removal of missing values from the raw data, calculation of AQI for each criteria pollutant month wise for each category of monitoring station and outlier removal. Missing value was obtained by taking average of the preceding and succeeding concentrations. Data visualization is done by plotting time series graph for the AQIs against time. For example, time series plot for NO_2 for industrial monitoring station (Fig. 3) shows that the data is cyclic in nature. The AQI was obtained by using Eq. 1. Each calculation has been shown for NO_2 industrial as a reference [28]:

$$\mathrm{Ip} = [\{(\mathrm{IHi} - \mathrm{ILo})/(\mathrm{BHi} - \mathrm{BLo})\} * (\mathrm{Cp} - \mathrm{BLo})] + \mathrm{ILo} \tag{1}$$

Here

BHi Breakpoint concentration greater or equal to given concentration.
BLo Breakpoint concentration smaller or equal to given concentration.
IHi AQI value corresponding to BHI.
ILo AQI value corresponding to BLO.
Cp Pollutant concentration.

The presence of some accidental or irregular data in the graph might be due to situations such as heavy rainfall and traffic jams which were removed by boxplot analysis and then replaced. The entire process had been done by considering the fact that the outlier removal does not alter the structure of the patterned data.

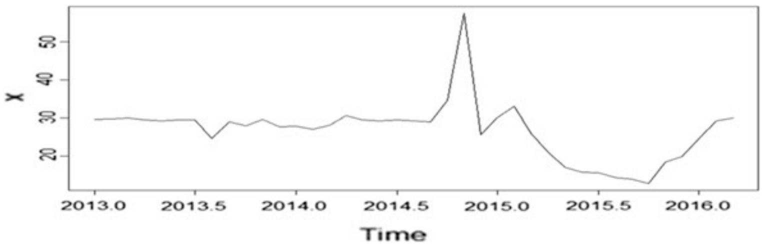

Fig. 3 Plot of NO_2 industrial (X-axis has time in year/month, i.e. January as 0 with scale as 0–5. Y-axis has AQI values.)

4.2 Model Selection Phase

Augmented Dickey-Fuller test (ADF) and null hypothesis are employed to identify whether a data set is stationary or non-stationary or not [29, 30]. A negative Dickey-Fuller and a higher error in case of NO_2 industrial show that the data is non-stationary.

Dickey-Fuller = –2.599
Lag order = 2
p value = 0.3442

Autocorrelation (ACF) and partial autocorrelation (PACF) were examined to determine the best combination order of ARIMA model for each data set. The ACF plot for NO_2 industrial monitoring station, for example, shows the presence of both moving average and autoregressive processes (Fig. 4).

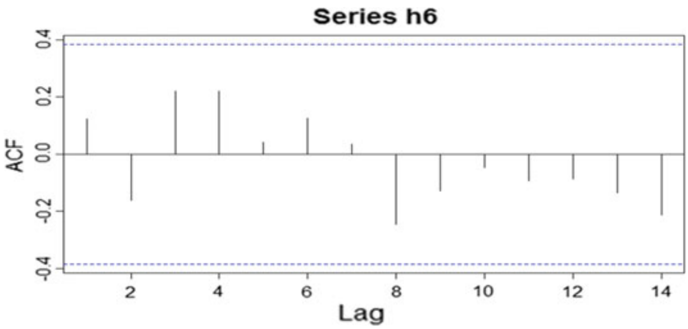

Fig. 4 ACF plot of NO_2 industrial

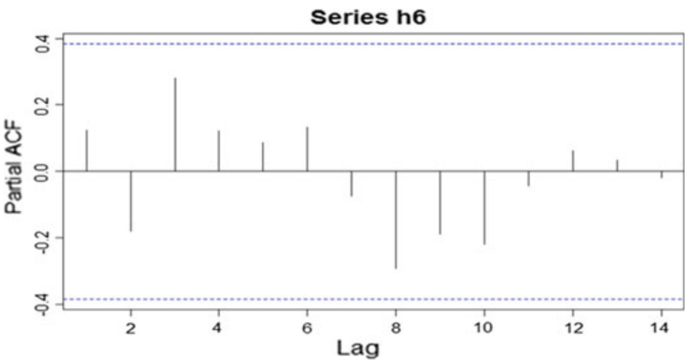

Fig. 5 PACF plot of NO_2 industrial

The partial autocorrelation function (PACF) in Fig. 5 indicates more than one order for ARIMA.

The selection of the best-suited order for prediction is done using Akaike information criterion (AIC) as the order which gives least AIC value that is considered as the degree of order for the prediction model.

4.3 Testing Phase

After model selection and fixing set of parameters, testing of achieved model has been done by analysis of two types of scenario. The data from January 2013 to November 2015 has been used for testing the model performance, and the data from December 2015 to March 2016 has been considered as unseen data to see the ability to predict in all scenario. Reasoning, analysis and conclusion about the model performance have been done in this subsection.

5 Experimental Results and Analysis

All pollutants have been analysed individually for industrial, residential and sensitive monitoring stations. Pollutant concentration for each type of monitoring station has been clubbed together; the monthly average concentration for each of them has been used to calculate the AQI. Computation for ARIMA has been done on R studio which utilizes R programming language. The model selection parameters are listed in Table 1 using proposed ARIMA model. Here, S stands for

Table 1 Model selection parameters

Pollutants	ADF statistic	P value	S/NS data	PACF order	AIC
NO_2 industrial	–2.599	0.3442	NS	3, 8	160.14, 166.02
NO_2 residential	–3.3297	0.08218	S	1, 4	150.12, 153.26
NO_2 sensitive	–1.6862	0.6963	NS	1, 3, 5	172.78, 173.29, 177.42
SO_2 industrial	–1.9065	0.61	NS	1	126.39
SO_2 residential	–1.7287	0.6796	NS	1	86.86
SO_2 sensitive	–1.9184	0.6054	NS	1	109.79
RSPM industrial	–2.6705	0.3108	NS	1, 3	321, 324.13
RSPM residential	–4.0746	0.01717	S	1, 3, 4	287.47, 290.15, 291.24
RSPM sensitive	–2.5135	0.374	NS	3, 4, 8, 10	345.07, 347.06, 349.01, 351.46

stationary, while NS for non-stationary. Other header of Table 1 is same as discussed in model selection phase. Prediction model based on ARIMA of NS may not be accurate but can still give important information like interval of predicted values and hence has been used for model development.

The prediction models for NO_2, SO_2 and PM_{10} is shown in plots for Industrial, Residential and Sensitive Areas in the next section.

5.1 SO_2 Result Analysis

Data is found to be non-stationary for each type of monitoring station. It is also evident from the plots that SO_2 contributes minimum to the pollution spectrum of Bengaluru. The boxplot analysis did not give any outliers and even if it did, it won't contribute anything to the plot since it is predominantly downward in trend (Fig. 6).

5.2 NO_2 Result Analysis

NO_2 industrial has non-stationary data which means the prediction might not exactly same as the actual NO_2 plot which is also reflected in the above plot. But the AQI interval on vertical axis for prediction is around 26–30 and actual is 25–35, so in hindsight, prediction interval is much smaller and accurate than the actual plot. NO_2 residential data according to ADF test stationary, which means the predicted plot, will be accurate to the actual plot which is reflected in Fig. 7b. The plots do not have much peaks and troughs as the industrial plot. NO_2 sensitive areas data was non-stationary; hence, there is difference in predicted last segment to the actual plot. From the three plots, we can infer that industrial and sensitive areas are more pollution-prone than residential as they have peaks and troughs, whereas residential which was stationary had a uniform data value distribution.

5.3 PM_{10} (RSPM) Result Analysis

RSPM industrial data is found to be non-stationary from the ADF, which implies that the predicted plot forecast to be different from the actual plot but they come under the same AQI value interval of 140–180. The industrial plots are pretty significant as they peak at 220 and a valley below 100. The RSPM residential data on the other hand was stationary; the prediction was expected to be similar. January 2106 in the prediction plot shows a dip in value when compared to the actual plot

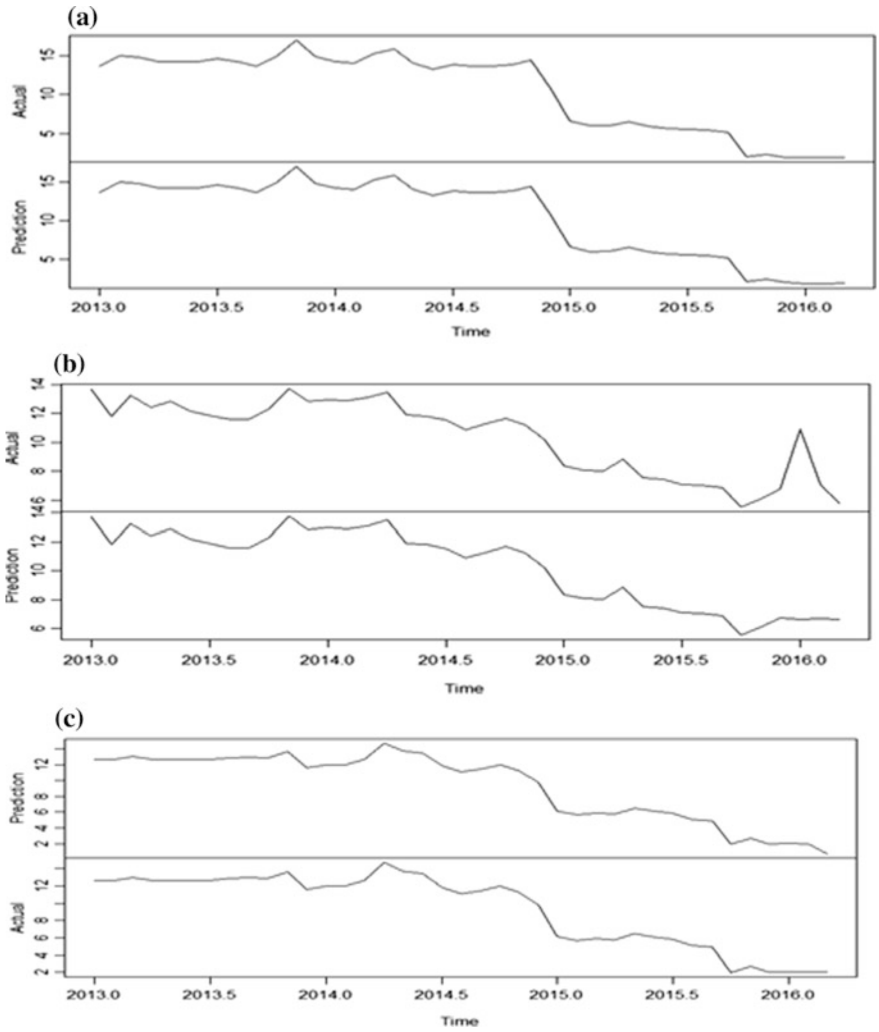

Fig. 6 **a** SO_2 industrial forecast. **b** SO_2 residential forecast. **c** SO_2 sensitive forecast

which can be attributed to the heavy adjustment of data done during the boxplot analysis. Non-stationary RSPM data for sensitive areas shows difference in last segment as expected. The predicted plot shows a little upward trend at the end, but the actual plot has a downward trend. RSPM from industrial area is found to be the highest contributor (Fig. 8).

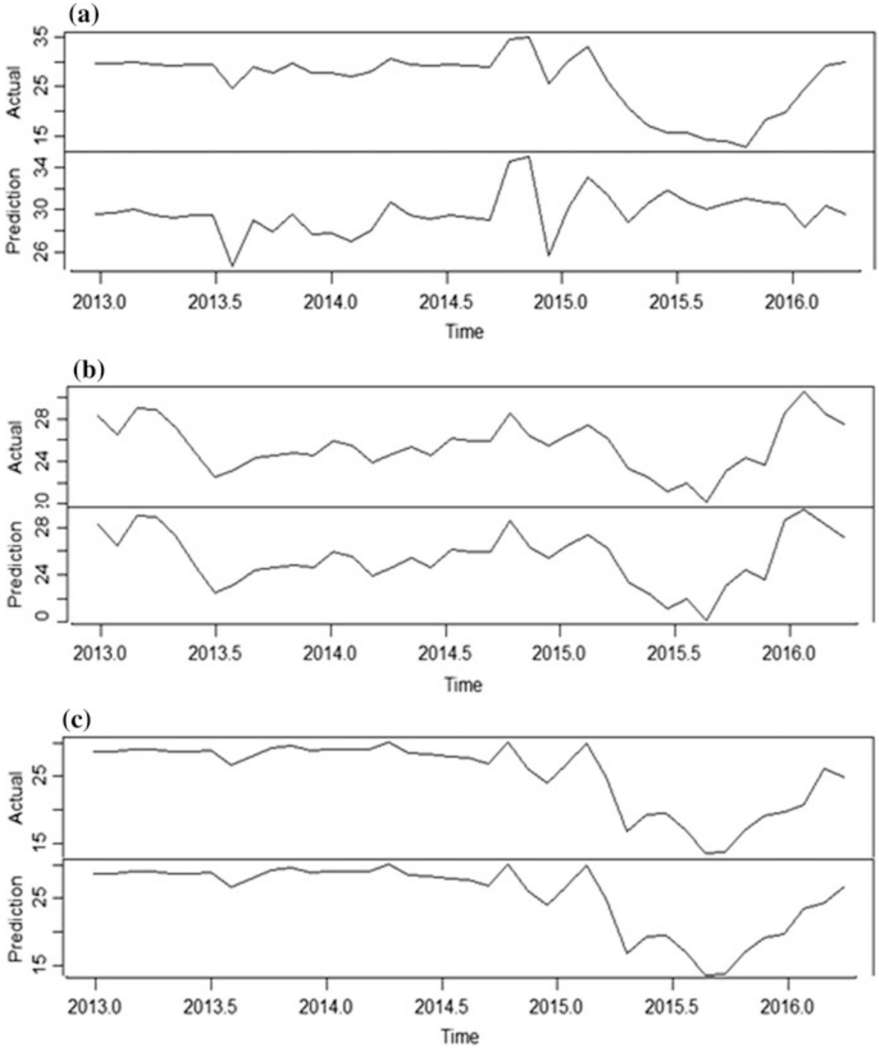

Fig. 7 **a** NO_2 industrial forecast. **b** NO_2 residential forecast. **c** NO_2 sensitive forecast

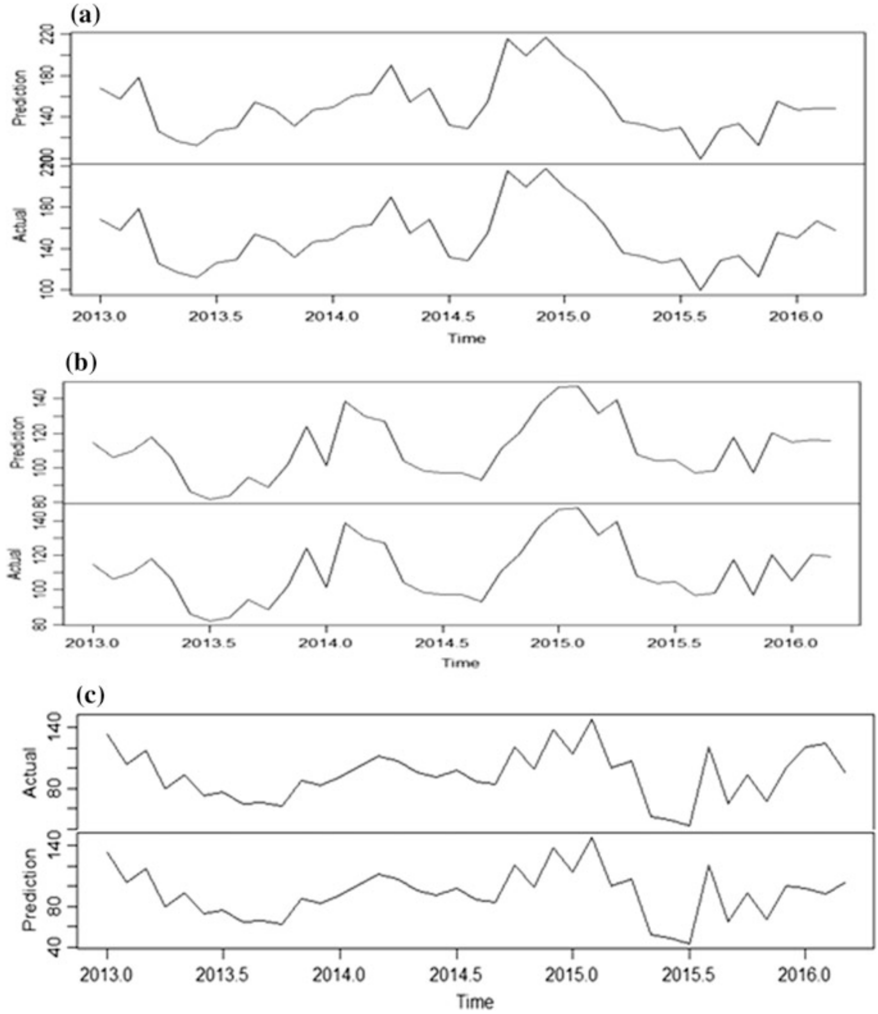

Fig. 8 **a** RSPM industrial forecast. **b** RSPM residential forecast. **c** RSPM sensitive forecast

6 Conclusion and Future Scope

RSPM residential and NO_2 residential are stationary data; hence, the prediction model perfectly fits with the actual plot. The rest of the plots are analysed on non-stationary data; hence, the forecasts are not matching but can still infer information such as trend and interval where the predicted values occur and match with the interval of the actual plot. RSPM is the dominant pollutant, and SO_2 contributes minimum. All the pollutants are analysed individually because even though their concentrations vary each of these pollutants has adverse health effects.

ARIMA model is suitable for short-term predictions because if the data is found to be stationary accurate predictions can be made.

Acknowledgements We thank Dr. Prashobhan Palakkeel, Assistant Professor, Amrita School of Business, Bengaluru, for the help he extended towards the application of statistical tools and Dr. Nagappa B, Scientific Officer, Karnataka State Pollution Control Board, for providing the air pollution data of Bengaluru.

References

1. WHO, Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide: report on a WHO working group. World Health Organization, Bonn, Germany 13–15 January 2003.
2. WHO, Air quality guidelines for Europe. World Health Organization, 2000.
3. WHO, Evaluation and use of epidemiological evidence for environmental health risk assessment: Guideline document. World Health Organization, 2000.
4. Brunekreef, Bert et al. “Air pollution and health”, The Lancet, Volume 360, Issue 9341, 1233–1242.
5. Usha Gupta, “Valuation of Urban Air Pollution: A Case Study of Kanpur City in India”, ISSN 1893-1891; 2006-WP17 SANDEE Working Papers, 2006.
6. National Ambient Air Quality Status & Trends In India-2010, NAAQMS/ 35/2011-2012, CPCB.
7. Box G.E.P. and Jenkins G.M., “Time Series Analysis, Forecasting and Control”, Holden-Day, San Francisco, CA, 1970.
8. Asha B. Chelani, D.G. Gajghate & M.Z. Hasan, “Prediction of Ambient PM_{10} and Toxic Metals Using Artificial Neural Networks”, Journal of the Air & Waste Management Association, 52:7, 805–810, 2010.
9. Theresa Hoang Diem Ngo, “The Box-Jenkins methodology for Time Series Models”, SAS Global Forums, 2013, Paper No, 454.
10. “Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach” Stochastic Environmental Research and Risk Assesment, May 2014, Volume 28, Issue 4, pp. 1045–1060.
11. Albuquerque Filho, Francisco S. de et al., “Time-series forecasting of pollutant concentration levels using particle swarm optimization and artificial neural networks” Quím. Nova, 2013, vol. 36 (6), pp. 783–789.
12. Mohd Zamri Ibrahim, Roziah Zailan, Marzuki Ismail and Muhd Safiih Lola, “Forecasting and Time Series Analysis of Air Pollutants in Several Area of Malaysia” American Journal of Environmental Sciences 5 (5): 625–632, 2009.
13. Taşpınar, F., “Time Series Models for Air Pollution Modelling Considering the Shift to Natural Gas in a Turkish City”, Clean Soil Air Water, 43: 980–988, 2015.
14. Tikhe Shruti S., Khare K. C., Londhe S. N., “Forecasting Criteria Air Pollutants Using Data Driven Approaches; An Indian Case Study”, IOSR-JESTFT, 2319–2399. Volume 3(5), pp. 1–8.
15. Inderjeet Kaushik and Rinki Melwani, “Time Series Analysis of Ambient Air Quality at ITO Intersection in Delhi (India)” Journal of Environmental Research And Development, Vol. 2 (2), 2007.
16. Amrita Thakur, “Study of Ambient Air Quality Trends and Analysis of Contributing Factors in Bengaluru, India”, Oriental Journal of Chemistry 2017, Vol. 33, No. (2): pp. 1051–1056.

17. Asha B. Chelani & Sukumar Devotta, “Nonlinear Analysis and Prediction of Coarse Particulate Matter Concentration in Ambient Air”, Journal of the Air & Waste Management Association, 56:1, 78–84.
18. A.L. Seetharam and B.L. Udaya Simha, “Urban Air Pollution – Trend and Forecasting of Major Pollutants by Timeseries Analysis”, World Academy of Science, Engineering and Technology International Journal of Civil, Environmental, Structural, Construction and Architectural Engineering Vol. 3(3), 2009.
19. A.K. Chinnaswamy, R.N.G. Naguib, Q.T. Nguyen, L.O. Olayanju, N. Trodd, I.M. Marshall, N. Yaacob, G.N. Santos, E.A. Vallar, M.C. Galvez, M.H. Shaker, T.N. Ton, “Air Pollution in Bengaluru, India: A Six-Year Trend and Health Implication Analysis”, IEEE Int. Conf. on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, Philippines, 2014.
20. Indian Meteorological Department, Meteorological centre, Bengaluru, 2008.
21. censusindia.gov.in
22. Ravindra, A.; Venkataraman, M.; Narayan, E.; Masta D. Centre of Excellence in Urban Governance, Centre for Public Policy, Indian Institute of Management, Bengaluru, 2012.
23. Venkataraman, M. Working Paper No: 464, Indian Institute of Management, Bengaluru, 2013.
24. Pollution Assessment monitoring and Survey, Mandate, CPCB.
25. Alan Pankratz, A Primer on ARIMA Models, John Wiley and Sons, 2012.
26. Javier Contreras, “ARIMA Models to Predict Next-Day Electricity Prices”, IEEE, Vol 18, No 3, August 2003.
27. Prapanna Mondal, Labani Shit and Saptarsi Goswami, “Study Effectiveness of Time Series Modeling (ARIMA) In Forecasting Stock Prices”, IJCSEA, Vol 4, No 2, April 2014.
28. National Air Quality Index, CPCB, 2014–15.
29. Artur C.B. da Silva Lopez, “Deterministic Seasonality in Dickey-Fuller Tests: Should We Care?”, ISEG-UTL and CEMAPRE, January 24, 2003.
30. Yoosoon Chang and Joon Y. Park, “On The Asymptotics of ADF Tests For Unit Roots”, Thesis, November 2001.