Journal of Computer Science and Applications.
ISSN 2231-1270 Volume 15, Number 1 (2023), pp. 1-9
© International Research Publication House
http://www.irphouse.com

Time Series Analysis Using Arima Model for Air Pollution Prediction in Cities of Rajasthan

Apoorva Verma[1*] and Dr. Leena Bhatia[2]

*[*1]Research Scholar, Department of Computer Applications,*
Rajasthan Technical University, Kota
E-mail: v.apoorva1995@gmail.com[*]
[2]Associate Professor, Department of Computer Applications,
S.S. Jain Subodh PG College, Jaipur
E-mail: leenabhatia@hotmail.com

Abstract:

The growth of industrialization has raised serious concerns about environmental pollution. Currently, the environment is facing a major issue related to air quality degradation. Air is a combination of gases that fills the atmosphere, providing life for humans, plants, and animals that make Earth a living planet. If the proportion of these gases becomes unbalanced, the quality of air is affected, and they become pollutants. The authors of this paper have analyzed these pollutants and predicted them using an Auto-Regressive Integrated Moving Average (ARIMA) model which is applied to the data set of air of various cities of Rajasthan which resulted in the prediction of the pollutants. ARIMA is a statistical model that uses stepwise regression, principal component regression (PCR), and multiple linear regression (MLR) to predict values. The present study finds out areas where the values of pollutants are exceeding the limits prescribed by World Health Organization (WHO) thus, help to create awareness among people and the government to take necessary actions to decrease the levels of such harmful pollutants. Based on the available data set, we measure the effectiveness and performance of the technique.

Keywords: Air Pollution, ARIMA, Time Series Analysis, Pollutants.

Introduction

Today, air quality is a major concern, especially in developing countries like India. The term "air pollution" refers to the introduction of particulates, biological molecules, or other harmful materials into the atmosphere resulting in disease, death, and damage to other living organisms. Thus, a clean, suitable atmospheric environment is necessary for human development. Prior to World Health Day (7th April), the World Health Organization (WHO) released its Air Quality Database 2022, which indicates that almost the entire global population (99%) breathes air that exceeds WHO's air quality standards. 11 of the 15 most polluted cities in Central and South Asia were located in India, according to the World Air Quality Report of 2021 published by the Swiss company IQAir [11]. Less than 1% of people now breathe air that complies with the WHO's toughest standards for air quality. Without taking into account the many other millions who suffer from air pollution-related chronic illnesses, the WHO estimates that there are 7 million preventable deaths annually, including around 600,000 children under the age of fifteen [14].

The major air pollutants present in air are particulate matter, suspended particulate matter, nitrogen dioxide, Sulphur dioxide, ozone, carbon dioxide and many more. If we look deep inside we will find that burning of fossil fuels emerges out to be one of the major reasons of increasing pollutants in the air. Sulphur dioxide is the major pollutant which is emitted in burning of fossils. Industrial processes such as extracting metal from ore, natural sources such as volcanoes and locomotives, ships and other vehicles, and heavy equipment burning fuel with high sulfur content are among the other smaller sources of emission of SO_2 [12]. Continue exposure to SO_2 may cause harm to human respiratory system. People are suffering from asthma particularly children are sensitive to SO_2. Thus it is requirement of the hour to control the activities which lead to emissions of SO_2 in the air.

Several models have been developed which forecasted the various pollutants in the air. But still there is a scope of improvement and there is gap as there are few major polluted cities in India and few pollutants which are not given a thought of. Through our experiments we are trying to forecast the concentration of SO_2 for the next coming 5 years in the major polluted cities of Rajasthan using ARIMA model.

Related Work

An accurate and reliable prediction of air quality is essential to public health and the environment [4]. It is difficult to predict air quality due to the dynamic nature of pollutants and particles, their volatility, and their variability over time and space. Due to the observed negative effects of air pollution on citizens' health and the environment, it is more important than ever before to model, predict, and monitor air quality [3].

There are three types of forecasting methods 1. Statistical models 2. Deterministic Models and 3. Hybrid Model. Deterministic Models includes major classes like Chemistry, Meteorological and Emission Models that can forecast the distribution, expulsion, and accumulation of pollutants but due to being not so cost effective and accurate Statistical models and Hybrid models are more used by the researchers for the prediction and forecasting of the time series data.

Many models have been developed by the researchers to measure the pollutant particle levels in air. The core of major of these models is analyzing historical time series data from various parts of the world. The major focus of the model selection was on statistical model including ARIMA model [24], SARIMA model [16], SARIMAX model to predict air quality in the Lahore city of Pakistan [2], Kalman Filter [17] and single variable Regression Model [7] However these models were not able to predict the air quality so efficiently, due to which researchers moved on to Machine Learning Techniques in combination with statistical models [20], [6], Deep Learning [5], Artificial Neural Network [15]. But in all these models the major drawback was that the dependency on the previous values of the pollutant was completely neglected. Various other methods which caught the focus and interest of researchers for air quality prediction were Random Forest, Support Vector Regression [22] and Exponential Smoothing [23].

Not only developing models and analyzing historic time series data, error findings and calculating the accuracy of the model was also a work of focus. A hybrid model was developed by Fan S. *et.al.* to predict air quality based on data decomposition. The root mean square error (RMSE), mean absolute error (MAE), and goodness of fit (R^2) are all reduced by 52%, 47%, and 18%, respectively, when the projected values for each sub-sequence are added to provide the final air quality prediction results [4]. A Hybrid Neural Network [19] can also be good choice when it comes to build an accurate model. Not only building new models but comparison of the existing models is also beneficial as it helps to choose the model and work for the betterment of that model in order to improve the forecasting accuracy. The same was done by Alireza Rahimpour *et.al.* According to the researchers, hybrid single decomposition (HSD) and hybrid two-phase decomposition (HTPD) models are also useful for air quality prediction, and HTPD models provide better results compared to HSD models in predicting AQIs [21].

Problem Analysis and Model Design:

Study Site

The state of Rajasthan is the largest in terms of area, covering 10.4% of the total land area of the country. It lies in the western part of the country between 23 30' and 30 11' North latitude and 69 29' and 78 17' East longitude [1]. In this study, we have chosen the major polluted cities of the Rajasthan State i.e. Jaipur, Kota, Alwar, Jodhpur, and Udaipur. There are more than a million people and industries in Jaipur, which is the capital and largest smart city in Rajasthan. Human activities are the primary reason for the continuously deteriorating air quality in this city. Kota and Udaipur are also smart cities. In Kota, air pollution is constantly getting worse due to the coal and thermal power plants that are most prevalent in the city. Human activities near the Chambal basin are also contributing to the problem. The city of Jodhpur ranks 45th among the most polluted in the world. Experts attribute Jodhpur's air pollution largely to dusky gales and traffic [8]. Air pollution in Udaipur is mainly caused by roadway dust, vehicular emissions, construction and demolition activities, and industrial emissions [9]. According to the latest survey IQAir, a swiss group Bhiwandi, an industrial town situated in the district of Alwar is the most polluted city of the world. The given below table shows the AQI of the various polluted cities of the Rajasthan:

Table 1: AQI of the various polluted cities of Rajasthan [10].

District	Status	AQI-US
Alwar	**POOR**	**181**
Jaipur	**POOR**	**169**
Jodhpur	**POOR**	**159**
Kota	**POOR**	**176**
Udaipur	**POOR**	**129**

Work Flow and ARIMA Model

Auto Regressive Integrated Moving Average (ARIMA) model is a statistical model which was developed by Box and Jenkins which is used for analysis of time series data and to forecast the future values. ARIMA is a mathematical model which uses historical data for the predictions. The major pre requisite for ARIMA model is that the data should be stationary which means the data should have a constant mean value, variance and correlation which should not be dependent on time. In figure 1 we have depicted the process of building ARIMA model using machine learning libraries and python language.

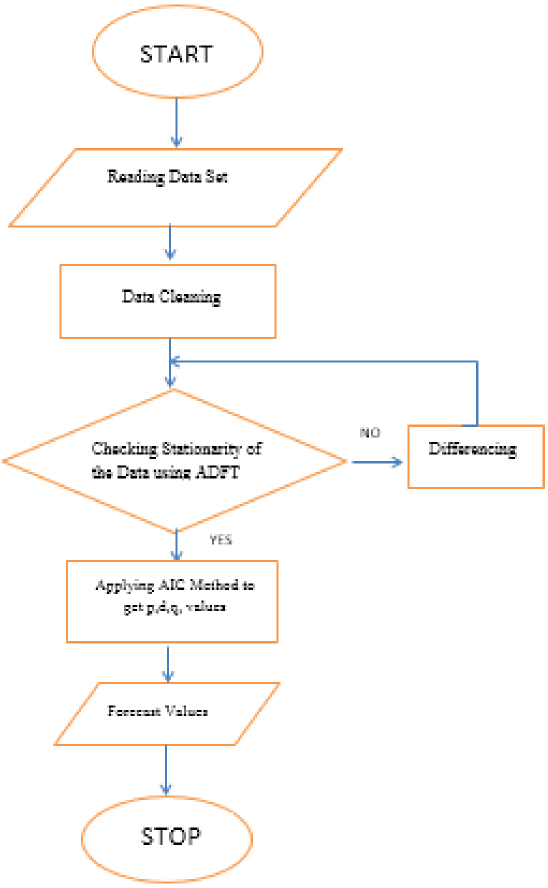

START
Reading Data Set
Data Cleaning
Checking Stationarity of the Data using ADFT
NO
Differencing
YES
Applying AIC Method to get p,d,q, values
Forecast Values
STOP

Figure 1: ARIMA MODEL.

Data Set

The data is an open dataset available on the Central Board of Pollution Control website under National Data Sharing and Accessibility Policy (NDSA) from the year 2010 to 2021 [13]. The dataset consists of the columns Stn_Code (Station Code), Sampling Date, State, Location, Agency, Type, SO_2 value, NO_2 value, RSPM value, SPM value, Location Monitoring Station, PM2.5 value and Date. In this study we have chosen SO_2 parameter of the air in the selected cities of Rajasthan. The dataset contains data from different stations from different districts like Kota-Regional Office, RJPB, Anantpura, Kota. From Jaipur data is collected from Adarsh Nagar, Police Commissionerate and Shastri Nagar Air Monitoring Stations. From Jodhpur we have data from air monitoring station situated at Collectorate, Ashok Nagar from Udaipur and Moti Dungri, Alwar.

Data Cleaning

For working on time series data we need to have only two column data i.e. Date and the Pollutant value for which we are doing the analysis and forecast. Thus our first step was to clean the data. By cleaning data, we ensure that it is accurate and free of any inaccurate records so that we can analyze it. In this process we drop the extra columns which are not required for the experiments, then in order to have clean data with no missing records we have used the Mean method to fill the missing records. After this process a new data set was created with only two columns Date and SO_2 value. Table 2 shows the sample dataset after cleaning of data containing two columns Date and SO_2 value.

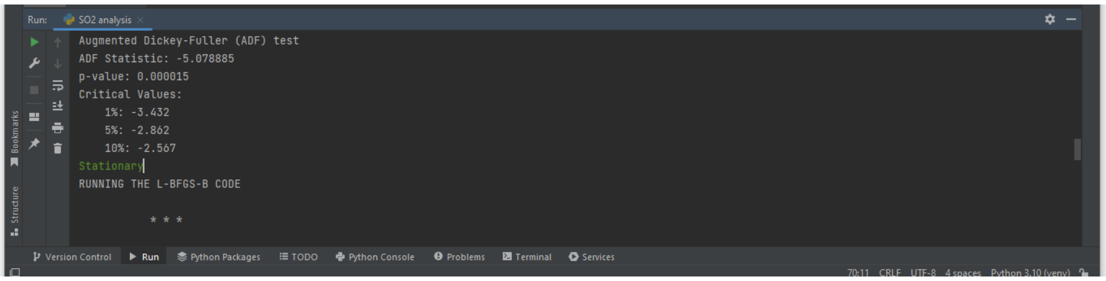
Table 2: Sample Dataset after Cleaning.

Date	**SO_2**
2010-01-02	13.9
2010-01-04	9.4
2010-01-05	8.0
2010-01-06	20.3
2010-12-28	52.0

Stationarity

To determine whether a model is stationary, there are two types of tests. The first is the Rolling Statistics (RS) and the second is the Augmented Dickey-Fuller Test (ADFT). In Rolling Statistics, observe the moving average and moving variance over time. The model is non-stationary if the answer is yes. Otherwise it's a stationary model. A nonstationary time series is believed to be the Null Hypothesis in an Augmented Dickey-Fuller Test. The results of the Dickey-Fuller Test indicate whether the model is stationary or not. The model is said to be stationary if the obtained p-value is very low and the critical values are greater than the test statistics.

When the model is not stationary, use the differencing technique to convert the time series into stationary. By comparing the current period with the previous period, we can perform the first differencing. If the model is still not stationary then go for second differencing and so on. In our data set after performing the Augmented Dickey-Fuller test we find the dataset to be stationary [6]. The test results are shown in Figure 2.



```
Augmented Dickey-Fuller (ADF) test
ADF Statistic: -5.078885
p-value: 0.000015
Critical Values:
    1%: -3.432
    5%: -2.862
    10%: -2.567
Stationary
RUNNING THE L-BFGS-B CODE

           * * *
```

Figure 2: Results of Augmented Dickey Fuller Test.

Forecasting Using ARIMA Model

After performing the data cleaning and checking the stationarity of the data set, the next step is to decompose the graph into its four components which are seasonal, trend, residual and observed using additive method of decomposition. Figure 3 shows the decomposition graph of the data. From the graph we can see that the trend graph shows that amount of SO_2 in air in Rajasthan is quite changing w.r.t. time while the seasonality remains the same throughout the duration.

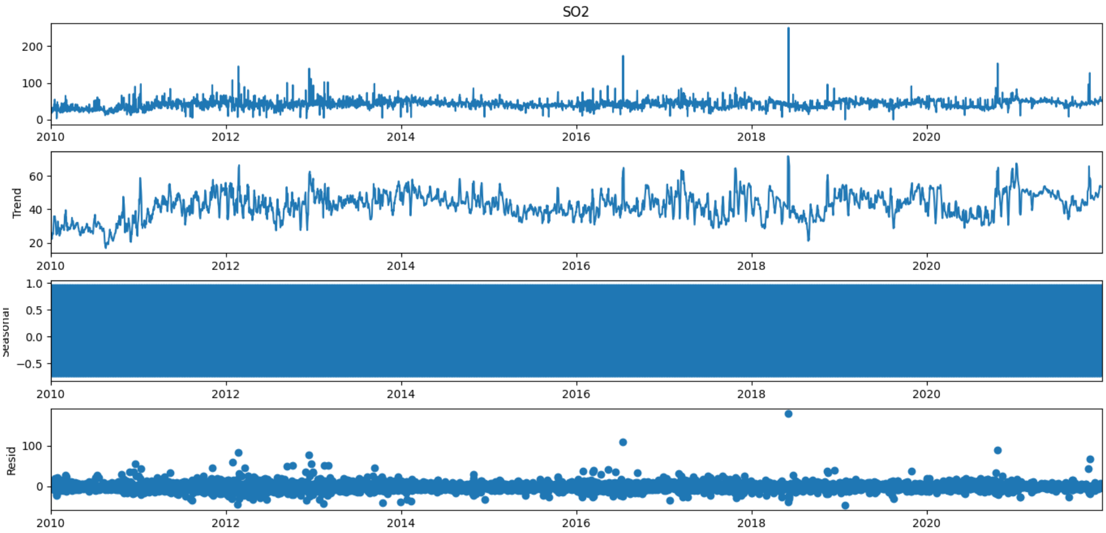


Figure 3: Decomposition Graph.

The most important part while building ARIMA model is to choose the p,d,q parameters for building the model where:

p: number of auto regressive lags
d: Required number of differencing to make time series stationary
q: The order of the moving average defines the significance of the time series' variance for earlier values used to forecast current values [18].

For getting the p,d,q parameters we have used AIC method of machine learning. Using statsmodel library of machine learning we implemented the AIC method. According to AIC method the parameters using which we get minimum AIC value we choose those p,d,q parameters. The parameters e got using this method are shown in Table 3.

Table 3: Parameters for ARIMA model.

Parameters	ARIMA
P	1
D	1
q	1

Experiment Results

After building the ARIMA model and carrying out the experiments, the results found are shown below in the figure 5. The orange line in the graph shows the forecasted values for the next 5 years from year 2021.

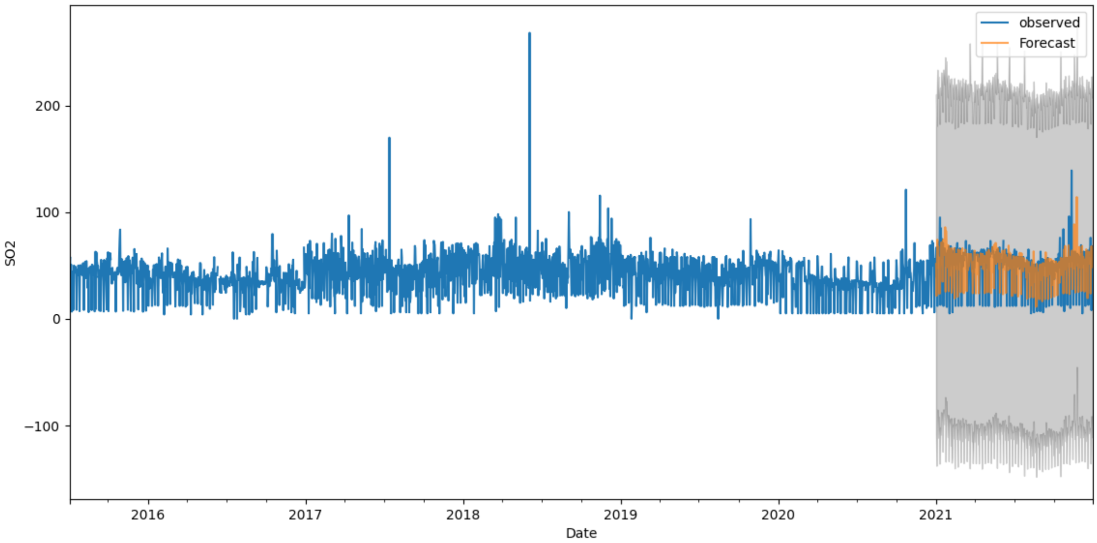


Figure 5: Forecasting using ARIMA model.

In order to check the accuracy of the forecasting done by the build ARIMA model there are various error rates which are calculated which indicate whether values is accurate or not. Some of them are MAPE (Mean Absolute Percentage Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and ME (Mean Error). The values calculated for the developed ARIMA model came out to be MAPE: 10.424506477389, ME: 0.09199609015564879, MPE: 0.5389951127880055.

Conclusion

The experiment successfully shows that the build ARIMA model (1,1,1) has forecasted the SO_2 concentration in the air of Rajasthan for next five years. The error rate shows that the accuracy achieved with the model is 89.58%. The future work includes working on other pollutants of the air, use of more advanced model and improving the accuracy of the previous build models. Working on data other than historical data can also result into more accurate results which can help the Govt. and other environmental organization to control the air pollution.

References:

[1] Barupal, T., Tak, P. K., Meena, M., Vishwakarma, P. K., & Swapnil, P. (2022). The Impact of COVID-19 Strict Lockdown on the Air Quality of Smart Cities of Rajasthan, India. The Open COVID Journal, 2(1).

[2] Bhatti, U. A., Yan, Y., Zhou, M., Ali, S., Hussain, A., Qingsong, H.,... & Yuan, L. (2021). Time series analysis and forecasting of air pollution particulate matter (PM 2.5): an SARIMA and factor analysis approach. IEEE Access, 9, 41019-41031.

[3] Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. Complexity, 2020.

[4] Fan, S., Hao, D., Feng, Y., Xia, K., & Yang, W. (2021). A hybrid model for air quality prediction based on data decomposition. Information, 12(5), 210.

[5] Freeman, B. S., Taylor, G., Gharabaghi, B., & Thé, J. (2018). Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association, 68(8), 866-886.

[6] Gopu, P., Panda, R. R., & Nagwani, N. K. (2021). Time series analysis using ARIMA model for air pollution prediction in Hyderabad city of India. In Soft Computing and Signal Processing (pp. 47-56). Springer, Singapore.

[7] Guo, Y., Tang, Q., Gong, D. Y., & Zhang, Z. (2017). Estimating ground-level PM2. 5 concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. Remote Sensing of Environment, 198, 140-149.

[8] http://timesofindia.indiatimes.com/articleshow/90531375.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

[9] https://cpcb.nic.in/Actionplan/Udaipur.pdf

[10] https://www.aqi.in/dashboard/india/rajasthan

[11] https://www.drishtiias.com/daily-updates/daily-news-analysis/air-quality-database-2022 who

[12] https://www.epa.gov/so2-pollution/sulfur-dioxide-basics

[13] https://www.kaggle.com/datasets

[14] https://www.unep.org/news-and-stories/statements/statement-chief-scientists-2022-international-day-clean-air-blue skies

[15] Huang, M., Zhang, T., Wang, J., & Zhu, L. (2015, September). A new air quality forecasting model using data mining and artificial neural network. In 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 259-262). IEEE.

[16] Jain, S., & Mandowara, V. L. (2019). Study on particulate matter pollution in jaipur city. International Journal of Applied Engineering Research, 14(3), 637-645.

[17] Lai, X., Yang, T., Wang, Z., & Chen, P. (2019). IoT implementation of Kalman Filter to improve accuracy of air quality monitoring and prediction. Applied Sciences, 9(9), 1831.

[18] Lin, W. Y., Hsiao, M. C., Wu, P. C., Fu, J. S., Lai, L. W., & Lai, H. C. (2020). Analysis of air quality and health co-benefits regarding electric vehicle promotion coupled with power plant emissions. Journal of Cleaner Production, 247, 119152.

[19] Mahajan, S., Liu, H. M., Tsai, T. C., & Chen, L. J. (2018). Improving the accuracy and efficiency of PM2. 5 forecast service using cluster-based hybrid neural network model. IEEE Access, 6, 19193-19204.

[20] Patil, R., Bedekar, G., Tergundi, P., & Goudar, R. H. (2022). An Efficient Implementation of ARIMA Technique for Air Quality Prediction. In Intelligent Data Communication Technologies and Internet of Things (pp. 441-451). Springer, Singapore.

[21] Rahimpour, A., Amanollahi, J., & Tzanis, C. G. (2021). Air quality data series estimation based on machine learning approaches for urban environments. Air Quality, Atmosphere & Health, 14(2), 191-201.

[22] Sanjeev, D. (2021). Implementation of machine learning algorithms for analysis and prediction of air quality. International Journal of Engineering Research & Technology (IJERT), 10(3), 533-538.

[23] Talamanova, I., & Pllana, S. (2022). Data-driven Real-time Short-term Prediction of Air Quality: Comparison of ES, ARIMA, and LSTM. arXiv preprint arXiv:2211.09814.

[24] Zafra, C., Ángel, Y., & Torres, E. (2017). ARIMA analysis of the effect of land surface coverage on PM10 concentrations in a high-altitude megacity. Atmospheric Pollution Research, 8(4), 660-668.