Feasibility Study of ARIMA Model for PM2.5 Prediction using Real-world Data Gathered from Pune Region

Aarohi Sudumbrekar
P.E.S Modern College of Engineering, Pune, India

Rutuja Kale
P.E.S Modern College of Engineering, Pune, India

Tanvi Kaurwar
P.E.S Modern College of Engineering, Pune, India

Vaishnavi Mule
P.E.S Modern College of Engineering, Pune, India

Anita Devkar
P.E.S Modern College of Engineering, Pune, India

Corresponding author: Rutuja Kale, Email: kalerutuja2017@gmail.com

Time and time again, air pollution has proven to be a formidable problem which needs to be tackled now more than ever. The increased levels of pollutants have significantly affected the respiratory health of Indian citizens. One such major pollutant is PM2.5. Consequently, it is pertinent that a good methodology be created to predict the PM2.5 values. In this paper, the ARIMA method of time series analysis is used to predict the values of P.M2.5. Thus, the viability of this approach has been assessed and the results have been put forth.

Keywords: ARIMA, PM2.5, time-series analysis, air pollution.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 105–111. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-13

1 Introduction

Today, Air Pollution prevention and management has become one in all the foremost difficult issues that a country will face. The impact of air pollution will be detrimental to the individuals, and even the infrastructure of any country. Air can be contaminated by a spectrum of various particles such as mud, soot, pollen smoke, and liquid droplets, several of which can be harmful to human health.

PM2.5 is considered to be a major factor responsible for air pollution. PM2.5 constitutes air pollution particles that have a diameter less than 2.5 micrometers. These particles tend to be suspended in the air for a longer period of time and can be responsible for both short-term and long-term health problems. A prolonged exposure to PM2.5 can cause severe and long-lasting respiratory issues like asthma, chronic bronchitis and heart disease. In order to mitigate and regulate the potential health effects of PM2.5, accurate prediction of PM2.5 is necessary.

In this paper, our aim is to discuss a probable algorithm for predicting the future values of PM2.5. For this, the time series analysis method that we have used is the autoregressive integrated moving average (ARIMA). This algorithm analyzes the sequence of historical data in a specific time frame to set up the forecasting model. This model aims to foretell a series that may not be deterministic due to the presence of an arbitrary component in the data set. Assuming this component is stationary, methods can be developed for accurate prediction of future values. Previously, the ARIMA model has been studied substantially and has proven to be efficacious for forecasting values [4]. Similar studies have been conducted previously, giving a detailed analysis of specific sources of pollution (domestic heating, factories, road traffic, etc.) [5]. In one such similar study[5], the area into consideration isa town in Bulgaria called Blagoevgrad and the data used is gathered by Executive Environment Agency, the study is mainly focused on time-series analysis and the development of univariate seasonal autoregressive integrated moving average (ARIMA) models. Whereas in our paper the study area is the city of Pune in Maharashtra, India. Currently Pune has 15 air monitoring stations controlled by Maharashtra Pollution Control Board (MPCB), IITM (through its System of Air Quality and Weather Forecasting and Research, SAFAR), and Pune Smart City Development Corporation. Although these organizations provide a fairly accurate overview of the air pollution levels and weather conditions, the data used by them for these purposes is not easily accessible by ordinary people for conducting their studies and research. That being the case, we have used data obtained from a citizen science project called BREATH2.

The main question considered in this study is : Using the ARIMA model, if simply the previous values of PM2.5 are used to predict its future values and all other factors affecting its concentration in the air (for example: traffic, weather, temperature, humidity, etc) are ignored, whether the ARIMA model gives us fairly accurate results.

2 Problem statement

In this paper, we are aiming to develop a solution to predict PM2.5 values based on real-world data collected from machines that are present at different locations across Pune city. The forecasting of PM2.5 values can be modelled as a classic univariate time series forecasting problem, where at any time T the objective is to predict PM2.5 values using the past values. In this paper, the aforementioned time-series forecasting problem has been explored using the ARIMA model.

3 Data

The data is collected using the Breathe2 air pollution (PM) devices. The five devices which are placed across Pune city (IMD Shivaji Nagar, Indradhanushya Hall Mhatre Bridge, Bapat Hospital FC Road, MIT Kothrud and SamuchitEnviroTechPvt.Lmt. Law college road) measure air pollution in a distributed networked format. Every 5-10 minutes the data gets collected and updated by the machines and is fed onto an IOT platform called Things peak. This data exists from the beginning of October 2019, however for our model we have considered the latest 8000 data entries with respect to each machine over the span of 2 years.

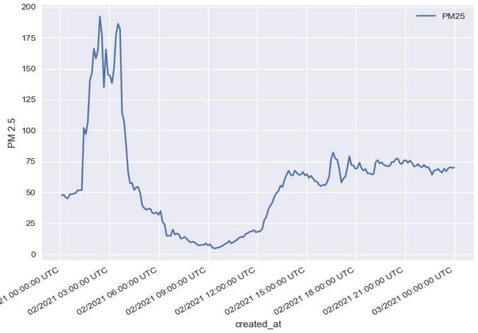

Fig. 1 Values of PM2.5 with respect to time for one day for Indira hall Region

4 Study Area

The area into consideration is the metropolitan city of Pune. Lying on the western margin of the Deccan plateau, at an altitude of 560 m (1,840 ft) above sea level, it is shielded by the Sahyadri mountain range, and its approximate location is 18° 32" north latitude and 73° 51" east longitude. Being a hilly city, it has a tropical wet and dry climate and average temperature range is between 20 to 28 °C. According to the 2011 census its population is estimated to be 3.12 million. Its various areas are characterized by heavy traffic and in the last five years, the air quality of Pune has exacerbated significantly.

5 Prediction Model

5.1 Importing Data: The data is stored in csv format on the device. To access this data for our application, a CSV module is used. The CSV module explicitly handles this task which makes dealing with CSV formatted files much easier. This becomes chiefly important when working with huge amounts of data, with various types of parameters.

5.2 Pre-processing of Data: After importing the data, the main objective is to make the data suitable enough to be used in a machine learning algorithm. For this, the outliers, inliers, NA values (the cells that are empty) are all to be found out and replaced with legible values, or simply removed. Such values can be replaced with the average value of that parameter.

5.3 Autocorrelation of Data. Autocorrelation is found by the ACF function. It gives us values of auto-correlation of any series with its lagged values. In simpler terms, it represents how well the present value of the series is related with the past values. A time series may have components like trend, seasonality, cyclic and residual. ACF takes into consideration the aforementioned components while finding correlations, hence, it is known as a complete auto-correlation plot. Figure 1 describes the ACF plot for the data being used. It can be concluded from the plot that the data is not white noise, about which more has been discussed in the further sections.

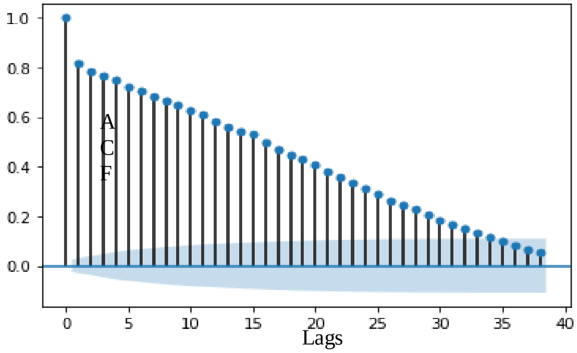

Fig. 2. Autocorrelation

5.4 Partial Autocorrelation of Data: Partial autocorrelation is found by PACF function. The partial autocorrelation function might be construed as a regression of the series against its past lags. So, instead of finding correlations of present with lags like what ACF does, it finds correlation of the residuals with the next lag value hence it is 'partial' and not 'complete'. So, in any case of hidden information in the residual which can be doctored by the succeeding lag, a good correlation can be achieved.

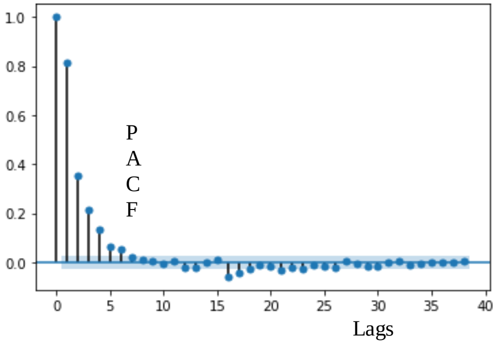

Fig. 3. PartialAutocorrelation

Figures 2 and 3 are examples of the ACF and PACF plots that were drawn with respect to the data we have used. These plots are with respect to only one region of the five regions that we have taken under observation.

5.5 White Noise. A time series is said to be white noise if the variables are independent and have an identical distribution with a mean of zero. It means that all variables have the same variance and every value has a zero correlation with all remaining values in the series. If the time series is white noise, then it is random. If it is so, then one cannot reasonably model and predict on such values. The data used here is not white noise, which is proved by the ACF plot.

5.6 Checking the Stationarity of data. A time series with constant mean and zero variance is considered to be a stationary data. To check if the data is stationary or not, the augmented Dickey–Fuller test (ADF) is used. This is used to test the null hypothesis whether a unit root is present in a time series sample or not. If the data is found to be stationary, the ARIMA model can use the differencing method to make the data non-stationary. The null hypothesis of the test says that the time series can be constituted by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (i.e. rejecting the null hypothesis) tells us that the time series is stationary. We can interpret this result using the p-value obtained from the test. A p-value below 0.05 conveys that the null hypothesis can be rejected, otherwise a p-value greater than 0.05 conveys that the null hypothesis cannot be rejected.

Table 1. Tests conducted on the gathered data

Name of Test	Value of the test
ADF statistics	-11.929742299906758
p-value	4.812554386574585e-22
Critical value at 1%	-3.4305433135162935
Critical value at 5%	-2.861625439704114
Critical value at 10%	-2.566815476885392

Running the test, the statistical value is -11. The more negative this value, the more likely we are to reject the null hypothesis, as we have a stationary dataset. In the above table, the statistic value of -11 is less than the value of -3.4305 at 1%. This propounds that the null hypothesis can be rejected with a significance of level less than 1%. In simpler words, there is a low probability that the result is a statistical fluke.This also means that the process has no unit root, and in turn that the time series is stationary in nature.

6 Machine Learning Module

ARIMA Models: An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of nonstationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity. ARIMA model is of the form: ARIMA (p,d,q): p is AR parameter, d is differential parameter, q is MA parameter.

Table 2. Comparison of actual and predicted values

DATE	PREDICTED VALUE	ACTUAL VALUE
Bapat Hospital (RMSE: 11.256)		
2020-05-06 23:49:18	3.519652	7.72
2020-05-06 23:41:57	3.518564	6.79
2020-05-06 23:34:36	3.517473	4.72
MIT (RMSE: 9.187)		
2020-02-12 02:39:31	37.757742	31.76
2021-07-01 00:26:43	37.899516	30.69
2021-07-01 00:19:03	38.099778	26.03
Samuchit Enviro Tech Pvt Lmt.(RMSE: 12.873)		
2021-02-06 21:25:58	19.628332	17.58
2021-02-06 21:16:20	19.875173	20.6
2021-02-06 20:51:36	20.111567	15.88
Indradhanushya Hall (RMSE: 1.515)		
2020-03-22 11:09:46	9.112476	9.86
2020-03-22 11:02:08	9.053799	9.68
2020-03-22 10:54:15	8.995122	9.15
IMD (RMSE: 0.469)		

2021-06-15 19:08:50	1.491459	1.34
2021-06-15 18:54:39	1.490703	1.7
2021-06-15 18:47:16	1.490078	2.94

7 Results and its interpretation

As seen from the table above, the results acquired are satisfactory, but not exactly accurate. It could be so because it has been found that the PM2.5 concentration depends upon a multitude of factors [1, 2]. These factors can be classified into categories such as human and natural upon which the PM2.5 concentration correlates either negatively or positively [2]. In India the majority of PM2.5 concentration is contributed by human activities, such as residential biomass combustion, powerplant and industrial coal combustion, burning of waste, transportation and distributed diesel [1]. Also, social factors, like burning firecrackers during the Diwali festival, play a significant part towards heightened concentration of PM2.5 levels [3]. With respect to the natural factors, it has been found that the levels of PM2.5 concentration increases with an increase in the factor (for example the relative air humidity), whereas decreases with a decrease in that factor (for example wind velocity, rainfall, air temperature, soil temperature and soil humidity) [1]. However, our univariate time series model considers the final concentration of PM2.5 in the atmosphere to be absolutely dependent on its former values and ignores its other interdependencies with the above-mentioned factors.

8 Conclusion

This paper aims to find whether future PM2.5 values can be predicted solely based on its own preceding values using a time series analysis model or not. In this study, the developed ARIMA model that uses the data collected from the Breathe2 device is taken under consideration. The anomalies from the data are removed during the pre-processing of the data. Autocorrelation and Partial autocorrelation, which determine whether the data is white noise or not, is found out. These relations determine that all variables do not have the same variance and every value has a non-zero correlation with the remaining values in the series. Then, after applying the ADF test on the intended dataset, stationarity of the datasets was found. Since, stationarity is desirable in the datasets, no further changes were required. The model was hence created using the pre-processed dataset, and values were predicted for PM2.5.

Root mean square error for the obtained data was calculated, and found out to be substantial enough to be not ignored. This suggests that the results obtained were not accurate enough. Therefore, we can conclude that when a prediction model is designed based on a univariate time series model, where only PM2.5 values are used to predict its future values, the results are not exactly accurate. This could suggest that the values of PM2.5 may depend upon other external factors, like anthropological and natural factors, as discussed in the previous sections.

Another conclusion that can be drawn from the above study is that the model alone cannot guarantee proper working of a prediction model. For a prediction model to give accurate results, one must also consider the complexity of data, for the data being used should be in accordance with the model and should also meet its needs.

Acknowledgement

We would like to thank our institution, P.E.S Modern College of Engineering, Pune for the opportunities provided. We would also like to extend our gratitude to our guide Prof. Anita Devkar for her continuous support and guidance during the course of our project.

References

[1] C. Venkataraman et al., “Source influence on emission pathways and ambient PM2.5 pollution over India (2015-2050)”, *Atmos. Chem. Phys.*, vol. 18, no. 11, pp. 8017-8039, 2018.

[2] Y. Zhang and W. Jiang, “Pollution characteristics and influencing factors of atmospheric particulate matter (PM2.5) in ChangZhu-Tan area”, *IOP Conf. Series: Earth and Environ. Sci.,* vol. 108, 2017.

[3] Y. Chena et al., “Local characteristics of and exposure to fine particulate matter (PM2.5) in four indian megacities”, *Atmos.Environ.: X*, vol. 5, 100052, 2019.

[4] R. Nimesh et al., “Predicting air quality using ARIMA, ARFIMA and HW smoothing”, *Model Assist. Stat. Appl.*, vol. 9, pp. 137-149, 2014.

[5] G. Ilieva et al., “Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach”, *Stoch. Environ. Res. risk assesst.,* vol. 18, no. 4, pp. 1045-1060, 2015.