Analysis Of Air Pollutants Affecting The Air Quality Using ARIMA

Akanksha Bhosale[1], Karuna Chaudhari[2], Komal Andhe[3], Farhat Naik[4], Sonal Chaudhari[5]

[1]Student, Computer Engineering, Datta Meghe College Of Engineering, Navi Mumbai, India [2]Student, Computer Engineering, Datta Meghe College Of Engineering, Navi Mumbai, India [3]Student, Computer Engineering, Datta Meghe College Of Engineering, Navi Mumbai, India [4]Student, Computer Engineering, Datta Meghe College Of Engineering, Navi Mumbai, India
[5]Assistant Professor, Computer Engineering, Datta Meghe College Of Engineering, Navi Mumbai, India

---

Abstract - *The reason for human life or any other living organism is air. However, this air is getting polluted because of a variety of factors, including traffic situations, industrial development and various construction businesses. Because of these factors, the quality of air is decreasing day by day, and all the life forms who depend on this air are gettingheavilyaffected. There are various factors who influence the quality of air, and it is measured by the Air Quality Index (AQI). NO2, CO, C6H6, SO2, and CO2 are some of the components which affect the quality of our air. The aim of our paper is to predict and forecast the AQI by using time series generalized models such as the Auto-Regressive Integrated Moving Average (ARIMA) model. The time series data collected has a lot of missing and corrupt values, and hence, it is subjected to cleaning, modification and aggregation as the requirement arose. The data is then checked for its stationarity, by performing various tests, and then the model is deployed. Prediction was performed on the aggregated data using ARIMA.*

***Key Words*: ARIMA, time-series data forecasting, Moving Average model, Auto-regressive**

1.INTRODUCTION

Air is one of the most crucial natural resources for all life on this planet's survival. Every life form depends on air for their existence and hence all the living beings require good air quality which is free from harmful gases for their existence. According to the Blacksmith Institute, two major pollution problems in the world are outdoor city air quality and indoor air pollution[1].

Air quality forecasting is conducted to obtain advanced knowledge of the air environment and to take preventative measures to avoid health problems. Pollution, which is found both indoors and outdoors, is causing the quality of air to deteriorate in emerging and even developed countries all over the world. Air pollution causes short-term and long-term health issues, mostly affecting the elderly and young children[2]. Short-term issues can include throat irritation, headaches, upper respiratory infections, and other short term but dangerous issues. Lung cancer, kidney damage, respiratory disease, heart disease, and brain damage are some of the long-term effects on health due to air pollution. Air pollution also causes depletion of the ozone layer, which is a major issue as it protects everyone from the sun's harmful UV rays [3]. Another harmful result of air pollution is acid rain, which affects rivers, trees, wildlife, and soils. Some of the environmental repercussions of air pollution include eutrophication, global warming, and haze. With the advent in technology, we aim to predict and forecast the Air Quality Index (AQI) by using unsupervised machine learning techniques. Primarily, Auto - Regressive Integrated Moving Average (ARIMA) model is used for air quality analysis.

The ARIMA model known as the auto-regressive integrated moving average is a model that includes two processes that are MA and AR. For developing the ARIMA(p,d,q) model both PACF and ACF auto- correlation functions are very useful.

1.1 Motivation

Air quality management has been acknowledged as a key issue at both national and local levels. In the past few years, research has been undertaken to identify the significant challenges because it is critical to ensure the health and cleanliness of the local surrounding and the community. It has been ascertained that different topic such as technological advancement, detail procedure regarding the measurement of air quality, determination of pollutant variable and its interconnection, identifying the causes and effects of air pollution and lastly forecasting of periodic and geographical variations in atmospheric levels are some of the topics that has been covered in regard to the study of air quality. Although, it has been investigated that developing countries are facing these air quality issues due to insufficient funds, and technological support.

Air pollution in a city requires immediate attention as a large number of people live in the city, and hence, air pollution may affect more people. There is an immediate need of strict laws and constant monitoring of air pollutants for air pollution management[4]. For this reason, our paper focuses on a prediction model that will help us with prediction of the air pollution.

2. LITERATURE REVIEW

A paper presented a neural calibration for benzene concentration prediction utilizing a gas multi-sensor system (solid-state) developed to measure urban pollution. The results are evaluated and analyzed using prediction error characterization during a 13-month period. The relationship between training duration and efficiency is also being

investigated. A neural calibration acquired with a modest number of measurement days were found to be competent at limiting the relative prediction error for more than 6 months, after which periodic impacts on prediction capabilities at low concentrations highlighted the need for a new calibration.

Another method was proposed for employing time- series data to study the air pollution indicator (API). The authors used the Box–Jenkins method to analyze the time series and focused on analyzing changing air pollution levels in Bangalore (INDIA) from January 2013 to March 2016 [4].A paper mentioned the use of a random forest classifier model for data forecasting in Beijing in 2019 to study and analyze the data and use it for prediction.

In another paper, air quality monitoring is classified by big data techniques into spatial, temporal, spatial- temporal models. Big data techniques that are needed in air quality forecasting are summarized into three folds, which are statistical forecasting model, deep neural network model, and hybrid model, presenting representative scenarios in some folds [2]. The air pollution traceability methods were analyzed and compared in detail, classifying them into two categories: traditional model combined with big data techniques and data driven model.

Interpolation, prediction, and feature analysis are three significant subjects in urban air computing used in a paper in 2019. Due to the costly cost and maintenance of city stations, little air quality data was obtained. They utilized the K-Means algorithm and supervised machine learning methods such as the Decision Tree algorithm and Multinomial Regression. This paper mainly contained the analysis of air pollution and prediction of air pollutants using machine learning algorithms. The proposed system was to build an algorithm in machine learning which would be a best fit model to analyze and predict the air pollutants [6]. The result told us that the MLR model was much better than the DT model by providing us with greater accuracy and being a best fit model.

3. REQUIREMENT ANALYSIS

The data set used in the project is a time series data. The data which is recorded over specific and consistent intervals of time is called as a time series data. The data which is obtained may be corrupted, and have missing values, which would affect the model. To prevent it, the data needs to be pre-processed. In case of any missing values, the mean value of the statistical measure is used to substitute with the missing values.

The data which is pre-processed, needs to be stationary before it is used. Stationarity means when the statistical characteristics of the time series data, such as mean and variance, do not change over a period of time. Stationarity helps us to understand the data better and choose a better range of values for model. If the data is not stationary, we perform various tests to achieve stationarity.

4. PROJECT DESIGN

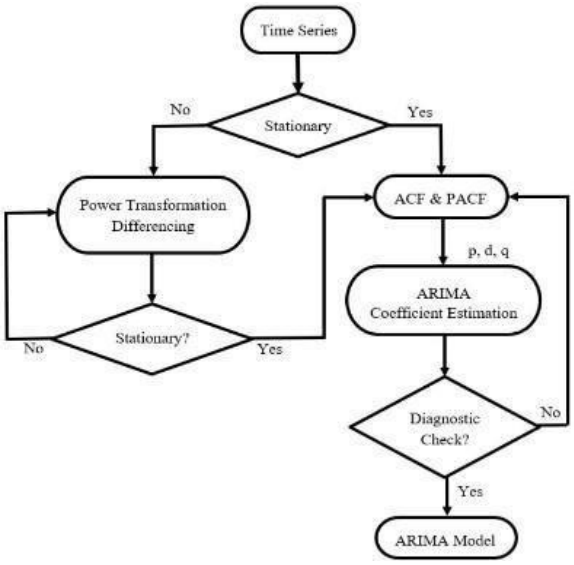

Chart-1: Flowchart of ARIMA

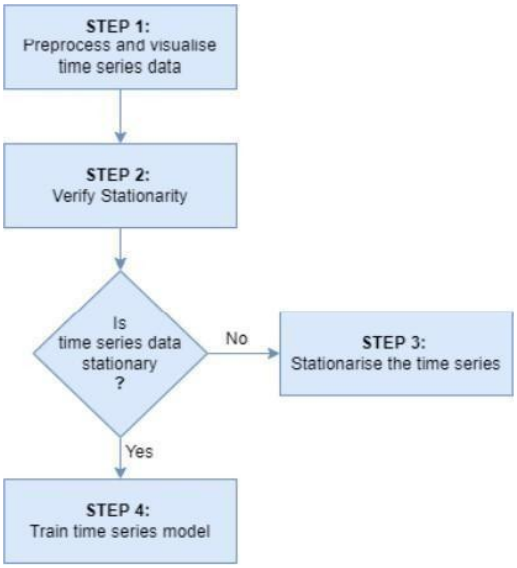

Chart-2: Flowchart of Time-Series Model Sequence

The process of project is as follows :

1. The data which is pre-processed, is our time series data. We check if this data is stationary or not.

2. The stationarity of data can be achieved by using techniques such as Rolling Statistics and Augmented Dickery- Fuller (ADF) test.

3. If the data is stationary, we can perform time series forecasting[7]. We use functions like ACF and PACF, which gives us the hyperparameters for ARIMA.

4. If the data is not stationary we can perform data differencing or log transformation to manage the data and to make it stationary as most time-series models assume that the data is stationary.

5. We use power transformation differencing to obtain stationarity of our data. The order of differencing depends on how many times we perform the test till we get stationary data.

6. The three hyperparameters of ARIMA are p,d and q. p stands for the moving average part which is obtained from the moving average component. This is acquired from the ACF graph. p is the autoregressive lag that comes from the autoregressive component, and is obtained from the PACF graph. d is the order of differentiation that we use to convert our non-stationary data to stationary data.

7. After our data is stationarized, we deploy our ARIMA model.

5. TECHNOLOGIES USED

Python :

Python is one of the widely used, and a user friendly programming language used for building websites and projects, as well as conduct data analysis. The unique design of this high level programming language is that it allows us to reuse code quite easily[8]. Python is not limited to any specific problems and can be used for making an array of unique and dynamic projects, and programs. It is dynamically-typed and garbage collected.

Google Colab :

Google Colab, also known as Colaboratory, is used to write and execute various programs and projects in the python programming language. Colab is a product of Google, included in the Google Research. It helps us to perform any machine learning algorithms, as well as any high level programming, along with data analysis[9]. Colab can be used to teach python as well, as it is free of charge and we can access it easily from our browser.

6.RESULTS

Fitting the model :

We have changed the frequency of our data into daily so as to make our model more fit. We can have hourly frequency as well but it has the disadvantage that the trends and stationary values are difficult to acquire.

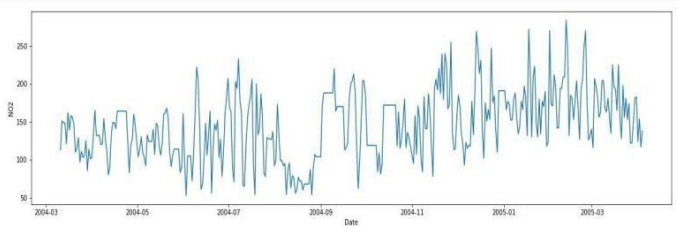

Fig-1: NO2 concentrations for daily averaged frequency

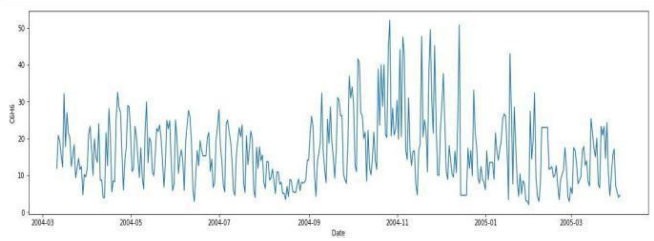

Fig-2: C6H6 concentrations for daily averaged frequency

After deciding the frequency and the decomposition of the data, we will next check for the stationarity of the data.

Stationarity :

In case of time series forecasting, there is a need to check if the time series data is stationary or not because the model requires stationary data for forecasting. Stationarity means that the mean and variance of the data should not vary over time, it should be constant. Stationarity of the data can be confirmed by making use of the Rolling Statistics test and Augmented Dickery Fuller (ADF) test.

Rolling Statistics Test :

In rolling statistics test, we create a window of a specific size and then the calculations are performed in this window. The calculations performed in the window are rolled through the data. This test is used to check the stability of the model over time.

Rolling Mean

Date	NO2(GT)
2004-03-10 18:00:00	NaN
2004-03-11 18:00:00	NaN
2004-03-12 18:00:00	NaN
2004-03-13 18:00:00	NaN
2004-03-14 18:00:00	NaN
2004-03-15 18:00:00	NaN
2004-03-16 18:00:00	140.428571
2004-03-17 18:00:00	146.857143
2004-03-18 18:00:00	147.571429
2004-03-19 18:00:00	147.142857
2004-03-20 18:00:00	141.714286
2004-03-21 18:00:00	141.000000
2004-03-22 18:00:00	136.285714
2004-03-23 18:00:00	130.285714

Fig-3: Rolling Mean of NO2

```
Rolling Standard Deviation
                     NO2(GT)
Date
2004-03-10 18:00:00      NaN
2004-03-11 18:00:00      NaN
2004-03-12 18:00:00      NaN
2004-03-13 18:00:00      NaN
2004-03-14 18:00:00      NaN
2004-03-15 18:00:00      NaN
2004-03-16 18:00:00  17.510541
2004-03-17 18:00:00  13.582201
2004-03-18 18:00:00  13.962535
2004-03-19 18:00:00  13.957418
2004-03-20 18:00:00  19.754445
2004-03-21 18:00:00  20.696215
2004-03-22 18:00:00  18.785760
2004-03-23 18:00:00  23.809762
```

Fig-4: Rolling Standard Deviation of NO2

We have taken the rolling window size as 7, which is a week value, but it can be adjusted accordingly. After taking the rolling mean, and rolling standard deviation of the remaining parameters, we perform the ADF test. We perform the ADF test to cross verify the statisticity of the data.

Augmented Dickery Fuller (ADF) Test :

In ADF test, we approach with a null hypothesis, assuming that our data is not stationary. We get a few values after performing the test. On the basis of those values, if the test result is less than critical value then we can reject our null hypothesis and state that our data is stationary.

```
ADF Statistic: -3.2302847715324456
p-value:  0.018297253960069745
Critical Values:
1% : -3.4479562840494475
5% : -2.869299109917524
10% : -2.57090345105665
```

Fig-5: Result of ADF test

As we can see above, we can see that the p-value is less than 0.05, which states that we can reject our null hypothesis and accept that our data is stationary. Once our data is stationary, we can deploy the ARIMA model. In case if the data is not stationary, we can perform data differencing or log transformation to manage the data and stationarize it since we require stationary data for our forecasting.

Deploying the ARIMA Model :

The ARIMA model has three main parameters - p – this is called the auto regressive lags which we get from the auto regressive component of the model. We can obtain this parameter using the PACF (partial autocorrelation function) graph. d – this is called order of differentiation which is the order of differentiation which is required to convert non stationary data to stationary data. q – this is called the moving average parameter which we get from the moving average component of the model, we can obtain this parameter using ACF (auto correlation function) graph. First, we use the predict function and get the following results -

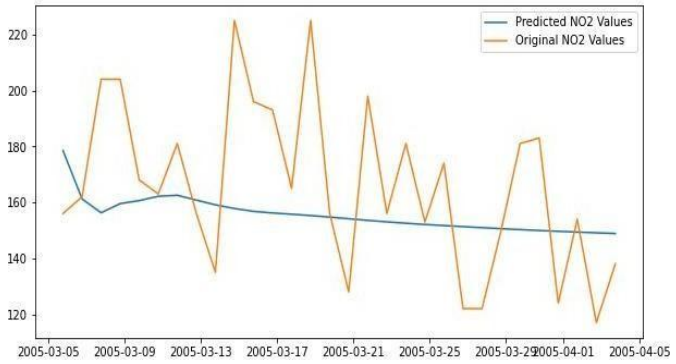

Fig-6: Predicted and Original NO2 values after applying ARIMA model without determining p, d, q values.

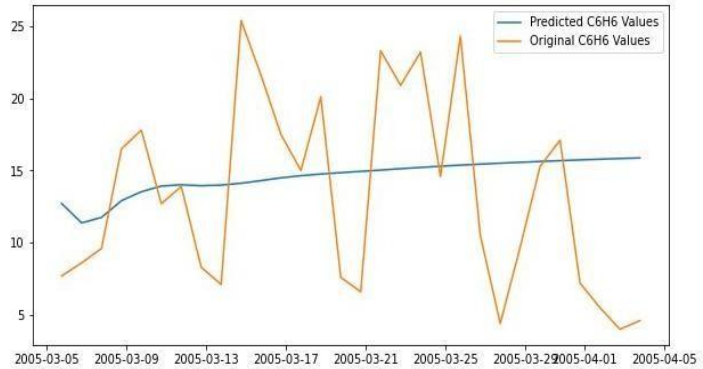

Fig-7: Predicted and Original C6H6 values afterapplying ARIMA model without determining p, d, q values.

We can see that the model is not performing well and the predicted values are very different with the actual values. As the model did not perform well, we need to analyse the accuracy metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

After recalculating the parameters and accuracy metrics we have a new graph with predicted values quite similar to original values with the p=5 d=0 and q=1 which are applied to all the parameters.

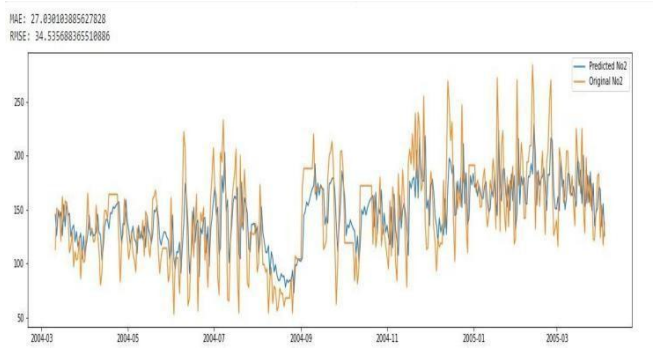

Fig-8: Predicted and Original values of NO2

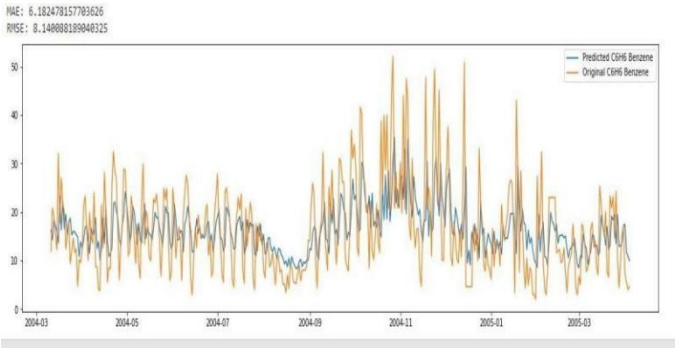

Fig-9: Predicted and Original values of C6H6

We can see that the accuracy of the model has improved drastically, once we use the p, d and q values, which are the hyperparameters of the ARIMA model. We can raise the accuracy by changing the parameters p, d, q accordingly as the data requires.

ARIMA Future Forecast :

On the basis of the train data, we can forecast the future values using ARIMA model. Using the parameter values and the accuracy metrics we can forecast for a whole year and use this forecasting to understand how this can impact the environment in the upcoming years. Currently, we have forecasted for only one year as we have only yearly amount of data.

```
2005-04-04 18:00:00    11.333033
2005-04-05 18:00:00    13.490868
2005-04-06 18:00:00    13.450862
2005-04-07 18:00:00    12.946229
2005-04-08 18:00:00    11.929580
2005-04-09 18:00:00    11.743363
2005-04-10 18:00:00    12.197085
2005-04-11 18:00:00    12.744811
2005-04-12 18:00:00    13.227583
2005-04-13 18:00:00    13.489576
2005-04-14 18:00:00    13.587067
2005-04-15 18:00:00    13.658820
2005-04-16 18:00:00    13.762566
2005-04-17 18:00:00    13.911602
2005-04-18 18:00:00    14.082667
2005-04-19 18:00:00    14.242757
Freq: D, dtype: float64
<Figure size 720x360 with 0 Axes>
```

Fig-10: Forecasting for NO2 concentration

```
2005-04-04 18:00:00    153.477264
2005-04-05 18:00:00    148.090201
2005-04-06 18:00:00    149.974908
2005-04-07 18:00:00    148.738107
2005-04-08 18:00:00    145.199518
2005-04-09 18:00:00    145.191795
2005-04-10 18:00:00    145.500353
2005-04-11 18:00:00    145.638852
2005-04-12 18:00:00    146.161360
2005-04-13 18:00:00    146.280792
2005-04-14 18:00:00    146.136716
2005-04-15 18:00:00    146.030059
2005-04-16 18:00:00    145.912900
2005-04-17 18:00:00    145.842850
2005-04-18 18:00:00    145.836065
2005-04-19 18:00:00    145.835529
Freq: D, dtype: float64
<Figure size 720x360 with 0 Axes>
```

Fig-11: Forecasting for C6H6 concentration

Final prediction for forecast and original values is done using ARIMA model with the predicted values. The p=5, d=0, q=1 and the parameters are set the same. To get more accuracy and exact forecasting of the data, the accuracy metrics of MAE and RMSE would have to be optimized using differentiation. We can further optimize the forecast by increasing the differentiation and accuracy metrics.

7.FUTURE SCOPE

The prediction model can be improved by strengthening the methods to forecast the concentration of air quality factors, majorly for O3, as O3 does not come from direct sources but due to multiple sources of emission and their reaction to each other. There are multiple time series models which can be used for this. The time series data can be collected for two or three years, or more than that, and we can work on that data in order to make more accurate predictions.

8.CONCLUSIONS

ARIMA model is suitable for short-term predictions because with the help of stationary data, accurate predictions can be made. Time series model used in forecasting is an important tool which helps us to control, analyse and monitor the air quality condition. It is useful to take quick action before the situation worsens in the long run [10]. For that reason, we need our model performance to be as accurate as possible so that good air quality forecasting can be achieved. Moreover, the pollutants must be considered in analysis of air pollution data.

REFERENCES

[1] Blacksmith Institute Press Release'. (October 21, 2008). [Online]. Available: http://www.blacksmithinstitute.org/the-2008-top- ten-list-of-world-s-worst-pollution-problems.html

[2] Huang, W., Li, T., Liu, J., Xie, P., Du, S. and Teng, F. (2021) An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability. Information Fusion 75, 28–40.

[3] Baralis, E., Cerquitelli, T., Chiusano, S., Garza, P. and Kavoosifar, M.R. (2016) Analyzing air pollution on the urban environment. 1464 1469https: //ieeexplore.ieee.org/abstract/document/7 522370 Accessed 14 May 2022.

[4] Abhilash, M.S.K., Thakur, A., Gupta, D., Sreevidya, B. (2018). Time Series Analysis of Air Pollution in Bengaluru Using ARIMA Model. In: Perez, G., Tiwari, S., Trivedi, M., Mishra, K.(eds) Ambient Communications

and Computer Systems. Advances in Intelligent Systems and Computing, vol 696. Springer, Singapore. https://doi.org/10.1007/978-981-10-7386-1_36

[5] De Vito, S., Massera, E., Piga, M., Martinotto, L. and Di Francia, G. (2008) On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. Sensors and Actuators B: Chemical 129 (2), 750–757.

[6] Nandini, K. and Fathima, G. (2019) Urban Air Quality Analysis and Prediction Using Machine Learning. 98–102 https://ieeexplore.ieee.org/document/9063845 Accessed 14 May 2022.

[7] Cheung, Y., and Xu, L. (2001). Independent component ordering in ICA time series analysis. Neurocomputing, 41, 145–152

[8].https://en.wikipedia.org/wiki/Python_(programming_language)

[9].https://research.google.com/colaboratory/faq.html [10]. Lee, M. H., Rahman, N. H. A., Suhartono, Latif, M. T., Nor, M. E. & Kamisan, N. A. B. (2012). Seasonal ARIMA for Forecasting Air Pollution Index: A Case Study. *American Journal of Applied Sciences*, *9*(4), 570-578. https://doi.org/10.3844/ajassp.2012.570.578.