Time Series Analysis and Forecasting of Air Quality in India

Vanshay Gupta
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
vanshaygupta7@gmail.com

Samit Kapadia
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
samitkk18@gmail.com

Chetashri Bhadane
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
chetashri.bhadane@djsce.ac.in

2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT) | 978-1-6654-9360-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICECCT56650.2023.10179673

***Abstract*—This paper aims to analyze the air quality in India and the effects of seasons and COVID-19 on the concentration of pollutants in the air and thereby their effect on the air quality index (AQI). The analysis is performed on a full scale, taking into consideration different levels of granularities such as daily, weekly and monthly data. This study performs extensive preprocessing of the time series data for air quality to make it output the best results. The results evidenced that particulate matter i.e., PM 2.5 and PM 10 have the greatest impact on air quality. Analysis of the effect of change in seasons on the overall air quality has been carried out, along with the impact of the nationwide lockdown due to COVID-19, which led to a substantial improvement in the AQI levels. Furthermore, we also use the state-of-the-art forecasting algorithm Prophet to predict the monthly average air quality index and compare it with the actual recorded values, giving us a highly accurate prediction. We also performed a comparative analysis of AQI for the cities of Delhi and Bengaluru, having different seasons and climates, which results in valuable insights on to what extent the environmental factors affect the air quality measures of that location.**

Keywords—Machine Learning, Time Series, Air Quality Index, COVID-19, Seasons, Prophet

I. Introduction

The air quality of metropolitan cities of India has worsened in the past decade due to the rapid increase in urbanization and industrialization. The heavy use of vehicles, machinery for construction, large factories, and manufacturing plants emit a wide range of particulate matter with traces of gases such as SO2, CO, NO2 and NH3. Thus, analysis of air quality is a very important and often researched topic. The trends and patterns obtained from the data are helpful in finding correlations of air quality measures with many daily activities. These variations of concentrations of pollutants and overall air quality measured at different granularities help researchers spot discrepancies in data. Further research also delves into specific factors such as seasons, festivals, and forecasting of AQI by giving a rough estimation of the AQI in the future which can be helpful in taking early precautions.

With the advent of COVID-19 and its rapid transmittable rate, the government initiated a nationwide lockdown, which resulted in a major halt in day-to-day activities. This stoppage resulted in the improvement of air quality in metropolitan cities, which has been proved by satellite images. It also helps us find new correlations of pollutants with daily activities, for example, due to very few cars on the roads the emission of carbon monoxide has reduced, thus showing a drastic change in its concentration during the COVID-19 lockdown. This data could help the government make policies to reduce carbon monoxide concentration by regulating the number of cars on the roads.

While air pollution is an issue all year round, different seasons see an increase in certain pollutants due to various factors such as temperature, climate, and human activities. Analysis with the season as a parameter gives us valuable insights. Winters are often accompanied by smog caused due to temperature inversion, this causes gaseous pollutants such as carbon monoxide, nitrogen oxides, particulate matter, and volatile organic compounds to be trapped at the ground level until the temperature changes. Such changes in temperature, wind flow, and velocity, rainfall patterns in different cities, duration of the season lead to disparity in the air quality data.

Forecasting of air quality is important for cities in helping them take preventive measures and plan ahead to protect their health. Predicting air quality helps increase awareness amongst people of the air they breathe in, the effect of pollutants on health as well as concentrations likely to cause adverse effects and actions to curtail pollution. When predicting air quality, there are many variables to take into account, some of which may be quite unpredictable.

II. Related Work

Analysis of air quality is very important and a lot of innovative work has been done on it before. [1] evaluates the concentration of pollutants using the Comprehensive Pollution Index Method, Euclid Approach Degree Method, and the Improved Grey Relational Degree Method based on the data obtained from the city of Shanghai.

Other Seasonal Trends are very essential for the concentration of pollutants and thereby necessary for the air quality index. [2] accounts for two seasons winter and summer and shows the variations in the concentration over the two seasons. [3] is on the same basis but takes it a step further, by dividing the year into four seasons. They plot regression curves of PM 2.5 and PM 10 mass concentration over a time series data.

COVID-19 has had a major impact on air quality. [3] is a study of time series data of India's four megacities from 2013 to 2019 and compares the concentration of pollutants with the data collected during the lockdown. [4, 5, 6, 7] show the positive impact the pandemic has on the air pollutants concentration, thus improving the air quality index and bringing it to ambient air quality level. [4] further dwells down into analyzing in-depth the impact different levels of lockdown have had on the concentration of the pollutants in the air.

Time series-based forecasting of air quality data can provide effective data on the concentration and patterns of pollutants and identification of uncertainties in advance. [8] puts forth an early warning system that uses a hybrid model based on fuzzy time series, which performs interval forecasting based on deterministic predictions to identify uncertainties in the concentration of air pollutants. The Deep Air Learning (DAL) system in [9] provides an integrated solution by utilizing the intrinsic characteristics of sapio-temporal information. [10] proposes a novel feature selection technique that integrates the decomposition techniques and the optimization algorithm to remove the noise and selects the optimal input structure.

The past research in this domain focuses largely on the analysis of one particular factor affecting the concentration of pollutants in the air and thereby the effects on the Air Quality Index (AQI). Time series data for air pollutants often has a lot of missing values due to faulty equipment or equipment under maintenance not being able to record the data. Though their works have valuable and novel insights on the topic, often dropping missing values causes discrepancies or discontinuity in the time series data, thus producing less accurate outputs.

III. Results and Analysis

A. Preprocessing

The dataset used in this paper has been formed from data that has been made publicly available by the Central Pollution Control Board of India. The methods involved in missing data imputation included forward fill and backward fill. The method with the least mean absolute error was selected to represent the data as accurately as possible. 200 known values were removed from each column in addition to the missing values, to create a validation set in order to understand which imputation technique would be ideal for the particular attribute.

TABLE I. Mean Absolute Error on Imputation of Missing Values in the Dataset

Pollutant	Forward fill	Backward fill
PM 2.5	25.178	**24.027**
PM 10	55.863	**51.196**
NO	14.935	**14.306**
NO2	9.989	**9.709**
NH3	6.144	**5.911**
SO2	2.702	**2.466**
O3	**8.455**	9.206
AQI	42.125	**39.215**

The data was analyzed at several different granularity levels. The data was resampled to obtain weekly, monthly and quarterly data to obtain data at the optimum frequencies to forecast future values and to reduce noise in the data.

B. Analysis of overall data and Comparison with Bengaluru

As seen in Fig. 1, the particulate matter, PM 2.5 and PM 10 correlates with AQI most strongly and has a significant impact on the overall air quality. Out of the gaseous pollutants, NO and NO2 have the highest correlations with AQI. These gases are primarily released by motor vehicles and industries dependent on fossil fuels.

In comparison to Delhi, Bengaluru has a much lower concentration of air pollution. In the Pearson correlation matrix (Fig. 2), particulate matter i.e., PM 2.5 and PM 10 has a much lower impact on the AQI than in Delhi, where the values are greater than 0.8. The gaseous pollutants to affect the air quality less in comparison to Delhi.

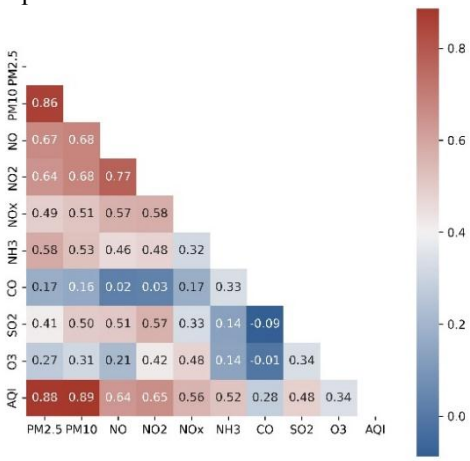

Fig. 1. Correlation Heatmap for Delhi

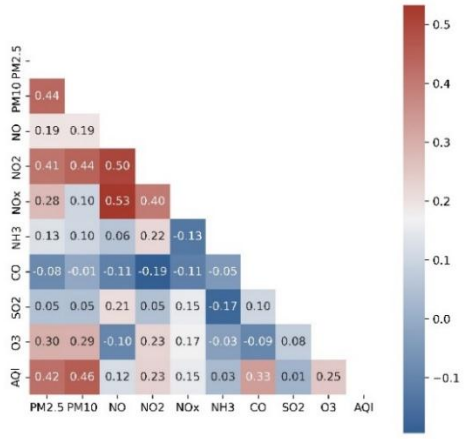

Fig. 2. Correlation Heatmap for Bengaluru

In the level plot (Fig. 3), the average monthly AQI values for Delhi and Bengaluru can be seen. There is a significant difference in the air quality of the two cities. Delhi has had notable changes in the air quality over the period of time, while the air quality in Bengaluru has remained largely constant. Delhi exhibits seasonal patterns, with the AQI increasing towards the end of the year as winter sets in, and dropping at the end of the extremely cold winter season. Bengaluru has less

pronounced changes in temperature with change in seasons due to its tropical nature and proximity to the ocean.

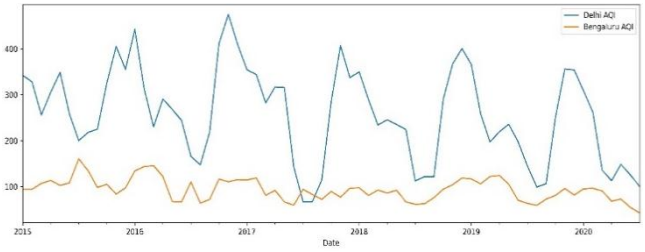

Fig. 3. Comparison of AQI values of Delhi and Bengaluru

C. Impact of Change in Seasons

In the plots below (Fig. 4, Fig. 5, Fig. 6), the correlations of various air quality parameters can be observed for summer, monsoon, and winter in Delhi. The correlation matrix for each season is plotted using the Pearson metric. During monsoon, the correlation of particulate matter i.e., PM 2.5 and PM 10 is the highest. The particulate matter has a strong link with the emission of gaseous pollutants like nitrogen and sulphur oxides, which also have the highest correlation with AQI at this time, emitted vehicular and industrial activities [11]. In summer, there is a significant drop in the correlation between AQI and Ammonia (NH3). Winter shows lowering of the correlation between AQI and carbon monoxide as well as ozone, while the correlation of particulate matter and nitrogen oxides with AQI increases again.

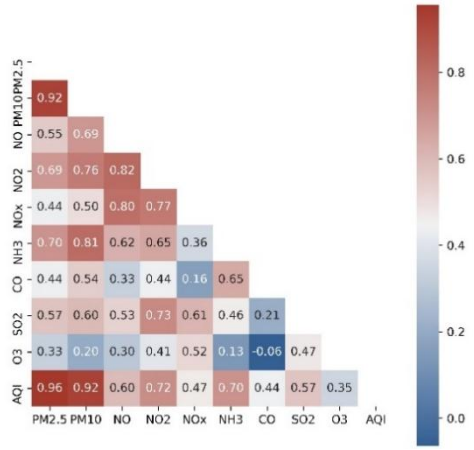

Fig. 4. Correlation Heatmap for Delhi Winter

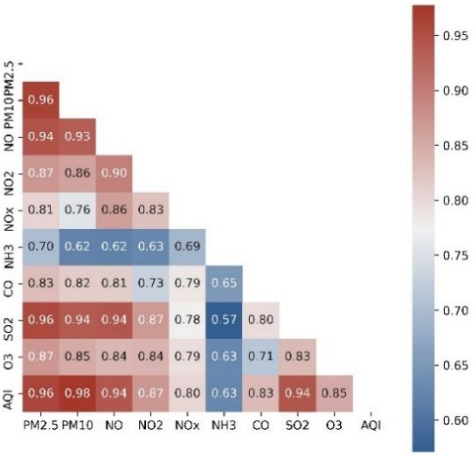

Fig. 5. Correlation Heatmap for Delhi Monsoon

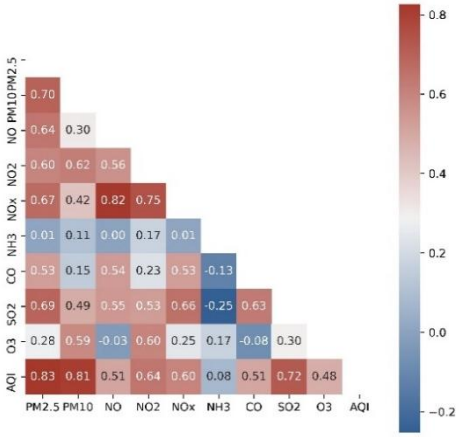

Fig. 6. Correlation Heatmap for Delhi Summer

The dataset has been split into the three major seasons in Delhi - summer, monsoon, and winter. The average values of each season over 5 years have been calculated and visualized. It can clearly be seen in Fig. 7 that the average AQI during winter is much higher than the other two seasons, especially at the peak of winter, with temperatures reaching negative on the Celsius scale. The AQI tends to be the lowest during the monsoon season and begins to increase with the onset of winter. The AQI during summer is generally around 200-250, in between the monsoon and winter values.

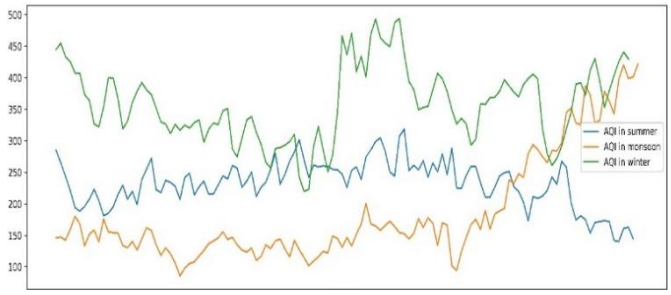

Fig. 7. AQI values for seasons in Delhi

D. Impact of COVID-19

COVID-19 has had a significant impact on the way of life around the world and also has substantially impacted the environment. Due to the imposition of a nationwide lockdown, a major dip in the emission of pollutants was seen, leading to a significant dip in the AQI levels. Fig. 8 shows the change in weekly AQI levels before and after lockdown.

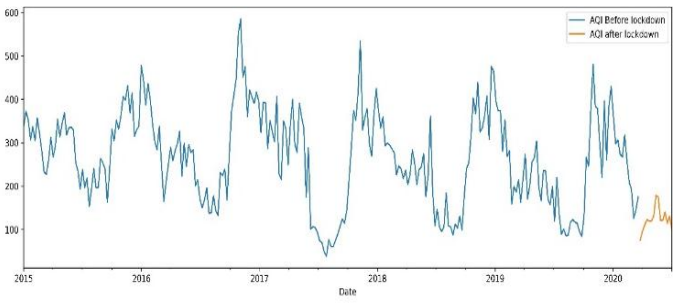

Fig. 8. AQI values in Delhi before and after lockdown

In the plots below (Fig. 9 and Fig. 10), the running minimum and running maximum AQI values have been found for periods before and after the national lockdown. It is calculated by comparing the current and previous values in the time series and returning the minimum and maximum values respectively.

Before the national lockdown, the running maximum was extremely high, peaking at 716, while after lockdown, the highest value was 238. The running minimum, on the other hand, shows a slight increase, from 29 before lockdown to 51 the period after it.

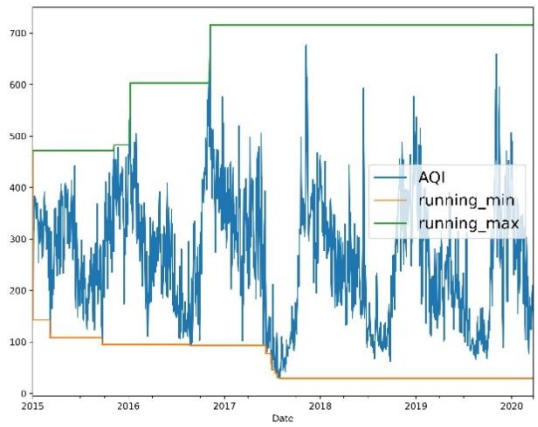

Fig. 9. Running minimum and maximum of AQI before lockdown

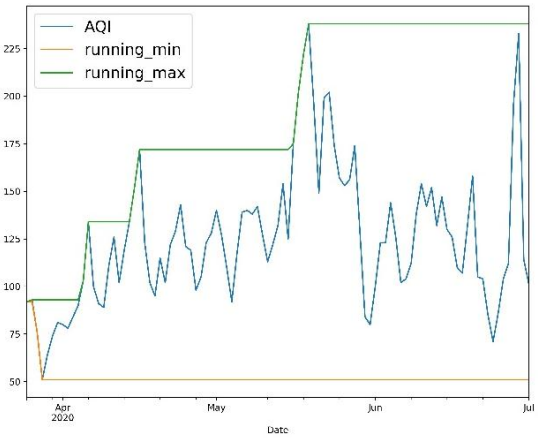

Fig. 10. Running minimum and maximum of AQI after lockdown

E. *Forecasting Future values*

Forecasting future values is essential to know the potential future implications of changes in air quality that will have a significant impact on policy changes and moving towards greener energy sources to reduce fossil fuel consumption. In this paper, two prediction models have been used to predict future air quality index values. The data was resampled into monthly average data to optimize the dataset for the prediction models, to predict AQI values for the next one year, i.e., August 2020 to August 2021. The monthly data was originally non-stationary and had to be made stationary by shifting the time series by one period and taking the difference between values of the original data and the shifted data.

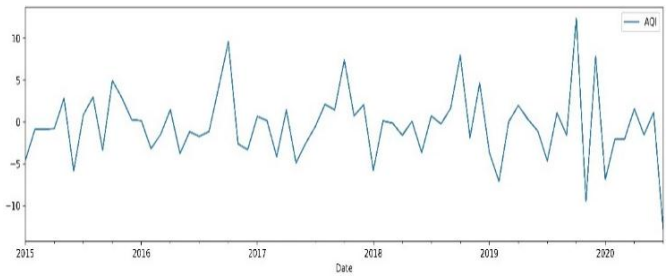

Fig. 11. Periodic difference in monthly AQI values in Delhi

The data was analyzed at several different granularity levels. The data were resampled to obtain weekly, monthly and quarterly data to obtain data at the optimum frequencies to forecast future values and to reduce noise in the data.

1) SARIMAX: SARIMAX, or Seasonal Auto-Regressive Integrated Moving Average with eXogenous Factors, is a model that extends ARIMA. The ARIMA model sees the current value as a weighted sum of previous values. Not being able to handle seasonality is a major downside. SARIMA takes care of the seasonality problem, as it adds seasonal AR and seasonal MA parameters. SARIMAX further builds on this by including the ability to tackle exogenous attributes [12].

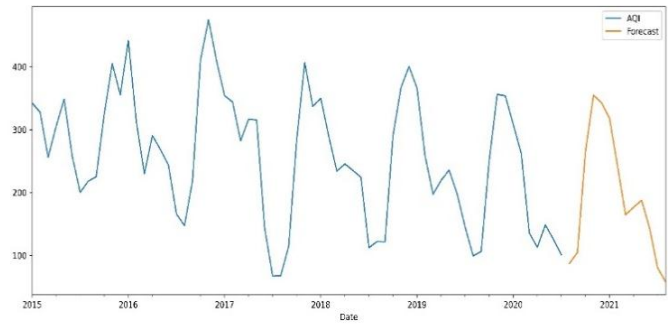

Fig. 12. Forecasting future AQI values of Delhi using SARIMAX

2) Prophet: Prophet is an open-source additivity model developed by Facebook, primarily for forecasting time series data. It is particularly good at forecasting highly seasonal data with protracted non-stationary patterns or missing values. The algorithm looks for trend changes caused by a variety of items, outliers, and seasonal influences such as weekly, monthly, and yearly cycles. MAPE enables us to forecast time series with great accuracy using basic intrinsic parameters. It attempts to fit numerous linear and nonlinear functions of time. Exponential smoothing uses a similar method to model seasonality as an additive element. Prophet can handle stationary data as well [13, 14].

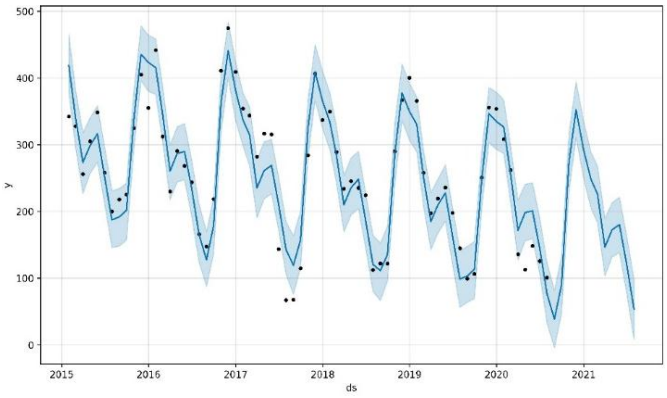

Fig. 13. Forecasting future AQI values of Delhi using SARIMAX

IV. Conclusion

This study performs an overall air quality analysis as well as in-depth analysis of how COVID-19 and seasons have an effect on the air quality measures. There is a significant improvement in the air quality post-lockdown. Particulate matter and pollutants such as carbon monoxide and ammonia have shown decrease in correlations with AQI after the pandemic. The AQI of Delhi during the months of monsoon tends to be the lowest and increases with the onset of winters, reaching highest average values during peak winter. The comparative study of Delhi with Bengaluru shows us that Delhi has a much higher and varied air quality round the year, whereas Bengaluru remains consistent throughout the year. Forecasting was carried out using SARIMAX and Prophet, that gave absolute mean errors of 37.922 and 26.284 respectively. These results can be used to create a system to daily record the air quality measures and feed it to the algorithm to help predict the AQI of a small or large window in the future. Further, data from other cities in India can be incorporated as an extension to this study, to better understand the correlations and the effect of geographical location in the air quality. The results from this study can be used in structuring environmental policies. The analysis and forecasts will help in taking preventive measures during predicted periods of very high AQI.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Y. Yan, Y. Li, M. Sun, and Z. Wu, "Primary pollutants and air quality analysis for urban air in China: evidence from Shanghai," Sustainability, 11(8), p.2319, April 2019.

[2] S. Jodeh, AR. Hasan, J. Amarah, F. Judeh, R. Salghi, H. Lgaz, W. Jodeh, "Indoor and outdoor air quality analysis for the city of Nablus in Palestine: Seasonal trends of PM 10, PM 5.0, PM 2.5, and PM 1.0 of residential homes," Air Quality, Atmosphere & Health, 11(2):229-37, Mar 2018.

[3] R. Yadav, N. Korhale, V. Anand, A. Rathod, S. Bano, R. Shinde, R. Latha, SK. Sahu, BS. Murthy, G. Beig, "COVID-19 lockdown and air quality of SAFAR-India metro cities," Urban Climate, 1;34:100729, Dec 2020.

[4] H. Gao, J. Wang, T. Li, C. Fang, "Analysis of Air Quality Changes and Influencing Factors in Changchun During the COVID-19 Pandemic in 2020," Aerosol Air Qual. Res. 21, 210055, May 2021.

[5] E Tello-Leal, BA Macias-Hernandez, "Association of environmental and meteorological factors on the spread of COVID-19 in Victoria, Mexico, and air quality during the lockdown," Environmental Research, 196, 110442, May 2021.

[6] A. Sannino, M. D'Emilio, P. Castellano, S. Amoruso, A. Boselli, " Analysis of air quality during the COVID-19 pandemic lockdown in Naples (Italy)," Aerosol and Air Quality Research. 1;21(2):1-5, Feb 2021.

[7] IA. Kangiwa, MI. Mohammed, "Impact of COVID-19 Induced Lockdown on Air Pollution and Remediation Measures," Asian Journal of Applied Chemistry Research. 7:43-52, Dec 2020.

[8] J. Wang, H. Li, H. Lu, "Application of a novel early warning system based on fuzzy time series in urban air quality forecasting in China," Applied Soft Computing. 1;71:783-99, Oct 2018.

[9] H. Liu, G. Yan, Z. Duan, C. Chen, "Intelligent modeling strategies for forecasting air quality time series: A review," Applied Soft Computing. 20:106957, Jan 2021.

[10] H. Li, J. Wang, R. Li, H. Lu, "Novel analysis–forecast system based on multi-objective optimization for air quality index," Journal of cleaner production. 20;208:1365-83, Jan 2020.

[11] G Ganesh, S Singh, Hemlata, P Raj, Air Quality of Delhi: An Analysis. http://cpcbenvis.nic.in/envis_newsletter/air%20pollution%20in%20delhi .pdf [Accessed September 10, 2021]

[12] B. Nguyen, End-to-End Time Series Analysis and Forecasting: a Trio of SARIMAX, LSTM and Prophet. https://towardsdatascience.com/end-to-end-time-series-analysis-and-forecasting-a-trio-of-sarimax-lstm-and-prophet-part-1-306367e57db8 [Accessed September 12, 2021]

[13] WX. Fang, PC. Lan, WR. Lin, HC. Chang, HY. Chang, YH. Wang, "Combine Facebook prophet and LSTM with BPNN forecasting financial markets: the Morgan Taiwan Index," In2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) 3 (pp. 1-2). IEEE, Dec 2019.

[14] A. Garlapati, DR. Krishna, K. Garlapati, U. Rahul, G. Narayana, "Stock Price Prediction Using Facebook Prophet and Arima Models," In2021 6th International Conference for Convergence in Technology (I2CT) 2 (pp. 1-7). IEEE, Apr 2021.