Analysing the Influence of COVID-19 Lockdown on Air-Quality in India

Anagha H M
Computer Science and Engineering
PES University
Bangalore,India
anu.hm0520@gmail.com

Anchal Sharma
Computer Science and Engineering
PES University
Bangalore,India
anchalsharma31@gmail.com

Ankita V
Computer Science and Engineering
PES University
Bangalore,India
ankita.v.2001@gmail.com

***Abstract*—The primary goal of this project is to use statistical analysis methods to gain insights about the air-quality data and the factors affecting it. This analysis is done using the data obtained from the Central Pollution Control Board, which is a statutory organization under the Ministry of Environment, Forest, and Climate Change. After an extensive review of related works, we learnt about multivariate testing and spatial analysis of air quality [2]. We also gained insight about the unexpected change in the air quality in the first few months of the COVID-19 pandemic lockdowns [1]. We also learnt the importance of visualization from different points of view and the effect it has on statistical analysis [3]. After literature review, we started data cleaning, pre-processing and exploratory analysis on the chosen dataset. We came up with early inferences about air quality in different cities and also a vague idea about how to proceed with our final solution.**

***Index Terms*—AQI, air-quality, time series, forecasting, seasonal decomposition, SARIMA, accuracy, RMSE, MAPE.**

I. Introduction

Air pollution is a critical problem all across the globe. We as humans, have ignored its consequences for so long. The natural disasters caused as a result dramatically changing climate, the rising temperature, untimely and fierce rains, all are consequences of our own ignorance. Trying to reduce world pollution is no easy task, and it can only be achieved by the co-operation of every single person in the world. Air pollution is defined as the presence of pollutants in the air that can prove to be detrimental to Earth and its creatures. The detrimental effects of air pollution in humans include short term effects like pneumonia, bronchitis, nausea, irritation and long term effects like heart disease, lung cancer, asthma and even certain birth defects. Apart from humans, air pollutants can also affect the environment and climate by damaging flora and fauna, soil, water etc. WHO estimates that 90% of humans currently breathe air that exceeds the WHO's limit for pollutants.

In an attempt to make air quality measurement easier to understand, the ministry of environment and forests launched a National Air Quality Index(AQI) - which runs from 0 to 500, where a higher value indicates higher pollution. Eight pollutants namely particulate matter PM10, PM2.5, Nitrogen Dioxide NO2, NOx, Ozone O3, Carbon Monoxide CO, Sulphur Dioxide SO2, Ammonia NH3, Nitrogen Monoxide NO, Benzene, Toluene and Xylene act as major parameters for deriving the AQI of an area. It also classifies the air quality as good,satisfactory, moderately polluted, poor, very poor and severe. Air Quality Index provides people with vital information about the condition of air in their location, using which they can find out how the air can impact their health. For example, elderly people, infants or people with respiratory issues would be advised not to leave their houses if the AQI is over 200.

India's air quality has deteriorated rapidly in the last few years as a result of rapid urbanisation and development. According to IQAir, in 2020, India ranked third amongst all countries in the world with the worst air quality. The northern regions alone made up 13 of the 15 most polluted cities in the world. Controlling air pollution and its effects is an urgent requirement.

The first step towards improving air quality is identifying the factors that affect air quality. The dataset used in this project was obtained via Kaggle, and contains information from the Central Pollution Control Board of India, a branch of the Indian Government. The dataset contains air quality data and AQI (Air Quality Index) at hourly and daily level of various stations across multiple cities in India. This is an example of Time Series - data that is collected when a sample is observed over a period of time. This kind of data allows us to analyse trends, and patterns, and the influence of certain factors on the sample. The goal of this project is to analyse data to come up with patterns and trends that will allow us to predict how much we will be affected by air pollution and the changes that would be required to prevent irreversible damage.

II. Literature Review

A. Impact of lockdown on AQI

This paper [1] talks about the air quality of major cities in India over the time period of the first lockdown. Air-quality determining factors like Particulate Matter, AQI (Air Quality Index), NO2 were compared to the values observed in 2019. The Central Pollution Control Board Dataset was analyzed from 15th March 2020 to 14th April 2020. The results obtained showed a notable decline in AQI, PM2.5 and, tropospheric NO2. PM2 levels in Northern India showed a larger decline as compared to those in Southern India due

to the crop burning done by farmers in the Indian Gangetic Plains(IGP). Before 2020 the sources of the air masses were recorded to be different as compared to those during 2020. Instead of the Bay of Bengal, air masses were coming from both the IGP and Bay of Bengal due to the impact of the long-range air masses. Abridged fossil fuel consumption also was a prominent factor in the decreased NO2 levels. Thus, thickly populated urban cities showed a slow improvement in air quality.

B. Correlation between pollutants

This paper [2] focuses on analysing air pollution in Madrid using multivariate and spatial analysis. The statistical methods used for multivariate analysis were Pearson's correlation coefficients, principle component analysis(PCA) and hierarchical cluster analysis . The spatial analysis was performed using topological, geometric and geographic properties. Their data set contained the annual average concentration of NO, NO2, PM10, and O3 recorded over a period of 7 years(2010-2017). After conducting the initial exploratory data analysis, a correlation between the 4 pollutants were found. PCA and hierarchical analysis allowed to establish correlation and to classify them based on the significance. Thus giving a better view of the sources that affect air pollution. The contour maps reflected the air-quality in each area and thus made it easier to propose elaborate air quality improvement plans.

C. Effects of air pollution

This paper [3] is mainly focused on statistical analysis of air pollution in some major cities of Karnataka. The authors first detail the causes and effects of air pollution. The method used for statistical analysis is simple and standard. The project began with Data Acquisition. The source of the data was the Central Pollution Control Board of India. Data has only been extracted from stations located in "most extreme dirtied zones". This was done to show heterogeneity in estimating the poison patterns. The next step was Data Exploration. Data was pre-processed and unwanted attributes such as locations, organizations and dates were removed. Hourly midpoints were used for plots of NO2, SO2, PM2.5, PM10. This was followed by Data Visualization. For each zone, bar plots were created for each of the pollutants. The final result of this paper details the findings from the analysis. The data was collected from 715 residential spots and 800 industrial areas across the state. The data for each pollutant was analysed, following a description of the harms of each pollutant. It was observed that Karnataka's average SO2 levels usually result in mild throat irritation. It also details that SO2 and NO2 levels have been on a decline since 2014. The visualization also found that Karnataka has higher PM levels than the country's average. The authors concluded the paper by putting forth their beliefs and reasons regarding the current trend of Air Quality in Karnataka.

D. Influence of COVID-19 on air quality

This paper [4] aims to discover the influence of the COVID-19 pandemic on air quality among populous sites of four major cosmopolitan cities in India which are namely Chennai, Delhi, Mumbai and Kolkata from January 1, 2020, to May 31, 2020, by analyzing PM2.5, PM10, nitrogen dioxide (NO2), ammonia (NH3), carbon monoxide (CO), sulphur dioxide (SO2), and ozone levels. Pearson's coefficient was used to determine the correlation between the features to find the most conspicuous pollutant. It was noticed that the AQI diminished and there was an overall increase in the Air Quality over the lockdown. Particulate Matter was found to affect the AQI the most.

```
city_hour.isna().sum()
```

```
City              0
Datetime          0
PM2.5        145088
PM10         296737
NO           116632
NO2          117122
NOx          123224
NH3          272542
CO            86517
SO2          130373
O3           129208
Benzene      163646
Toluene      220607
Xylene       455829
AQI          129080
AQI_Bucket   129080
dtype: int64
```

Fig. 1. Nan values in city_hour

```
city_day.isna().sum()
```

```
City             0
Date             0
PM2.5         4598
PM10         11140
NO            3582
NO2           3585
NOx           4185
NH3          10328
CO            2059
SO2           3854
O3            4022
Benzene       5623
Toluene       8041
Xylene       18109
AQI           4681
AQI_Bucket    4681
dtype: int64
```

Fig. 2. Nan values in city_day

III. Proposed solution

The goal of this project is to discover the impact that the COVID-19 lockdown had on the air of this country, here represented by the Air Quality Index and the density of various pollutants in the air. The approach towards achieving this goal was compound.

The first part approach involved comprehensive visualizations over various pollutants, cities, days, etc. Data from pre-covid was compared visually with data post covid.

The second part of the approach involved the use of one of the various models specifically designed to analyse Time Series Data. Models were fit on pre covid data and were used for predicting post covid conditions.

A. Dataset

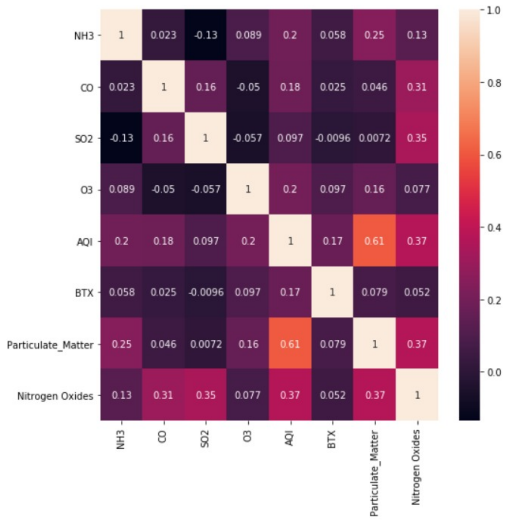
The project commenced with the selection of the dataset. The dataset that was chosen for our problem statement was Air Quality Data in India (2015 - 2020). It was acquired via Kaggle. The data is made up of 5 different comma-separated value files, which are city_day, city_hour, station_day, station_hour

and stations. For the purpose of our project, only city_day and city_hour were used. The city_day.csv consisted of 29531 rows and 16 columns, namely - City, Date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI and AQI_Bucket. The city_hour.csv had 707875 rows and 16 columns, with the only difference being the DateTime column instead of just Date as was in city_day.csv.

The column City consisted of the names of the cities whose Air Quality were being measured. There were a total of 26 unique cities. The column Date (in city_day.csv) stored dates from 01/01/2015 to 01/07/2020 and the column Datetime (in city_hour.csv) stored not only the dates but also the hour of the day for which the data was being recorded. The values in columns PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene represented the amount of each of these components in the air, measured in $\mu g/m^3$. The final columns AQI and AQI_Bucket provided insight about the air quality index of the given cities at given times. AQI gave the numerical value associated with Air Quality Index which ranges from 13 to 2049, while AQI_Bucket gave us a description regarding the air quality, ranging from Severe to Good. Both files had quite a few NaN values in each column as seen in Fig. 1 and Fig. 2, which were dealt with during pre-processing.

B. Data Preprocessing

An important part of the project involved preprocessing the data in order to bring it to an appropriate form that could then be subjected to time series models. The dataset Air Quality 2015-2020 contained multiple missing values which cannot be removed as they are of importance. Since the Air Quality does not greatly differ from one day to another, bfill and ffill were used with the help of pandas to fill in the missing values.



The Date was converted from an object type to a DateTime type. Similar features in the dataset were merged to form a combined column for example PM2.5 and PM10 were combined to form Particulate matter, NO , NO2 and, NOx were combined to form Nitrogen Oxides. A heat map was formed to find the correlation of the attributes with respect to the AQI.

C. Data Visualization

Analysis and visualization of the pollutants was performed for the Air Quality in India from 2015 to 2020. For ease of execution, the dataset was divided into a pre-lockdown and post-lockdown subset.

1) A heatmap was plotted shows the relation among each of the features in the featureset.
2) The concentration for the various pollutants were plotted against their respective cities. There is noticeable decline in the PM2.5 values for the Mumbai and Delhi.
3) A table generating the precovid and postcovid values for SO2 was generated showing the declining trend in metropolitan cities.
4) Similarly tables were generated for PM2.5, PM10 and, NO2 which had varying trends.
5) A bar plot for SO2, NO2, PM2.5, PM10 and, AQI for major cities of Delhi, Ahmadabad, and Mumbai were plotting showing the impact of the lockdown on the air.

D. Fitting the Models

The main objective of our project was analysing the time series data available to us. In an effort to make the data more accessible, the dataset was pivoted, with the columns representing cities, and rows representing the dates for which the pollutant is being recorded. The cells contained the amount of pollutant for given city at a given time.A dataframe was made for each of the pollutants. The data was then further resampled to contain the monthly averages of all pollutants for India, rather than individual cities. Seasonal decomposition was performed on each column of this resampled matrix.

Superficially, all the column followed a seasonal and trend pattern. To get more details about their trend and seasonality, Augmented Dickey Fuller Test was performed for each pollutant. The p value was obtained. pvalue greater than 0.05 implies non-stationary data while that less than 0.05 implies stationary data. The pollutants that showed a high pvalue were differenced until the p value was lower than 0.05. Differencing can stabilise the mean of a time series by reducing its trend and seasonality. The columns only required first order differencing to stabilize their pvalues. Using the newly differenced columns, ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) graphs were plotted for them to find the corresponding p and q values. Not all graphs presented a well defined answer for p and q. In such a scenario, closest values were considered. The p,d,q values were pushed as parameters to the SARIMAX function from the statsmodel.tsa.api library.

To find out the accuracy of the SARIMAX model, it was only trained on a subset of the dataset (From '2015-01-01' to '2018-01-01'). The model was made to predict the values for the subsequent 4 months. The predicted values were compared with the actual values to generate Root Mean Square Error and Mean Absolute Percentage Error. The training data was

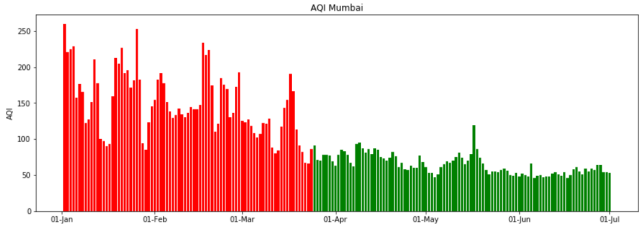


Fig. 3. AQI values for Mumbai(a Metropolitan City)

chosen as such to avoid any unprecedented changes caused by the COVID-19 pandemic. The calculated accuracy serves as a control for our project. New models were trained on the subset of the dataset (exculding data from June, 2019) and were made to predict values for the time duration March 2020 to July 2020. This would allow us to draw comparisons between how the Air Quality of the country was affected due to the COVID-19 lockdowns.

IV. Results

From the data visualizations, it was clear that the lockdown had had a significant impact on the air quality of the country. Pre-Covid, Ahemadabad had the highest overall AQI, but it was quickly overtaken by Patna during the pandemic. Pollutant concentrations pre and post covid were compared as seen in Fig 3, and as expected, most concentrations went down as the lockdown hit India. There were a few un-explainable phenomenons, but it lead us to insights such as an increased pollutant density for Talcher during the pandemic which could be atrributed to the fact that the city houses a major power plant.

For fitting the model, each pollutant was analysed. A seasonal decomposition was performed, and the graphs were analysed. Most pollutants showed trends and seasonality. To verify this non-stationary pattern, an Augmented Dickey-Fuller Test was performed, and as expected, high p-values were observed. Running a SARIMA model on such data would lead to incorrect predictions, therefore the data was first order differenced wherever required. The new differenced values were again put through the ADF test, and produced low p-values.

ACF and PACF graphs were created for each attribute. The p and q values were inferred from the same. These values were used for p,d,q parameters of the SARIMAX function being used. After fitting the model on the training set, the model was tested on the test set. The accuracy of the predictions was calculated using Root Mean Square Error and Mean Absolute Percentage Error. For AQI only, the RMSE = 21.65681701412844 and MAPE = 10.678253931666394. When the model was made to predict data during the pandemic, it had higher RMSE = 44.965194507031235 and MAPE = 37.666664680909015. This was because the pandemic was an unexpected, un calculated occurrence. Therefore the values predicted by the model would be closer to the values if the pandemic had not occurred.

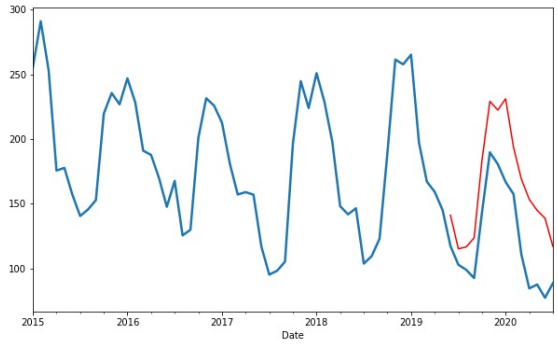


Fig. 4. Actual vs predicted values of AQI

V. Inferences

The lockdown had a significant impact on the air quality in India. Due to the lack of transportation many of the metropolitan cities saw a decline in the AQI. There were some exceptions to this. Cities like Talcher, Bajarangnagar have a Thermal Power Plant, thus causing the AQI to follow the previous trend. From Fig. 4 we can clearly see that without the lockdown the AQI values would have been significantly higher. The higher RMSE and MAPE values very higher since the lockdown is neither a seasonal nor a predictable trend.

VI. Future Prospects

- The model can be built with a larger dataset containing more values.
- Building a model on the dataset containing air quality from all over the world, so as to analyse the worsening conditions of our planet.
- Values for the future can be predicted and the most uninhabitable cities can be found.

VII. Contributions

- Anagha H M Compiling the report, literature survey, visualizations of the model, metrics of the model.
- Anchal Sharma Compiling the report, literature survey, SARIMA model, metrics of the model.
- Ankita V Data visualizations, Exploratory Data Analysis, effect of the lockdown on the air quality.

Acknowledgement

We would like to thank Dr.Gowri Srinivas for her continued support and engagement, and the Data Analytics team for their invaluable guidance throughout this experience. We would also like to thank the PES Computer Science Department for this wondrous opportunity and constant support with our research endeavors

REFERENCES

[1] Singh, R.P., Chauhan, A. Impact of lockdown on air quality in India during COVID-19 pandemic. Air Qual Atmos Health 13, 921–928 ,2020.

[2] David Nuñez-Alonso, Luis Vicente P rez-Arribas, Sadia Manzoor and Jorge O. Caceres, "Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region", 2019.

[3] S. Bhat, G. C. B, S. N. Anil, S. H. P and P. Shetty, "Data Analytics based Statistical Analysis of Air Pollution in the Major Cities of Karnataka," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021.

[4] Pant, G., Alka, Garlapati, D. et al. Air quality assessment among populous sites of major metropolitan cities in India during COVID-19 pandemic confinement. Environ Sci Pollut Res 27, 44629–44636 (2020).