# Analysis of Air Pollutant Levels in India

Bharani Ujjaini Kempaiah
*Computer Science Dept.*
*PES University*
Bangalore, India
ukbharani@gmail.com

Bhavya Charan
*Computer Science Dept.*
*PES University*
Bangalore, India
bhavya.charan.edu@gmail.com

Ruben John Mampilli
*Computer Science Dept.*
*PES University*
Bangalore, India
rubenjohn1999@gmail.com

***Abstract*** **: Air pollution, one of the most serious problems in the world, is a major environmental risk to health. Through this project, we try to focus on analysing the air quality trends by means of Air Quality Index (AQI). We focus on the pollutant levels in India, over the years of 2016-2018. AQI values are calculated across multiple cities. This project also aims at identifying and analysing the various external factors that could contribute to the changing air quality. We also check the feasibility of constructing a forecasting model to predict the future air quality in India**

***Keywords— Air Pollution, Air Quality Index***

## I. Introduction

Air pollution in India is emerging as a major factor that seems to contribute to many of the health hazards that people are facing. Air pollution occurs when harmful or excessive quantities of substances including gases, particles, and biological molecules are introduced into the Earth's atmosphere. With the increasing number of vehicles, factories and crop residue burning, the quality of air we breathe is deteriorating every second and it is playing a harmful role in the overall balance of the ecosystem. Thus, it is of utmost importance to analyse the trend of air pollution and attribute the changing levels to external factors. The large scale data monitoring of pollutant levels of $SO_2$, $NO_2$, CO, $PM_{10}$, $PM_{2.5}$ and $O_3$ are used to arrive at conclusions about the air quality. A parameter often and effectively associated with air pollution is the Air Quality Index (AQI). AQI is an overall scheme that transforms weighted values of individual air pollution related parameters ($SO_2$, CO, visibility, etc.) into a single number or set of numbers.

## II. Related Work

1. Greenpeace India [1]

The Greenpeace India organisation has performed analysis regarding pollution levels throughout the country. The published report states and supports the fact that air pollution is not an issue just in the National Capital Region of New Delhi but it is a national problem that is costing the economy an estimated 3% of GDP. They worked on data collected from the SPCB(State Pollution Control Board) and examined the annual average $PM_{10}$.Their work consisted of focus on specific cities in the states of Andhra Pradesh, Bihar, Chandigarh, Chhattisgarh, Gujarat, Haryana, Jharkhand, Karnataka, Madhya Pradesh, Maharashtra, Odisha, Punjab, Rajasthan, TamilNadu, Telangana, Uttar Pradesh and Uttarakhand.

2. Forecasting air pollution load in Delhi using data analysis tools [2]

This study focuses on the air quality in the Delhi region, conducting a detailed analysis from 2009-2017. Descriptive analysis and predictive analysis has been used to study the trends of various pollutants like $SO_2$, $NO_2$, Particulate matter, $O_3$, CO and benzene and predict the future trends. The data has been collected from SPCB. Collected data has been pre processed using steps like parsing of dates, noise removal, cleaning, training and scaling. Further, descriptive analysis has been carried out on two different platforms- Rstudio and Tableau. For observing the forecasted results, predictive analysis has been done from previously observed values.

## III. Problem Statement

Air is an invisible substance surrounding the earth and providing us all with breathable oxygen and performs a vital role in supporting life on planet Earth. But with the passage of time, with increasing population, extensive use of fossil fuels and increased levels of combustion, pure air is gradually getting contaminated. Air pollution is now a global phenomenon which needs to be addressed at the earliest. The purpose of this study is to analyse the overall trend of pollutant concentrations in India throughout 2016-18. We are working with data obtained from OpenAQ[3] which has been pre-processed by us as per our needs. OpenAQ reports data from recognised sources such as Central Pollution Control Board, State Pollution Control Board and others. The data reported by openAQ is untouched and remains as is from the original source. Thus there are large amounts of noise, negative values and repetitive values which are hurdles that have to be crossed to put the data to good use. Through this study we try to expose the common trends observed in the pollution levels and attribute it to its root cause and produce results that depict the deteriorating air quality in India. Through suitable visualisations we hope to convince the reader of the declining levels of air quality and bring in awareness of how alarming the condition is. We also make an attempt to build a forecasting model which predicts the Air Quality Index for Delhi. Though the pollutant levels depend on dynamic variables like temperature, humidity, precipitation, this is a naive attempt to make use of existing data and to put the forecasting models to test. Modelling dynamic systems is a tough task and the essence of this study lies in the visualisations presented henceforth.

## IV. PROPOSED WORK

There are 2 main components to this project, the first being visualisations and the second being forecasting.

Shown below is a basic outline for the workflow

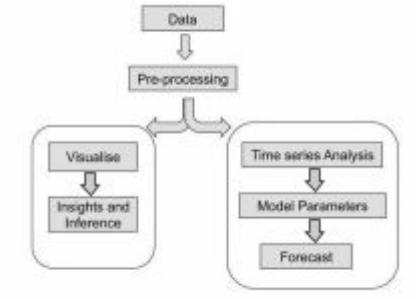


Fig 1 : Flowchart of proposed approach

### A. *Data*

The data was run through an extensive phase of pre-processing[4] which includes parsing the dates to convert them to the right format, cleaning of negative concentrations of pollutants and filling in large amounts of missing values for latitude and longitude. This was performed by making use of entries in other parts of the dataset based on city to ensure minimal loss of information. Data was also aggregated to monthly/weekly data depending on the appropriate frequency for the visualisation being dealt with. Throughout this study, the geological granularity has been restricted to the city level to obtain better insights which a reader can relate to and also for ease of work.

| | city | utc | parameter | value | latitude | longitude |
|---|---|---|---|---|---|---|
| 0 | Hyderabad | 2016-01-03T18:30:00.000Z | pm25 | 61.0 | 17.385070 | 78.455439 |
| 1 | Chennai | 2016-01-03T18:30:00.000Z | pm25 | 28.0 | 13.051966 | 80.235423 |
| 2 | Mumbai | 2016-01-03T18:30:00.000Z | pm25 | 127.0 | 19.068058 | 72.896524 |
| 3 | Kolkata | 2016-01-03T18:30:00.000Z | pm25 | 337.0 | 22.588437 | 88.368451 |
| 4 | Delhi | 2016-01-03T18:30:00.000Z | pm25 | 374.0 | 28.633446 | 77.174244 |

Fig 2 : Dataset after pre-processing

### B. *Components*

Major portion of this study focuses on searching for patterns in the data and analysing the pattern to investigate the cause and how the problem can be remedied.

All inferences were derived from a basic dropdown visualisation which can be tailored as per one's choice to observe a particular city and a particular pollutant as shown below[Fig 3].

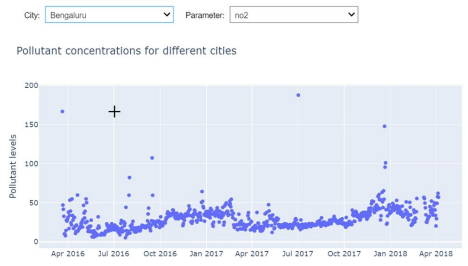


Fig 3 : An interactive drop-down visualisation

The drop down was used to uncover some patterns in the pollutant level across seasons and throughout the day. Further investigations of the inferences obtained from the basic dropdown will be presented in the next section.

The next portion of the study focuses on the AQI calculations which was performed based on the linear segmented principle described below.

$I_P = [\{(I_{HI} - I_{LO})/(B_{HI} - B_{LO})\} *(C_P-B_{LO}) ] + I_{LO}$ where

$I_P$ : sub-index for a given pollutant

$C_P$: pollutant concentration

$B_{HI}$ : Breakpoint concentration greater than or equal to given conc

$B_{LO}$ : Breakpoint concentration smaller than or equal to given conc.

$I_{HI}$ : AQI value corresponding to $B_{HI}$

$I_{LO}$ : AQI value corresponding to $B_{LO}$

Finally :

$AQI = Max(I_p)$ (where; p = 1,2,..n; denotes n pollutants)

The Central Pollution Control Board has released a standard[5] with all the corresponding breakpoints per pollutant which were used to perform the calculations in this study.

The last portion of the study focuses on forecasting the AQI values for the national capital of Delhi. The signal was first decomposed into its components and was concluded to be an additive model. Further analysis was performed to obtain the ACF, PACF plot, tests of stationarity indicated the non-stationarity of the signal and first order differencing was performed to convert it into a stationary signal. Statistical tests such as Dickey fuller were also performed to cross check the stationarity. The model parameters were obtained through the plots and finally the forecasting was done and suitable error metrics were calculated.

## V. RESULTS

Since there is data collected almost every half an hour, we start by looking at the variation of levels of different pollutants during the span of a day.

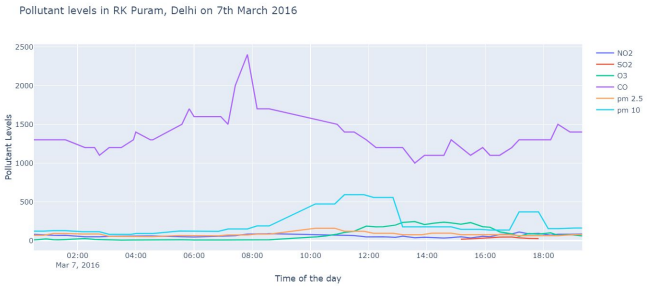


Fig 4 : Variation of all pollutants throughout the day

Some of the inferences we can make by looking at the individual traces for each pollutant are as follows. We observe that the $NO_2$ values increase drastically at around

10 AM and then again at around 5 PM. $SO_2$ levels also increase during the evening. This is the time when people commute heavily as they start from their homes and offices/schools respectively and both $NO_2$ and $SO_2$ are released during fuel combustion.

Next, a drop down was created allowing us to choose the city and the pollutant, showing the variation of that pollutant over the entire span of two years. Some interesting insights can be from the graph for $PM_{10}$. For most of the cities, the levels of $PM_{10}$ during the monsoon months of July-November are considerably low as compared to the neighboring seasons. This is due to wet deposition and air scrubbing by rainfall.


Fig 5 : $PM_{10}$ levels in Delhi

$PM_{2.5}$ also follows a similar trend to that shown by $PM_{10}$


Fig 6 : $PM_{2.5}$ levels in Delhi

To gain an insight about how the AQI values varied in different regions of the country, a geospatial plot was constructed. Different AQI levels are represented using different colors. We observe that the AQI has been increasing which emphasises the fact that the quality of air has been deteriorating with time.

The AQI range to color mapping is as follows-


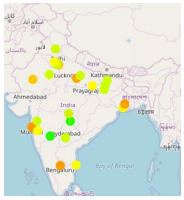

Fig 7 : 03-05-2016

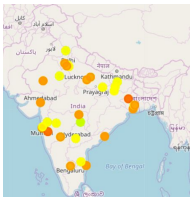

Fig 8 : 03-11-2016

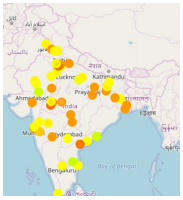

Fig 9: 03-12-2017

Next we tried to look at the pollution levels during the Indian festival Diwali. Since people burst a lot of crackers during this time, we expect the pollutant levels to rise in this span of five days. A line graph was plotted for the years 2016 and 2017 with the five days highlighted.

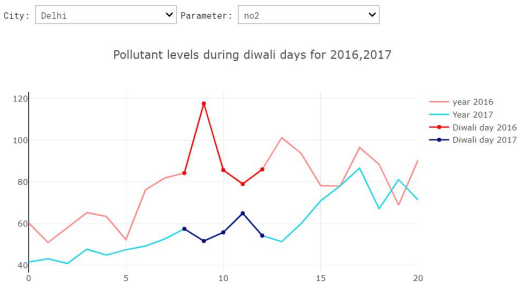


Fig 10 : $NO_2$ levels during Diwali for 2016, 2017

We notice that there is a sharp rise in the $NO_2$ levels in 2016 for the second day as compared to the days before and after. This is in accordance with what we expected and saw from the base dropdown. Also, the pollutant levels are higher in 2016 than in 2017 which could suggest that people are becoming more sensitive towards the environment now.

We now present to you the "Burning issue of crop burning"

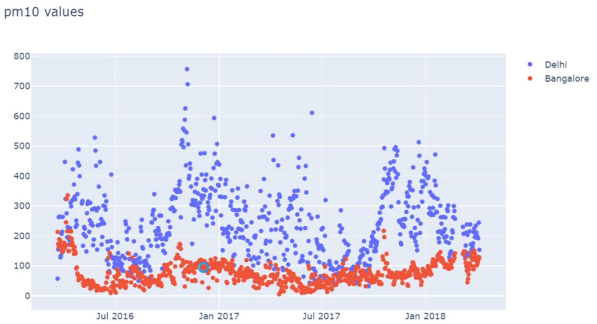


Fig 11 : $PM_{10}$ levels in Delhi and Bangalore

Farmers of Punjab and Haryana burn their rice crop stubble to quickly prepare their field for rabi crop wheat. This proves to be an inexpensive alternative to farmers compared to utilising mechanical tilling machines. The smoke resulting from this burning activity moves to surround Delhi in winter months of November-January which results in a sudden increase of the particulate matter levels as observed from the U shaped trend in the plot above(Fig 11). We observe that the pollutant levels peak during the "post harvesting" season. The reason for farmers to choose this alternative in spite of being aware of the repercussions is that the tilling machine costs almost 10,000 rupees to rent per day whereas burning requires just a 1000 rupees and the plot is ready the very next day. This is one of the major reasons for air pollution and the government should emphasise the need for ecologically safer alternatives by providing subsidies for farmers to rent the machine and also to explore clean energy fuels and gradually phasing out fossil fuels, employing zero waste technology will curb air pollution menace in due course.

We also wanted to see if there was any difference in pollutant levels between weekends and weekdays. The graph shows that there isn't any significant difference between the two.

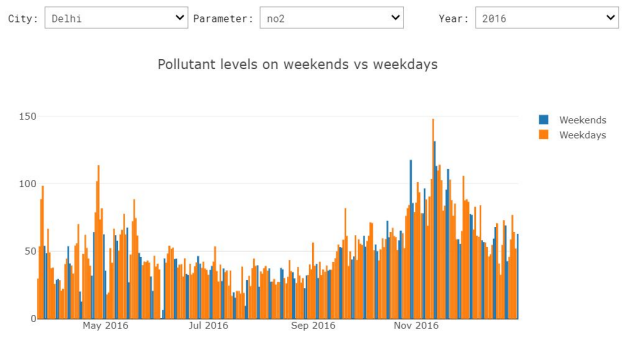


Fig 12 : NO2 values for 2016 split based on weekends/weekdays

We also attempted to observe the impact on declaring "Bandh" in our country and if it had any impact on the pollutant levels.


Fig 13(a) : Effect of Bandh on pollutant levels in Delhi


Fig 13(b) : Effect of Bandh on pollutant levels in Bengaluru

From the graphs (Fig 13(a),13(b)) we can observe that there appears to be no significant trend. The visualisations provides the user with options to choose the city as well as the pollutant. The dot indicates the exact day of the Bandh and the line represents the neighboring concentration levels to get a perspective. There appears to be no decrease or increase in concentrations and thus the general public's idea of a Bandh being a "low activity" or almost zero activity day seems to be false.

Moving on to analyzing the external factors which could influence the pollutant levels in a region, we took the aid of two external datasets - the monthly rainfall in India[6] and the average monthly temperature for 2016 - 2017[7]. To check how rainfall or temperature of a region influenced the AQI values, the following graphs were plotted.

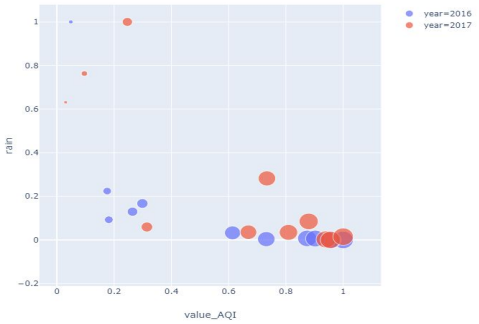


Fig 14 : Rainfall vs AQI value for 2016, 2017

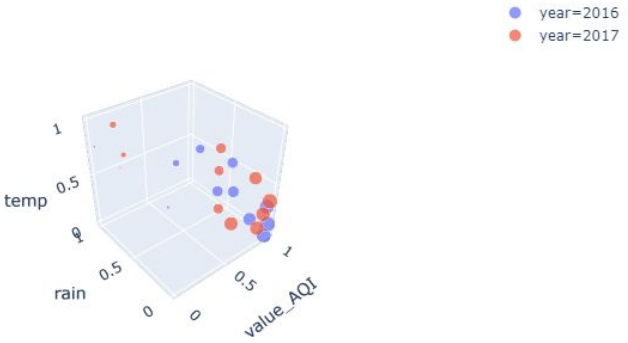


Fig 15 : Rainfall vs AQI vs Temperature for 2016, 2017

The cluster of points around high AQI and low rainfall areas of the plot (in Fig 14) show that regions that have less rainfall have a higher AQI. This could suggest that rainfall helps clear out the air thus reducing the AQI. From Fig15, we observe that during the cold months of winter, the AQI value is higher because the cooler and denser air traps pollutants.

We now move to the forecasting part of this study. We attempt to forecast the AQI values for Delhi. The initial step was to decompose the signal to its constituents (Fig 16) and identifying whether its an additive or multiplicative model. From the graphs below we can conclude that the signal is in fact additive (Fig 17).

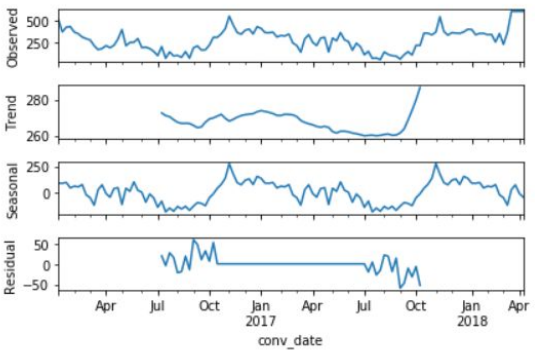


Fig 16 : Decomposed Delhi AQI data

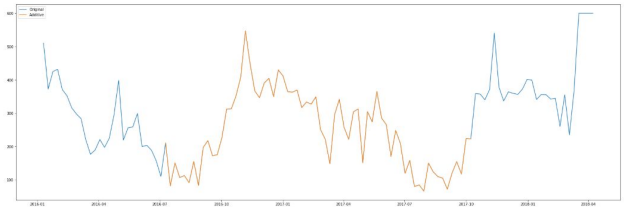

Fig 17 : Additive data overlaps with the original data

The next step is to identify whether the AQI data is stationary. We first plot the ACF and PACF curves and we observe that the ACF curve is gradually decreasing and this sinu-soidal shape indicates non-stationarity(Fig 18).

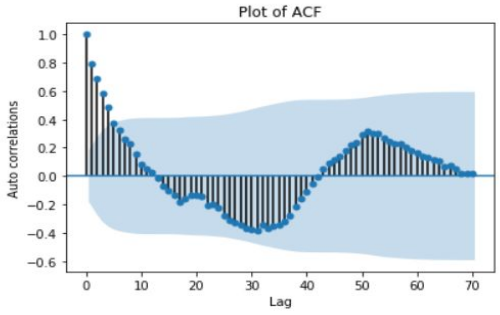


Fig 18 : ACF prior to differencing

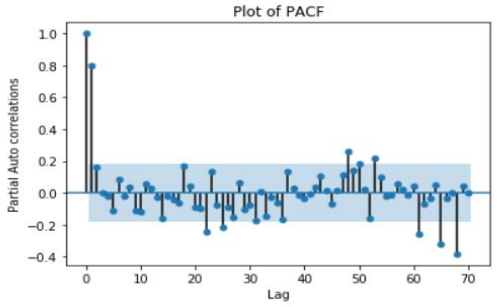


Fig 19 : PACF prior to differencing

To cross-verify, we also perform the Dickey fuller test and from the results we observe that the data is indeed non-stationary because the test statistic is greater than the critical value and thus its plausible that the null hypothesis which states that the data is stationary (Fig 20) and first order differencing was performed post which we recomputed the statistic for the Dickey fuller test, the data is now stationary(Fig 21) since the test statistic is lower than the critical value and we safely reject the null hypothesis and conclude that the data is now stationary and ready to be fit into a SARIMA model.

| | |
|---|---|
| Test Statistic | -1.993112 |
| p-value | 0.289601 |
| #Lags Used | 4.000000 |
| Number of Observations Used | 113.000000 |
| Critical Value (1%) | -3.489590 |
| Critical Value (5%) | -2.887477 |
| Critical Value (10%) | -2.580604 |

Fig 20 : ADF statistics prior to differencing

| | |
|---|---|
| Test Statistic | -5.566274 |
| p-value | 0.000002 |
| #Lags Used | 3.000000 |
| Number of Observations Used | 114.000000 |
| Critical Value (1%) | -3.489058 |
| Critical Value (5%) | -2.887246 |
| Critical Value (10%) | -2.580481 |

Fig 21 : ADF statistics post differencing

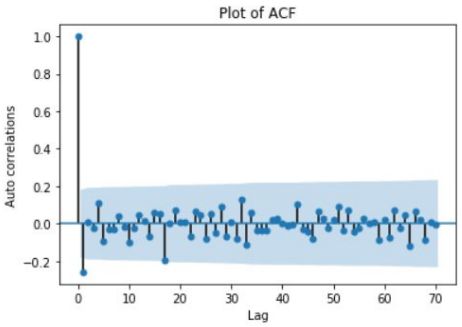


Fig 22 : ACF post differencing

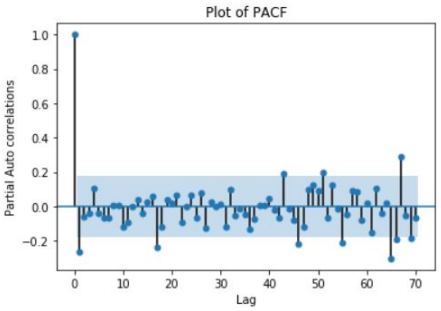


Fig 23 : PACF post differencing

From the decomposed data, it is evident that there is seasonality involved and a SARIMA model will fall perfectly in place. For the stationary data, we plot the ACF and PACF curves to obtain the parameters of the SARIMA model (Fig 23). The chosen parameters for p,d,q are (1,1,0) and consequently grid search was performed to obtain the seasonal parameters P,D,Q and the parameters with the lowest AIC criterion was chosen. The final model is SARIMA(1,1,0)(2,1,0)(12) where 12 is for yearly seasonality.

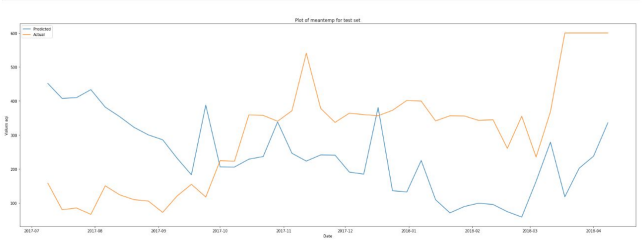

Fig 24 : Predicted Vs Actual for SARIMA(1,1,1)(1,1,2)(12)

```
The Mean Squared Error is 55185.75
The Root Mean Squared Error is 234.92
```

Fig 25 : RMSE value for the SARIMA model

The data was split into training and testing, the model was built on training whereas the test data was used to evaluate the model. The RMSE value obtained for the test set = 234.92. The model does not seem to be doing a good job. This forecast is not accurate since the AQI depends on instantaneous variables which cannot be accounted for in a simple SARIMA model. Parameters such as rainfall, temperature, instantaneous weather parameters determine the Air Quality. Thus, a univariate time series like the one we have done will prove to be less useful.

## VI. Conclusion

Through this project, we gained a deeper understanding of the current situation of pollution in our country. It helped us see the alarming concentrations of pollutants we are exposed to every single day and how the situation is worsening day by day. We were able to interpret how various factors like bursting crackers, burning crops were responsible for increasing pollutant levels. And how external factors like rainfall, temperature also play a big role in influencing the AQI values. We would like to conclude by saying that air pollution is one of the most pressing problems our country is facing today and unless something is done about it soon, the overall growth of the country will be impacted in a negative way. The government has to take up sufficient measures to encourage usage of cleaner, safer and healthier resources. We citizens must also cooperate to achieve this feat.

## VII. References


[1] Sunil Dahiya, Lauri Myllyvirta and Nandikesh Sivalingam, GreenPeace, India.
https://secured-static.greenpeace.org/india/Global/india/Airpoclypse--Not-just-Delhi--Air-in-most-Indian-cities-hazardous--Greenpeace-report.pdf

[2] Nidhi Sharma, Shweta Taneja, Vaishali Sagar, Arshita Bhatt
https://www.sciencedirect.com/science/article/pii/S1877050918307555

[3] OpenAQ dataset *https://openaq-data.s3.amazonaws.com/index.html*

[4] Link to our dataset on kaggle
https://www.kaggle.com/ruben99/air-pollution-dataset-india20162018

[5] National Air Quality Index, Central Pollution Control Board, Ministry of Environment, Forestry and climate change.
http://www.indiaenvironmentportal.org.in/files/file/Air%20Quality%20Index.pdf

[6] Monthly rainfall in India
https://www.kaggle.com/rstogi896/rainfall-in-india

[7] Monthly temperature in India.
https://www.kaggle.com/bhavyacharan/montly-temperature-india-19012017