Analysing the Influence of Various Pollutants 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 wirn our final solution.**

***Index Terms*—air-quality, air-quality index(AQI), India, air pollution, NO2, PM2.5**

I. Introduction

As a result of the Industrial Revolution, Urbanization and Globalization, air pollution has become a prominent term in recent years. 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.

Measurement of Air Pollution is done using the Air Quality Index - 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. 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

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 lockdown due to COVID-19 caused major changes to be implemented in India. Vehicular transport, Domestic flights, and trains were stopped throughout the nation. 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). An HYSPLIT Model was drawn to study the sources of the air masses reaching the cities, how distant and in what direction the air pollutants travel, and how it affected the Air Quality. 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.

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.

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. They also mention the instruments that are in use to measure air pollution and the drawbacks of using such methods. The paper also further elaborates the pollutants that are usually measured to give an estimate of air quality. 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 intention of representation was to obtain helpful visual discoveries. Multiple points of view were considered before coming up with any sort of conclusion. Statistical Analysis was done to verify relations between columns. Scatter plots and heat maps were used for the same. 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.

III. Problem Statement

To find the leading causes of air pollution in India and the external factors that lead to the same. To find which regions in India that have the highest pollutants. An attempt to understand the impact of the COVID19 pandemic.

IV. Exploratory Data Analysis and Visualization

We performed the analysis and visualization of data pertaining to air-quality from Indian cities for a period of 2 years, 2019-2020.

The data was stored in 5 separate files - city_hour(hourly measures for each city), city_day(daily measures for each city), station_hour(hourly measures for each station), station_day(daily measures for each station), station(city the station belongs to).

Apart from city and state, all columns had a significant number of null values. Out of 707875 approximately 18-19% data was missing. These values were dealt by deleting the rows containing the values.

A bar plot showing the average NO2 values for the cities in India was visualized and is shown in Fig 1. From this plot, Ahmadabad has a significantly higher value of average NO2 concentration.

A bar plot showing the average AQI values for the cities in India was visualized and is shown in Fig 2. From this plot we can see that, Ahmadabad has the highest average AQI concentration.

A bar plot showing the average PM2.5 values for the cities in India was visualized and is shown in Fig 3. Delhi has the highest average value of PM 2.5 concentration.

We can also observe that pollutant concentrations are usually higher for northern cities when compared to southern cities.

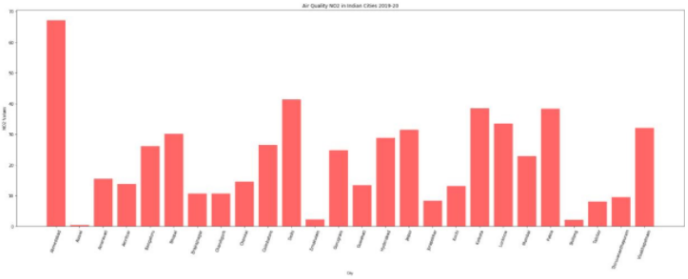

Fig. 1. Average NO2 ug/m3.

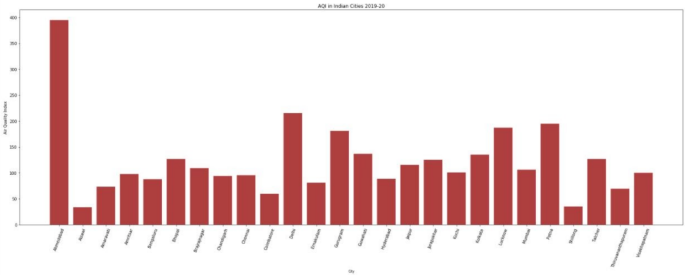

Fig. 2. Average AQI.

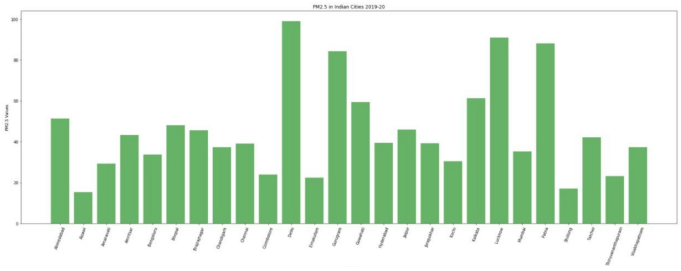

Fig. 3. Average PM2.5.

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

Fig. 4. Missing values per column.

Data	Type
City	String
Datetime (Date - Hour)	Datetime
PM2.5 (Particulate Matter 2.5-micrometer in ug / m3)	Float
PM10 (Particulate Matter 10-micrometer in ug / m3)	Float
NO (Nitric Oxide in ug / m3)	Float
NO2 (Nitric Dioxide in ug / m3)	Float
NOx (Any Nitric x-oxide in ppb)	Float
NH3 (Ammonia in ug / m3)	Float
CO (Carbon Monoxide in ug / m3)	Float
SO2 (Sulphur Dioxide in ug / m3)	Float
O3 (Ozone in ug / m3)	Float
Benzene (Benzene in ug / m3)	Float
Toluene (Toluene in ug / m3)	Float
Xylene (Xylene in ug / m3)	Float
AQI (Air Quality Index)	Integer
AQI_Bucket (Air Quality Index bucket)	String (categorical)

Fig. 5. Description of the dataset.

V. Plans for upcoming weeks

In the coming weeks we would like to perform a more thorough pre-processing which includes a more effective way of data cleaning.We would aim to find correlations between various attributes, to derive trends and patterns. Going forward, we would like to experiment with various or more elaborate methods of visualization. Our main goal is to subject the data to various regression and prediction models to gain insight about the deteriorating nature of air in India.

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.