Visual Analysis of Air Pollution in India

RAJANI MOHAN JANIPALLI

INM433 Visual Analytics

City University of London

Abstract- **This paper presents the visual analysis of air pollution levels in India, for a period of 4 and 6 years. The data consists of measurements various pollutants, that enable the analysis from different points of view. The analysis aids to answer research questions, which try to find different patterns based on these different points of view and also try to indicate the parameter(s) from the given data, that could be a possible factor due to which those patterns are observed. The analysis was done in a Jupyter notebook and Tableau.**

I. PROBLEM STATEMENT

An estimate seven million people worldwide die due to air pollution [1]. As per WHO data, about 99% of the global population inhales the air which has pollutant levels above the safe limits given by WHO. Particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulphur dioxide are the pollutants that majorly affect public health.

India, which has world's second highest population [2], also happens to be one of the top five countries with highest number of deaths due to air pollution [3] as well as most polluted countries [4]. This paper tries to answer the following questions in this respect:

1. How drastically the spatiotemporal pattern of air pollution has changed in India in the recent years?

2. Which region in India is the most polluted?

3. Is there any one particular pollutant that can be attributed for this change?

Although there are many sources of air pollution, energy production is one of the major sources of air pollution, a lot of which comes from electricity generation using coal as fuel [5]. In India, half of the power generation is from coal-based power plants [6]. In this respect, this paper tries to answer another the following question:

4. Are thermal power plants in India a possible factor for such high air pollution levels?

The main data set analysed for this paper had plenty of observations of the above-mentioned pollutants, with 108,035 observations. Another data set with 4945 observations was used to analyse coal-based power generation in India.

II. STATE OF THE ART

The primary paper referred for this study was published by Lina Ren and Ken'ichi Matsumoto [7]. The target of the paper was to use provincial level data in China to understand the spatial effects and factors affecting air pollution between the years 2011 and 2017 with focus on SO2 and NOx as target air pollutants. First, a spatial autocorrelation analysis was done using Moran's I to measure the spatial correlation of SO2 and NOx. The results of Moran's I were also reflected through the Moran scatter plot, where it was found that the slope of linear smooth line of the scatterplot was consistent with the Moran's I value. Then, spatial pain models were made such that the spatial econometric model mainly analyses the interaction and interdependence of spatial regions. This was achieved by making simple pooled linear regression models with time as independent variable and pollutant (that is SO2 or NOx) as dependent variable, and this was done first without spatial interaction effects and then with spatial interaction effects.

The motives of the above referred paper are closely similar to the motives of this study. The idea of using provincial level data was taken to perform spatiotemporal analysis in this study. The approach of making linear regression model was utilised to investigate the pollutant that might have influenced air pollution the most.

The secondary paper referred for this study was published by Álvaro Gómez-Losada and others [8]. The objectives of the paper were the estimation of background pollution and the spatial and temporal analysis of the background pollution. The data analysed for the study in this paper were hourly time series for each year from 2001 to 2017, of NO2, O3, PM10 and SO2, for Madrid City. The estimation of background air pollution concentration was done independently on annual time series of NO2, O3, PM10 and SO2 pollutants at hourly resolution and summarized as an average annual average concentration. A Hidden Markov Model was used for estimation, to obtain groups of hourly observations of air pollutants in each annual time series, forming different clusters of concentration values that are assumed to represent profiles of pollution. Within each cluster the hourly observations were averaged and these average values were used to summarize each of the pollution profiles. The cluster with lowest of these average values was considered to be the annual background air pollution average concentration. The spatial distribution of the background air pollution was estimated based on average estimates of a non-Geo statistical and a Geo statistical method.

One of the objectives of the above referred paper is in line with one of the aims of this study. The approach of averaging the concentrations of pollutants for analysing the time series was implemented, considering the big size of the data set.

III. PROPERTIES OF THE DATA

The primary data set sourced from Kaggle [9], consists of 108035 observations and 16 features. The data are emission values of different pollutants across 21 provinces/states in India between the 2015-01-01 and 2020-07-01. The features include station ID, date, emission values of different pollutants, AQI and AQI Bucket. The Station ID has unique IDs of pollution measuring stations where the emissions were recorded. The date column had date, month and year of observation. The pollutants whose emissions have been taken are viz., particulate matter of 2.5 Micron size and above (PM2.5), particulate matter of 10 Micron size and above (PM10), nitrogen monoxide (NO), nitrogen dioxide (NO2), nitrogen oxides (NOx), ammonia (NH3), carbon monoxide (CO), sulphur dioxide (SO2), ozone (O3), benzene toluene and xylene. AQI or air quality index is a Quantitative representation of quality of air in terms of pollution. AQI Bucket is a qualitative representation of quality of air in terms of pollution, ranging from “severe” for the worst condition to “good” for the best condition. Another data from the same source was merged with this primary data using station ID as the common feature and as a result the data now has columns of city and state. Since the spatial analysis is planned to be done on provincial level, as inspired from one of the reference papers, the state column was kept whereas the station ID and city columns were dropped along with other unnecessary columns. One more datum from a second source [10] was merged with the primary data using state as the common feature. With this merging, latitudes and longitudes of the states were added to the primary data set. The primary data set then has all the features that are required for analysis to answer the first three research questions.

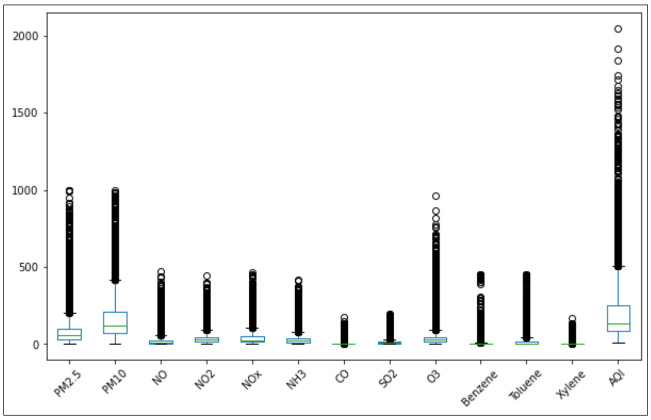


Figure 1: Boxplot of Primary data set

In order to identify the outliers, a common plot of all the numeric columns, except for latitude and longitude, was plotted. The plot indicated the presence of lot of outliers in all the columns. Combining further analysis with the fact that the measurements were obtained from stations of pollution Control Board, it was decided not to remove any apparent outliers. There were many missing values in various columns, which were filled with median values of the respective columns.

The secondary data set sourced from Kaggle [11] consists of 4945 observations and 9 features. The data are values of power generated through various methods, across different regions in India between 2017-09-01 and 2020-08-01. Except for the three useful columns, viz., date, region and thermal generation actual, rest of the columns were dropped as they were unnecessary further analysis. The secondary data set then had all the features required for the analysis along with the primary data set to answer the fourth research question.

There were no missing values in the columns used for analysis. Based on a plot of thermal power generation, it was decided not to remove any outliers as the distribution appeared to be uniform.

IV. ANALYSIS

A. Analysis Approach

The first point planned at the start of the study was to use Python and its libraries along with Tableau, and use them interchangeably so as to utilise the visual analytic approaches and visual plots that would give the clearest output necessary for rational human reasoning and judgment.

The analysis started with pre-processing of primary data set in Python. The first step of this was checking the data types of the features and changing them wherever required, based on human judgment of its need in further steps. The second step was identification and removal of outliers as these may bias the analysis. Visual plots of data points and computational methods were used for evaluating descriptive statistics that are required for identifying Outliers. Visual analysis by human, human knowledge of statistical rules for removal of outliers and human understanding of validity of data points were key in deciding whether or not to remove the outliers. The third step was filling missing values, in which human judgment of what to fill missing values with played key roll on how the data distribution will be affected. The fourth step was dropping unnecessary columns and adding necessary columns by merging other data, for which human understanding of requirement for analysis is required.

After completion of pre-processing, the analysis was started to find answer to the first research question. First computational methods were used to plot temporal visual and spatiotemporal visual of AQI. Based on the visual analysis and judgement of both plots jointly by human, the first research question as well as the second one was answered.

To find the answer to the third research question, temporal plot of AQI and other pollutants was prepared using computational methods. Human analysis and judgment were used to find out the pollutants that were close to the pattern of AQI. To verify the observations from this step, a plot of correlations among the columns of the data was plotted.

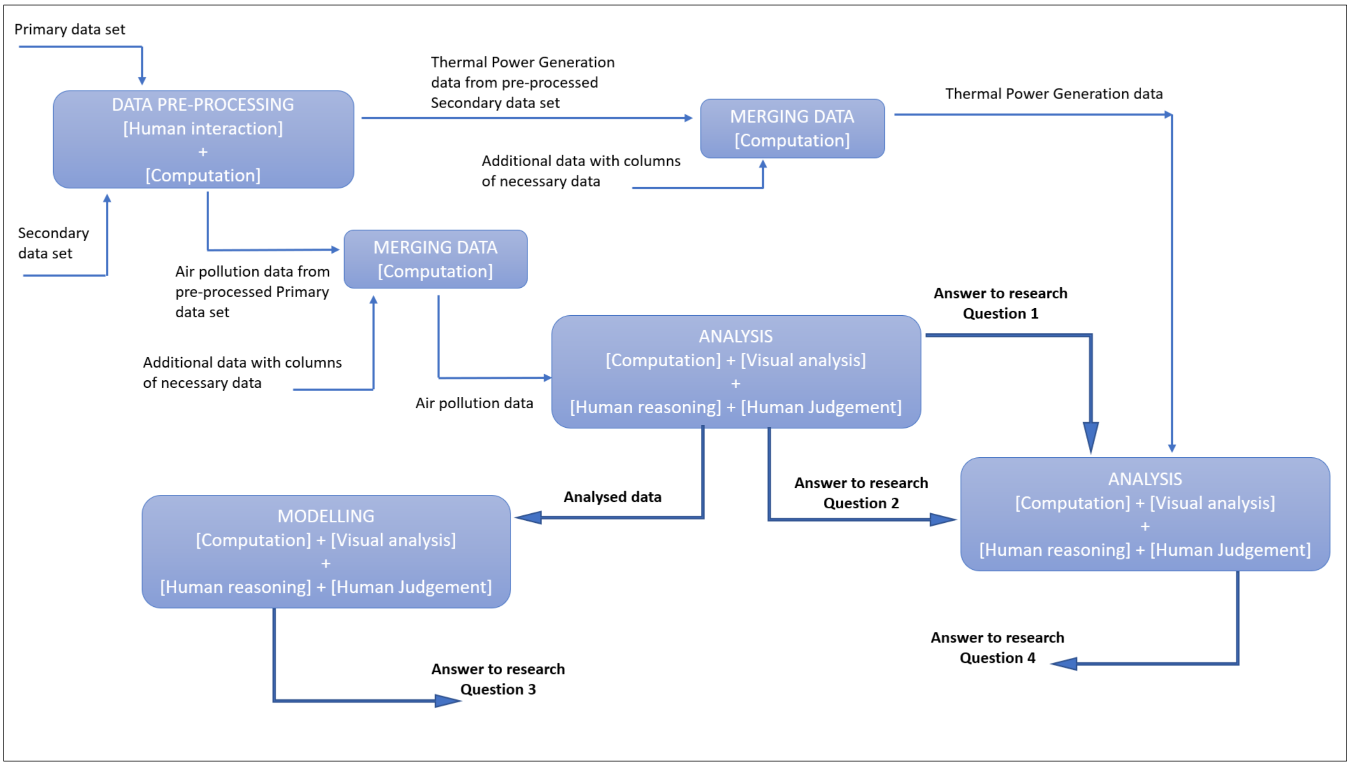

Primary data set
DATA PRE-PROCESSING [Human interaction] + [Computation]
Thermal Power Generation data from pre-processed Secondary data set
MERGING DATA [Computation]
Additional data with columns of necessary data
Thermal Power Generation data
Secondary data set
MERGING DATA [Computation]
Air pollution data from pre-processed Primary data set
Additional data with columns of necessary data
Air pollution data
ANALYSIS [Computation] + [Visual analysis] + [Human reasoning] + [Human Judgement]
Answer to research Question 1
ANALYSIS [Computation] + [Visual analysis] + [Human reasoning] + [Human Judgement]
Analysed data
Answer to research Question 2
MODELLING [Computation] + [Visual analysis] + [Human reasoning] + [Human Judgement]
Answer to research Question 3
Answer to research Question 4

Figure 2: Schematic workflow of analysis

Based on the human judgment of the previous two steps, the pollutant that was most closely related to AQI was fit into a model. Visual analysis and judgement of the model by human, concluded the answer to the third research question.

For further analysis, the secondary data set was pre-processed in the same manner as the primary data set was done. Then, temporal plot and spatiotemporal plot of thermal power generation were prepared using computational methods. Similarly temporal plots and spatiotemporal plots of main pollutants emitted by thermal power generation were prepared. Comparison, visual analysis and judgment of the plots of thermal power generation and the pollutants, by human, deduced the answer to the fourth research question.

B. Analysis Process

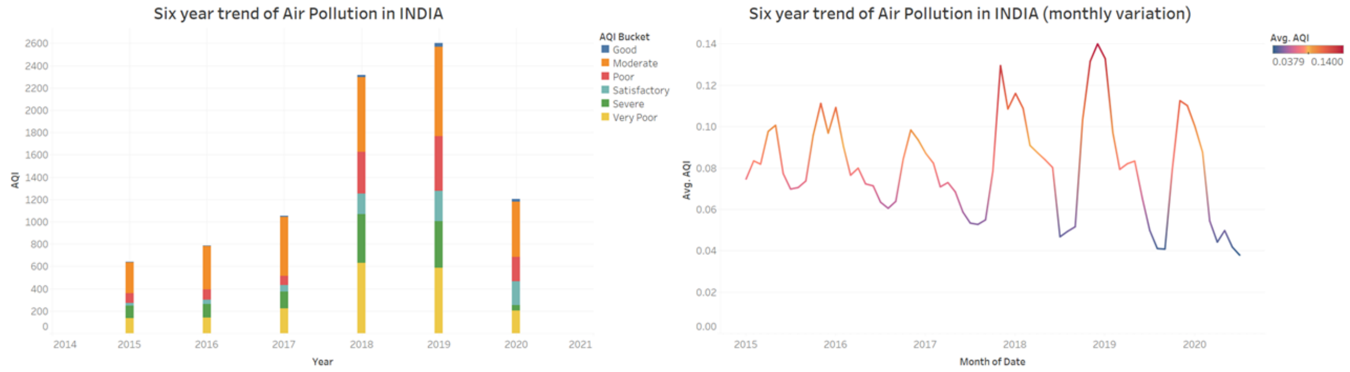
The first step of data preparation was checking the data types of the columns of the primary data set. Except for the date column, rest of them were appropriate and hence the date column was converted from object to datetime data type.

After analysing the box plot of all the numeric columns for identification of outliers, scatter plots of all those columns were plotted for further analysis. The scatter plots showed a different story from that of the boxplot, as there weren't many outliers visibly apparent on the scatter plots. For further clarity, descriptive statistics of all the numeric columns were evaluated. It was observed that for some columns the standard deviation was less than the mean whereas for others it wasn't. It was also observed that using the thumb rule of removing data points greater than twice the mean or removal of outliers using Z-score method would remove a lot of useful data that weren't apparent outliers in the scatter plots. Combining all these above observations with the fact that the data are emission values of pollutants recorded at pollution control board authorised stations it was decided not to remove outliers.

To fill the missing values in the numeric columns either mean or median are the most widely used values. Scatter plots of these columns were plotted with indication of both mean and median. From Visual analysis of these scatter plots it was felt that filling missing values with median would have less impact on the distribution of the data than with mean. Hence, missing values in all the numeric columns were filled with medians of the respective columns. The AQI Bucket, which is a categorical column, has values based on the values of the AQI column. Computation showed that for all the observations of AQI having median values, corresponding value of AQI Bucket was moderate. Therefore, all the missing values in AQI Bucket were filled with moderate.

From the descriptive statistics of all the numeric columns, it was also observed that the values of different pollutants were of different range and hence were not comparable directly. So, scaling of data was done and standardization method was used to convert the values of date to be in the range 0 to 1.

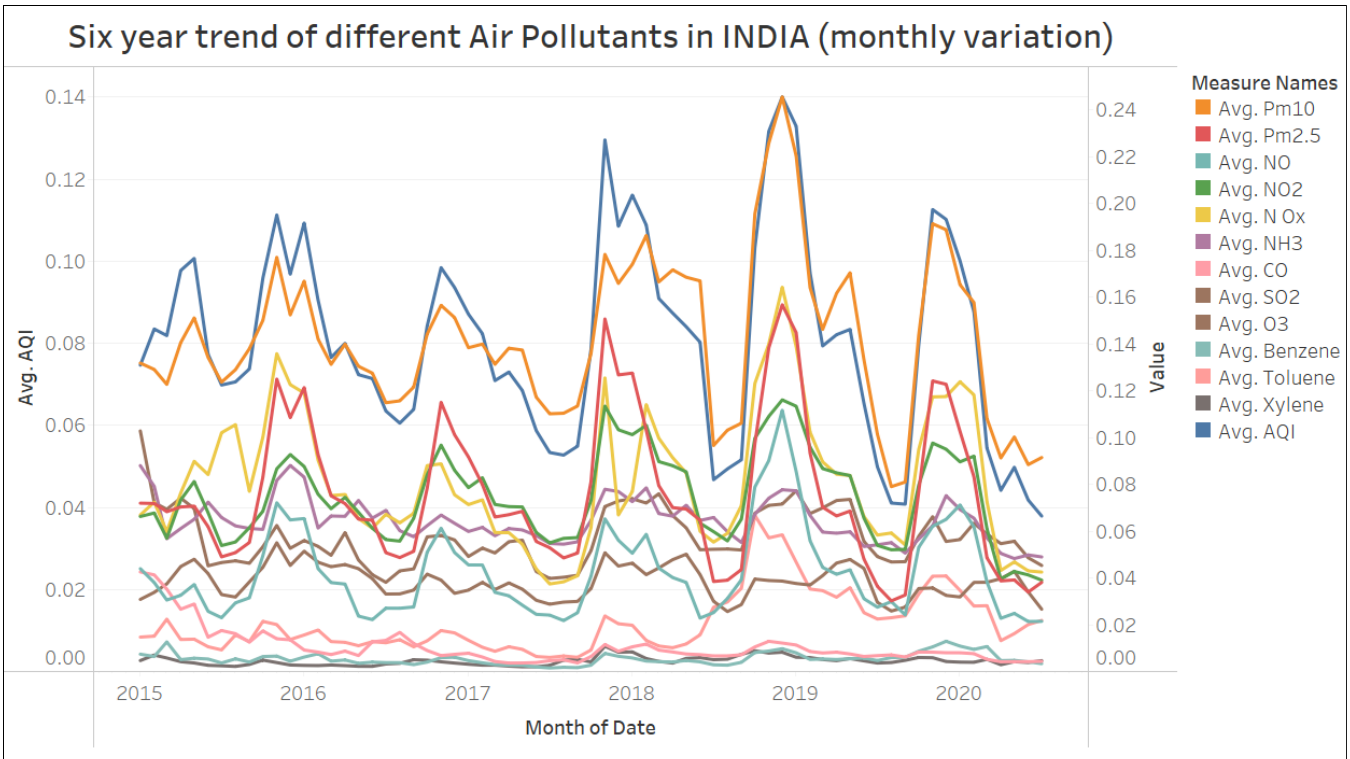
The analysis of the primary data was started with a bar plot of a AQI for each year, with each bar indicating sum of AQI for that year, in Tableau [12]. This plot indicated a steady increase in pollution from 2015 to 2019 with a sharp increase in 2018 and 2019, after which there was a sudden drop in 2020, which may be due to Covid-19 pandemic. To investigate further, a line plot of AQI as a monthly time series was plotted, with average values of AQI on monthly basis, inspired from the secondary reference paper. Only difference in this plot from that of the bar plot was that the year 2017 appeared to be the year with lowest pollution levels, otherwise the trend of remaining years was the same as the tough bar plot. So, temporal pattern of air pollution in India can be concluded.



The plot of sum of AQI for Year. Color shows details about AQI Bucket.

The trend of average of AQI for Date Month. Color shows average of AQI.

Figure 3: Bar plot of yearly sum of AQI & Line plot of monthly time series of average AQI



The trends of Avg. AQI, Avg. AQI, Avg. Pm10, Avg. Pm2.5, Avg. NO, Avg. NO2, Avg. N Ox, Avg. NH3, Avg. CO, Avg. SO2, Avg. O3, Avg. Benzene, Avg. Toluene and Avg. Xylene for Date Month. Color shows details about Avg. AQI, Avg. Pm10, Avg. Pm2.5, Avg. NO, Avg. NO2, Avg. N Ox, Avg. NH3, Avg. CO, Avg. SO2, Avg. O3, Avg. Benzene, Avg. Toluene and Avg. Xylene.

Figure 4: Line plots of monthly time series of average of AQI & pollutants

The latitude and longitude values of the states were used to plot a map of India in Tableau. Then a map based yearly time series off AQI was plotted with yearly average of AQI for each state available in the data, such that there were maps of India for all the six years in the data and each state in the data was coloured as per the average AQI value of that state for that year. Visual analysis of this plot showed that there wasn’t a huge variation in air pollution levels of most of the states, except for those in northern region and in the western region. Especially, the westernmost state of India was the only location which followed the pattern as that indicated by the temporal analysis as well. So, spatial pattern of air pollution in India can be conclude, with a note that it was heavily influenced by the westernmost region of India.

So, visual analysis which was most part of analysis till now, played a major role in producing most apt visuals suitable for human reasoning and judgement, in finding answers for the first two research questions.

To find the answer to the third research question, the analysis started with comparative line plot of all the pollutants with AQI, in Tableau, for initial identification of pollutants that could have possibly influenced AQI the most. This was followed by evaluation of a correlation matrix of all the numeric columns in the data, by using inbuilt method of pandas data frame in Python, which is it computational method. To ease the analysis of this correlation matrix and to make it more interactive to the human, a heatmap of this correlation matrix was plotted. Visual analysis of this correlation heat map indicated that PM2.5 and PM10 where the most correlated pollutants with AQI. To investigate further, scatter plots of both these pollutants were plotted with AQI as the dependent variable and put the pollutants as independent variables. From the scatter plots it was seen that both the pollutants had a linear relation with AQI for majority of the data points, with apparent slope of PM2.5 higher than that of PM10. To get further clarity, regression plots of zero order, first order and second order were plotted for both these pollutants [13]. From the plots it was observed that most of the data points for PM2.5 were skewed above the regression line for all the three orders of regression. In case of PM10, it could be seen that most of the data were equally distributed above and below the regression line for all the three orders of regression. So, the correlation of PM 10 with AQI seemed to be more consistent than that of PM 2.5. Also, the zero-order regression with clearly the best fit further data of PM 10.

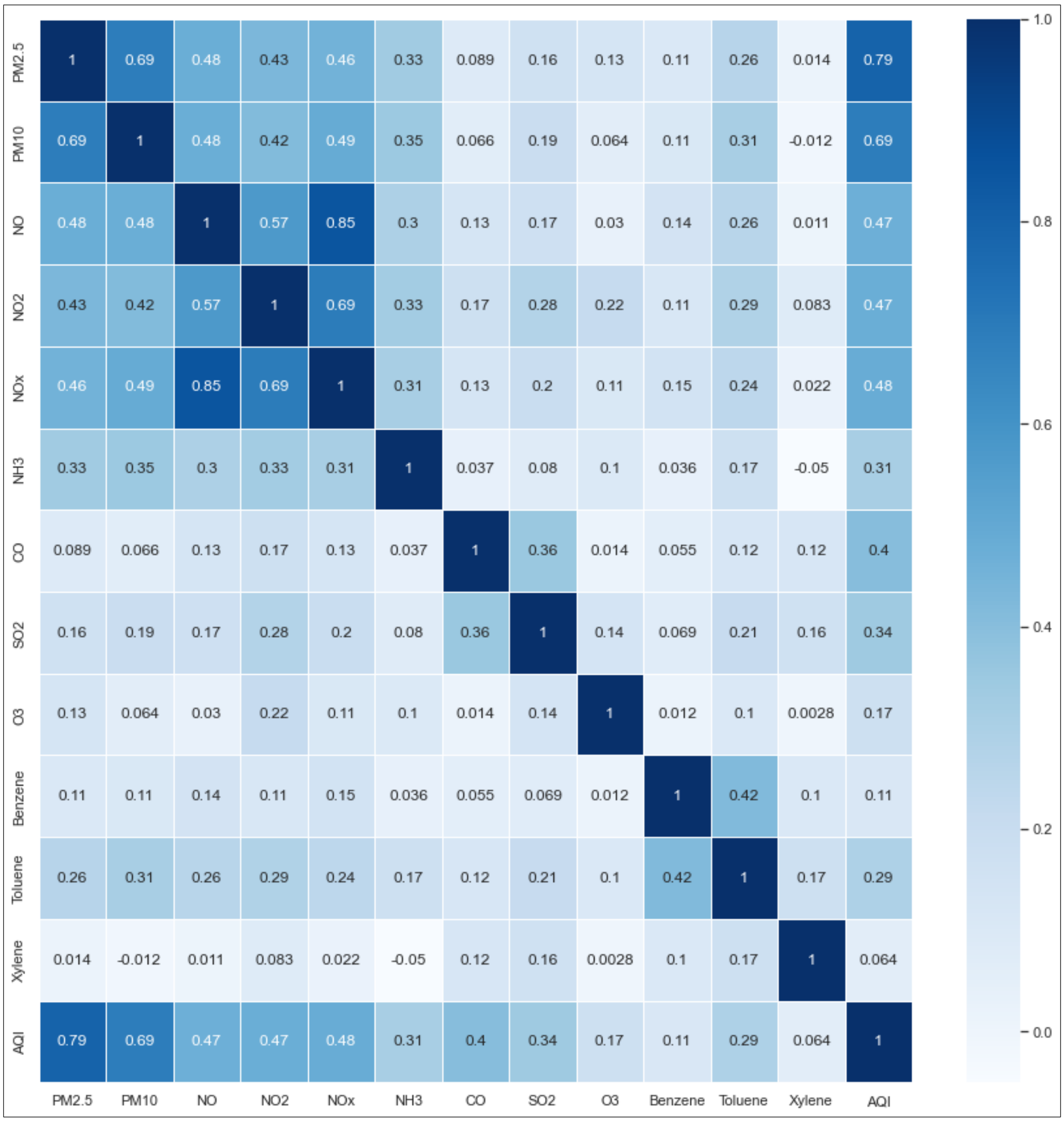


Figure 5: Heatmap of correlations among columns of air pollution data

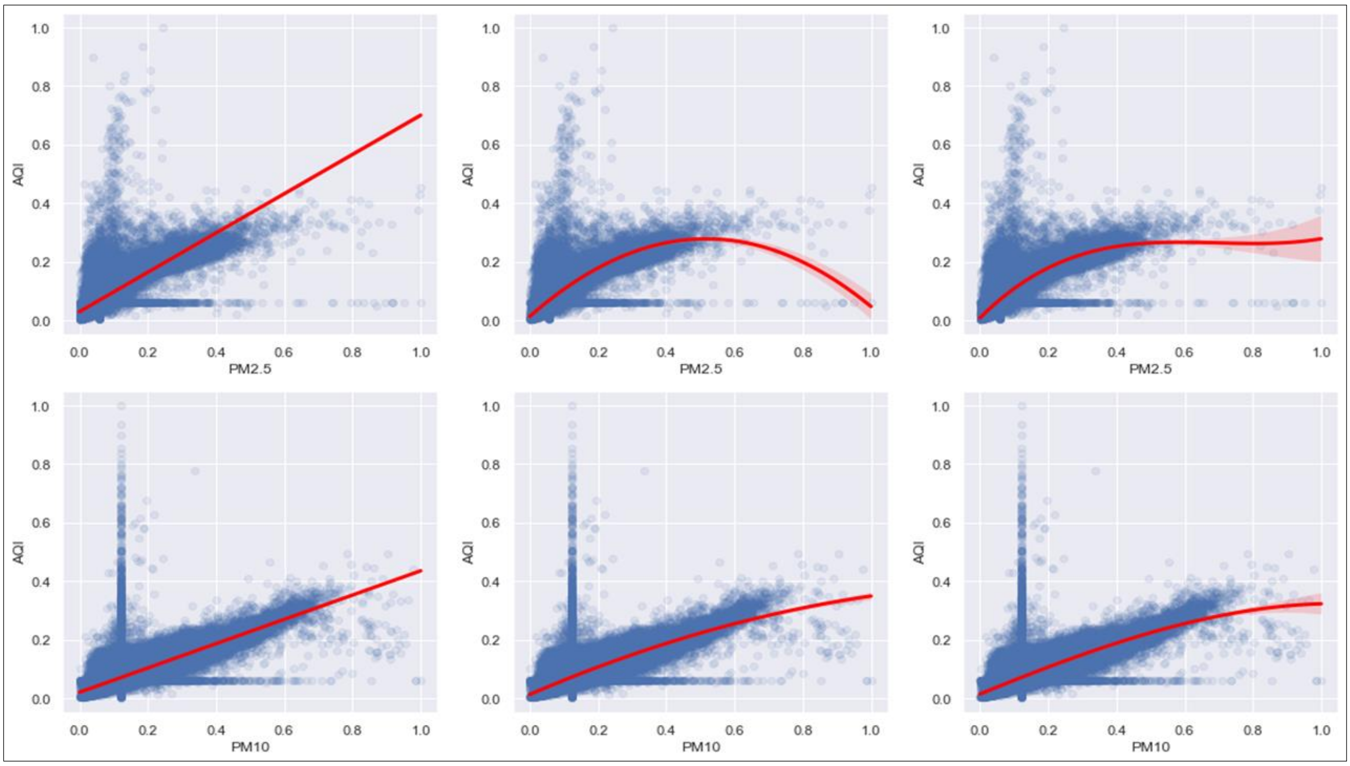
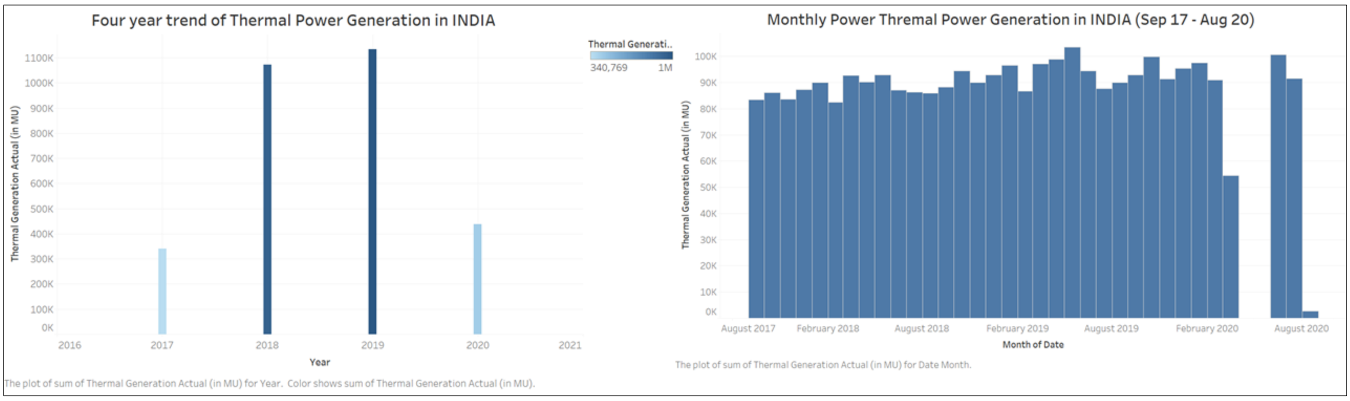


Figure 6: Regression plots of PM2.5 & PM10

Hence, combination of computational methods, visual analytic approaches and human reasoning and judgment has led to the answer for the third research question.

After dropping all the unnecessary columns in the secondary data set as a part of data preparation, a monthly time series of thermal power generation as a bar plot, with each bar representing the sum of power generation across India for that month, was plotted. The bar plot showed a uniform distribution of the data and hence it was decided not to identify and remove outliers, also considering the size and availability of the data.



The plot of sum of Thermal Generation Actual (in MU) for Year. Color shows sum of Thermal Generation Actual (in MU).

The plot of sum of Thermal Generation Actual (in MU) for Date Month.

Figure 7: Bar plots of yearly and monthly time series of thermal power generation

Further was started by with a yearly time series of power generation as a bar plot, with each bar indicating this sum of power generated in that year. This plot showed up sudden surge in the year 2018 and 2019 followed by a drop in 2020. But as the data for the year 2017 was available only from the month of September, the sudden surge here is a false indication. Comparing the bar plots of monthly time series and yearly time series, it was clear that undoubtedly 2018 and 19 had the highest power production.

Unlike the primary data set, the longitudes and latitudes for this data for were available only on a regional basis i.e., eastern, western, northern, north-eastern and southern regions. So, spatiotemporal plots on the regional basis were prepared and analysed, in the same manner as it was done on state basis to answer the second research question. Coincidently, like the analysis for second question, this analysis showed that the western region of India had the highest power generation throughout the time frame of the data.

As per the guidelines of Indian government on emission limits from thermal power plants, PM, NOx and SO2 are the most crucial pollutants [14]. From primary data set, a map based spatiotemporal plot of yearly time series for PM2.5, PM10, NOx and SO2 were plotted to understand the emission pattern of these particular pollutants. Since the time frame for the secondary data set of thermal power generation was between the years 2017 and 2020, a subset of primary data set was created for the same time frame, to make these spatiotemporal plots.

Visual analysis of these spatiotemporal blots showed some interesting results. Both PM2.5 and PM10 consistently had highest emissions in the northern region of India for all the four years, with the western and eastern regions having second highest emissions almost equally. The NOx emission was almost equally high for both northern and western regions from 2017 to 2019, but in 2020 there was a drop in the northern region. The SO2 emission was however always high in the western region. Considering the fact that during lockdowns in 2020 only essential services were running, the emission of all types of pollutants in that year can be linked to emission from power plants. So, through visual analysis and human reasoning and judgment of all the above comparisons, the answer to the fourth research question was concluded.

C. Analysis Results

The spatiotemporal analysis of AQI has shown that the pollution levels have very drastically shot-up on a temporal basis in the years 2018 and 2019, with an overall increase over the six years. Whereas on special basis, it hasn't changed to a great extent. These two results answer the first research question.

However, some states in north and West regions of India had clear variations in pollution levels, especially, the western most state of India consistently had the highest pollution levels and its variation over the years matched the temporal pattern of the country as a whole. This answers the second research question.

Based on the regression model visualised, it can be said that PM10 is the pollutant which might have greatly influenced the trend of pollution levels in India. This answers the third research question.

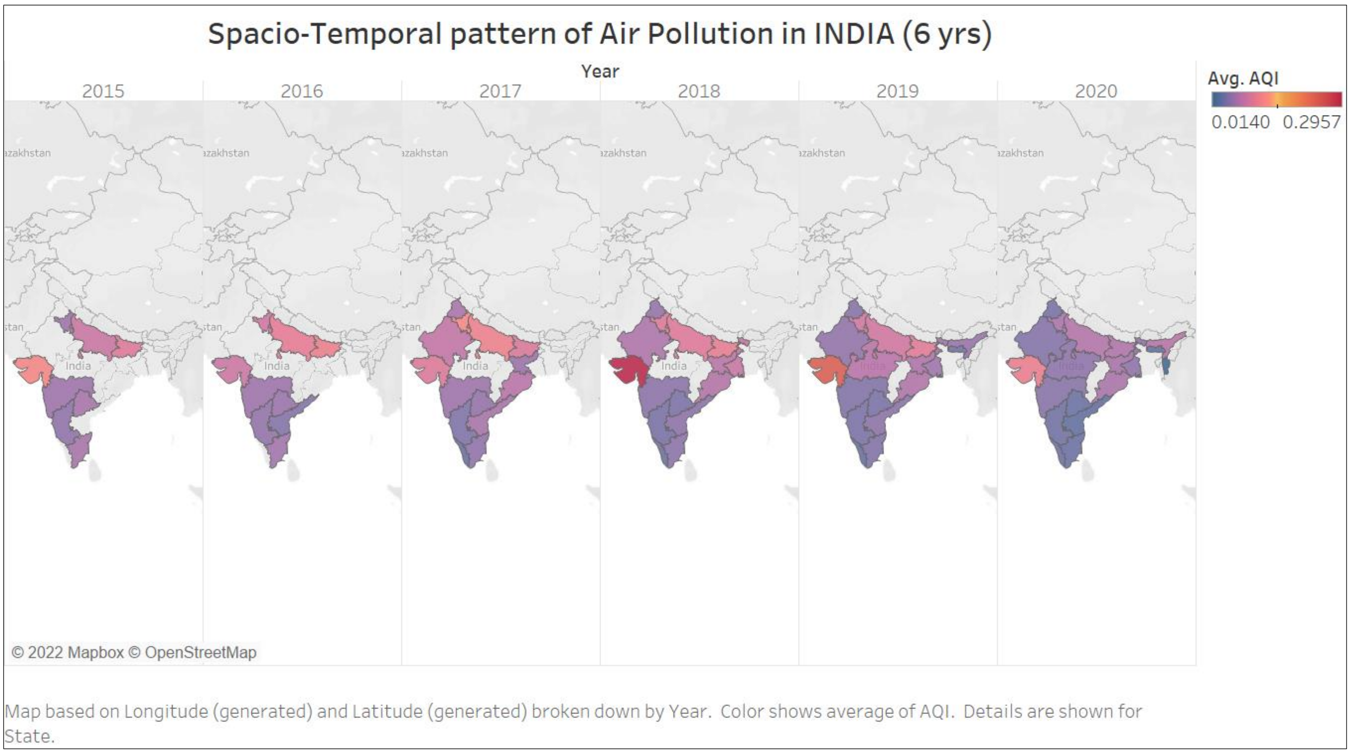

Spacio-Temporal pattern of Air Pollution in INDIA (6 yrs)
Year
2015
2016
2017
2018
2019
2020
Avg. AQI
0.0140
0.2957
azakhstan
India
© 2022 Mapbox © OpenStreetMap
Map based on Longitude (generated) and Latitude (generated) broken down by Year. Color shows average of AQI. Details are shown for State.

Figure 8: Spatiotemporal plot of air pollution across India

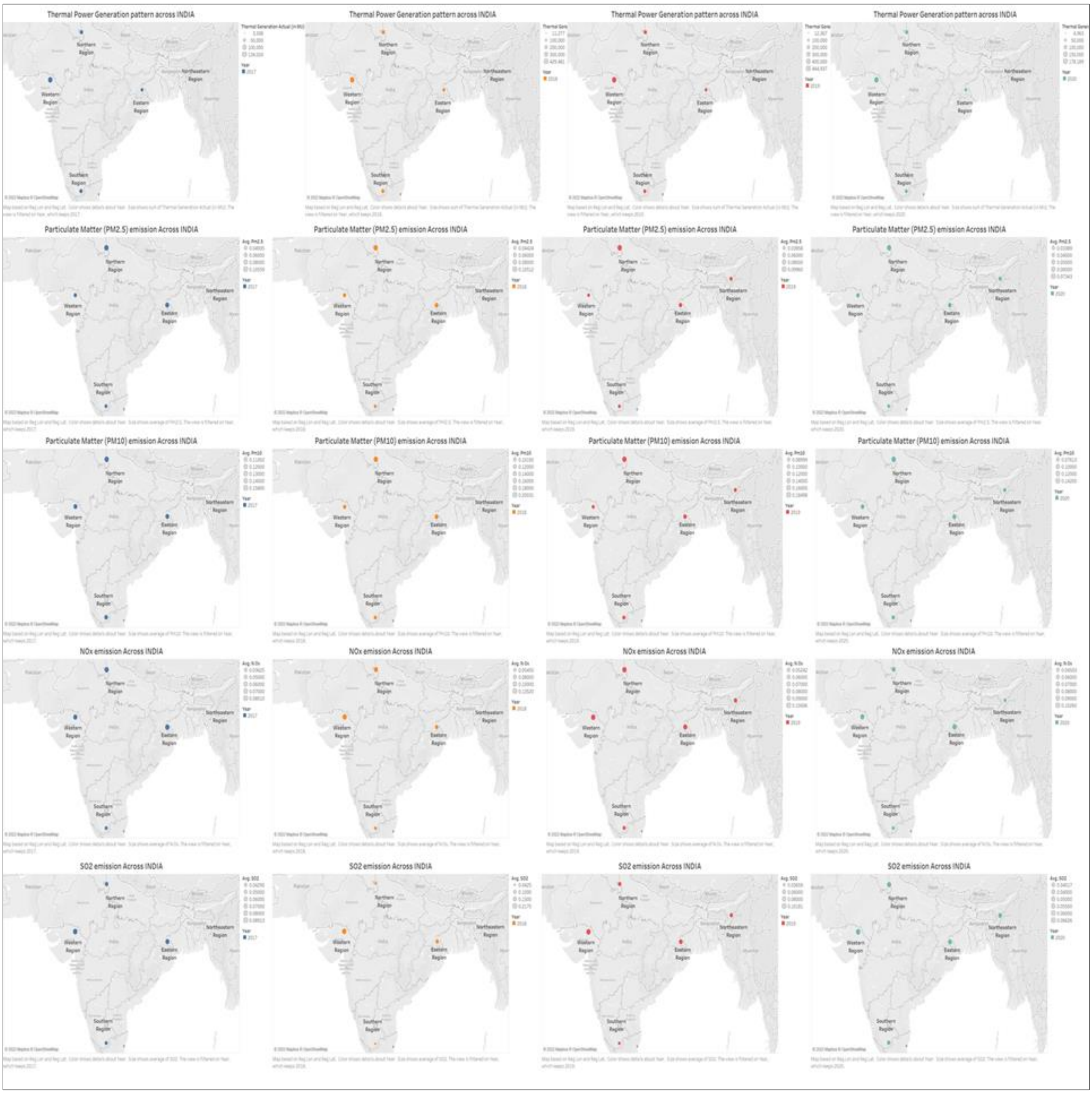

Thermal Power Generation pattern across INDIA
Particulate Matter (PM2.5) emission Across INDIA
Particulate Matter (PM10) emission Across INDIA
NOx emission Across INDIA
SO2 emission Across INDIA
Northern Region
Western Region
Eastern Region
Northeastern Region
Southern Region
Year
2017
2018
2019
2020

Figure 9: Comparison of spatiotemporal plots of thermal power generation with PM2.5, PM10, NOx & SO2

The comparison of spatiotemporal plots of power generation and particular pollutants has indicated that the thermal power generation has greatly contributed to air pollution in the western part of India, the most polluted region of the country that influenced the AQI pattern of India. This answer to the fourth research question.

V. CRITICAL REFLECTION

Without visual analytic approaches, this study wouldn’t have progressed to the point it was concluded.

In the process of identification of outliers in the primary data set, without the scatter plots, use of only computational techniques and boxplot would have led to deletion of a great amount of valuable data points.

In the analysis to find the answer to the first research question, the bar plot showed a consistent pattern which raised a doubt and led to the idea of checking the pattern through a monthly time series of average values of AQI. A comparison of these two plots had given rather an unbiased conclusion.

With regards to analysis for the second research question computational methods solely would have never given such a clear picture of changes in spatial patterns of air pollution in India. Undoubtedly, map-based plot had given appreciable conclusions with variations of colours for states, based on their pollution levels.

From the correlation matrix of numeric columns, PM2.5 appeared to be influencing AQI the most. Its only after the plotting of scatter plots and regression plots of both PM2.5 and PM10, it was realised that as a linear model PM10 would be a more consistent model and hence would influence AQI the most. This could also be seen in the line plot of monthly time series of average values of all the pollutants compared with that of AQI.

Without spatiotemporal comparison of pattern of power generation and that of pollutants emission, understanding the contribution of pollutants generated through power generation towards the overall air pollution to such an extent, would not have been possible simply through computational techniques.

One of the shortcomings, without which the analysis of contribution of pollutants from power generation towards the overall air pollution, would have been much better, was the unavailability of data at a state or provincial level rather than regional level. Also, the availability of the power generation data for the same time frame as that of the air pollution data, would have given more confident results.

Availability of air pollution data from different sources would have given much precise results then what have been achieved in this study for the fourth research question.

In any case, the results of this study are useful to the Government of India, the power plants in India, the public in India and other stakeholders who are on the receiving end as well as the producing end of air pollution. Based on the results of this study, further research can be done two clearly figure out if thermal power plants are the major source of air pollution in India and if so, then taking immediate measures to control and neutralise them would decrease the heavy number of deaths taking place in India due to air pollution.

TABLE 1

Table of Word Counts of different sections		
Section No.	***Section Title***	***No. of Words***
	Abstract	88
I	Problem Statement	241
II	State of the Art	459
III	Properties of the data	488
IV	Analysis	
A	Analysis Approach	439
B	Analysis Process	1489
C	Analysis Results	187
V	Critical Reflection	462
TOTAL		3853

REFERENCES

[1] “Air pollution.” https://www.who.int/westernpacific/health-topics/air-pollution (accessed Jan. 05, 2022).

[2] “Current Population.” https://www.census.gov/popclock/print.php?component=counter (accessed Jan. 05, 2022).

[3] “Global deaths due to air pollution by country 2019,” *Statista*. https://www.statista.com/statistics/830953/deaths-due-to-air-pollution-in-major-countries/ (accessed Jan. 05, 2022).

[4] “Most Polluted Countries 2021.” https://worldpopulationreview.com/country-rankings/most-polluted-countries (accessed Jan. 05, 2022).

[5] O. US EPA, “Sources of Greenhouse Gas Emissions,” Dec. 29, 2015. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions (accessed Jan. 05, 2022).

[6] “Power Sector at a Glance ALL INDIA | Government of India | Ministry of Power.” https://powermin.gov.in/en/content/power-sector-glance-all-india (accessed Jan. 05, 2022).

[7] Lina Ren, Ken’ichi Matsumoto, “Effects of socioeconomic and natural factors on air pollution in China: A spatial panel data analysis | Elsevier Enhanced Reader.” https://reader.elsevier.com/reader/sd/pii/S0048969720336767?token=8908A763F18B9A690729FDCC5B692EFC448DDDEADCECEBD9FCC2B0599A46B972B1132CE0CED73B4C62F4B8AE122EB60C&originRegion=eu-west-1&originCreation=20220102160849 (accessed Jan. 02, 2022).

[8] Álvaro Gómez-Losada, Francisca M. Santos, Karina Gibert, José C.M. Pires, “A data science approach for spatiotemporal modelling of low and resident air pollution in Madrid (Spain)_ Implications for epidemiological studies | Elsevier Enhanced Reader.” https://reader.elsevier.com/reader/sd/pii/S0198971518304447?token=0D1E1C93FB73816F9259E8C7614F50B14DB233F40904EC6083FBFA52A447D2A351095090CCFA7A5FB3219CBFB6652F64&originRegion=eu-west-1&originCreation=20220102155920 (accessed Jan. 02, 2022).

[9] “Air Quality Data in India (2015 - 2020).” https://kaggle.com/rohanrao/air-quality-data-in-india (accessed Jan. 06, 2022).

[10] “Power consumption in India(2019-2020).” https://kaggle.com/twinkle0705/state-wise-power-consumption-in-india (accessed Jan. 06, 2022).

[11] “Daily Power Generation in India (2017-2020).” https://kaggle.com/navinmundhra/daily-power-generation-in-india-20172020 (accessed Jan. 06, 2022).

[12] "INM433_PRD1_A_2021-22: Lab Week 02." https://moodle.city.ac.uk/mod/page/view.php?id=2016150 (accessed Jan. 07, 2022).

[13] "INM433_PRD1_A_2021-22: HTML version of the notebook." https://moodle.city.ac.uk/mod/resource/view.php?id=2016216 (accessed Jan. 07, 2022).

[14] Central Electricity Authority of India, "A Paper on Plant Location Specific Emission Standards." https://cea.nic.in/wp-content/uploads/tprm/2020/12/Review_of_Plant_Emission_Standards_29.pdf