Indian Institute of Technology, Kanpur
Department of Computer Science and Engineering

Analysis of Increasing Air Pollution Levels in India

Mid-term Report
CS685: Data Mining
Academic Year 2020 - 2021

Ankita Dey (20111013)
ankitadey20@iitk.ac.in

Deeksha Arora(20111017)
deeksha20@iitk.ac.in

Sambhrant Maurya(20111054)
samaurya20@iitk.ac.in

Sharvari Oka(20111055)
okasharvar20@iitk.ac.in

Tamal Deep Maity(20111068)
tamalmaity20@iitk.ac.in

Contents

Abstract

Air pollution is a complex term to define, but any release of toxic gases or harmful particles into the atmosphere that are detrimental to human health can be termed as Air Pollution.There exist about 200 known pollutants, the major pollutants found in air include Carbon Dioxide, Sulphur Dioxide, Nitrogen Oxides, Soot, Smog, Dioxins, Polycyclic aromatic hydrocarbons, Chlorofluorocarbons and Methane.

The average human being inhales $10000 - 15000$ litres of air everyday. When polluted air is inhaled, pollutants enter our lungs; they can enter our bloodstream and be carried to our internal organs such as the brain. This can cause severe health problems such as lung diseases, cardiovascular diseases and even cancer. New studies have also found that air pollution affects every organ in the body and reduces the quality as well as number of years of life. The well known annual smog of India's capital-Delhi has been known to irritate the eyes, throat and lungs. Long exposure to severe pollutants like smog can affect an individual's IQ and the ability to learn. Smog aggravates heart problems, bronchitis, asthma, and other lung diseases. Sulphur Dioxide- a major pollutant, can cause respiratory problems like bronchitis and asthma attacks and has been linked to cardiovascular diseases as well.

WHO estimates that 9 out of 10 people in India are exposed to polluted air everyday. Air Pollution was among the top 5 risk factors for deaths in India for 2019. A study has found that India suffers most pollution related deaths in the world, where about 2 million deaths are linked to Air Pollution annually. Hence, being aware about air pollution, it's causes and it's effects is of utmost importance.

Chapter 1

Introduction and Broad Aims of the Project

1.1 Introduction

In today's era, there are many potential sources of Air Pollution. Air Pollution mostly comes from energy use and production. In India, air pollution is believed to have grown more than the birth rate in the country. The number of vehicles running on the roads is on the rise in India and has possibly contributed to the much of Air Pollution in India. With booming industrialization in the country, pollution is seemingly unavoidable because people are interested in making more money, but are not equally concerned about the environment. The number of factories is on the rise, and factories are known to emit tremendous amount of pollutants into the atmosphere. Deforestation is also on the rise, because the need for rapid development demands clearing of forests for more available land. Rapid deforestation has possibly fueled the rise in air pollution, because trees inherently prevent pollutants from settling around. The ever increasing population of India is also possibly correlated with the rise in the Air Pollution levels in the country.

This project will use 4 input features indicative of possible causes of Air Pollution- *Number of Vehicles per state*, *Number of Factories per state*, *Population Density* and *Forest Cover for each state*.

1.2 Broad aims of the project

This project aims to use data mining techniques to accomplish the following:

1. To scrape Air Pollution data for each state and finding trends in the Air Pollution levels in various states of India over the years $2008 - 2014$.

2. To find the most polluted states in India with respect to SO_2, NO_2 SPM and RSPM concentrations.

3. To find the hotpots of Air Pollution by comparing the Pollution levels in each state with it's neighboring states using z-score.

4. To find a correlation between the trends in Air Pollution of each state with these possible factors- Number of Vehicles in the state, Number of Factories operating in the state, Forest Cover and Population Density of the state.

Chapter 2

Datasets Required

The following datasets are required for the analysis:

1. Air Quality Data of India
2. Statewise Recorded Forest Area of India
3. Statewise Total Registered Motor Vehicles in India
4. Statewise Total Number of Industries in India
5. Statewise Population Enumeration Data

Chapter 3

Data Preparation

This chapter includes the source of datasets and how to clean these datasets for our analysis.

3.1 Data Collection

- **Air Quality Data of India**: The air quality data of India has been obtained from *Kaggle*, available at India Air Quality Data. This dataset is available as a **csv** file. It contains SO_2, NO_2, RSPM, SPM and PM 2.5 for all the states of India from 1990 to 2015.
- **Recorded Forest Area of India**: The state-wise forest area of the country has been taken from *www.mospi.nic.in* available at Statewise Forest Cover as **.xls** file.
- **Statewise Total Registered Motor Vehicles in India**: The dataset of statewise total registered motor vehicles has been taken from *www.mospi.nic.in* available at Total Registered Motor Vehicles as **.xlsx** file.
- **Statewise Total Number of Industries in India**: The dataset for total number of industries in each state has been taken from *www.mospi.gov.in* available at Total Number of Industries as **.xlxs** file.
- **Statewise Population Enumeration Data**: The population data of India has been taken from Indian census 2001 and census 2011. The 2001 census data is available at 2001 Census Data webpage and 2011 census data is available at 2011 Census Data as **.xls** file.

3.2 Data Cleaning

The data collected needs to be cleaned to make it ready for any further processing. The various steps to clean each dataset are:

- **Air Quality Data of India**: This dataset contains SO_2, NO_2, *RSPM*, *SPM* and *PM* 2.5 values recorded for different cities from 1990 to 2015. Some insights about the data — the number of entries in each column, the type of entry in each column, etc are shown in the below figure, where we see that we have 435742 entries in our dataset. There are very few non-null values present for PM 2.5 and it might not be able to contribute much. So, only SO_2, NO_2, SPM and RSPM levels will be used in our analysis from 2008 to 2014. Also, for Arunachal Pradesh, air quality data is not available for 2008-2014, so Arunachal Pradesh is not considered for analysis.

```
In [4]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   stn_code                    291665 non-null  object
 1   sampling_date               435739 non-null  object
 2   state                       435742 non-null  object
 3   location                    435739 non-null  object
 4   agency                      286261 non-null  object
 5   type                        430349 non-null  object
 6   so2                         401096 non-null  float64
 7   no2                         419509 non-null  float64
 8   rspm                        395520 non-null  float64
 9   spm                         198355 non-null  float64
 10  location_monitoring_station 408251 non-null  object
 11  pm2_5                       9314 non-null    float64
 12  date                        435735 non-null  object
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

Figure 1: Insights of Air Quality Data of India

To get state-wise values of SO_2, NO_2, SPM and RSPM for each year from 2008 to 2014, the mean of data available for cities of the corresponding state for that particular year will be calculated. The missing values will be replaced by mean value of that particular state for that particular year.

- **Recorded Forest Area of India**: This dataset contains state-wise Total Forest Area, Reserved Forest Area, Protected Forest Area and Unclassified Forest Area in square km for 2009, 2011, 2013 and 2015. The forest cover for missing years will be obtained by taking the mean of data from the preceding and succeeding years. For example, forest area for a particular state for 2010 can be found by taking mean of forest area of that state in 2009 and 2011.

- **Total Registered Motor Vehicles in India**: This dataset contains state-wise total registered motor vehicles from 2001 to 2015. The data for the years 2008 to 2014 will be used for this project.

- **Total Number of Industries in India**: This dataset contains state-wise distribution of factories, fixed capital, working capital, productive capital, invested capital, number of workers, total persons engaged, wages to workers, total emoluments, fuel consumed, materials consumed, total input, products and by products, total output, depreciation, net value added, rent paid for fixed assets, interest paid, gross fixed capital formation and value of addition in stock for 2008 to 2014. The information about number of factories in a state from this dataset will be used. Also, for Mizoram no information is available about number of factories, so we intend to drop this feature during analysis of air pollution trends for Mizoram.

- **Statewise Population Enumertion Data**: Total population of a state in 2001 is obtained from 2001 census data available at the Census India webpage. The data will be scraped from this webpage using python's BeautifulSoup library. The census 2011 data contains area of state(in sq. km), male population, female population, total population, rural population and urban population. The total population and area of a state(in sq. km) will be used from this dataset. Using the data obtained from 2001 and 2011 census of India, the population for intervening years and years after 2011 can be estimated for each state. For estimation, we intend to use a constant rate of growth r per year given by $(1+r)^{time\ duration} = \frac{population\ in\ 2011}{population\ in\ 2001}$. Using this rate of growth, the population for any year can be estimated using $(population\ in\ 2001) \times (1+r)^{year\ -\ 2001}$.

Chapter 4

Methodology

This project intends to use Python with the support of *numpy, pandas* libraries for dataset manipulation and *matplotlib* for plotting various plots for the analysis.

After the data pre-processing stage,the state-wise population density will be estimated from the state-wise population and area of each state (extracted from the 2001 and 2011 census datasets). A mapping of the neighbouring states for each state will also be constructed manually from the official Indian map for the year 2014.

The cleaned data will be used to generate a $35 * 4 * 7$ $3D$ input matrix where the 1^{st} dimension represents the State or Union Territory, 2^{nd} dimension represents the input features i.e. Number of Motor Vehicles, Number of Factories, Statewise Population Density and Forest Cover, and the 3^{rd} dimension represents the year (2008-2014).

The output of the analysis can be modeled as a $3D$ matrix of size $35 * 4 * 7$ where the 1^{st} dimension represents the State or Union Territory, 2^{nd} dimension represents the 'Four Air Pollution level indicators' i.e. the SO_2, NO_2, SPM and RSPM levels for each state, and the 3^{rd} dimension represents the year (2008-2014).

First, compute the mean and standard deviation of SO_2, NO_2, SPM and RSPM levels for each state over the period of 7 years. Following that, compute the z-score values of each Air Pollution level indicator for each state with respect to its neighbouring states. The output of this phase will be used to find the hotspots of Air Pollution in India.

Next, try to look for a correlation between the input features and the output features. Initially, *Pearson's correlation coefficient* will be used, assuming a possibly linear relationship between these variables. If this step does not provide satisfactory results, it will be assumed that the relationship between the variables is non-linear and the use of *Spearman's rank correlation coefficient* will be employed. Hopefully this computational phase would provide a clear idea as to how the input features are correlated with the air pollution levels in each state.

The results of the analysis done so far can then be used to plot the correlation matrix, bar plots and heat plots which can be further studied to draw results and conclusions.

Chapter 5

Expected results and their evaluation metrics

1. **Most polluted states in India with respect to SO_2, NO_2, SPM and RSPM concentrations**
   Evaluation metrics:
   - Bar plots of SO_2, NO_2, SPM and RSPM vs all states
2. **Trends in Air Pollution levels in various states of India**
   Evaluation metric:
   - Heatmaps for SO_2, NO_2, SPM and RSPM with state and year attributes
3. **Possible correlation between the input features and Air Pollution levels in various states**
   Evaluation metrics:
   - Plots of input features vs pollution levels over the years $2008 - 2014$ for the top 5 hotspots (where input features are Number of vehicles, Number of factories, Forest cover and Population density)
   - Correlation matrix
4. **Hotspots: Performance of each state in comparison to its neighboring states:**
   Evaluation metric: A Leaflet-based heatmap of India for visualizing the pollution levels in each state and the state hotspots found using z-score.