

Understanding Pollution Pattern of Maharashtra, India

Maxim Rohit

MSC Data Science Research Proposal

Proposed Research

- Indian Government on 8th December 2018[1] released a first of its kind report stating few disturbing facts about how the increase in air pollution had a degrading impact on Indian Living conditions.
 - Air pollution can be attributed to 12.5% of deaths in India[1].
 - 1.24 Million deaths in India attributed to air pollution in 2017, of which 50+% were in individuals younger than 70 years.[1]
 - Life expectancy in India has gone down by 1.7 years due to health loss caused by higher than accepted air pollution level. [1]
- Similar study for New York & Madrid attributed Seasonal Human & Natural activities influencing concentration level of pollutants.
- The main objective of the research is to build an understanding of air pollution pattern in the state of Maharashtra. The research is novel in nature, as being the first for the region. Since little is known about the current state it will be exploratory in nature.
- We will evaluate the pollution impact in novel light w.r.t features like elevation from sea level, Longitude, latitude, industrial area coverage spread among others non-novel one's like rainfall, season, population, etc.

Previous Work

- The aforementioned report is the first comprehensive study of air pollution impact in India.[1]
- Few Highlights from the reports are as follows:-
 - In 2017, annual exposure of $90 \mu\text{g}/\text{m}^3$ of $\text{PM}_{2.5}$ in India was one of the highest in the world. It's particularly high in the states of Delhi, Uttar Pradesh, Bihar and Haryana.[1]
 - 77% of Indian population is exposed to $\text{PM}_{2.5}$ of $40 \mu\text{g}/\text{m}^3$ and above, which is the recommended limit by the National Ambient Air Quality Standards.[1]
 - Air pollution attributable cases of major non-communicable diseases are equal to those attributable to tobacco for the year 2017.[1]
- Study for New York city; where Increased Domestic heating, shorter photoperiod month among other variables were linked to the seasonal behaviour(during Winter's Nov-Feb) of peak Pollutant concentration levels.[2]
- Another study revealed that annual average of pollutant concentration in metropolitan area of Madrid exceeded the annual average threshold for human health protection of $40 \mu\text{g}/\text{m}^3$ established by the European and Spanish regulations.[3]

Motivation

- In the release of Indian Report, Prof. Balram Bhargava, Secretary to the GOI, Department of Health Research, Ministry of Health & Family Welfare, and Director General, ICMR, said **“It is important to have robust estimates of the health impact of air pollution in every state of India in order to have a reference for improving the situation.”**[1].
- The research is dedicated to answer one such question from pollution angle for the state of Maharashtra, State I Live in!
- Categorization of cities via clustering algorithm will be validated against R-squared value.
- For timeseries prediction for future pollutant levels, MAPE will be used as acceptance criteria.
- For PCA the loading on each component and overall variance coverage will be considered to validate & measure interaction between features.

Research Question

- Identify various subdivisions within Maharashtra, India based on Air pollutant like SO₂, NO₂, SPM and other features.
 - Separating out the high risk zone will enable better refraintments.
- What role does the below mention feature play in concentration of different kind of pollutants?
 - I.Population, II.Elevation from Sea level, III.Seasonal, IV.Rainfall levels, V.Longitude, VI.Latitude, VII.Industrial Area, VIII.Others.
 - Understanding if a variable has major say in the pollution trends, may help us prioritize our action plan.
- How does cities like Pune (An IT Hub in India) & Mumbai (An Overall commercial Hub) fair against neighbouring upcoming towns (Nasik, Nagpur) versus the less progressive one like Ahmednagar, Solapur over the last couple of decades?
 - Understanding the impact of city progression, will help plan future development better.
- How does concentration of NO₂ impact SO₂,SPM and vice versa?
 - Understanding pollutants correlation will help us modulate them better.
- Predicting by when the air pollution levels will go beyond acceptable standards?
 - Gives us a time frame to strategize to slow down/avoid higher pollution level.

Proposed Methodology

- Analyse & categorize the rate at which air pollution is increasing in various cities & towns of Maharashtra using clustering techniques like hierarchical clustering and K-mean.
- Understand how various pollutants and other features co-relates to each other using PCA.
- Evaluate time based (seasonal & otherwise) variation in the spread of various pollutant, via EDA.
- Use timeseries analysis to project the pollution growth pattern in the coming year and predict when the air pollutant level goes beyond breathable standards.
- Data is being collected from the following government repositories:-
 - **Pollutant**:-<https://data.gov.in/catalog/historical-daily-ambient-air-quality-data>
 - **Vehicle Registration**:-<http://mospi.nic.in/statistical-year-book-india/2017/189>
 - **Rainfall**:-<https://www.indiawaterportal.org/>
 - **Elevation**:-<https://en.wikipedia.org/wiki/<CityBasedURL>>
 - **Industrial Area**:- <http://dcmsme.gov.in/<CityBasedURL>>
 - **Population**:-<https://mahasdb.maharashtra.gov.in/population1.do>

Proposed Methodology

- Data Preparation : EDA (Univariate & Bivariate analysis), Outlier normalization, Missing data removal/imputations.
- Machine learning methods:-
 - Clustering techniques k-means & hierarchical for identifying grouping of cities within Maharashtra state.
 - PCA to analysis to understand the interaction between pollutant and various features.
 - Time series analysis to predict pollutant trends in the future.
- Coding/Development: R(R-Studio)/Python(Jupyter notebook)
- Visualization: Tableau
- Softwares are already installed on personal laptop.

Project Plan + Risk or Contingency Plan

- We are using Environmental & Census data made publicly available by the government
- All required software & library are readily available.
- If the pollutant spread is similar then we might not be able to categorize?
 - Then Similar dispersion pattern of Pollutant will be established.
- There's a possibility that novel attributes being examined don't impact the pollutant concentration.
 - Then we would conclude by saying that these attributes are irrelevant w.r.t pollutant concentration.

Timelines

- **Sep, 2019:** Exploratory data Analysis and Data Cleansing
- **Oct, 2019:** Clustering to categorize cities
- **Nov, 2019:** PCA to understand interaction between the variables
- **Dec, 2019:** Time series analysis to predict future pollution patterns
- **Jan, 2020:** Consolidation & Report Writing

Project Planner

[illegible]

References

- [1] Public Health Foundation of India, New Delhi, 6 Dec 2018, Press Release
<https://phfi.org/wp-content/uploads/2018/12/first-comprehensive-estimates-of-the-impact-of-air-pollution-india.pdf>
- [2]Emami, Fereshteh & Masiol, Mauro & Hopke, Philip. (2017). Air pollution at Rochester, NY: Long-term trends and multivariate analysis of upwind SO₂ source impacts. The Science of the total environment. 612. 1506-1515. 10.1016/j.scitotenv.2017.09.026.
- [3]Núñez-Alonso, David & Pérez-Arribas, Luis & Manzoor, Sadia & Caceres, Jorge. (2019). Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region. Journal of Analytical Methods in Chemistry. 2019. 1-9. 10.1155/2019/9753927.

Additional References

- Fotopoulou, Eleni & Zafeiropoulos, Anastasios & Papaspyros, Dimitris & Hasapis, Panagiotis & Tsiolis, George & Bouras, Thanassis & Mouzakitis, Spiros & Zanetti, Norma. (2015). Linked Data Analytics in Interdisciplinary Studies: The Health Impact of Air Pollution in Urban Areas. IEEE Access. 4. 1-1. 10.1109/ACCESS.2015.2513439. [10]
- Shaban, Khaled & Kadri, Abdullah & Rezk, Eman. (2016). Urban Air Pollution Monitoring System With Forecasting Models. IEEE Sensors Journal. 16. 1-1. 10.1109/JSEN.2016.2514378.[11]
- Geng, Zhaowei & Chen, Qixin & Xia, Qing & Kirschen, D.s & Kang, Chongqing. (2017). Environmental Generation Scheduling Considering Air Pollution Control Technologies and Weather Effects. IEEE Transactions on Power Systems. 32. 127-136. 10.1109/TPWRS.2016.2544851.[12]
- Qu, Huamin & Chan, Wing-Yi & Xu, Anbang & Chung, Kai-Lun & Lau, Alexis & Guo, Ping. (2007). Visual Analysis of the Air Pollution Problem in Hong Kong. IEEE transactions on visualization and computer graphics. 13. 1408-15. 10.1109/TVCG.2007.70523. [13]