Understanding Air Pollution trends in various Part of Maharashtra, India

maxim rohit

Thesis Report

February 2020

ABSTRACT

Indian Government released a first of its kind report on air pollution pattern & related health impact for the country on 8th December 2018. Few of the disturbing facts presented were that 12.5% of death in India are attributable to air pollution along with the average life going down by 1.7 years owing to health loss due to high concentration level pollutant including suspended particulate matter (SPM). Previous researches directed at understanding pollutants patterns in localize area like Madrid, New York, Hong Kong internationally and Kolkata, India. These researches cover the pollution pattern in the respective Cities from various angles starting with the time of day, the season of area, impact of nearby industrial area & rural & metropolitan separation within the city among others. This Novel exploratory research is aimed at answering the question from Pollutant’s point of view for the State of Maharashtra, India. This Research will establish the relationship between four major pollutants SO₂, NO₂, SPM2.5, and SPM10. Further Extending into finding the natural grouping of cities concerning pollutants and other factors impacting them. A few examples of novel factors under consideration are Elevation from Sea level, Forest & Industrial Area Distribution among other non-novel features like population density, seasonal patterns. We found that Mumbai a commercial hub of India still maintains low levels of NO₂ and SO₂, whereas Pune being the 2nd most progressive city in Maharashtra is doing much worse but within acceptable Indian standard. Other major upcoming cities like Nagpur, Nashik & Amravati has the lowest level of pollutant concentration, signifying that the progression of the city has an impact on them. Hence proper consideration needs to be put in place while planning upcoming cities to not repeat this pattern. During PCA analysis SPM10 & NO₂ Pollutant levels were found to be highly correlated on all percentile levels signifying their common source of origin. On the other hand, SO₂ & SPM2.5 show a correlation in decreasing order from P10 till P90, signifying that spikes are caused by different sources, but latent levels are maintained by a common source. We observed a negative trend for all pollutants but one that is SO₂ during the economic recession period of 2008-10. India’s fossil fuel-based power generation attributes to 15% of SO2 pollutant based hotspot in the world, reported by Green Peace in 2019. With SPM10/2.5 levels already above the acceptable standard, time-series analysis was concentrated towards SO2 & NO2 pollutants for major cities of Maharashtra that is Pune & Mumbai. Primary observation was that rainy season has the lowest pollutant levels. These levels tend to rise in post-monsoon season raising to the peak in mid-winter and the reducing in quantum as we move towards summer.

LIST OF TABLES

[Table 1. Indian Government Pollutant Permissible Concentration Level.[6] 25](#_Toc31291460)

[Table 2. India’s Season Month wise Breakdown 31](#_Toc31291461)

[Table 3. STL model details 34](#_Toc31291462)

LIST OF FIGURES

[Figure 1. Residential pollutant levels for years 2004-2010 25](#_Toc31549699)

[Figure 2. Residential pollutant levels for years 2011-2015 25](#_Toc31549700)

[Figure 3. NO₂ & SPM10 co-relation for years 2004-2010 28](#_Toc31549701)

[Figure 4. Co-relation SO2 & SPM2.5 2004-10 29](#_Toc31549703)

[Figure 5. Co-relation SPM10, SPM2.5 & NO2 --- P10 & P90 2004-10 29](#_Toc31549704)

[Figure 6. Co-relation NO2(P10-P50), SO2 & SPM2.5 P90 2004-10 30](#_Toc31549705)

[Figure 7. Pune - SO2 Seasonality 31](#_Toc31549706)

[Figure 8. Pune - NO2 Seasonality 32](#_Toc31549707)

[Figure 9. Mumbai - SO2 Seasonality 32](#_Toc31549708)

[Figure 10. Mumbai - NO2 Seasonality 33](#_Toc31549709)

[Figure 11. Co-relation NO2, SO2 & SPM2.5/10 @ Median(P50) level, 2004-10 35](#_Toc31549710)

LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| SPM | Suspended particle concentration in air that can be inhaled is considered as air pollutants |
| SPM2.5 | Suspended particulate matter of diameter below or equal to 2.5 μm. Our nasal hair cannot prevent their inhalation and they reach our lungs and blood circulation directly. |
| SPM10 | Suspended particulate matter of diameter between 2.5 μm and 10 μm.  Inhalation is prevented by our nasal hair |
| SO₂ | Sulphur dioxide, its concentration in air is representative of the sulphur oxide family’s concentration. |
| NO₂ | Nitrogen dioxide, its concentration in air is representative of the nitrous oxide family’s concentration. |
| P10 | Monthly percentile 10 value for the pollutant. |
| P50 | Monthly percentile 50 values for the pollutant. |
| P90 | Monthly percentile 90 values for the pollutant. |
| PCA | Principal Component Analysis is a technique to uncover hidden patterns in data by collecting similar variables in orthogonal components. |
| STL | Seasonal- Trend Decomposition using Loess regression, Timeseries analysis method capable of handling multiple seasonality. |
|  |  |

ACKNOWLEDGEMENT

I would like to thank the faculty of LJMU for explaining the nuance of report writing.

Also, faculty of IIIT-B for in-depth teaching of data science concepts.

Amandeep my thesis supervisor for all the guidance & ideas during the process.

Rutuja a colleague for proofreading.

Last but not the least my wife Morien for going through endless non-user-friendly government sites in search of relevant data.

TABLE OF CONTENTS

[CHAPTER 1 18](#_Toc32332957)

[INTRODUCTION 18](#_Toc32332958)

[1.1 Background of the Study 18](#_Toc32332959)

[1.2 Problem Statement 19](#_Toc32332960)

[1.3 Aim and Objectives 20](#_Toc32332961)

[1.4 Structure of the Study 20](#_Toc32332962)

[CHAPTER 2 23](#_Toc32332963)

[BACKGROUND AND LITERATURE REVIEW 23](#_Toc32332964)

[CHAPTER 3 27](#_Toc32332965)

[ANALYSIS 27](#_Toc32332966)

[2.1 Data Preparation 27](#_Toc32332967)

[2.2 Clustering 27](#_Toc32332968)

[2.2.1 Hierarchical clustering 27](#_Toc32332969)

[2.2.2 K-Means 28](#_Toc32332970)

[2.2.3 Clustering Execution 29](#_Toc32332971)

[2.3 Principal Component analysis aka PCA 29](#_Toc32332972)

[2.3.1 PCA background 29](#_Toc32332974)

[2.3.2 PCA Execution 30](#_Toc32332975)

[2.4 Timeseries analysis 30](#_Toc32332976)

[2.4.1 Autoregressive integrated moving average aka Arima 30](#_Toc32332977)

[2.4.2 Classical decomposition Manual 31](#_Toc32332978)

[2.4.3 Seasonal-Trend Decomposition STL 31](#_Toc32332979)

[2.4.4 Time Series Execution 32](#_Toc32332980)

[CHAPTER 4 33](#_Toc32332981)

[RESULTS AND DISCUSSION 33](#_Toc32332982)

[4.1 Part I Clustering: 33](#_Toc32332983)

[4.1.1 Cluster 2004-2010 33](#_Toc32332984)

[4.1.2 Pune cluster 34](#_Toc32332985)

[4.1.3 Mumbai cluster 36](#_Toc32332986)

[4.1.4 Other Cluster 37](#_Toc32332987)

[4.1.5 Exceptional Observations 37](#_Toc32332988)

[CHAPTER 5 45](#_Toc32332989)

[CONCLUSION 45](#_Toc32332990)

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Air pollution is an ever-increasing concern for a developing nation like India, Greenpeace (the NGO) release an report[8] on 19th August 2019 stating that India is world’s number one producer for a major air pollutant SO2, around 15% of world total SO2 hotspot detected in India by NASA ‘Ozone Monitoring System’. The report highlights fossil fuel-based power generation & other Industrial use cases to be the primary source of SO2. The report highlights several power plants like Singrauli, Neyveli, & Chennai among many other as a primary contributor for India, from these Chandrapur based power plant is of special consideration for us due to its presence within Maharashtra state.

The seriousness of the air pollution situation can be comprehended by the action of the Indian government, in the budget for the current financial year that is 2020 (spanning April’20 – March’21) Indian government allocated INR 4400 crores to implementation of air pollution control measures. The ‘National Clean Air Program’ program was initiated in the 2019 budget targeted at 102 cities with a population of over 1 Million to reduce air pollution by 20 - 30% by the year 2024. The current allocation of INR 4400 crores is 10 times that that of INR 460 crores allocation in the program inception year of the Financial year of 2019.

Indian government’s first comprehensive report[1] around air pollution and its impact stated ‘1.24 million deaths in India are attributable to air pollution in 2017, of which 50+% were in individuals younger than 70 years’. A more recent example of the seriousness of this issue is the government school’s remaining closed for 2 days in Nov’2019 due to unbreathable living conditions post-Diwali celebration.

In the same report Department of Health Research representative, said ‘It is important to have robust estimates of the health impact of air pollution in every state of India to have a reference for improving the situation’. Our research is in response to the callout covering the state of Maharashtra.

1.2 Problem Statement

The research aims to understand pollutant spread, correlation and seasonal pattern over the region of Maharashtra, India. With ever-increasing sources of pollutant like-new vehicles, construction and power sources, we need better awareness of the environmental cost incurred, in the scope of our problems the environmental cost is represented by the level of air Pollutant.

Building an understanding of the pattern behind Air Pollutants is the first step towards being able to prepare the control measure. The patterns of the pollutant will be explored & explained in term of co-relation among the four pollutants (NO2, SO2, SPM 2.5/10) and additional variables under study. Seasonality is one variable that will be studied in isolation. Clustering the cities into different sets is the exercise to find commonality, rather expose differentiation between the various cities. All the above exercise will allow us to establish a pattern and co-related them with possible sources.

1.3 Aim and Objectives

1. Identify various subdivisions within Maharashtra, India based on Air pollutants like SO2, NO2, SPM and other features.
   1. Separating the high-risk zone will enable better refrainment’s.
2. Understanding the impact of the feature on the pollutant concentration.

This enables us if the feature has a major say in the pollution trends, may help us prioritize our action plan.

1. Impact of progressiveness of the city; 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.
   1. Understanding the impact of city progression will help plan future development better.
2. How does the concentration of NO2 impact SO2, SPM and vice versa?
   1. Understanding pollutants correlation will help us modulate them better.
3. Understanding the Seasonal behaviour of the pollutant concentration
   1. Enable us to strategize to the varying seasonal concentration of the pollutants.

1.4 Structure of the Study

**Part I, Clustering:**

Natural grouping of cities will be explained via clustering in terms of pollutants and properties of cities like elevation from sea level, population-density, total area, industrial area, number and type of industries among others. For analysis, Pollutants are aggregated monthly to their percentile’s values of 10, 50 & 90. The properties data is gathered from various resources and brought to the monthly level to enable merging with pollutant data.

The data was will be scaled to remove the impact of differences in the unit. Scaling is the process of bringing the distribution to mean zero and standard deviation of one. The scaled data will then be ready to run through clustering algorithms.

The similar records will be clustered together in the same set, thus creating clusters representing different pattern. The cities distribution in the cluster will allow us to analyse the similarities and more importantly the dissimilarities between them. Even the movement of one city from one cluster for a certain year or month to another for a different time frame will us help us understand the changing nature of pollutants in the city.

This will enable us to compare the progression of the cities through the years and answer questions like how do major industrial and IT hub based cities - Mumbai and Pune compare against the smaller town of Kolhapur, Nashik, and Nagpur among others.

**PART II, Principal Component Analysis:**

We will be studying the interrelation between various pollutant in this second part of the study. The analysis will again be carried across the range of pollutants percentile values of 10, 50 & 90. We will be using Principle Component Analysis (PCA) to study the underlying themes between these pollutant ranges.

PCA highlights these themes by binding the features into mutually orthogonal components which are eigenvectors. Each component tries to cover as much of residual variance as possible. By analysing features combined loading on individual components, we will be able to identify relationships among variables both positive and negative.

For example, with the knowledge of the impact of SO₂ on NO₂ concentration’s median level on a particular principal component X, we will be able to identify their (un)common trend and possible sources, eventually helping us to plan better control measures for both.

**PART III, Time Series:**

In this phase of the study, we will concentrate on analysing the seasonal component of the pollutant series. This would enable us to co-relate the possible reasoning behind spikes or dips of the pollutant with annual season pattern. This will further allow us to highlight the season which is expected to have pollutant concentration beyond an acceptable government standard.

We would be using various time series analysis techniques to quantify the seasonality and trend in our pollutant data. Arima methodology is one common method for Time series analysis we try to estimate the impact of previous week(s) value points on the next week, the impact is accessed from two angles actual value (AR component) and noise term (MA term).

The remaining I term represents the differencing required to curve to be linear.

Alternatively using the classical decomposition method, we will try to separate season & trend component using a multiplicative or additive combination of sin or cos curve. We will evaluate other techniques like Seasonal Decomposition Of Time Series By Loess (STL) as well for the 2nd part.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

Most of the studies of similar nature are targeted at explaining these variable concentrations at City level, observing the pattern over the Season and Source change/control (New York [2]) and Nature of Area (Madrid’s Metropolitan vs Rural part [3]).

New York saw a decline in sulphur-based pollutant by switching to non-diesel based fuel and using distillate oil for commercial/domestic heating. Coal-based power plants shut down and recession further helped the cause. Madrid based study, on the other hand, highlighted that the metropolitan area of Madrid was observing NO2 levels 62.9 µg/m3, 50% higher than the acceptable level of 40 µg/m3. PM10 levels, however, are within the limit for the city. The report highlights the co-relation of NO2 & O3 levels as a photochemical reaction of NO2 produced by road traffic resulting in Ozone(O3). Although the sources are co-related NO2 levels are reported higher on the Metropolitan side whereas Ozone levels are higher towards the rural part of the region. Madrid Community’s air quality improvement plan mainly focused on motor traffic-related emission reduction stands as a proof of concept that if we behave responsibly then we can overturn the spread of these pollutants.

Another study based on the city of Kolkata, India [4], attempts to quantify pollutants in different parts of the city and during different time frames within the day. Out of the 17 ambient air quality monitoring stations in Kolkata, all reported high or critical level of NO2 and SPM10 particles. The study highlighted that respiratory disease outweighs water-born counterparts 5.6:1. The report also highlighted Slum-dweller over susceptibility to air-born diseases due to indoor cooking practices. Results show morning and evening office rush hours traffic causes a spike in pollutants level.

Our study uses approaches applied in these analyses on a broader level for a set of neighbouring cities i.e. State Level while comparing the cities distinguishable attributes impact on pollutant concentrations.   
One of the innovative attempts towards explaining the pollutant behaviour via visual representation for Hong Kong city[5] also elaborates on the Time Series nature of pollutant concentrations. The study highlighted that SO2 pollutant family contribution was negligible as compare to Suspended Particulates and Ozone levels, which are the major pollutant for the city. The study also pointed out the impact of wind which is bringing in pollutant from the heart of manufacturing of china that if factories located on the Peral river delta located north-west of Hong Kong. On the contrary, Kwai Chung Pollution centre sees it’s pollution elevated with Kwai Tsing Container Terminal located towards the southwest of the city. there were other internal factors like vehicles and monopolistic power plants which contribute to an extent other than the wind-based source.

While this study concentrated on exploring the interaction between the pollutants over time, we will be concentrating on studying the seasonal pattern of pollutant patterns.

Seasonal impact of natural factors like temperature, humidity, wind speed/direction on the coastal city of Chennai [7] showed that SO₂ & NO₂ were negatively co-related to temperature during summer & monsoon season. During the post-monsoon season this co-relation changes to positive. One of the primary reason for this behaviour during summer can be attributed to uneven heath of land and sea which results in sea breeze flow towards land and reduces pollutant concentration. The rainy season, on the other hand, washes away the pollutant from the air. The post-monsoon season, however, lacks any of these advantages and hence see the temperature to be positively co-related to SO2 and NO2 concentration levels.

The same study showed that both SPM types have a positive correlation with all seasons except post-monsoon. Due to reduction in rainfall levels during the post- Monsoon season, scrubbing process of rain is reduced in quantum and as temperature increase, the wind flow & air mixing at lower height levels cause the SPM2.5/10 concentration increase, explaining the positive co-relations.

The transport of air pollutant via aerosols specifically for India were covered by in the paper by R. Ravi Krishna[9], the study highlighted that the presence of aerosols in the atmosphere affects the climate change and health of people due their presence near earth’s atmosphere. It aims at understanding the co-relation between gathering the data based on location and time and helps to understand the cause of air pollution and its effect on the health of people. It also reveals that the environmental data that India has is also very sparse as compared to other countries like the USA. However now some large-scale companies have come forward in collecting reliable data and making it available on widely so that it can be used with other studies to study the impact on air pollution.

The study also emphasizes the need for gathering reliable data and how upgrading the existing methods will help in getting the data online and better understand the chemical composition of aerosols. It highlights that sophisticated tools are still not used in India like AMS (aerosol mass spectrometer) which is an instrument that gathers real-time data and studies the chemical composition and widely used in other parts of the world. The study concludes that the awareness in India relating to the aerosols science and its studying is very minimal and can be increased by introducing some academic courses in the field of chemistry and physics focusing in environmental study.

Weisi Lin[10] & others have proposed an alternative way to measure air quality, which plays a major role in determining the air pollution levels and how it affects the health of people. This method is Heuristic recurrent air quality predictor (RAQP) to infer the air quality by studying meteorology and pollution-related variables to infer the concentration of air pollutants like PM 2.5. The paper establishes that there is a strong correlation between meteorological factors and air pollutants concentrations (APCs) which helps in predicting the air quality indices. It also highlights the challenges due to nonlinear and chaotic reasons the co-relation between the two falls with the increase in the time interval, therefore unable to predict air quality after a long period. To solve this, RAQP method suggests a 1-h prediction model which determines the air quality after 1 hour by analysing current data, which in turn is used to infer the air quality after several hours.

The study concludes that RAQD model is more effective than the other advanced technologies used for air quality prediction. It also proposes to enhance this model using ensemble learning and deep learning for better prediction.

On similar lines, another study by Khaled, Abdullah & Eman[12] focuses on a system which helps in forecasting and monitoring urban air pollution. This monitoring system consists of motes which have meteorological and gaseous sensors. They receive the data and then send it wirelessly to a platform which stores, analyse and convert the data into useful information to forecast the pollutants. They help in determining the levels of 03 Ozone, nitrogen dioxide (No2) and Sulphur dioxide (s02). The study highlights that the ML algorithms like support vector machines, M5P model trees, and artificial neural networks (ANN) are used in this system which does univariate and multivariate modelling and performance is evaluated using prediction trend accuracy and root mean square error (RMSE). Multivariate modelling with different features using MP5 algorithm showed much accurate forecasting as compared to other models and ANN was determined to give the worst results.

It also highlighted that there are many dependencies between the external factors like temperature., day and presence of humidity and evaluating the data for the presences of gases in the atmosphere.

The study concluded that the system can be enhanced to consider factors like humidity, temperature etc over time and so that it can automatically consider the changes in the atmosphere to predict more accurate data real-time data.

The indoor aspect of air pollution was discussed in a report published in 2008[11] by Kazuo, Takeki & Manisha, it highlights the ways to control the indoor air pollution and for air purification using microplasma filter as a device. It compares the general method of generating non-thermal plasma with the more efficient microplasma filter and states that while the former uses more voltage, the latter uses voltage as less as 1KV to generate the microplasma making it more efficient and cost-saving. The discharge voltage is passed between two electrodes with a gap less than 100um. This gas is used for treating indoor air pollution and removes pollutants like formaldehyde and nitric oxides. During the study, it was observed that the generation of ozone is more if the gap is reduced between electrodes whilst passing the same voltage. The microplasma removes formaldehyde and releases nitric oxide, ozone, water and C02 as resultant. It concludes that as compared to corona discharge which is one general method of generating non-thermal plasma, microplasma filter stands out with less energy consumption and low cost.

CHAPTER 3

ANALYSIS

2.1 Data Preparation

Pollutants data for SO₂, NO₂, SPM10 & SPM2.5 was aggregated for each city at monthly levels and percentile 10, 50 & 90 were calculated to represent the range. We calculated month on month increase for these percentile values for tracking their seasonal pattern. Percentile 50 values were used for the final comparison.

Other features that are merged with the data are:

1. Population & Population Density
2. Elevation from sea level
3. Total/Forest/Industrial area spread and respective percentages
4. Roads
   1. national/state/district/rural
   2. A representative of automobile circulations & count
5. Rail line
6. Industrial
   1. services/manufacturing
   2. micro/mini/medium/large

Few interaction variables were calculated like

1. Rainfall per Area
2. Industrial area per Forest area among others.

2.2 Clustering

Clustering is a technique used to partition data points into sub-sets based on its similarities.

Data points that look statistically closer to each other are combined into.

We used two algorithms K-Means & Hierarchical clustering for our problem.

2.2.1 Hierarchical clustering

There are two basic approached top-down aka divisive wherein a singles cluster is broken into smaller set over each iteration. The bottom-up approach is called agglomerative were each data point start as its cluster and is them combined.

The distance between individual data point within a group is called intra-cluster distance. One part of the algorithm optimization is to minimize intra-cluster distance, representing the fact that the closer the data points are within the cluster the more similar they are.

Intra-cluster can be calculated in the following ways: -

1. Complete Diameter Distance: It is calculated as the maximum distance between any two points within a cluster.
2. Average Diameter Distance: It is calculated as the average distance between all the pairs of points within a cluster.
3. Centroid Diameter Distance: It is calculated as the double average distance between all the points from the centroid of the cluster.

The other part of the algorithm emphasizes maximizing the inter-cluster distance that is the distance between two clusters. Greater distance between the cluster ensure higher contrast between its data points to that of another cluster.

The inter distance calculation can happen in multiple ways, variants are explained below:

1. Single Linkage Distance: The distance between the cluster is calculated as minimum distance between any two points belonging to them.
2. Complete Linkage Distance: The distance between the cluster is calculated as maximum distance between any two points belonging to them.
3. Average Linkage Distance: The distance between the cluster is calculated as the average distance between all points belonging to them.
4. Centroid Linkage Distance: The distance between the centroid of two clusters.
5. Average Centroid Linkage Distance: It’s the average distance between all point of one cluster from the centroid of the other cluster calculated for both the clusters.

The distance calculation could itself be Euclidian, Manhattan, spearman among others.

2.2.2 K-Means

K-Means Clustering originally designed for signal processing is a vector quantization technique. It was proposed by Stuart Lloyd at bell labs in 1957[13] and in Edward W. Forgy published the same method 8 years later in 1965[14], hence the algorithm is referred to as Lloyd-Forgy as well.

K-means algorithms try to divide n-observation into k clusters. The algorithm has two basic steps assignment & update. In Assignment step distance from the k- centroid is calculated for each data point, the data points are allocated to the closed centroid point and form the new cluster. In the update step, a new centroid is calculated according to the new cluster. Then the assignment step is executed and the cycle continues until no data points re-assigned during subsequent assignment steps, at this movement the algorithms are said to have converged. Convergence to global optimum is not guaranteed, the discussion around this point is beyond the scope of this study. The common choices of initial centroid can be done at randomly choosing clusters (random partition) or points (forgy), this is called the Initialization step for the process.

2.2.3 Clustering Execution

We performed two-levels of clustering, first was all variables from 2004-10 and the next set was limited to pollutant information from 1987-2015, purely due to lack of availability of data. The natural grouping found in the first exercise was validated against the longer range of the second cluster and found to be similar. All variables were scaled for consistency and equal weightage. Scaling canceled out the impact of varying units of features, for example, μg/m3 (unit for pollutant concentration) is not similar square meters(unit of measurement for land) but once it’s scaled the values ranges between 0 – 1(or -1 to 1) within a features represent its relative weight in the context.

Silhouette Analysis[15] and Elbow curves were used to decide the ideal number of clusters.

Silhouette Analysis is a graphical technique of deciding the optimal number of cluster, Silhouette represents an individual cluster’s inter & Intra distance, its value ranges between -1 to 1 with a higher value representing that the data points are well-matched in their cluster and poorly matched in other clusters.

Elbow curves are plotted by plotting the total sum of the square of intra-cluster distance for all clusters on the y-axis and number for the cluster on the x-axis. The distance drops drastically with the initial increase in the number of clusters, the drop slowly decreases in quantum until it becomes immaterial. The ideal number of clusters is chosen as the point after which the drop in the intra-distances sum of squares is considered too small.

2.3 Principal Component analysis aka PCA

2.3.1 PCA background

PCA is the process new creating linearly un-correlated features called principal components while covering maximum variance from the set of original co-related features employing orthogonal transformation. The first component tries to cover maximum variance, the 2nd tries to cover maximum variance of what remained after the 1st component and so on. Each subsequent component is orthogonal to the previous one. Each principal component is a linear combination of original features, the coefficient associated with each original feature is called the loading of the feature on the principal components. To keep these loading values representative we convert the feature individual values to its Z-score that is moving their distribution to have mean zero and standard deviation one. Calculation wise eigendecomposition and singular value decomposition are the mathematical concepts at play behind the statistical procedure, more on the implementation can be read in the paper by Tharwat, Alaa[16].

2.3.2 PCA Execution

PCA analysis was used to understand the underlying themes of the pollutant data. The relationship of pollutants can be analyzed by studying their loading on the same principal component. The negative or positive loading on a component represents how similar or dissimilar its loaded attributes are to each other. Once the relationship is highlighted by the PCA we will be studying these attributes in greater detail to list out their patterns (dis)similarities.

PCA was done on 12 variables which represent the four pollutants and their range in term of monthly percentile value of 10, 50, 90. With 4 principal components inline with four pollutant types, we were able to cover only 87% of the variation of data. This highlights that the percentile values that signified the monthly range of pollutants are following a different pattern since 13% of the variance is still unexplained. We needed all 12 PCA component to get 100% variance coverage.

2.4 Timeseries analysis

We modelled the pollutant levels using Arima & Classical decomposition methodology.

2.4.1 Autoregressive integrated moving average aka Arima

Auto-regressive part aka AR component of the variable is calculated via regressing the variable against its own lagged values. The calculation is representative of the impact of the variable on itself in the future, for example, a high pollutant concentration caused by wildfire will continue to remain high for the couple one week. Moving Average represents regression error aka residual noise is a linear combination of noises from the past, an example would be the launch of new Mobile variant not in line with annual release cycle it will raise an unexpected spike in the time series that will continue to have its impression over the next couple of weeks. Integrated aka I represents that the values subtracted from its previous values and the process can be repeated multiple times.

A non- seasonal Arima model is usually denoted as ARIMA(p,d,q), where p,q represents the order of lag of AR and MA components respectively and d represents the level of differencing. The seasonal components get accommodated in the representation as additional component (P, D, Q)m where m represents the number of the period each season, and remaining upper case character and seasonal equivalent for their lower case no-seasonal counterpart.

Arima levels were graphically evaluated with the aid of ACF( MA ) & PACF (AR) plots.

2.4.2 Classical decomposition Manual

The first step was to apply moving average smoothing on time series to remove the fine-grained variation aka noise and expose the smoother variant for actual modeling.

We used a combination of sin and cos curves to estimate the seasonal and trend component for the time series model. The combination can be additive or multiplicative as demonstrated below:-

Additive model :

*y ~ poly( period\_order , polynomial\_degree ) + sin( sin\_value \* period\_order)*

*+ cos((1- sin\_value)\* period\_order)*

*Multiplicative model :*

*y ~ poly( period\_order, polynomial\_degree\_sin\_term ) \* sin( sin\_value\_weightage \* period\_order)*

*+ poly(period\_order , polynomial\_degree\_cos\_term) \* cos((1- sin\_value\_weightage) \* period\_order)*

y = represents the dependent variables, we are predicting

Independent variables:

|  |  |
| --- | --- |
| period\_order | incremental variable tracking the duration , starting with 1 |
| polynomial\_degree (sin/cos) | degree of the polynomial being fitted (between 1-3) |
| sin\_value | weightage for period for the respective curve(.1-1) |

The residual from noise was again put through ARIMA modeling to capture any remaining pattern.

2.4.3 Seasonal-Trend Decomposition STL

STL: Seasonal-Trend Decomposition procedure based on Loess [STL reference], is an advanced algorithm that allows the trend & seasonal components gradients to change over time. With basic classical decomposition we assumed the trend and seasonality to be consistent across time, STL overcomes this shortcoming.

* + We limited our selection to additive models.
  + The forecast is obtained by applying a non-seasonal forecasting method on seasonally adjusted data.
  + With STL modelling we will be able to map the changes in our seasonal & trend pattern over the years.
  + STL flexibility to specify the window for trend & seasonality will allow us to estimate the rate at which the pattern of above component change.

KPSS and Dicky fuller tests we re-used to test the stationary of the original time series and the final residual.

2.4.4 Time Series Execution

We are concentrating on Pune & Mumbai as major cities of Maharashtra in this phase of the research. Data availability limiting our analysis to NO2 & SO2, along with that fact that SPM2.5 & SPM10 are already above acceptable levels of 40 and 60 µg/m3. We will be analysing the Residential area’s monthly median level of these two pollutants in term of time-series while concentrating on seasonality. Understanding the Seasonal behaviours will guide modularization of efforts for each season.

Other important aspect considered during time series modelling:

* We will use MAPE as quantifying measure for the fit of our model.
* We evaluated with quarterly and trimester based moving average smoothing for classical decomposition.
* We ran a grid to optimize
  + weightage and polynomial values during classical decomposition.
  + Estimating STL seasonal & trend change period
* We evaluated multiple time duration between the years 2004-2015 to get the best results.

CHAPTER 4

RESULTS AND DISCUSSION

****Table 1. Indian Government Pollutant Permissible Concentration Level.[6]****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pollutant** | **Time Weighted Average** | **Indian standard (**inµg/m3**)** | **WHO**  **(**inµg/m3**)** | **Indian to WHO ratio** |
| **Sulphur Dioxide (SO₂), µg/m3** | Annual\* | 50 | 20 | 2.5:1 |
| **Nitrogen Dioxide (NO₂), µg/m3** | Annual\* | 40 | 40 | 1:1 |
| **Particulate Matter (size less than 10 µm) or SPM10 µg/m3** | Annual\* | 60 | 20 | 3:1 |
| **Particulate Matter (size less than 2.5 µm) or SPM2.5µg/m3** | Annual\* | 40 | 10 | 4:1 |
| Note: \* Annual arithmetic mean of minimum 104 measurements in a year at a particular site taken twice a week 24 hours at uniform intervals. | | | | |

We start with a comparative analysis of acceptable levels for the four pollutants under study by the Indian Government, to set the context for our observations. Indian standard is very relaxed as compared to WHO standards specifically when it comes to the suspended particulate matter where Indian levels very relaxed 3 to 4 times when compared to WHO standards. We will discuss our finding in light of both these standards.

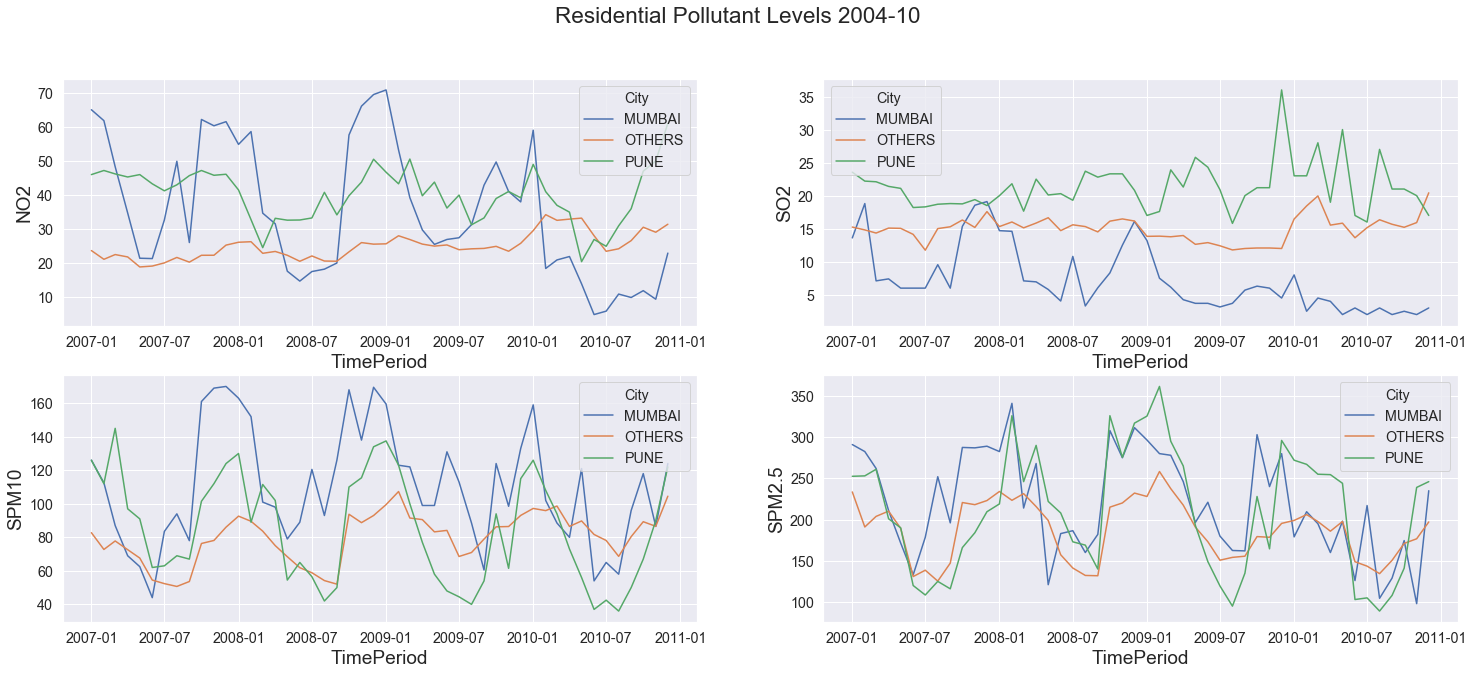
4.1 Part I Clustering:

4.1.1 Cluster 2004-2010

During clustering, we observed that the best clustering statistic was being observed when we divided the data into 6 clusters. These six clusters were set for two to represent things two different levels of pollutants for the same set of cities, That is a cluster representing Higher pollutant level for Mumbai city and another composed on Mumbai city-data point alone but with lower pollutant levels. We will continue our discussion with 3 clusters (set of two each) for elaboration. The behavior was consistent across the three area wise division names Residential, Industrial & rural areas within the set of cities.

Figure 1 represents the monthly level of pollutants of Residential area for three clusters representing Pune (Green), Mumbai (Blue) & all other cities (orange). It covers the time frame between 2004-2010 which was used for the first level of clustering with all variables.

**Figure 1. Residential pollutant levels for years 2004-2010**



Three major clusters were observed one for Mumbai & Pune each and the third one composed of all the remaining cities.

4.1.2 Pune cluster

Pune cluster has the highest level of average monthly levels of SO₂ @29µg/m3 & NO₂ @39µg/m3 over the years, still lower than acceptable levels of NO₂ 40µg/m3 & SO₂ 50µg/m3 as per Indian standard. WHO standard wise Pune did cross the acceptable standard of 20µg/m3 almost constantly post 2008.

* + The **Residential** **area** of Pune has shown an increasing pattern for months Oct, Dec, Jan post the year 2006 till 2009 for SPM2.5/10.
  + The **Industrial** **area** replicates the same behaviour and decreases during the remainder of the year. Even monthly Percentile 10 values for SPM2.5 is above the acceptable level of 40µg/m3, which when put under prospective of who standard of 10µg/m3 represents an extremely dire state of affair.
* Attribute wise Pune cluster stands out with the highest elevation i.e. 34% higher than others.
  + Highest Roads length including national, state highways and internal district and rural roads are also stand out features. These highways correspond to higher transportation mostly via diesel-based heavy vehicle that explains the higher SO2 levels.
  + Micro-industries are also highest for this cluster in terms of manufacturing & services both, questions can be asked about them following effective measures to control their share of pollutants.
  + Rainfall per area is also lowest for Pune along with the low population density and lowest percentage of forest area.

Pune’s behaviour over the clustering exercise for years 1987-2015 saw that the residential area consistently had higher values as compared to other cities whereas industrial area has seen a decline in NO₂, SO₂, and SPM2.5 levels. This behavior signifies that effective measures are being taken towards pollution control over susceptible zones like around manufacturing industries, whereas the internal residential area is suffering cause of the ever-increasing construction & commercialization of land represented by the lowest percentage of the forest area of the cluster. Lowest rainfall compared to other parts of Maharashtra allows for less natural scrubbing of pollutant, hence higher pollutional pattern are relatable.

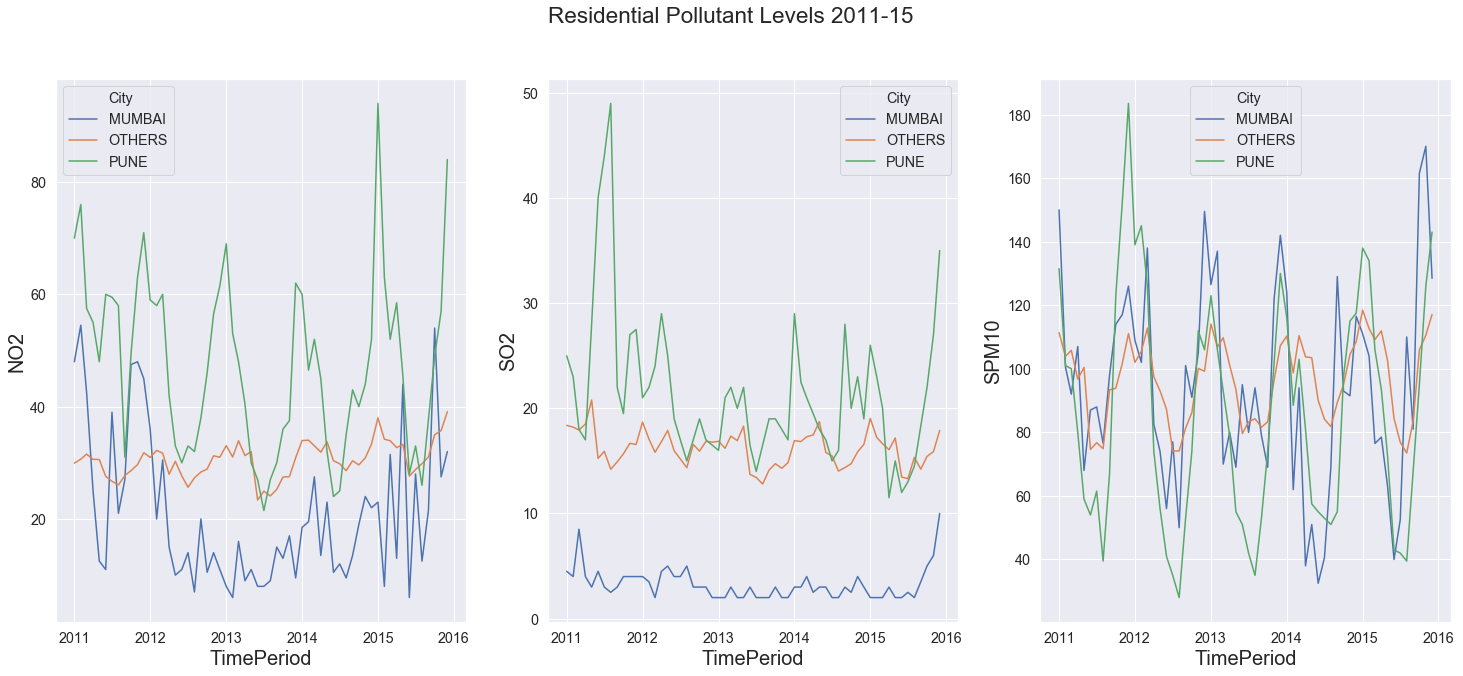
Figure 2. Residential pollutant levels for years 2011-2015

Figure 2. represents the state of the pollutants post-2010. SPM2.5 data is sparingly present for this period hence dropped from the analysis for this period.

4.1.3 Mumbai cluster

This cluster stands out for its lowest SO₂ level.

* For **Industrial** **area** level of NO₂, SPM2.5&10 decreases during Feb (Mid Summer) till Aug (Mid rainy season) and is within the acceptable standards accompanied by higher rainfall levels.
* Also, the range of all the pollutants except SPM2.5 remains under acceptable levels. During the remainder of the year the level of the three pollutants increases, specifically SPM2.5 level rises dangerously to @229 µg/m3 and SPM10 goes beyond acceptable level @124 µg/m3.
* **Residential** **area** observes the increasing phase during Aug, Oct & Dec (Post Monsoon and starting of Winter) and decreases during the remainder of the year.

Attribute wise Mumbai site on the opposite end of the spectrum with the lowest elevation & highest rainfall per area when compared to the Pune cluster. Other then the natural scrubbing via higher rainfall and the lowest elevation which alternatively can be interpreted as its being on sea level has a mazor role to play in the lower pollutant levels. As pointed out in the Coastal Urban Area [7] study the land faster heating & cooling from sea intiate a air current that helps scrubbing out the pollutants.

Another standout feature of the cluster is the largest industrial area and the highest number of large manufacturing firms all that is packed in the lowest overall area and forest portion-wise among the cities. This symbolises the effective pollution control measure that these industries have put it place, which is commendable.

1987-2015 clustering also shows similar results w.r.t pollutant concentration behaviour as compared to the other clusters. The only exception being NO₂ whose levels starts decreasing from the year 2010 and continues dropping into the next decade.

4.1.4 Other Cluster

This cluster displays a similar pattern as in Mumbai cluster’s decreasing phase but at 20-30% lower intensity, except for SO₂ concentration which is 230% higher. The higher SO2 can be attributed to 18 coal-based power generation plants[17] located around Maharashtra.

Attribute wise these cities are placed in between Mumbai & Pune in terms of elevation, rainfall per total area. These cities stand out in terms of the highest forest area per total area, the lowest population density, industrial area and count of large industries. All these factors can be considered to be having a negatively co-relation with the pollutants(other than SO2).

While validating the behaviour over 1987-2015 few exceptional observations were made:

1. Aurangabad is seeing a rise in all four pollutant levels over the years, and continuoes to do so in 2020[18], unplanned development and ever increasing count of vehicles being one of the primary cause of the poor state of affairs. Special attention based action plan is needed in order to curd the increasing rate of pollution in Aurangabad.
2. Solapur and Mahad are the only cities reducing the level of SPM2.5, the pollutant with the most extreme values in all other cities.

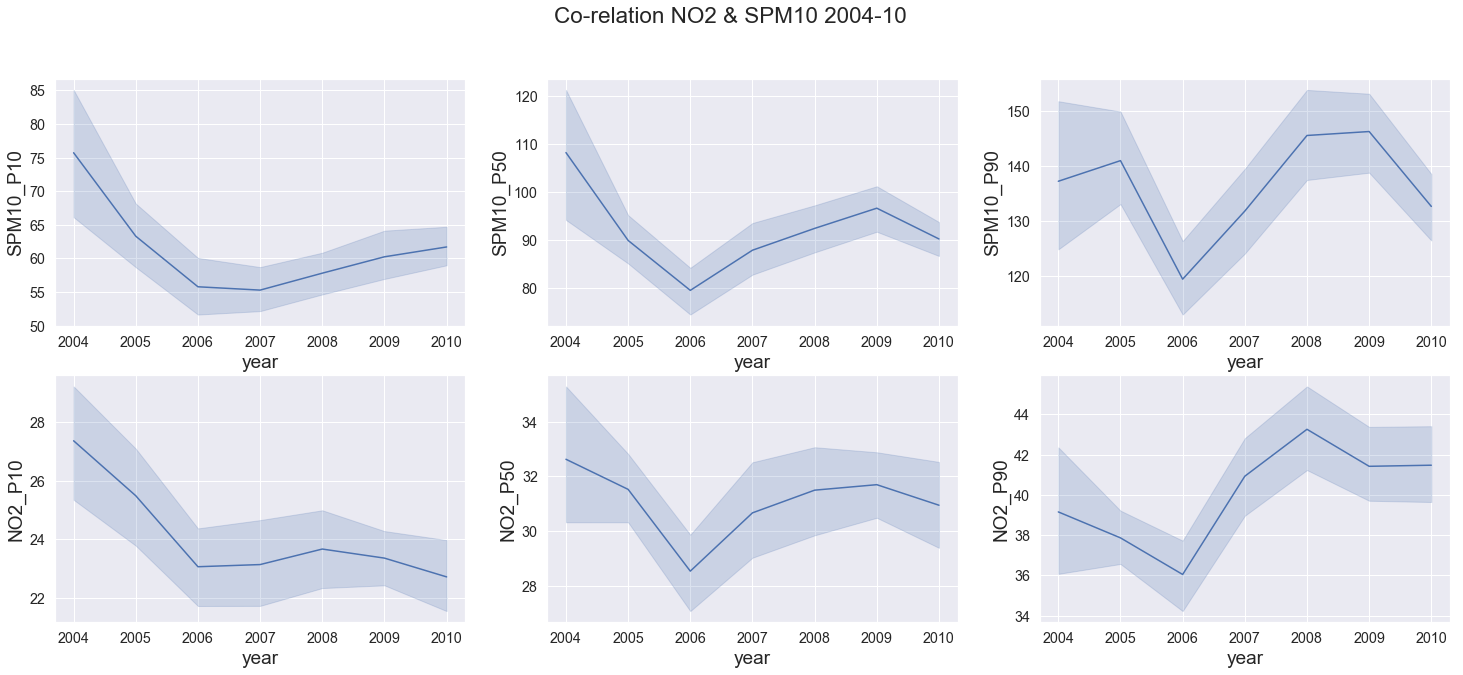
4.1.5 Exceptional Observations

Chandrapur is an exception to these clusters and shows the highest value for SO₂ @24.3µg/m3 (still acceptable), RSPM2.5 & 10 (extremely high) from all the clusters with continous monthly increase in RSPM2.5 & 10 values over the years 2004-2010. Chandrapur Super Thermal Power Station, the coal-based powerplant is one of the major sources for SO2 levels in the region.

For the years 2006-2010 and months Sep, Oct & Dec Nagpur displayed similar behaviour, again it can be attributed to another coal-based Koradi Thermal Power Station.

**PART II PCA:**

Figure 3. NO₂ & SPM10 co-relation for years 2004-2010



First principle component has a high loading for SPM10 & NO₂, figure 3 displays the annual tread across Maharashtra, the relation is highlighted at al1 three-level P10, P50 & P90 with co-relation being 0.49, 0.52 & 0.56 increasing with percentile.

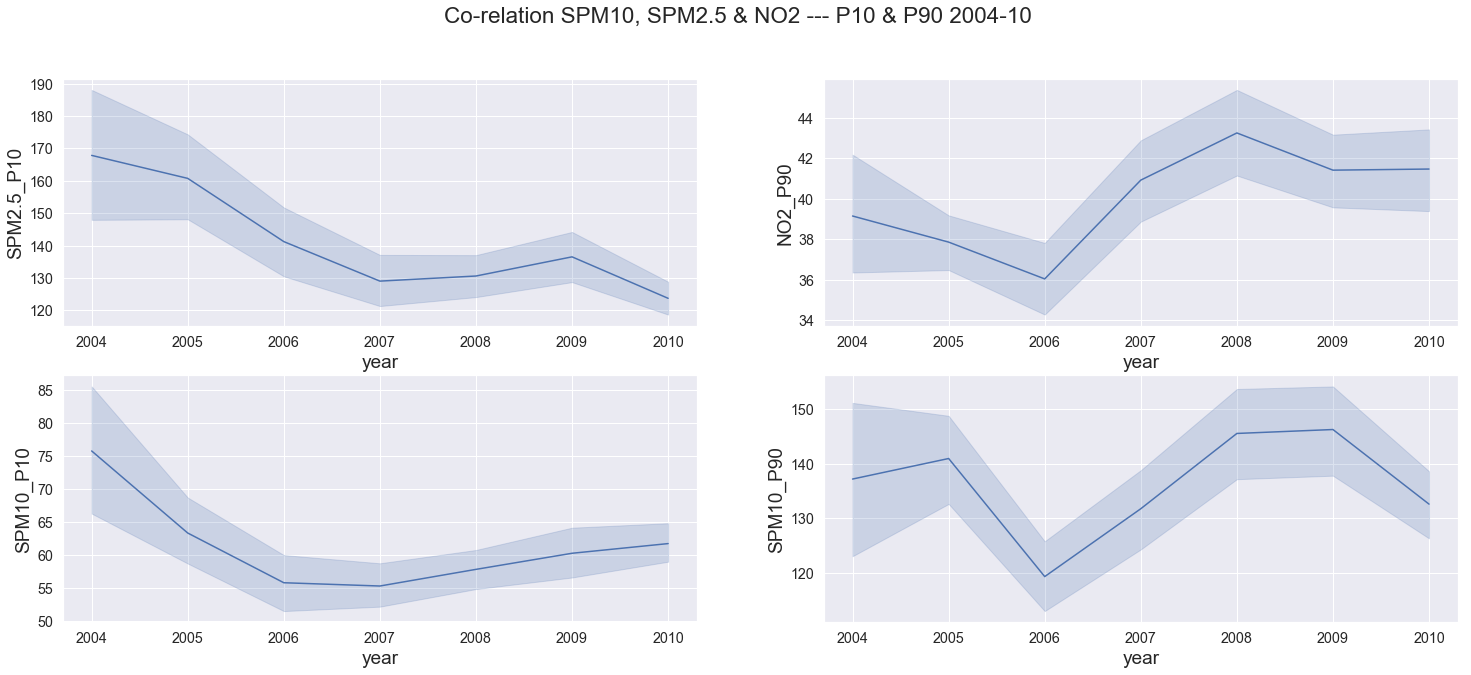
Except for SMP10 P90 growth vs drop of NO₂ P90 between 2004-05, the trend matches for all other years. The matching trends signify a common source mostly fossils fuels combustions (commercial & individual) and industrial activities, Maharashtra is also known as the industrial capital of India.

Figure 4. Co-relation SO2 & SPM2.5 2004-10****

On the 2nd principle component, SO₂ has high positive loading and inversely SPM2.5 has negative loading. Their behaviour across the year as depicted in figure 4 confirm similar movement before 2007 (SPM2.5 has greater quantum) and 2007 onwards both variables start to move in inverse directions.

One exception prominent was between 2004-05 for P90 values of SPM2.5 which saw an increase while all other values decreased. This signifying that sources of SPM alone were at play like construction work, natural wind-based debris.

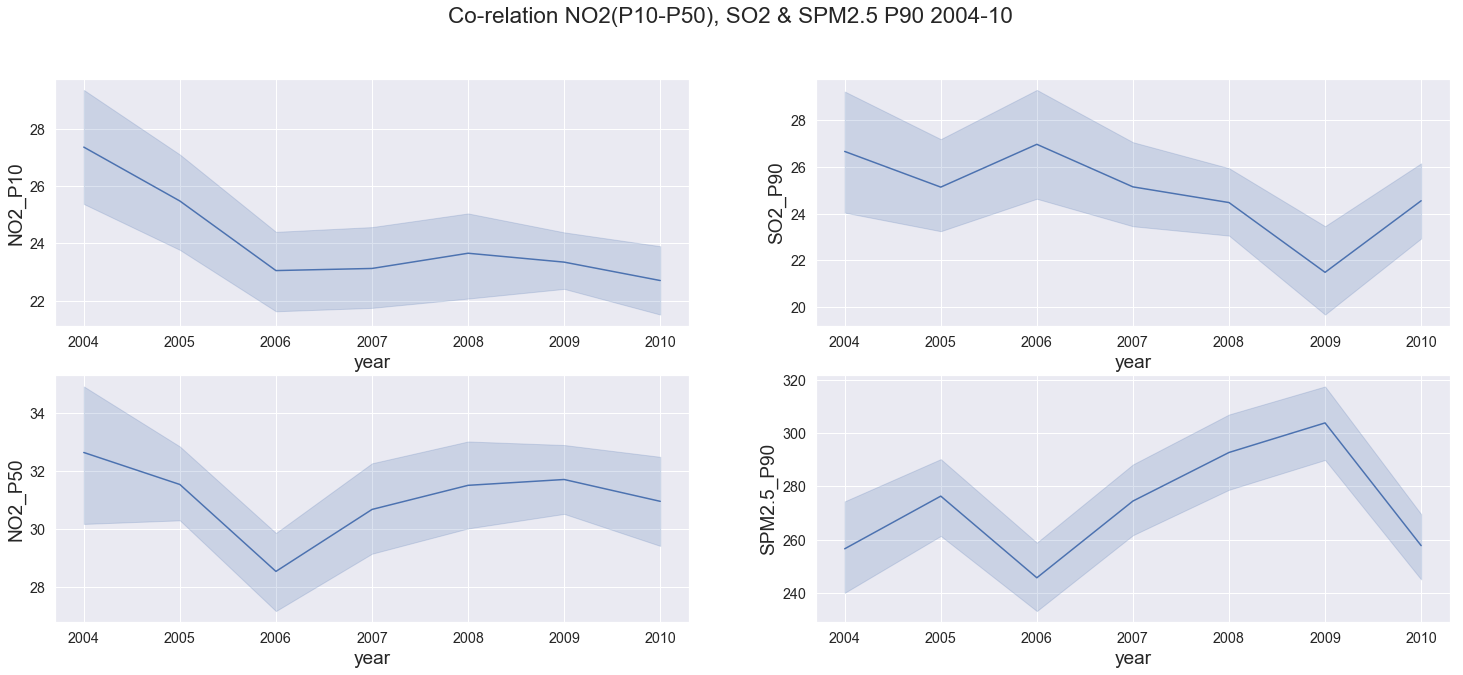
As depicted by their corresponding negative and positive loading in the component the co-relation is non-existence below 0.07.

Figure 5. Co-relation SPM10, SPM2.5 & NO2 --- P10 & P90 2004-10****

SPM2.5 & SPM10 are expected to be co-related, their P10 values loading on the 3rd principal confirm that relationship. An exception is observed for the year 2009-10 where both variables move in the opposite direction. Their correlation is above .66 at all percentile levels.

P90 values for SPM10 & NO₂ have negative loading on this component and drastically increase between 2006-08. SPM10 percentile ranges depicted by P10 & P90 show very different behaviour, highlighting the fact that the range is increasing owing to more days with higher pollutant concentration during the years between 2006-2008.

A consistent observation in all this variable expects SO₂ is a negative change in trend between 2008-10 coinciding with the impact of Great Recession [19] that affected the USA between Dec07-June09 and whose after effect was felt in India from Sep’08-Sep’09. SO2 levels souce being the coal-based power plant remians the prominent reason for the variables not dropping in during recension.

Figure 6. Co-relation NO2(P10-P50), SO2 & SPM2.5 P90 2004-10

Fourth principle component has NO₂ positively loaded and SPM2.5 & SO₂ negatively loaded.

The raise in NO₂ P50 levels (while P10 remained constant) post year 2006 stretches the range of the pollutant signifying that its level went up overall and are persistence.

On the other hand, the movement of SPM2.5, SPM10 & NO₂ are similar and SO₂ seems to be following a different pattern observed here and overall.

**Part III- Time Series analysis:**

STL outperformed Arima and manual classical decomposition techniques to give the best results. We are going to discuss the seasonality component in further details as trend components for both pollutant that is NO2 and SO2 for both the cities Pune & Mumbai is on a slow decline.

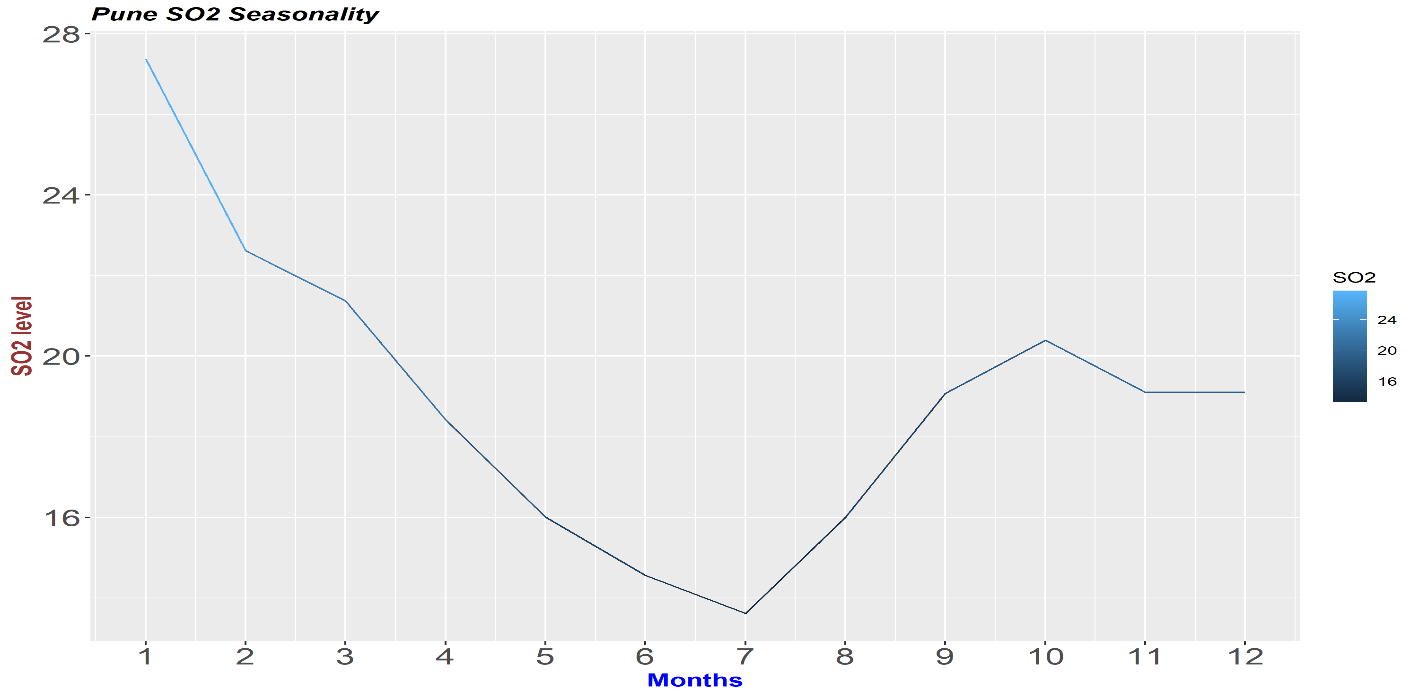
Table 2. India’s Season Month wise Breakdown

|  |  |
| --- | --- |
| **Season** | **Month Range** |
| **Winter** | **December - February** |
| **Summer** | **March-May** |
| **Rainy(Monsoon)** | **June - September** |
| **Post-Rain** | **October - November** |

For the discussion scope, we have divided the year into four reference seasons as shown in

**Table 2.**

Figure 7. Pune - SO2 Seasonality

****

End of summer in April with Rainy season between June (6)-August (8) sees the lowest levels of SO2 concentration below 16 µg/m3.

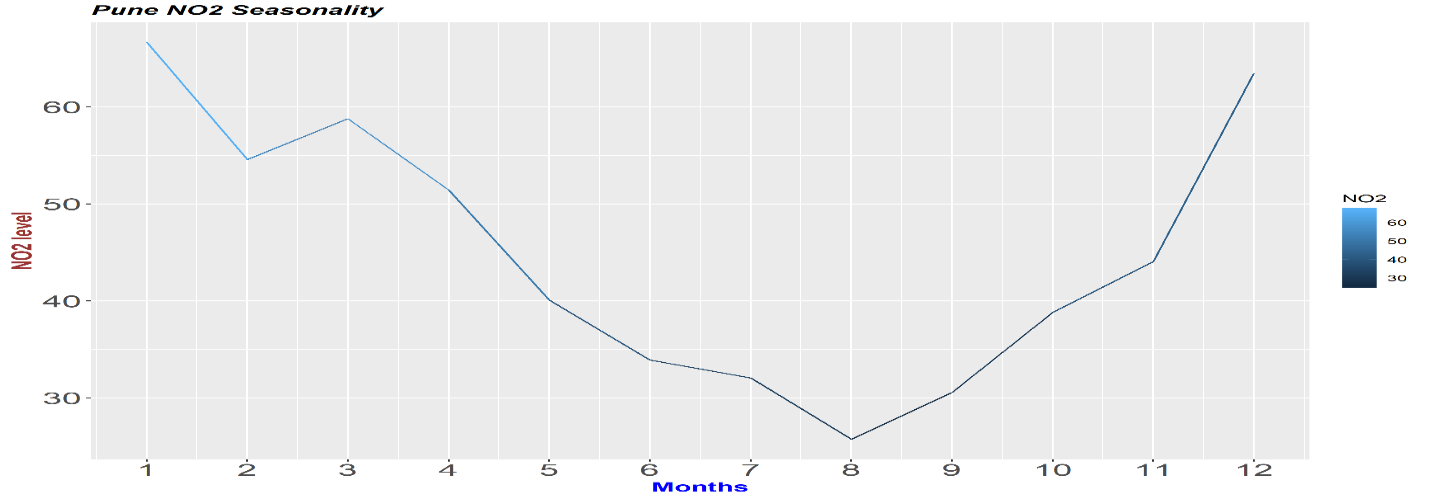
Post monsoon rainy season there is a 25% increase in SO2 levels by October (10), followed by another stagnation with the beginning of winters in December. Winter sees the highest level in Jan & Feb followed by continues decline till the start of Monsoon Season again. Jan & Feb are the only duration when the SO2 levels goes ablve WHO set standard of 20 µg/m3.

This pattern is consitant with Pune cluster observation where we observed spikes for Oct , Dec & Jan for SPM2.5/10 for residential area.

One of the cause for winter spike is the increased vehicular activity caused due to the festivity in india during this time. Diwali being one of the biggest festival in india almost all auto manufacture come out with special offers , followed by Christmas and year end deals, greater sales represents more vehicle to cause pollution. Maharashtra have almost always posted highest vehicular sales among other states of india, only recently to be bettered by Utar Pradesh only for few quarters.

Other forms of public & private trsnportation also reaches its peak for personnel and instruments specially with e-commerce era. Festival of Diwali itself caues a two day massive spike during Oct specificaly for SPM2.5 concetration due to cracker busting tradition.

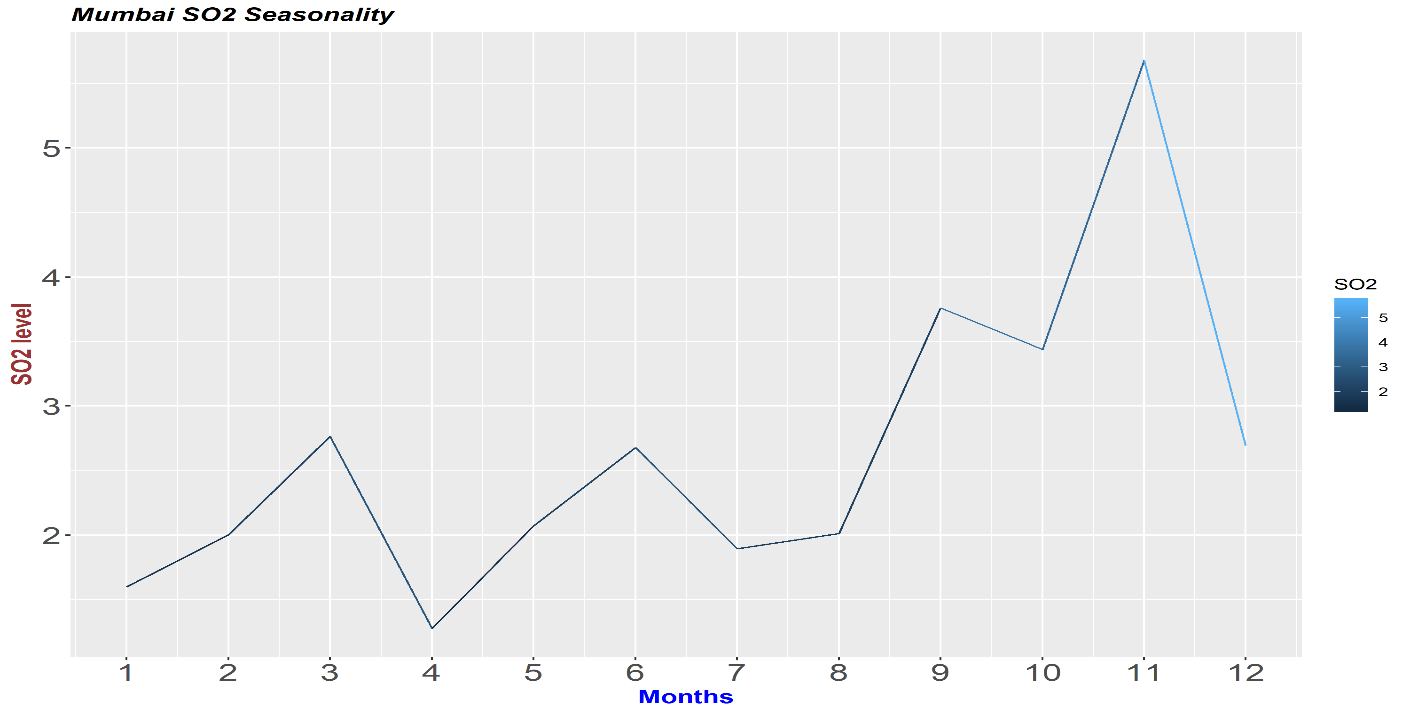
The reasonsiening applied to all pollutnats as End of year spikes trends are observed in all pollutant time series discussed below.

Figure 8. Pune - NO2 Seasonality

Again during the Monsoon season lasting from June(6)-September(9) we see the lowest level of NO2 between 25-35 µg/m3, acceptable under Indian and WHO standards. Post monsoon season NO2 level start to increase reaching its peak midway in winter in January(1) around 65 µg/m3. Start dropping down post-January winter till it reaches in a low point in monsoon again.

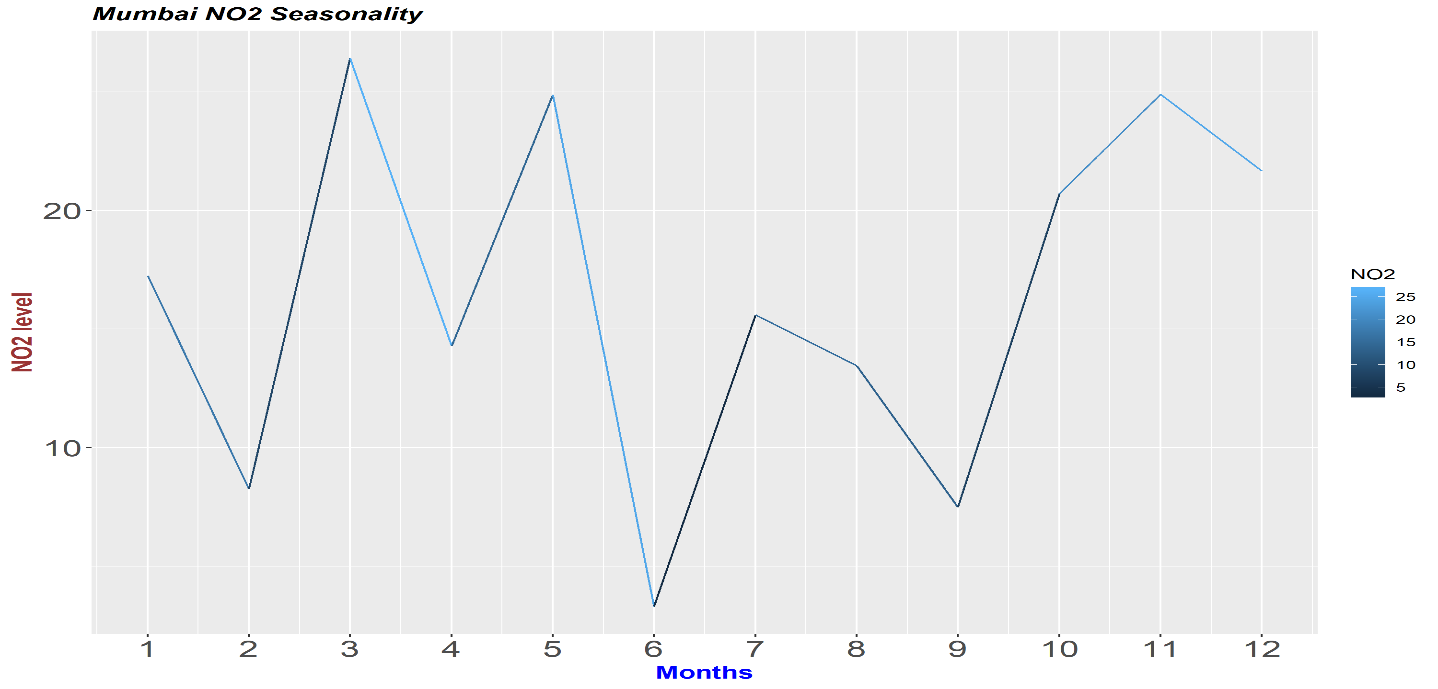
November to April which represents a half year phase for Pune observes above Indian & WHO acceptable levels for the pollutant, in this case, NO2 of 40 µg/m3.

Figure 9. Mumbai - SO2 Seasonality

****

The first 2/3rd of the year sees very low SO2 levels below 3 µg/m3 with minima being below 1.5 µg/m3 in April. As reported in the clustering exercise Mumbai maintains the lowest So2 levels which are confirmed in these finding. End of monsoon rainy season with September sees a sharp rise which reaches its pick by post-monsoon season end that is November with SO2 levels doubling to 6 µg/m3.

Even the spikes are very low within 6 µg/m3 for Mumbai well within the acceptable range of 50 µg/m3 & 20 µg/m3 fro Indian and WHO standards respectively.

Figure 10. Mumbai - NO2 Seasonality

Lowest NO2 Levels are observed at the start of monsoon season lasting till the entirety of the season. The post-monsoon season sees a sharp spike with NO2 levels almost doubling at the start of post -monsson season in October. Winter season see slow decline afterwards with start of summer bringing a similar spike with the season change, were in NO2 levels triple in March as compared to February. Summer sees a zig-zag pattern with march spikes followed by a comparative drop in April, similar to SO2.

Overall Mumbai & Pune sees the lowest level in rainy season in its pollutant level followed by highest levels in the post-monsoon season. Quantum wise Mumbai maintains accpratbel standard with repsetc to India & WHO guidelines, whereas Pune NO2 pollutant concentration are higher the accpetabel standard for almost half the year cober post-monsoon & winter season. SO2 levels for pune although are under Indian standards but WHO standards are not met for January and feurary months.

Table 3. STL model details

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Pollutant** | **Seasonal** | | | **Trend** | | | **Looes** | | | **Mape** |
| **window** | **jump** | **degree** | **window** | **jump** | **degree** | **window** | **jump** | **degree** |
| **Pune** | **SO2** | **9** | **1** | **1** | **1** | **9** | **1** | **9** | **9** | **1** | **14.04** |
|  |  |  |  |  |  |  |  |  |  |  |
| **NO2** | **7** | **7** | **1** | **1** | **17** | **1** | **5** | **17** | **0** | **13.76** |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **Mumbai** | **SO2** | **9** | **5** | **1** | **5** | **17** | **1** | **1** | **17** | **1** | **24.6** |
|  |  |  |  |  |  |  |  |  |  |  |
| **NO2** | **7** | **1** | **1** | **5** | **1** | **1** | **1** | **25** | **1** | **30.24** |
|  |  |  |  |  |  |  |  |  |  |  |  |

Loess window for seasonal extraction that gave best results for NO2 & SO2 were 7,9 respectively, indicating that SO2 patterns are slower to change. The slow nature of change might again be subject to its source as coal-based power generation being one of the primary producer control its levels.

On one side we see the seasonal behaviour to be pollutant specific whereas trend follows a city-based pattern. Mumbai showing best results with 5-month window smoothing for trend change whereas Pune displays best results with no smoothing.

We can establish that their a change in trend every 5 months for Mumbai whereas Pune trend is stagnant.

CHAPTER 5

CONCLUSION

**Clustering**

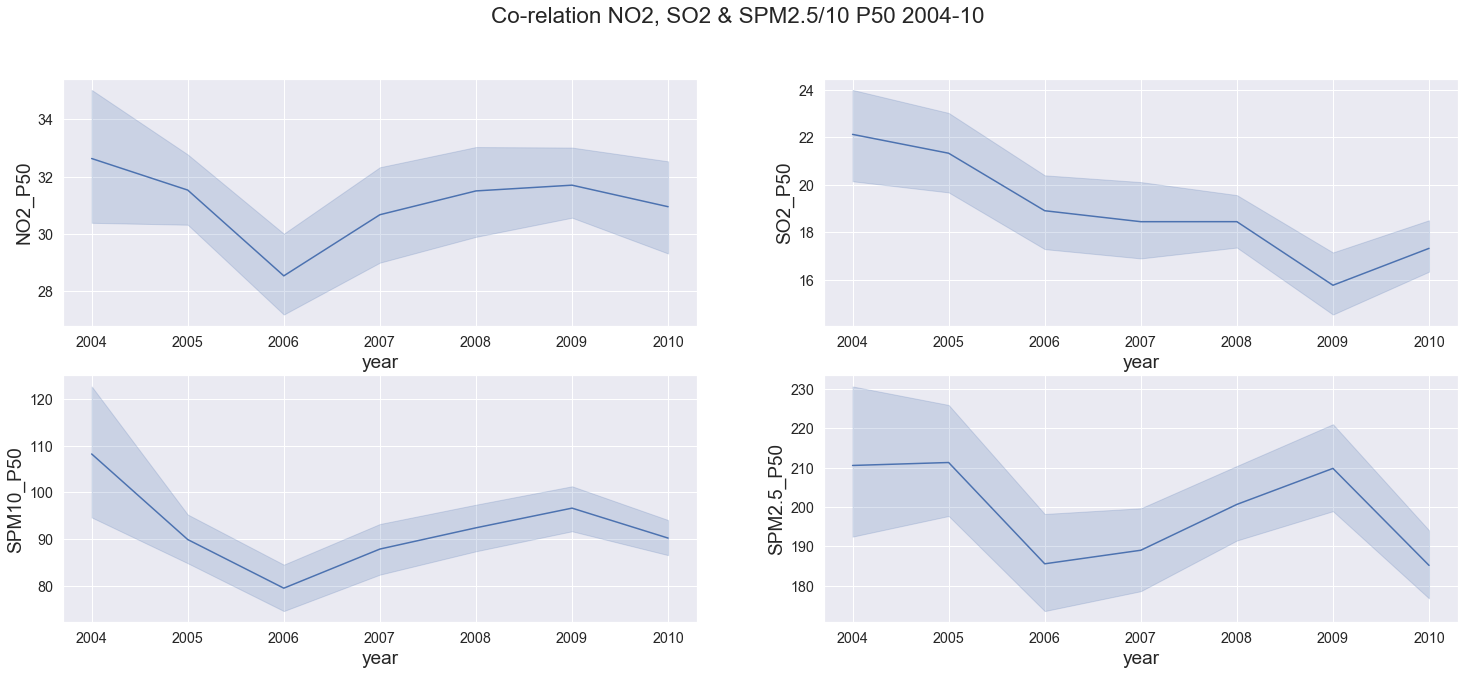
Mumbai although heavy on the industrial area and large industries count still registered smaller pollutants concentration numbers, this can be co-related to its properties of lower elevation and high rainfall. Pune is on the other end of the spectrum with higher elevation and lower rainfall and shows the highest pollutant concertation levels.

All other less progressive cities of the state fall under the same umbrella. They exhibit the least pollutant concentration (other than SO₂) correlated with the higher forest area percentage, lower industrial area and large industries counts. This behaviour answers our question about the city’s progression resulting in higher pollutant level, which should point us in the direction of better planning for future development keeping measure to control air pollutant in mind.

Lowest SO₂ level observed in Mumbai warrants further research, its relationship to humidity may be a factor here as Mumbai has highest humidity number owing to its closeness to the Arabian Sea.

**PCA**

Figure 11. Co-relation NO2, SO2 & SPM2.5/10 @ Median(P50) level, 2004-10



PCA highlighted that SO₂ concentration level followed a different trend as compared to other pollutants.

**NO₂ & SPM2.5/10:** We observed that post-2006 the increasing trend of all pollutant reduced in amplitude yearly and became negative by 2009.

SO₂, on the other hand, has seen a decreasing trend throughout the years only to increase post-2010**.**

SO₂ decreasing co-relation with NO₂ ranges from 0.55, 0.41 & 0.37 at P10, P50 & P90 level respectively. This highlights that their spikes are caused by uncommon sources, whereas the common sources like industrial/commercial fossil fuels combustion are maintaining the relationship at lower percentile levels.

**Time-series**

Overall both cities see a spike in its pollutant level in post rainy season continuing into the winter season.

Mumbai patterns are more erratic as compared to Pune. Pune mostly seems decline starting from summer reaching the lowest point in Monsoon and start spiking during winters. Mumbai follows a similar pattern but tends to spike at the starting month of the Monsoon season June for SO2 and July for NO2.

From these only NO2 levels of Pune are above acceptable Indian pollution standards that too for a good part of the year from December to April are above Indian acceptable levels, we need to put in measure to control the post-monsoon and winter spikes.

RSPM2.5/10 levels are also very high throughout Maharashtra and urgent effective measure needs to be taken to bring their levels under control.

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-pollut ion-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.

[4] Haque, M.S. & Singh, R.B.. (2017). Air Pollution and Human Health in Kolkata, India: A Case Study. Climate. 5. 10.3390/cli5040077.

[5] 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.

[6] Government of India, Ministry of Environment, Forest & Climate change, Permissible Level for Pollutants: <http://www.indiaenvironmentportal.org.in/files/file/Permissible%20Level%20for%20Pollutants.pdf>

[7] Ramasamy Jayamurugan, B. Kumaravel, S. Palanivelraja, and M. P. Chockalingam, “Influence of Temperature, Relative Humidity and Seasonal Variability on Ambient Air Quality in a Coastal Urban Area,” International Journal of Atmospheric Sciences, vol. 2013, Article ID 264046, 7 pages, 2013. <https://doi.org/10.1155/2013/264046>.

[8] Dahiya, S. and Myllyvirta, L. (2019). India largest SO2 emitter in the World, says Greenpeace’s new analysis - Greenpeace India. [online] Greenpeace India. Available at: https://www.greenpeace.org/india/en/press/4015/india-largest-so2-emitter-in-the-world-says-greenpeaces-new-analysis/ [Accessed 3 Feb. 2020].

[9] Krishna, R.. (2012). Current atmospheric aerosol research in India. Current Science. 102. 440-451.

[10] Gu, Ke & Qiao, Junfei & Lin, Weisi. (2018). Recurrent Air Quality Predictor Based on Meteorology- and Pollution-Related Factors. IEEE Transactions on Industrial Informatics. PP. 1-1. 10.1109/TII.2018.2793950.

[11] Shimizu, Kazuo & Sugiyama, Takeki & Samaratunge, Manisha. (2008). Study of Air Pollution Control by Using Micro Plasma Filter. Industry Applications, IEEE Transactions on. 44. 506 - 511. 10.1109/TIA.2008.916738.

[12] 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.

[13] Lloyd, Stuart P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, Stuart P. (1982). "Least squares quantization in PCM" (PDF). IEEE Transactions on Information Theory. 28 (2): 129–137. CiteSeerX 10.1.1.131.1338. doi:10.1109/TIT.1982.1056489. Retrieved 2009-04-15.

[14] Forgy, Edward W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics. 21 (3): 768–769. JSTOR 2528559.

[15] Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. Journal of Computational and Applied Mathematics. 20. 53-65. 10.1016/0377-0427(87)90125-7.

[16] Tharwat, Alaa. (2016). Principal Component Analysis (PCA) : An Overview.

[17] “Category:Coal-fired power stations in Maharashtra,” Wikipedia, 20-Sep-2019. [Online]. Available: https://en.wikipedia.org/wiki/Category:Coal-fired\_power\_stations\_in\_Maharashtra. [Accessed: 10-Feb-2020]

**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]
* 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]

APPENDICES

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](http://dcmsme.gov.in/%3cCityBasedURL)>
* Population:-https://mahasdb.maharashtra.gov.in/population1.do