# STUDY ON AIR QUALITY OF SAFAR CITIES OF INDIA

*Project submitted to the University of Calicut in partial fulfilment of the Requirements for the Award of the Degree of Master of Science in Statistics*

**PRAISY THANKACHAN**

**(Register No.DVAUMST015)**


**DEPARTMENT OF STATISTICS**

**ST. JOSEPH'S COLLEGE DEVAGIRI,CALICUT-8,**

**JULY 2022**

# STUDY ON AIR QUALITY OF SAFAR CITIES OF INDIA

Submitted by

Praisy Thankachan

Register.No: DVAUMST015

# DEPARTMENT OF STATISTICS

# ST. JOSEPH'S COLLEGE (AUTONOMOUS)

# DEVAGIRI, CALICUT

## CERTIFICATE

This is to certify that the dissertation entitled "**STUDY ON AIR QUALITY OF SAFAR CITIES OF INDIA**" is a genuine record of the Project done by **PRAISY THANKACHAN**. Under my guidance and supervision during the year 2020-2022. The dissertation submitted to the Department of Statistics, St. Joseph's College Devagiri for the partial fulfilment of the requirement for the award of Master of Science degree in Statistics.

**Dr. P. ANIL KUMAR**

**Head of the Department**

**Department of Statistics**

**St. Joseph's College (Autonomous) Devagiri, Calicut**

**Calicut**

**July 2022**

# DECLARATION

I **PRAISY THANKACHAN**, do hereby declare that this dissertation entitled **STUDY ON AIR QUALITY OF SAFAR CITIES OF INDIA** submitted to St. Joseph's College Devagiri affiliated to the University of Calicut in partial fulfilment of the requirement for the award of master of science is bona fide record of research work carried out by me during 2020-2022,under the supervision and guidance of Dr P. Anil Kumar

**PRAISY THANKACHAN**

**(Register No.DVAUMST015)**

**Department of Statistics**

**St. Joseph's college Devagiri**

**Devagiri**

**Calicut**

# ACKNOWLEDGEMENT

First of all, I express my heartfelt gratitude to God, the almighty, without whose blessing this endeavour would not have been a reality in time.

It is a matter of great pleasure for me to express a deep sense of gratitude to my supervisor who enlightens my thoughts through his valuable guidance throughout the period of this dissertation work.

I express my sincere thanks to the Principal, **Dr. SABU K THOMAS**, for providing facilities to me.

I would like to thank, **Dr**. **P.ANILKUMAR,** Head of the department of Statistics for his helpful advices and the facilities in the department to carry out the dissertation work.

I wish to express my special thanks to Mr Jomon Jose and Mrs. Aswani Muraleedharan M for guiding throughout the dissertion work.

I extend my sincere thanks to all their encouragement and also to the library staff for their timely assistance.

I remember my family for always providing me with moral and mental support throughout the work. I also wish to thank my friend for their assistance, support and words of goodwill during the period of my dissertation

# CONTENTS

# CHAPTER 1

# ABSTRACT

Pollution is defined as the presence of impurities or pollutant substances in sufficient concentration levels, causing harmful effects on human beings, animals, plant life or material resources when exposed for a sufficient duration of time, thus reducing the quality of life in the environment. Pollutants include solid, liquid or gaseous substances present in greater than natural abundance, produce due to human activity, which have a determined effect on our environment. Environmental pollution is one of the greatest challenges that the world is facing today.

Air is an invisible substance surrounding the Earth and providing us all with the breathable oxygen and performs a vital role in supporting life on Earth. But with the passage of time the fresh and pure air is gradually getting contaminated due to increase in air pollution. Air pollution is the presence of one or more substance at a concentration above their natural levels, with the potential to produce an adverse effect. Air pollution can be human-made or occur naturally in the environment. Human-made pollutants are caused by fossil fuel combustion, industrial manufacturing, waste-burning, dust from traffic, smoke, and exhaust from vehicles, ships and airplanes.

Air pollution is a major global public health risk in cities across the world. It is one of the highest-ranking environmental health challenges in the world, especially in developing countries like India. The increasing level of pollutants in ambient air has deteriorated the air quality of Indian cities at an alarming rate. This brought us to focus our study on air quality to SAFAR cities. The prediction of future air quality has been carried out by analysing the pollutants through data analysis techniques. Statistical models and methods can be used to advance knowledge and understanding of different aspects of the pollution. Multivariate analysis and time series analysis are employed for the analysis of the data. Statistical software such as SPSS and Microsoft Excel and R Programming are used to analyse data.

# CHAPTER 2

# INTRODUCTION

According to United Nations Environmental Program (UNEP), presence of substances and heat in environmental media (air, water, and land) force nature, location or quantity produces undesirable environmental effects. The rate at which urban air pollution has grown across India is alarming. A vast majority of cities are caught in the toxic web as air quality fails to meet health-based standards. Almost all cities are reeling under severe particulate pollution while newer pollutants like oxides of nitrogen and air toxics have begun to add to the public health challenge. In the above context, we felt, if we closely study the Air Quality Data for major cities in India, we should be able to identify patterns (spike in air pollution levels), identify correlating factors on key levels of Air Pollution across key locations of India. Through this study we hope to develop some insights that can help organizations (State/Central Pollution Control Boards and NGOs and general awareness among public) to advocate more stringent policy frame work to control air pollution.

In this project we basically look into the SAFAR cities of India. The SAFAR cities refers to the metropolitan cities of India. They are: AHMEDABAD, PUNE, MUMBAI AND NEWDELHI .Levels of air pollutants in Indian cities, including national capital New Delhi, are on the rise. India has the distinction of releasing the largest volumes of pollutants into the air after China. Majority of the cities with the most polluted air are in India and WHO goes so far as to call them death traps. Air pollution is the fifth largest killer in the country, according to Global Burden of Disease. Most vulnerable will be the old, children, homeless and poor sections of the society. Indoor air pollution caused by the burning of fuel wood and coal leads to formation of $CO_2$ and CO along with hydrocarbons. Chronic lung disorders, cancer, prenatal deaths and low birth weight are a common occurrence due to air pollution. Industrial air pollution from petroleum refineries, chemical industry, paper and dye industries is causing severe damage to the ecology as well as several man-made structures. The losses caused due to mortality and morbidity in humans due to industrial pollution, when accounted for, would run into crores. Vehicular pollution will trigger many respiratory ailments, as traffic speed has come down considerably due to congested roads. India is providing low-sulphur diesel in only some of its cities, and the rest of the country uses high-sulphur diesel for its buses and trucks that spew noxious sulphur and nitrous oxides into the air. Therefore air pollution must be a matter of discussion.

# 2.1 SAFAR CITIES

The System of Air Quality and Weather Forecasting And Research (SAFAR) is a national initiative introduced by the Ministry of Earth Sciences (MoES) to measure the air quality of a metropolitan city, by measuring the overall pollution level and the location-specific air quality of the city.The system is indigenously developed by the Indian Institute of Tropical Meteorology (IITM), Pune and is operationalized by the India Meteorological Department (IMD).It has a giant true colour LED display that gives out real-time air quality index on a 24x7 basis with color-coding (along with 72 hours advance forecast).

The ultimate objective of the project is to increase awareness among the general public regarding the air quality in their city so that appropriate mitigation measures and systematic action can be taken up. SAFAR is an integral part of India's first Air Quality Early Warning System operational in Delhi. It monitors all weather parameters like temperature, rainfall, humidity, wind speed, and wind direction, UV radiation, and solar radiation. The Pollutants monitored are $PM_{2.5}$, $PM_{10}$, Ozone, Carbon Monoxide (CO), Nitrogen Oxides ($NO_2$), Sulphur Dioxide ($SO_2$), Benzene, Toluene, Xylene, and Mercury. The World Meteorological Organization has recognized SAFAR as a prototype activity on the basis of the high-quality control and standards maintained in its implementation. SAFAR system would benefit cost savings to several other sectors like agriculture, aviation, infrastructure, disaster management, tourism, etc. which directly or indirectly gets affected by air quality and weather. The SAFAR cities are namely:

1. AHMEDABAD
2. PUNE
3. MUMBAI
4. NEWDELHI

## 2.2 SIGNIFICANCE OF STUDY

The analysis of the level of AQI of the metropolitan cities of India can be an eye opener to the level of air pollution that we suffers and can also shows us the main factor behind air pollution being the fifth largest killer of Indian population.The SAFAR cities under the project are:

- **AHMEDABAD**

Ahmedabad is one of India's largest and fastest growing cities with a population over 7.3 million. The World Health Organization (WHO) urban air quality database , and several international and Indian studies have identified Ahmedabad as one of the most polluted cities in the world. In an effort to protect local communities from rising air pollution levels, the Ahmedabad Municipal Corporation (AMC) is developing an air quality index (AQI) with the technical expertise of the Indian Institute of Tropical Meteorology, Pune (IITM) and SAFAR (System of Air Quality and Weather Forecasting And Research). With rising air pollution levels and deadly health risks, leading cities have developed clean air programs using the AQI. To support the AQI in protecting citizen health in Ahmedabad, the Indian Institute of Public Health Gandhinagar (IIPH-G) and the Natural Resources Defence Council (NRDC) are working with the AMC on information, education, and communication strategies for the new AQI launched in Ahmedabad. The combined efforts of government agencies, health professionals, and community leaders can serve to effectively inform the public about rising air pollution health risks in India, and how to take steps to protect community and individual health. This issue brief serves to make key recommendations for the AMC in

developing a city-wide health risk communication plan based on the AQI. This issue brief has two parts.

The first part focused on Ahmedabad and the AQI system with three sections:

- the first section covers air pollution and associated health impacts in Ahmedabad,
- the second section covers the AQI,
- the third section covers the elements of applying a successful AQI.

The second part focuses on the health impacts of air pollution and international practices on health risk communication.

## ➢ PUNE

The air quality in Pune is deteriorating fast, with pollution levels in the city crossing the 'moderate' mark of the Indian National Ambient Air Quality Standards, as per a recent studies .The study — titled Anthropogenic fine aerosols dominate over the Pune region, Southwest India — was jointly taken up by the Indian Institute of Tropical Meteorology (IITM), the department of geology at the Savitribai Phule Pune University and Stockholm University. It tracked levels of particulate matter (PM) 2.5 and 10 over Pune during 2016 and found that the highest contributor to the city's deteriorating air quality was carbonaceous fine particles.

## ➢ MUMBAI

Almost half of the city's population is at risk of suffering from serious, long-term health problems. The pollution level in the air in the city has risen to 136 micrograms, which is a dangerous indication of rise in pollution in the city. The level has exceeded the permissible limits set by the world health organisation .Mumbai was at 12$^{th}$ position in the WHOs Global Air Pollution Report of year 2014 ,while India ranked 9$^{th}$ out of 95 countries. Nearly half of Mumbai's population faces 2.5 times higher risk of air pollution than the levels recommended by WHO. According to the Maharashtra Pollution Control Boards 2010 reports ,the major contributors of air pollution include vehicular emission and power plants, followed by road dust .the level of suspended particulate matter per cubic metre of air is about 172 micrograms sometimes in Mumbai while any level above 100 micrograms is harmful to health as per doctors

## ➢ DELHI

The concentration of $PM_{2.5}$ on 5th November 2017 was averaged close to 700 micrograms per cubic meter, which is 12 times higher than government norms and a whopping 70 times higher than WHO standards. The visibility was less than 50 meters. This has become an annual event in Delhi. To ameliorate the situation this year the Supreme Court had banned the sale of firecrackers in NCR region. Delhi has been tagged as one of the most polluted capital cities of the world with an unhealthy Air Quality Index (AQI) swinging mostly between 'poor' and 'very poor' categories. This quality of air is experienced by people throughout the year, but the condition becomes worse in winters when fog envelops the city, converting into toxic smog. Over the years, Delhi's problem of air pollution has become more complex. The population of Delhi is basically living in a gas chamber with lethal air quality. A lot of factors contribute to the poor AQI in Delhi. The condition just happens to worsen during a few months. The majority of the blame is attributed to burning of paddy straw by farmers and fire crackers during the Diwali festival in months of October and November. Air pollution in Delhi is a complex reaction which involves various industries as catalysts, including the transport industry, industrial emissions, biomass burning, and dust. According to the Ambient Air Pollution (AAP) report for the year 2018, Delhi had one of the highest $PM_{2.5}$ pollution levels in the world.

# CHAPTER 3

# OBJECTIVES OF STUDY

The purpose of this study is to explore the use of statistical tools and methods in the analysis of air quality data. The major objectives are listed below:

- Analysis of air quality of SAFAR cities using air quality index.

- To identify unobservable factors of air pollution using factor analysis.

- To identify a statistical model for $PM_{2.5}$ level of capital city Delhi and use it for prediction of future indices.

# CHAPTER 4

# DATA DESCRIPTION

The data used is a secondary data from the official website data.gov.in

We have two datasets:

- The first one comprises of air quality parameters (variables under study) of the mentioned SAFAR cities (Ahmedabad, Delhi, Pune, Mumbai) at different stations during the year 2019 .The variables under study are $SO_2$, $NO_2$,$PM_{10}$ and $PM_{2.5}$

- The second one comprises of the monthly data of $PM_{2.5}$ level of city Delhi at the during 2015-2019 .

# 4.1 VARIABLES UNDER CONSIDERATION

The different variables under the study are as follows:

- **SULPHUR DIOXIDE($SO_2$)**

Sulphur dioxide is highly reactive gas with a pungent irritating smell. It is formed by fossil fuel combustion at power plants and other industrial facilities .Natural processes that release sulphur gases include decomposition and combustion of organic matter, spray from the sea, and volcanic eruptions. It contributes to the formation of particulate matter pollution. Sulphur dioxide irritates the lining of the nose, throat and lungs and may worsen existing respiratory illness especially asthma. It has also been found to cause cardiovascular diseases.

**Potential health effects from exposure to sulphur dioxide:**

- Narrowing of the airways leading to wheezing, chest tightness and shortness of breath
- More frequent asthma attacks in people with asthma
- Cause of cardiovascular diseases

- **NITROGEN DIOXIDE($NO_2$)**

Nitrogen dioxide is a highly reactive gas formed by emissions from motor vehicles, industry, unflued gas-heaters and gas stove tops. High concentrations can be found especially near busy roads and indoors where unflued gas-heaters are in use. Other indoor sources can be from cigarette smoke or from cooking with gas. Outdoors, nitrogen dioxide contributes to the formation of ground-level ozone (O3) as well as particulate matter pollution. Nitrogen dioxide is a respiratory irritant and has a variety of adverse health effects on the respiratory system.

**Potential health effects from exposure to nitrogen dioxide:**

- Increased susceptibility to lung infections in people with asthma
- Increased susceptibility to asthma triggers like pollen and exercise
- Worsened symptoms of asthma – more frequent asthma attacks
- Airway inflammation in healthy people

- **PARTICULATE MATTER ($PM_{10}$ AND $PM_{2.5}$)**

Particulate matter, also known as particle pollution or PM, is a term that describes extremely small solid particles and liquid droplets suspended in air. Particulate matter can be made up of a variety of components including nitrates, sulfates, organic chemicals, metals, soil or dust particles, and allergens (such as fragments of pollen or mould spores). Particle pollution mainly comes from motor vehicles, wood burning heaters and industry. During bushfires or dust storms, particle pollution can reach extremely high concentrations

The size of particles affects their potential to cause health problems:

- $\boldsymbol{PM_{10}}$ (particles with a diameter of 10 micrometres or less): these particles are small enough to pass through the throat and nose and enter the lungs. Once inhaled, these particles can affect the heart and lungs and cause serious health effects.
- $\boldsymbol{PM_{2.5}}$ (particles with a diameter of 2.5 micrometres or less): these particles are so small they can get deep into the lungs and into the bloodstream. There is sufficient evidence that exposure to $PM_{2.5}$ over long periods (years) can cause adverse health effects. Note that $PM_{10}$ includes $PM_{2.5}$

**Potential health effects from exposure to particulate matter:**

There are many health effects from exposure to particulate matter. Health effects can occur after both short and long-term exposure to particulate matter. Short-term and long-term exposure is thought to have different mechanisms of effect. Short-term exposure appears to cause pre-existing diseases while long-term exposure most likely causes disease and increases the rate of progression.

Short-term exposure (hours to days) can lead to:

- irritated eyes, nose and throat
- worsening asthma and lung diseases such as chronic bronchitis (also called chronic obstructive pulmonary disease or COPD)
- heart attacks and arrhythmias (irregular heart beat) in people with heart disease
- increases in hospital admissions and premature death due to diseases of the respiratory and cardiovascular systems.

Long-term exposure (many years) can lead to:

- reduced lung function
- development of cardiovascular and respiratory diseases
- increased rate of disease progression
- reduction in life expectancy.

# 4.2 SOFTWARE TOOLS

The following software are used for analysis:

- **IBM SPSS**

This software package is used for statistical analysis. The software name originally stood for Statistical Package for the Social Sciences (SPSS). It can perform standard analyses including descriptive statistics, exploratory data analysis, correlations, a variety of regression, general linear modelling (including ANOVA), Time series modelling and analysis, forecasting etc.

- **MICROSOFT EXCEL**

Excel is an electronic spreadsheet program that can be used for storing, organizing and manipulating data. It has many inbuilt formulas and functions for statistical analysis including charts and graphs.

- **R PROGRAMMING**

R acts as an alternative to traditional statistical packages such as SPSS, SAS, and Stata such that it is an extensible, open-source language and computing environment for Windows, Macintosh, UNIX, and Linux platforms .R performs a wide variety of basic to advanced statistical and graphical techniques at little to no cost to the user. These advantages over other statistical software encourage the growing use of R in cutting edge social science research. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others.

# CHAPTER 5

# METHODOLOGY

Any data analysis study needs the use of statistical and computational methodologies. Since the data under study include multivariate data and time series data, we make use of multivariate analysis and time series analysis techniques for analysing the data.

## 5.1 AIR QUALITY INDEX

An air quality index is defined as an overall scheme that transforms the weighed values of individual air pollution related parameters (for example, pollutant concentrations) into a single number or set of numbers (Ott, 1978). The result is a set of rules (i.e. set of equations) that translate parameter values into a more simple form by means of numerical manipulation.

### Structure of an Index

Primarily two steps are involved in formulating an AQI:

(i) formation of sub-indices (for each pollutant)

(ii) aggregation of sub-indices to get an overall AQI. Formation of sub-indices ($I_1, I_2, ...., I_n$) for n pollutant variables ($X_1, X_2...., X_n$ ) is carried out using subindex functions that are based on air quality standards and health effects. Mathematically;

[1] $\boldsymbol{I_i = f(X_i), i=1, 2,...,n}$

Each sub-index represents a relationship between pollutant concentrations and health effects. The functional relationship between sub-index value ($I_i$) and pollutant concentrations ($X_i$) is explained later in the text. Aggregation of sub-indices, $I_i$is carried out with some mathematical function (described below) to obtain the overall index (I), referred to as AQI.

[2] $\mathbf{I = F(I_1, I_2,...., I_n)}$

The aggregation function usually is a summation or multiplication operation or simply a maximum operator.

### Sub-indices

Sub-index function represents the relationship between pollutant concentration $X_i$ and corresponding sub-index $I_i$. It is an attempt to reflect environmental consequences as the concentration of specific pollutant changes. It may take a variety of forms such as linear, non-linear and segmented linear. Typically, the I-X relationship is represented as follows:

$$\mathbf{I = \alpha X + \beta}$$

Where,
α =slope of the line,
β = intercept at X=0.

The general equation for the sub-index ($I_i$) for a given pollutant concentration (Cp); as based on 'linear segmented principle' is calculated as:

$$\mathbf{I_i = [\{(I_{HI} - I_{LO})/(B_{HI} - B_{LO})\} * (Cp - B_{LO})] + I_{LO}}$$

where,
$B_{HI}$ = Breakpoint concentration greater or equal to given concentration.
$B_{LO}$ = Breakpoint concentration smaller or equal to given concentration.
$I_{HI}$ =AQI value corresponding to BHI
$I_{LO}$ = AQI value corresponding to BLO
Cp = Pollutant concentration

## AQI Calculation

1.The Sub-indices for individual pollutants at a monitoring location are calculated using its 24-hourly average concentration value (8-hourly in case of CO and O3) and health breakpoint concentration range. The worst sub-index is the AQI for that location.

2. All the eight pollutants may not be monitored at all the locations. Overall AQI is calculated only if data are available for minimum three pollutants out of which one should necessarily be either PM2.5 or PM10. Else, data are considered insufficient for calculating AQI. Similarly, a minimum of 16 hours' data is considered necessary for calculating subindex.

3. The sub-indices for monitored pollutants are calculated and disseminated, even if data are inadequate for determining AQI. The Individual pollutant-wise sub-index will provide air quality status for that pollutant.

4. The web-based system is designed to provide AQI on real time basis. It is an automated system that captures data from continuous monitoring stations without human intervention, and displays AQI based on running average values .

5. For manual monitoring stations, an AQI calculator is developed wherein data can be fed manually to get AQI value.

## Indian Air Quality Index (IND-AQI)

Air quality standards are the basic foundation that provides a legal framework for air pollution control. An air quality standard is a description of a level of air quality that is adopted by a regulatory authority as enforceable. The basis of development of standards is to provide a rational for protecting public health from adverse effects of air pollutants, to eliminate or reduce exposure to hazardous air pollutants, and to guide national/local authorities for pollution control decisions. With these objectives, CPCB notified a new set of Indian National Air Quality Standards (INAQS) for 12 parameters

[carbon monoxide (CO) nitrogen dioxide ($NO_2$), sulphur dioxide ($SO_2$), particulate matter (PM) of less than 2.5 microns size ($PM_{2.5}$), PM of less than 10 microns size ($PM_{10}$), Ozone ($O_3$), Lead (Pb), Ammonia ($N_{H3}$), Benzo(a)Pyrene (BaP), Benzene ($C_6H_6$), Arsenic (As), and Nickel (Ni)] .

The first eight parameters (Table 3.1) have short-term (1/8/24 hrs) and annual standards (except for CO and $O_3$) and rest four parameters have only annual standards.

Table 3.1: Indian National Air Quality Standards (units: µg/m3 unless mentioned otherwise)

| **Pollutant** | **$SO_2$** | **$NO_2$** | **$PM_{2.5}$** | **$PM_{10}$** | **$O_3$** | | **CO (mg/m$^3$)** | | **Pb** | **$NH_3$** |
|---|---|---|---|---|---|---|---|---|---|---|
| Averaging time (hr) | 24 | 24 | 24 | 24 | 1 | 8 | 1 | 8 | 24 | 24 |
| Standard | 80 | 80 | 60 | 100 | 180 | 100 | 4 | 2 | 1 | 400 |

BaP, $C_6H_6$, As, and Ni have annual standards

The objective of an AQI is to quickly disseminate air quality information (almost in real-time) that entails the system to account for pollutants which have short-term impacts. It is equally important that most of these pollutants are measured continuously through an online monitoring network. Consequently, in the AQI system, the following pollutants are considered CO, $NO_2$, $SO_2$, $PM_{2.5}$, $PM_{10}$, $O_3$, $N_{H3}$ and Pb.

Figure 3.1 shows the operational scheme of AQI system based of maximum operator function (i.e. maximum sub-index being the overall index).

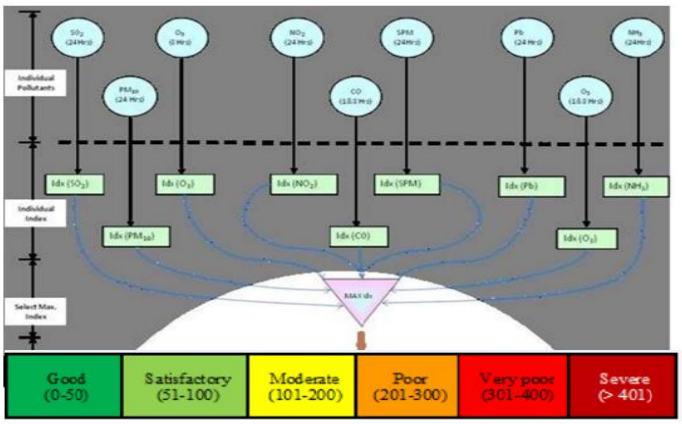



**Figure 3.1 Overall AQI System**

To present status of the air quality and its effects on human health, the following description categories have been adopted for IND-AQI (Table 3.2):

**Table 3.2: IND-AQI Category and Range**

| AQI Category | AQI Range |
|---|---|
| Good | 0 – 50 |
| Satisfactory | 51 – 100 |
| Moderate | 101 – 200 |
| Poor | 201 – 300 |
| Very Poor | 301 – 400 |
| Severe | 401 - 500 |

These categories/AQI ranges should map to key references (breakpoints) of concentration of each pollutants through a segmented linear or a nonlinear function.

## 5.2 MULTIVARIATE ANALYSIS

The term multivariate analysis refers to the analysis of many variables, which effectively includes all of the statistical samplings (Seber (1938)). Multivariate statistics are an extension of univariate and bivariate statistics. If your design has many variables, multivariate techniques often let you perform a single analysis instead of a series of univariate or bivariate analysis.

## 5.3 FACTOR ANALYSIS

The broad purpose of factor analysis is to summarize data so that relationships and patterns can be easily interpreted and understood. It is normally used to regroup variables into a limited set of clusters based on shared variance. Factor analysis operates on the notion that measurable and observable variables can be reduced to fewer latent variables that share a common variance and are unobservable, which is known as reducing dimensionality (Bartholomew, Knott, & Moustaki, 2011). These unobservable factors are not directly measured but are essentially hypothetical constructs that are used to represent variables.

Suppose we have a set of p observable random variables $X_1, X_2, \ldots, X_p$ with means $\mu_1, \mu_2, \ldots, \mu_p$. Suppose for some unknown constants $l_{ij}$ and k unobserved random variables $F_j$, where i € 1,2,….,p and j € 1,2,…,k ,where k < p, we have

$$\mathbf{X_i - \mu_i = l_{i1}F_1 + \ldots\ldots + l_{ik}F_k + \varepsilon_i}$$

Where, $\varepsilon_i$ are unobserved stochastic error terms with zero mean and finite variance, which may not be the same for all i.

In matrix terms, we have

$$X - \mu = LS + \varepsilon$$

If we have n observations, then we will have the dimensions, $X_{p\times n}$, $L_{p\times k}$ and $F_{k\times n}$. .Each column of X and F denote values for one particular observation, and matrix L does not vary across observations.

Also, we will impose the following assumptions on F:

- F and ε are independent
- E(F)=0
- $(F) = I$ ( so that factors are uncorrelated)

Any solution of the above set of equations following the constraints for F is defined as the factors and L as the loading matrix.

Suppose, $(X - \mu) = \Sigma$ . Then note that from the conditions just imposed on F, we have

$$\mathbf{Cov}(X - \mu) = \mathbf{Cov}(LF + \varepsilon)$$

Or

$$\mathbf{\Sigma = LCov}(F)\ \mathbf{L^T} + (\varepsilon)$$

Or

$$\mathbf{\Sigma = LL^T + \Psi}$$

Note that for any orthogonal matrix Q, if we set L=LQ and F=$Q^T$F, the criteria for being factors and factor loadings still hold. Hence a set of factors and factor loadings is unique only up to orthogonal transformation.

# FACTOR ANALYSIS USING SPSS

There are three main steps in a factor analysis:

## 1.Calculate initial factor loadings.

This can be done in a number of different ways; the two most common methods are described very briefly below:

- **Principal component method**

As the name suggests, this method uses the method used to carry out a principal 1 components analysis. However, the factors obtained will not actually be the principal components (although the loadings for the k th factor will be proportional to the coefficients of the k th principal component).

- **Principal axis factoring**

This is a method which tries to find the lowest number of factors which can account for the variability in the original variables that is associated with these factors (this is in contrast to the principal components method which looks for a set of factors which can account for the total variability in the original variables). These two methods will tend to give similar results if the variables are quite highly correlated and/or the number of original variables is quite high. Whichever method is used, the resulting factors at this stage will be uncorrelated.

## 2. Factor rotation

Once the initial factor loadings have been calculated, the factors are rotated. This is done to find factors that are easier to interpret. If there are 'clusters' (groups) of variables — i.e. subgroups of variables that are strongly inter-related — then the rotation is done to try to make variables within a subgroup score as highly (positively or negatively) as possible on one particular factor while, at the same time, ensuring that the loadings for these variables on the remaining factors are as low as possible. In other words, the object of the rotation is to try to ensure that all variables have high loadings only on one factor. There are two types of rotation method, orthogonal and oblique rotation. In orthogonal rotation the rotated factors will remain uncorrelated whereas in oblique rotation the resulting factors will be correlated. There are a number of different methods of rotation of each type. The most common orthogonal method is called varimax rotation

### 3. Calculation of factor scores

When calculating the final factor scores (the values of the m factors, $F_1$, , $F_2$, . . . , , $F_m$, for each observation), a decision needs to be made as to how many factors to include. This is usually done using one of the following methods:

• Choose m such that the factors account for a particular percentage of the total variability in the original variables.

• Choose m to be equal to the number of eigenvalues over 1 (if using the correlation matrix). [A different criteria must be used if using the covariance matrix.]

• Use the scree plot of the eigenvalues. This will indicate whether there is an obvious cut-off between large and small eigenvalues.

In some statistical packages (e.g. SPSS) this choice is actually made at the outset. The second method, choosing eigenvalues over 1, is probably the most common one. The final factor scores are usually calculated using a regression-based approach.

## 5.4 TIME SERIES MODELLING

A time series is a collection of observations made sequentially through time. The main objectives of time series analysis are to provide a compact description of the data, to explain seasonal factors and to use the model to forecast future values of the time series. The various factors which effects the values of time series data is known as components of time series data are :

• Trend
• Seasonal Variations
• Cyclic Variations
• Random or Irregular movements

The initial step in any time series analysis is to plot the observations against time. This graph, called a time plot, will show up important features of the series such as trend, seasonality, outliers, and discontinuities. A graph will reveal any outliers that do not appear to be consistent with the rest of the data. An outlier may be a perfectly valid, but extreme, observation, which could, for example, indicate that the data are not normally distributed. Alternatively, an outlier may be a freak observation arising, for example, when a recording device goes wrong or when a strike severely affects sales. In the latter case, the outlier needs to be adjusted in some way before further analysis of the data. Seasonal and Cyclic Variations are the periodic changes or short-term fluctuations. Many time series exhibit a seasonal pattern that has the tendency to repeat itself over a certain fixed period of time called seasonality and the length of the cycle as seasonal period “s”.

A time series is said to be stationary if there is no systematic change in mean (no trend) if there is no systematic change in variance and if strictly periodic variations have been removed. Much of the probability theory of time series is concerned with stationary time series, and for this reason, time series analysis often requires one to transform a non-stationary series into a stationary one so as to use this theory.

## Augmented Dickey-Fuller test (ADF)

ADF test is used to test whether the time series is stationary or not.

H0: Time series data is non stationary.

H1: Time series data is stationary

.

The test statistic for ADF test is defined as follows:-

$$\mathbf{Q_m = \sum_{k=1}^{n} n(n+2)}$$

$$\mathbf{DE_T = \hat{\gamma} / S.E(\gamma)}$$

## Autocorrelation Function (ACF)

Autocorrelation is the correlation between observations of a variable taken at different time point. Autocorrelation plot is a commonly used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. Correlogram is the plot of the autocorrelation coefficient as a function of the lag k.

## Partial Autocorrelation Function (PACF)

In time series analysis, the partial autocorrelation function (PACF) plays an important role in identifying the extent of the lag in an autoregressive model. Given a time series $\{Z_t;$ t=1,2,..$\}$ the partial autocorrelation of lag k, is the correlation between $Z_t$ and $Z_{t+k}$ with the linear dependence of $Z_{t+1}$ through to $Z_{t+k-1}$ removed; equivalently, it is the autocorrelation between $Z_t$ and $Z_{t+k}$ that is not accounted for by lags 1 to k − 1, inclusive.

## ARMA (Auto-Regressive Moving Average) Model

Given a time series of data $\{Z_t\}$ , the ARMA model is a tool for understanding and predicting future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving

average (MA) part. The model is usually then referred to as the ARMA (p, q) model. The notation ARMA (p, q) refers to the model with p autoregressive terms and q moving average terms. This model contains the AR (p) and MA (q) models.

$$\mathbf{Z_t} = C + \mathbf{a_t} + \sum_{\mathbf{i=1}}^{\mathbf{p}} \boldsymbol{\phi}_\mathbf{i} \mathbf{Z_{t-i}} + \sum_{\mathbf{i=1}}^{\mathbf{q}} \boldsymbol{\theta}_\mathbf{i} \mathbf{a_{t-i}}$$

The notation AR (p) indicates an autoregressive model of order p.

The AR (p) model is defined as

$$\mathbf{Z_t} = C + \mathbf{a_t} + \sum_{\mathbf{i=1}}^{\mathbf{p}} \boldsymbol{\phi}_\mathbf{i} \mathbf{Z_{t-i}}$$

Where $\phi_1, \phi_2, \ldots\ldots, \phi_p$ are the parameters of the model, C is a constant and $a_t$ is white noise.

The notation MA (q) refers to the moving average model of order q.

The MA (q) model is defined as

$$\mathbf{Z_t} = C + \mathbf{a_t} + \sum_{\mathbf{i=1}}^{\mathbf{q}} \boldsymbol{\theta}_\mathbf{i} \mathbf{a_{t-i}}$$

Where $\theta_1, \theta_2, \ldots..,\theta q$ are the parameters of the model and $a_t, a_{t-1}, \ldots\ldots$are the white noise error terms.

## ARIMA (Auto Regressive Integrated Moving Average) model

An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series. They are applied in some cases where data shows evidence of non-stationary, where an initial differencing step (corresponding to integrated part of the model) can be applied to remove the non-stationary. This model is generally referred to as an ARIMA (p, d, q) model, where p, d, and q are integers greater than or equal to zero. If we get ARMA (p, q) series after differentiating the given series d times the original one is referred as ARIMA (p, d, q) series.

## Residual Analysis

When a model has been fitted to a time series it is advisable to check that the model really does provide an adequate description of the data. This is usually done by looking at the residuals which are defined by

**Residual = observed value – fitted value**

For a univariate time series model, the fitted value is the one step ahead forecast so that the residual is the one step ahead forecast error.

## Kolmogorov-Smirnov test

Let $x_1, x_2, \ldots, x_m$ be observations on continuous i.i.d randomvariables $x_1, x_2, \ldots, x_m$ with a c.d.f. F. We want to test the hypothesis that,
$H_o$: $F(x) = F_o(x)$ for all x

$H_1$: $F(x) \neq F_0(x)$ for at least one value of x

Where $F_0$ is known as the c.d.f.

The Kolmogorov-Smirnov test statistic $D_n = \sup_{x \in R} | F(x) - F_o(x)|$,

Which is an empirical cumulative distribution. We reject $H_o$ when $D_n > D_{n,\alpha}$, $D_{n\alpha}$ is the value corresponding to 100% significance level obtained from the table. We usually use Kolmogorov - Smirnov test to check the normality assumption in the analysis of variance.

## Forecasting

Forecasting is defined as an activity to calculate or predict some future event or condition usually as result of rational study or analysis of pertinent data. Once a time series model has been obtained, the minimum means square error forecasts are easily calculated from the difference equation of the model

**TIME SERIES MODELLING USING R**

- Reading Time Series Data
- Plotting Time Series
- Decomposing Time Series
- Seasonally Adjusting
- ARIMA Models
- Differencing a Time Series
- Selecting a Candidate ARIMA Model
- Forecasting Using an ARIMA Model

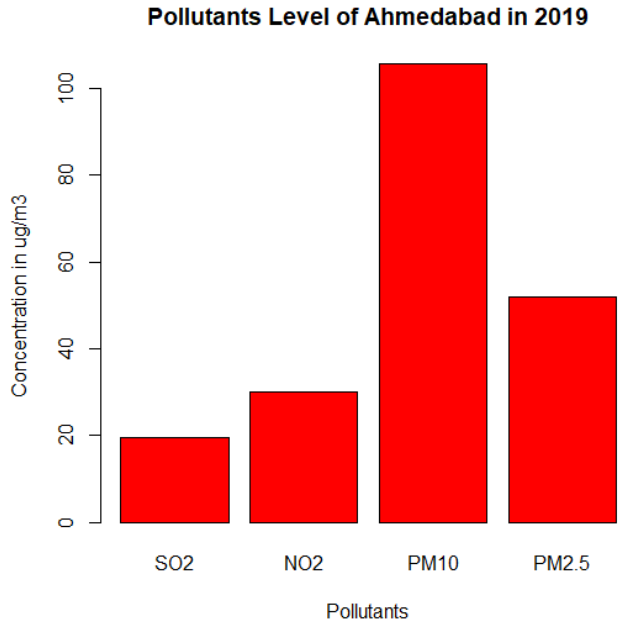
# CHAPTER 6

# DATA ANALYSIS

## AIR QUALITY INDEX OF SAFAR CITIES

We use the dataset comprising of the 4 air quality variables for each city during the year 2019. We calculate the average AQI for each city

## AHMEDABAD




**R code**

>aqi=c(19.58,30.08,105.5,51.91)

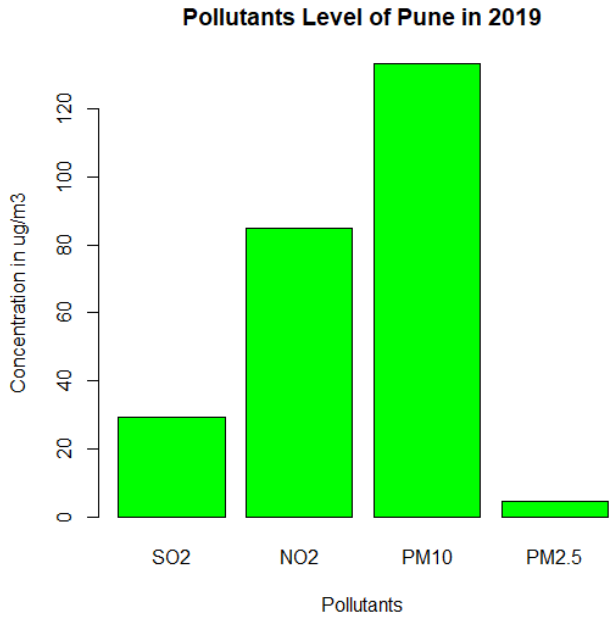
> barplot(aqi,main="Pollutants Level of Ahmedabad in 2019",xlab="Pollutants",ylab="Concentration in ug/m3",names.arg=c("SO2","NO2","PM10","PM2.5"),col="red")

**INFERENCE**

The above graph shows that

- the concentration of $PM_{10}$ is high as its above 100
- the concentration of $PM_{2.5}$ is less than standard level and its safe
- the concentration of $NO_2$ is less than standard level and its safe
- the concentration of $SO_2$ is less than standard level and its safe

# PUNE




**R code**

>aqi=c(29.33,84.75,133.16,4.5)

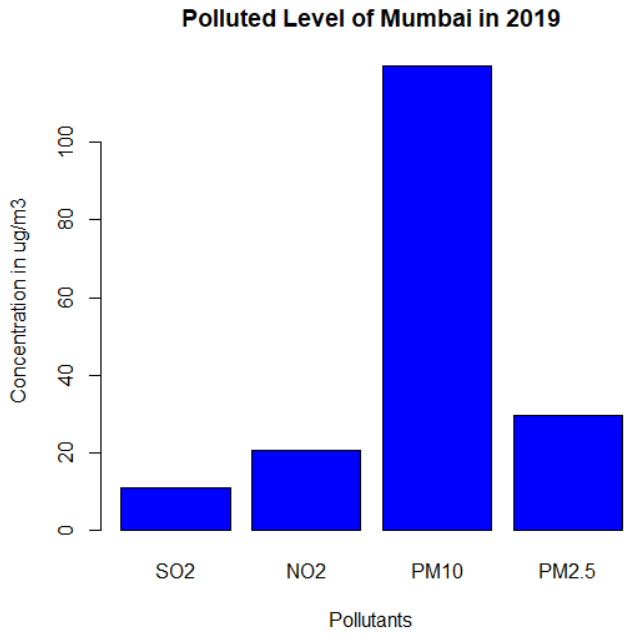
> barplot(aqi,main="Pollutants Level of Pune in 2019",xlab="Pollutants",ylab="Concentration in ug/m3",names.arg=c("SO2","NO2","PM10","PM2.5"),col="green")

**INFERENCE**

The above graph shows that

- the concentration of $PM_{10}$ is high as its above 100
- the concentration of $PM_{2.5}$ is less than standard level and its safe
- the concentration of $NO_2$ is high as its above 80
- the concentration of $SO_2$ is less than standard level and its safe

# MUMBAI




**R code**

>aqi=c(11,20.5,119.66,29.5)

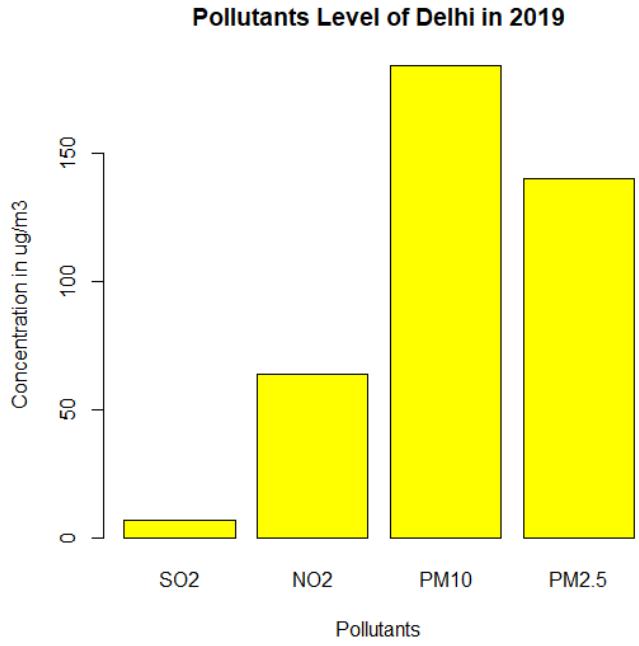
> barplot(aqi,main="Polluted Level of Mumbai in 2019",xlab="Pollutants",ylab="Concentration in ug/m3",names.arg=c("SO2","NO2","PM10","PM2.5"),col="blue")

**INFERENCE**

The above graph shows that

- the concentration of $PM_{10}$ is high as its above 100
- the concentration of $PM_{2.5}$ is less than standard level and its safe
- the concentration of $NO_2$ is less than standard level and its safe
- the concentration of $SO_2$ is less than standard level and its safe

# DELHI




**R code**

>aqi=c(7.08,63.66,184.08,140.05)

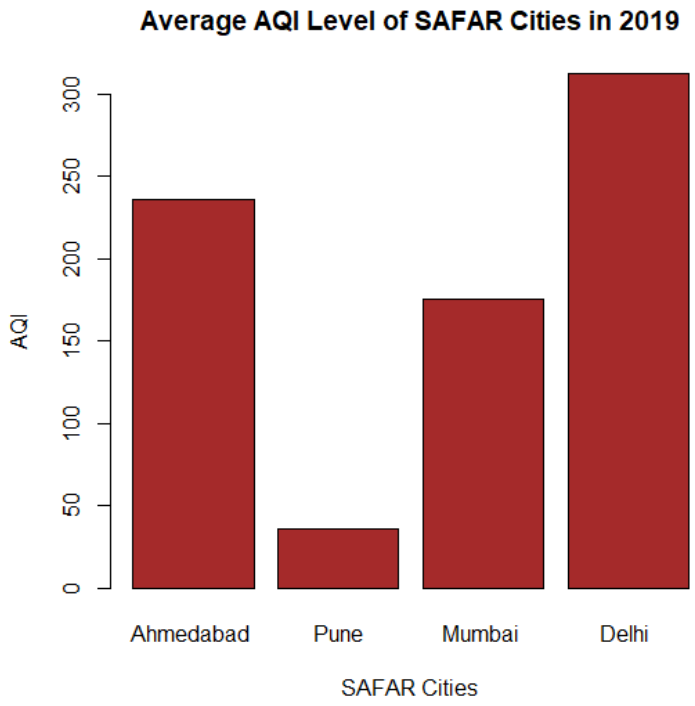
> barplot(aqi,main="Pollutants Level of Delhi in 2019",xlab="Pollutants",ylab="Concentration in ug/m3",names.arg=c("SO2","NO2","PM10","PM2.5"),col="yellow")

**INFERENCE**

The above graph shows that

- the concentration of $PM_{10}$ is above 150 which is high
- the concentration of $PM_{2.5}$ is above 100 which is very high
- the concentration of $NO_2$ is below than standard level so its safe
- the concentration of $SO_2$ is below the standard level and its safe

# 6.1 COMPARISON OF AQI OF SAFAR CITIES




**Rcode**

aveAverageAQI=c(236,35.7,175,312)

barplot(aveAverageAQI, main="Average AQI Level of SAFAR Cities in 2019",xlab="SAFAR Cities", ylab=" AQI",names.arg=c("Ahmedabad ","Pune","Mumbai ","Delhi "), col="brown")

**INFERENCE**

The above graph shows that

- Delhi has the highest AQI,which means it’s the most polluted city
- Ahmedabad is the second polluted city
- Mumbai and Pune is the third and fourth polluted SAFAR cities simultaneously

# 6.2 FACTORS AFFECTING AIR QUALITY OF SAFAR CITIES

Factor analysis is applied to the air quality data of the considered Cities during the year 2019. This method has been applied to identify the important underlying factors of air pollution.

## AHMEDABAD

The analysis results shows that:

| Component Matrix[a] | | |
|---|---|---|
| | Component | |
| | 1 | 2 |
| SO2 | -.187 | .965 |
| NO2 | .825 | -.098 |
| PM10 | .887 | .006 |
| PM2.5 | .834 | .307 |
| Extraction Method: Principal Component Analysis. | | |
| a. 2 components extracted. | | |

| TOTAL VARIANCE | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2.198 | 54.940 | 54.940 | 2.198 | 54.940 | 54.940 | 2.176 | 54.394 | 54.394 |
| 2 | 1.036 | 25.898 | 80.838 | 1.036 | 25.898 | 80.838 | 1.058 | 26.444 | 80.838 |
| 3 | .458 | 11.441 | 92.279 | | | | | | |
| 4 | .309 | 7.721 | 100.000 | | | | | | |
| Extraction Method: Principal Component Analysis. | | | | | | | | | |

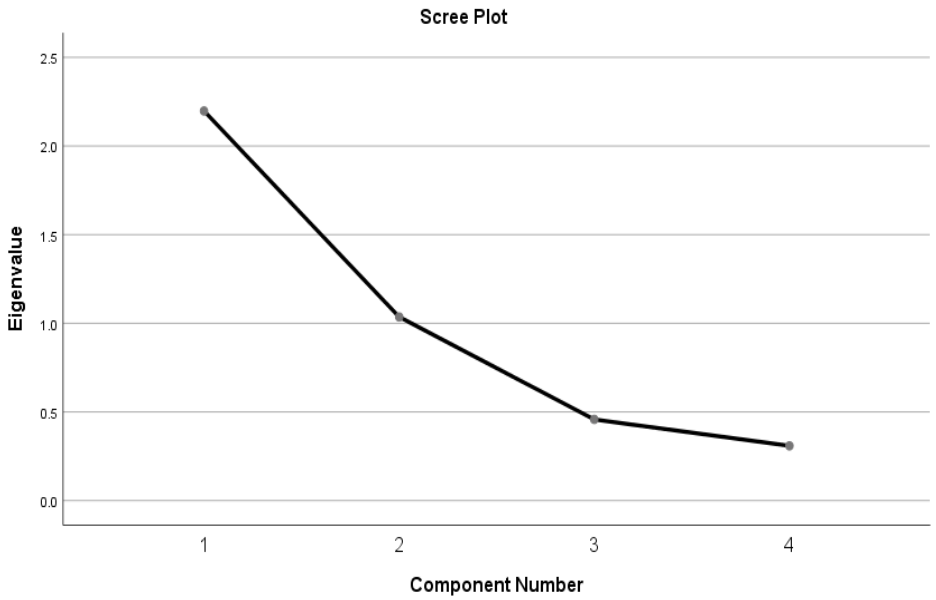
**Rotated Component Matrix[a]**

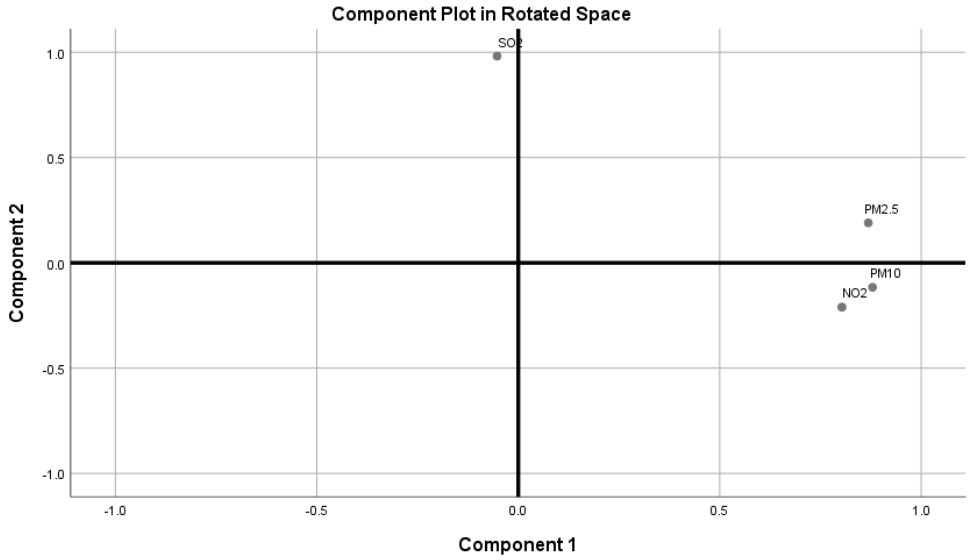
| | Component | |
|---|---|---|
| | 1 | 2 |
| SO2 | -.053 | .982 |
| NO2 | .803 | -.211 |
| PM10 | .879 | -.116 |
| PM2.5 | .869 | .190 |

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Rotation converged in 3 iterations.






## INFERENCE

We have identified two underlying factors from the data, two components explain 80.83% of information. To make a clear picture of underlying factors we use varimax factor rotation technique and obtained factor loading matrix.

From the factor loading matrix, we arrive at the following conclusions,

- $SO_2$ is loaded strongly on Factor 2, it is identified as Chemical Factor
- $PM_{10}$ is loaded strongly on Factor 1, it is identified as Particulate Factor

## PUNE

The analysis results shows that:

| Total Variance Explained | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 1.514 | 37.856 | 37.856 | 1.514 | 37.856 | 37.856 | 1.484 | 37.099 | 37.099 |
| 2 | 1.186 | 29.643 | 67.499 | 1.186 | 29.643 | 67.499 | 1.216 | 30.400 | 67.499 |
| 3 | .861 | 21.536 | 89.036 | | | | | | |
| 4 | .439 | 10.964 | 100.000 | | | | | | |
| Extraction Method: Principal Component Analysis. | | | | | | | | | |

| Component Matrix[a] | | |
|---|---|---|
| | Component | |
| | 1 | 2 |
| SO2 | .581 | .194 |
| NO2 | .870 | -.017 |
| PM10 | .529 | -.734 |
| PM2.5 | .373 | .780 |
| Extraction Method: Principal Component Analysis. | | |
| a. 2 components extracted. | | |

**Rotated Component Matrix[a]**

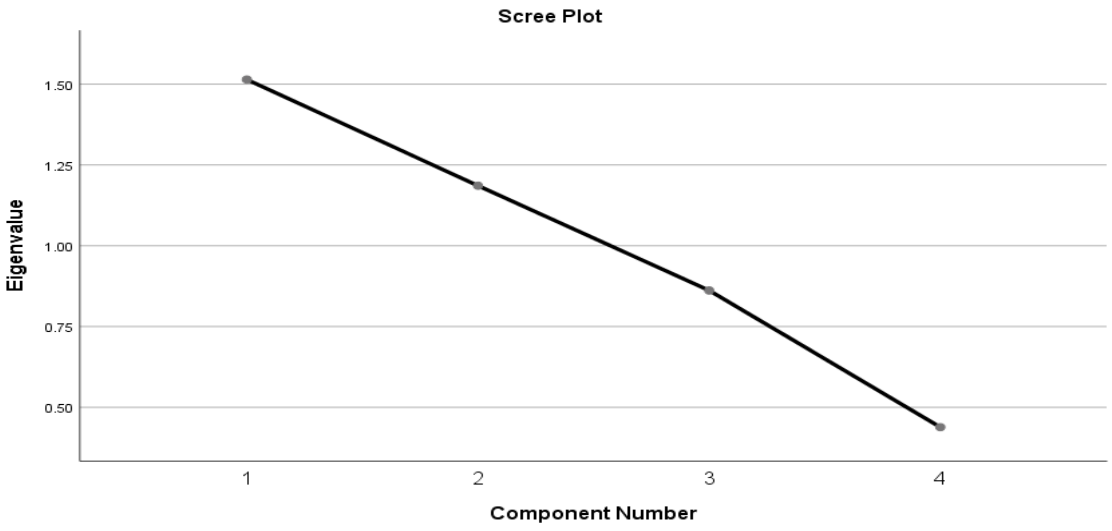
| | Component | |
|---|---|---|
| | 1 | 2 |
| SO2 | .613 | -.009 |
| NO2 | .824 | .281 |
| PM10 | .281 | .860 |
| PM2.5 | .592 | -.630 |

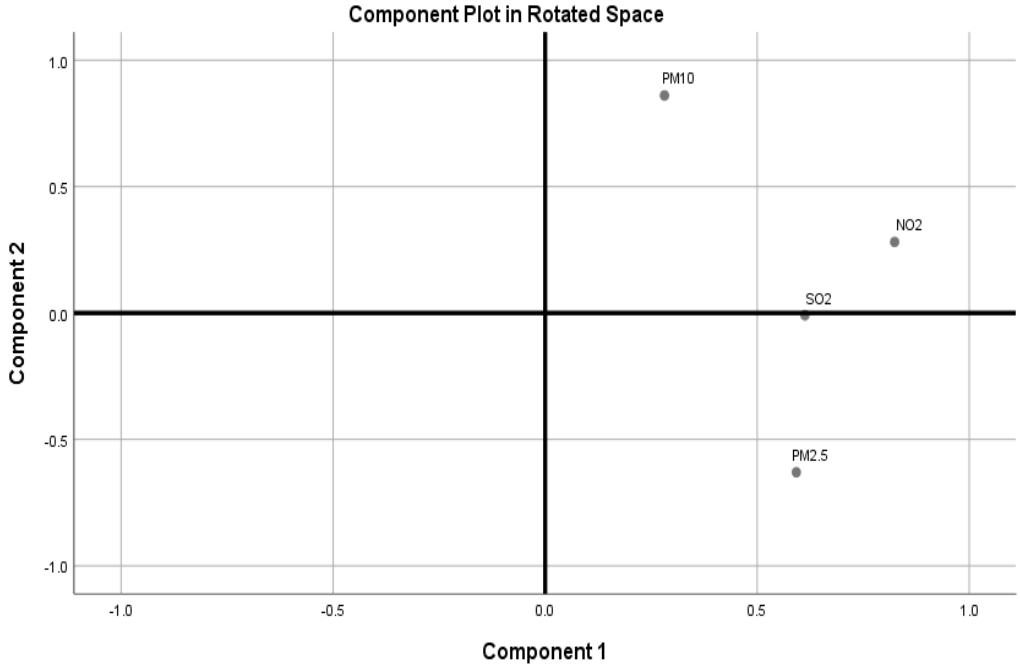
Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

.






## INFERENCE

We have identified two underlying factors from the data, two components explain 67.49% of information. To make a clear picture of underlying factors we use varimax factor rotation technique and obtained factor loading matrix.

From the factor loading matrix, we arrive at the following conclusions,

- $NO_2$ are loaded strongly on Factor 1, it is identified as Chemical Factor
- $PM_{10}$ is loaded strongly on Factor 2, it is identified as Particulate Factor

## MUMBAI

The analysis results shows that:

| Total Variance Explained | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2.383 | 59.587 | 59.587 | 2.383 | 59.587 | 59.587 | 2.308 | 57.700 | 57.700 |
| 2 | 1.006 | 25.148 | 84.735 | 1.006 | 25.148 | 84.735 | 1.081 | 27.035 | 84.735 |
| 3 | .541 | 13.528 | 98.263 | | | | | | |
| 4 | .069 | 1.737 | 100.000 | | | | | | |
| Extraction Method: Principal Component Analysis. | | | | | | | | | |

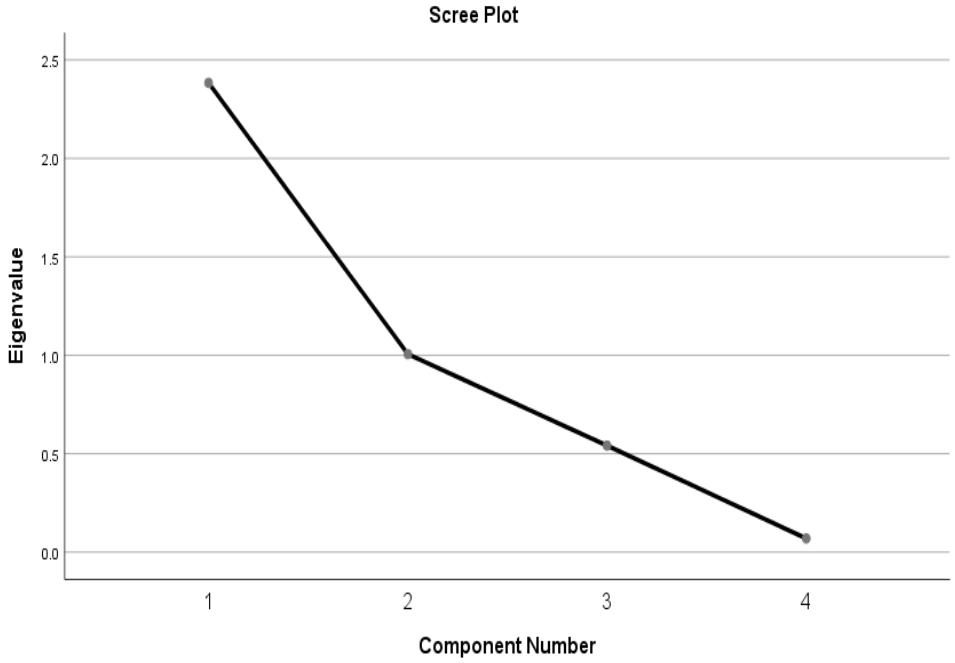
| Component Matrix[a] | | |
|---|---|---|
| | Component | |
| | 1 | 2 |
| SO2 | .774 | -.186 |
| NO2 | .943 | .252 |
| PM10 | -.283 | .934 |
| PM2.5 | .903 | .189 |
| Extraction Method: Principal Component Analysis. | | |
| a. 2 components extracted. | | |

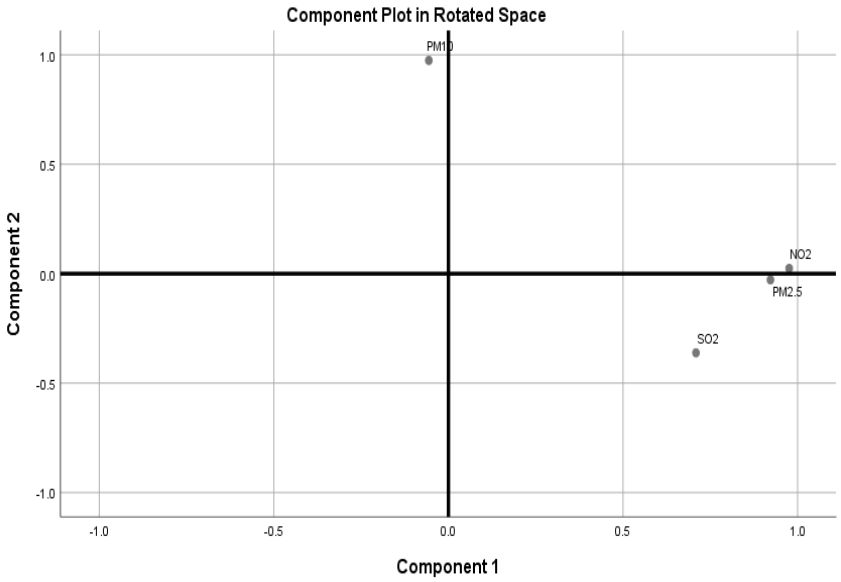
| Rotated Component Matrix[a] | | |
|---|---|---|
| | Component | |
| | 1 | 2 |
| SO2 | .709 | -.362 |
| NO2 | .976 | .024 |
| PM10 | -.056 | .974 |
| PM2.5 | .922 | -.028 |

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.






## INFERENCE

We have identified two underlying factors from the data, two components explain 84.73% of information. To make a clear picture of underlying factors we use varimax factor rotation technique and obtained factor loading matrix.

From the factor loading matrix, we arrive at the following conclusions,

- $NO_2$ are loaded strongly on Factor 1 it is identified as Chemical Factor
- $PM_{10}$ is loaded strongly on Factor 2, it is identified as Particulate Factor

## DELHI

The analysis results shows that:

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 1.986 | 49.653 | 49.653 | 1.986 | 49.653 | 49.653 | 1.882 | 47.061 | 47.061 |
| 2 | 1.021 | 25.522 | 75.176 | 1.021 | 25.522 | 75.176 | 1.125 | 28.114 | 75.176 |
| 3 | .709 | 17.729 | 92.905 | | | | | | |
| 4 | .284 | 7.095 | 100.000 | | | | | | |

Extraction Method: Principal Component Analysis.

**Component Matrix[a]**

| | Component | |
|---|---|---|
| | 1 | 2 |
| SO2 | .913 | .008 |
| NO2 | -.552 | .643 |
| PM10 | .806 | .000 |
| PM2.5 | .446 | .779 |

Extraction Method: Principal Component Analysis.

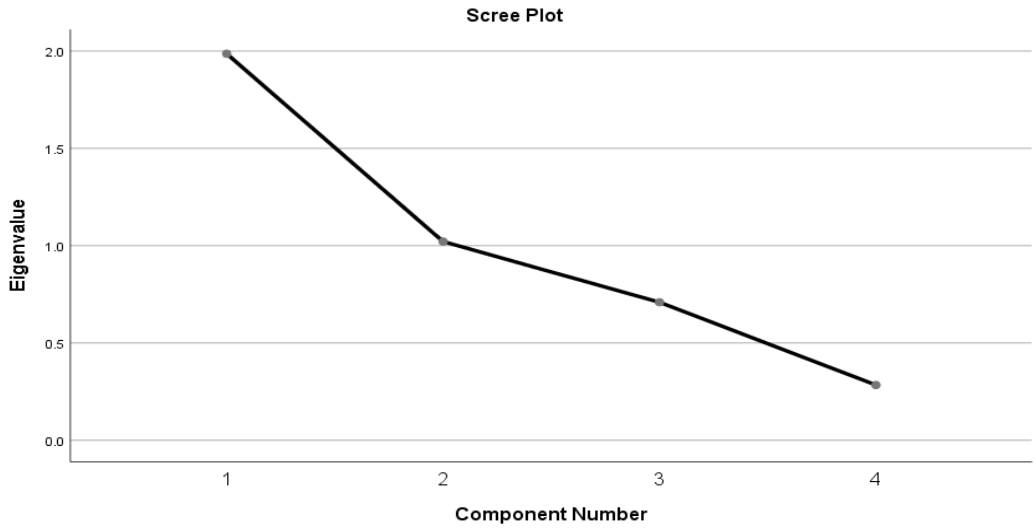
a. 2 components extracted.

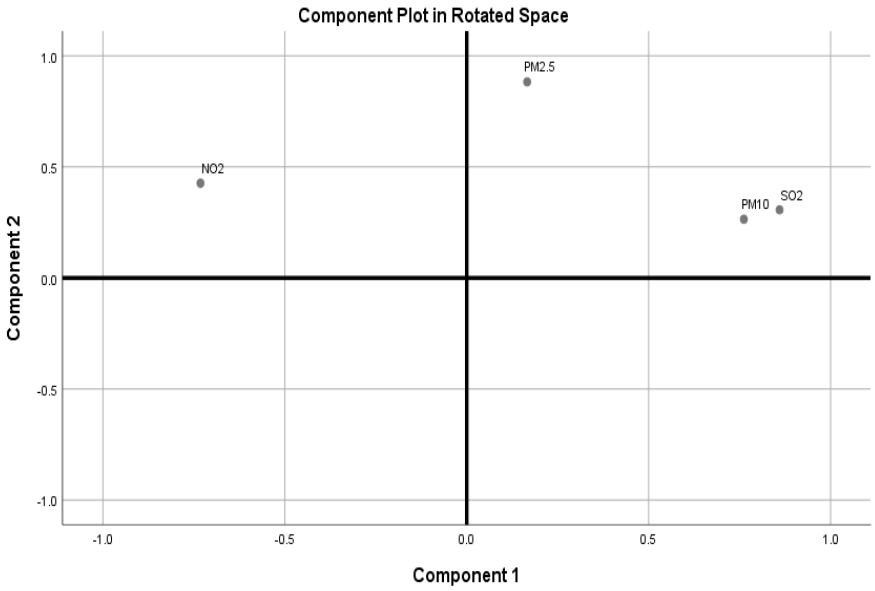
| Rotated Component Matrix[a] | | |
|---|---|---|
| | Component | |
| | 1 | 2 |
| SO2 | .860 | .306 |
| NO2 | -.732 | .427 |
| PM10 | .762 | .264 |
| PM2.5 | .166 | .883 |

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.






## INFERENCE

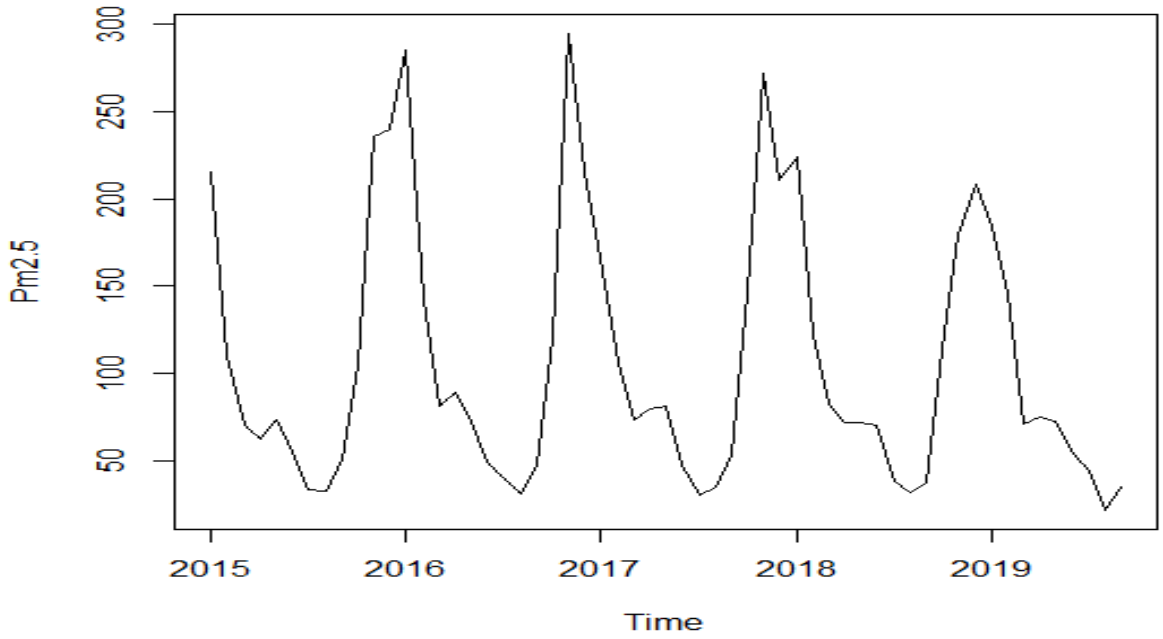
We have identified two underlying factors from the data, two components explain 75.17% of information. To make a clear picture of underlying factors we use varimax factor rotation technique and obtained factor loading matrix.

From the factor loading matrix, we arrive at the following conclusions,

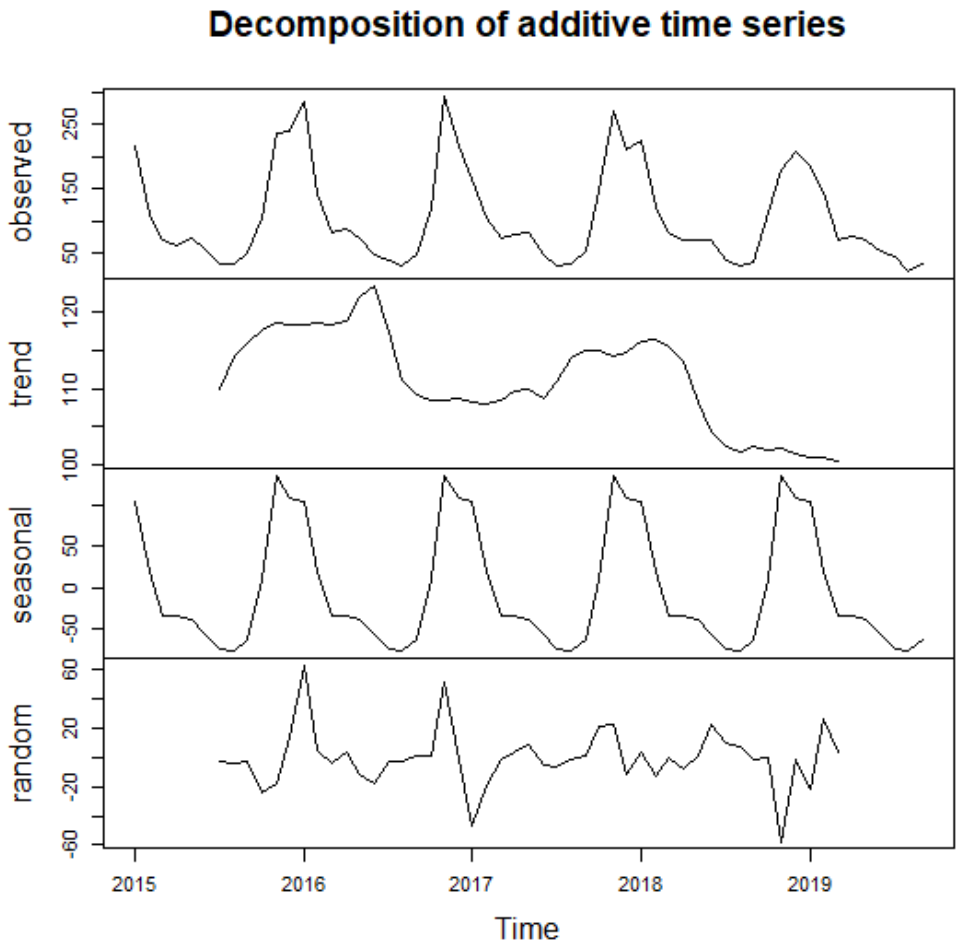
- $SO_2$ are loaded strongly on Factor 1 it is identified as Chemical Factor
- $PM_{2.5}$ is loaded strongly on Factor 2, it is identified as Particulate Factor

## 6.3 PREDICTING THE $PM_{2.5}$ VALUE OF MOST POLLUTED SAFAR CITY

In 2019, the $PM_{2.5}$ concentrations in Delhi violated the annual average standards by about three times. Transport (23 per cent), industries including power plants (23 per cent), and biomass burning (14 per cent) were the major contributors to prevailing winter time $PM_{2.5}$ concentrations in Delhi during 2015-19. In the past few years, levels of smog have increased throughout Delhi resulting in the deterioration of air quality, raising worldwide concerns. $PM_{2.5}$ (particles less than 2.5 micrometers in diameter) can penetrate deeply into the lung, irritate and corrode the alveolar wall, and consequently impair lung function. Hence it is important to investigate the impact of $PM_{2.5}$ on the coming years and then to help combat the current air pollution problems. We use the time series analysis for fitting a model for average monthly AQI of New Delhi during the years 2015-2019.For modelling AQI data, we consider data from January 2015 to September 2019 and keep the rest of the data as validation period(ie. for validating the effectiveness of the proposed model in forecasting).

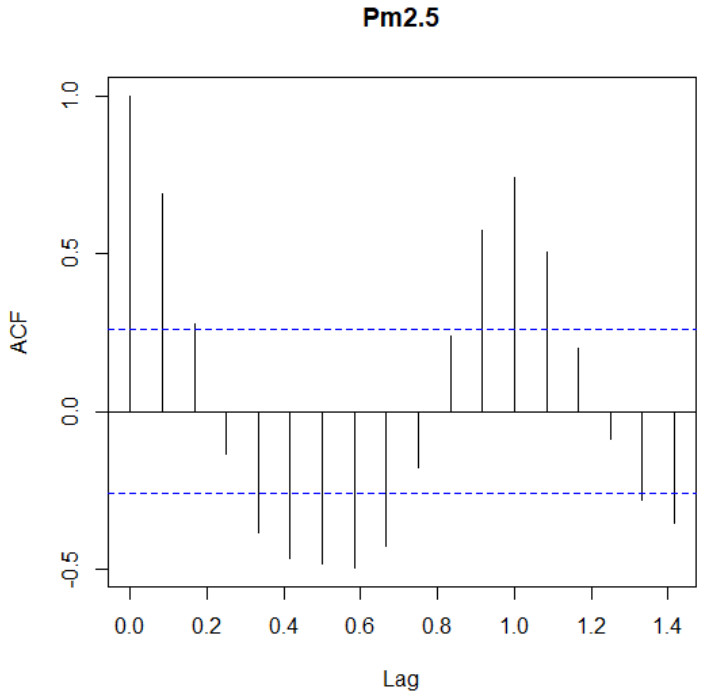


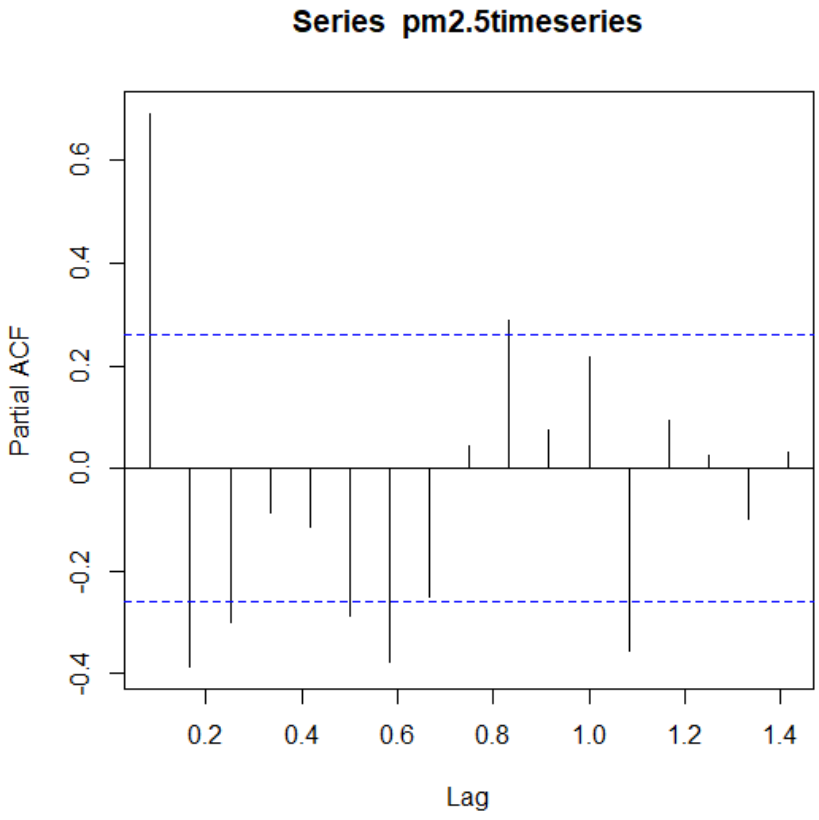

**Time series plot of the data**




For checking the stationarity of the series, we carried out Dickey-Fuller Unit-root Test. For the above data series, p-value of the unit root test obtained is less than 0.05, which ensures the stationarity of the data. Now from the ACF and PACF of the data, we fitted the model ARIMA(0,0,1)(1,1,0)[12]

The ACF and PACF Plots are obtained as follows:



From the ACF and PACF of the residuals, it is clear test the residuals are uncorrelated. Since p- value greater than 0.05, we do not reject the null hypothesis. ie. The residuals are uncorrelated. For checking the normality assumption on residuals, we consider Kolmogorov-Smirnov (K-S) test Since p-value = 0.897> 0.05, we do not reject the null hypothesis, that mean the residuals are normal

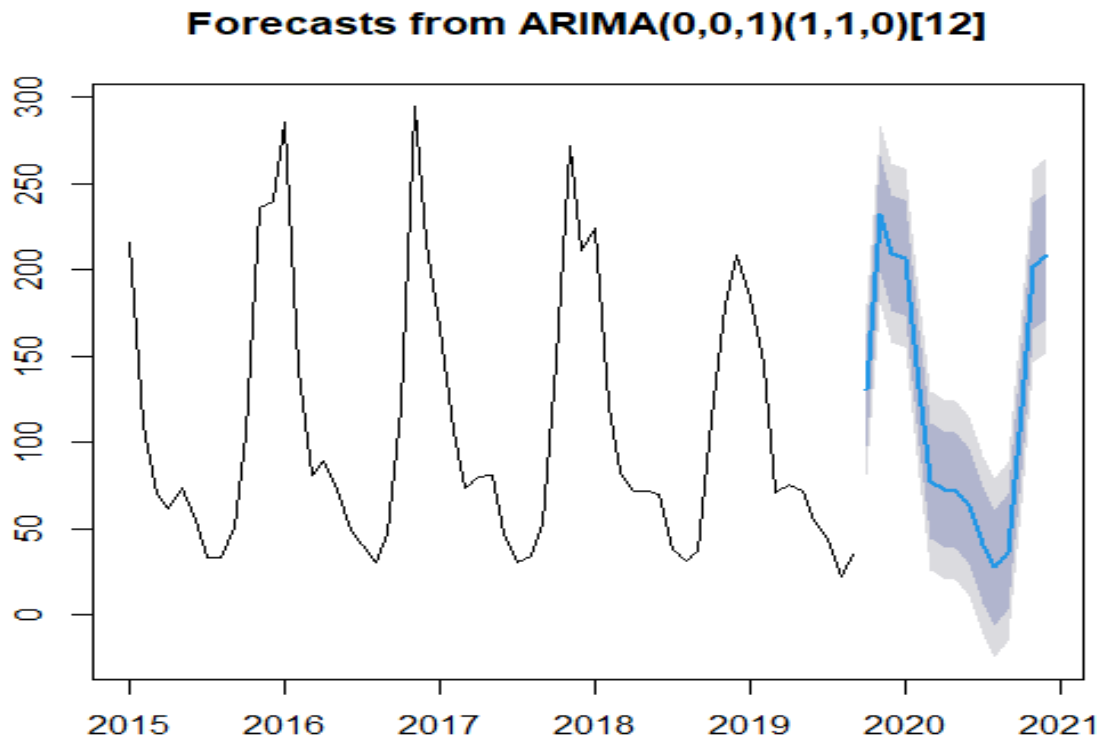
## FORECASTING

To check the validity of the proposed model in forecasting, we consider the data up to 2019 September and forecast the monthly AQI for the next three months ahead. Forecasted and observed values of AQI

| DATE | ACTUAL | FORECAST | LCL | UCL |
|---|---|---|---|---|
| OCTOBER | 127 | 130 | 97.815821 | 162.79839 |
| NOVEMBER | 240.2 | 232.12 | 198.360669 | 265.88750 |
| DECEMBER | 215.8 | 209.7 | 209.74314 | 175.979728 |

We can infer that forecasted and observed values are almost near. That is the model is effective in forecasting. Now we augment the remaining data points to analyse the complete sample. Analysis of the complete data yields the same model as before. All model diagnostic checking procedures are done. So by making use of the proposed model for the complete data, we forecast monthly AQI for the next year,

| DATE | FORECAST |
|---|---|
| **Jan 2020** | **206.96138** |
| **Feb 2020** | **131.81179** |
| **Mar 2020** | **77.70119** |
| **Apr 2020** | **73.14248** |
| **May 2020** | **72.08569** |
| **Jun 2020** | **63.24455** |
| **Jul 2020** | **40.74225** |
| **Aug 2020** | **27.40104** |
| **Sep 2020** | **36.40022** |
| **Oct 2020** | **120.47265** |
| **Nov 2020** | **201.87615** |
| **Dec 2020** | **208.86118** |




**Rcode**

```
pm2.5=read.table("C:/Users/hp/Desktop/pm2.5.txt",header=T)

pm2.5

pm2.5timeseries=ts(pm2.5,start=c(2015,1),frequency=12)

pm2.5timeseries

plot(pm2.5timeseries)

pm2.5timeseriescomponents=decompose(pm2.5timeseries)

pm2.5timeseriescomponents

plot(pm2.5timeseriescomponents)

library(tseries)

adf.test(pm2.5timeseries)
```

```
acf(pm2.5timeseries,lag.max=17.5)

acf(pm2.5timeseries,lag.max=17.5,plot=FALSE)

pacf(pm2.5timeseries,lag.max=17.5)

pacf(pm2.5timeseries,lag.max=17.5,plot=FALSE)

library(forecast)

fit=auto.arima(pm2.5timeseries)

fit

forecast(fit,15)

plot(forecast(fit,15))
```

## INFERENCE

For the calculated AQI data of Delhi, we fitted the time series model ARIMA(0,0,1)(1,1,0)[12] , the effectiveness of the model in forecasting is confirmed as we can see that forecasted and observed values are almost near. The difference can be due to some external factors. Then we can conclude that, the model is effective in forecasting. Then the fitted model is used to forecasts AQI for the next year (Jan 2020- Dec 2020).

# CHAPTER 7

# CONCLUSION

According to the WHO, air pollution is the fifth largest killer in India. There are a variety of ways in which the air pollution of an area can be measured. One of the ways is the measurement of particulate matter in air. Particulate matter is a mixture of extremely small particles and liquid droplets like acids, chemicals, gas, water, metals, soil dust particles, etc. These particles cause major health hazard in India. The changing temperature and slowing winds trap soot, dust and fine particulate matter. The statistical methods like multivariate analysis, factor analysis, and time series analysis were employed on our data in this study which gives the following conclusions:

- Considering the pollutants:
  - Ahmedabad has high level of $PM_{10}$
  - Pune has high level of $NO_2$ and $PM_{10}$
  - Mumbai has dangerously highest level of $PM_{10}$ in SAFAR cities
  - Delhi has dangerously very high particulate matter $PM_{2.5}$ and $PM_{10}$,and with highest level of $PM_{2.5}$ in SAFAR cities which is highly problem causing particulate matter.

- Considering the AQI:
  - Delhi is the most polluted SAFAR city,ie average AQI is the highest
  - Ahmedabad is the second polluted city
  - Mumbai and Pune is the third and fourth polluted SAFAR cities simultaneously

- Using factor analysis, we identified the underlying unobservable factors and the highly contributing factors of pollution from the air pollution data for each SAFAR city.
  - For Ahmedabad, $SO_2$ is chemical factor and $PM_{10}$ is particulate factor
  - For Pune, $NO_2$ is chemical factor and $PM_{10}$ is particulate factor
  - For Mumbai, $NO_2$ is chemical factor and $PM_{10}$ is particulate factor
  - For Delhi, $SO_2$ is chemical factor and $PM_{2.5}$ is particulate factor

- For the calculated AQI data of Delhi, we fitted the time series model , the effectiveness of the model in forecasting is confirmed . Then the fitted model is used to forecasts AQI for the next year (Jan 2020- Dec 2020).

# 7.1 SUGGESTIONS

**Air quality monitoring**

a. The number of air quality monitoring stations should be increased from the present to keep pace with city's growth. More stations are needed particularly in hot spots where the vehicular and industrial emissions are high. Similarly, monitoring stations at areas with low background air pollution level may be considered.

b. Besides increasing the number of monitoring stations, the performance of the existing stations should be overhauled and streamlined ensuring strict quality control. The collected air quality data should be comprehensively and statistically analyzed to get an insight into temporal and spatial trends.

c. More recent epidemiological studies have identified that $PM_{10}$ and $PM_{2.5}$ are the principal mediators of health effects of air pollution. Therefore, a review is needed as to which pollutant should be monitored at a regular basis. Monitoring of $PM_{2.5}$ at regular basis is recommended.

d. Attention has been focused almost entirely on controlling emissions from road traffic. But the contributions of other significant sources of pollution, for example, small stationary sources like refuse burning, emissions from small-scale industries, and household use of biomass are not known with any degree of certainty. Efforts should be made at all levels to control these sources through an effective urban air quality management strategy.

e. Since people on average spend two third of their daily time indoor, indoor air quality has profound effect on human health. In fact 'sick building syndrome' is a growing concern worldwide. Besides smoking, indoor air quality in Indian households is affected by emissions from burning biomass and kerosene during cooking, emissions from mosquito-repellants, from molds, and from cooking oil vapors during cooking at high temperatures. Reports are scanty on air pollution level in indoor air and contribution of each potential source in Indian households. Therefore indoor air quality needs to be monitored periodically.

**Reduction of vehicular emissions**

a. The use of CNG and LPG in all classes of vehicles, both private and public, should be encouraged in Delhi and in other cities with high level of vehicular pollution.

b. The authority should explore the avenues of improvement in quality of automotive fuels, lubricating oils, and vehicle types.

c. The use of cetane enhancer and detergent additive in diesel fuel may be encouraged to combat the build-up of detrimental deposits of fuel injectors that is expected to reduce emissions and increase fuel economy and extended component life.

**Health impact**

a. An important problem is the absence of facilities for regular monitoring of public health in relation to air pollution exposures despite the fact that protection of public health is the ultimate goal of air quality monitoring. Regular monitoring of public health may be carried out.

b. The concentrations of air-borne pollutants in different parts of India have been measured by various organizations but in most cases these studies were not linked to health of the exposed subjects. So much better studies must be done.

**Public education and awareness**

a. Mass awareness campaigns involving local bodies, voluntary organizations, students, trade unions and others may be initiated educating people about the health impact of air pollution. Moreover, air quality management and air pollution mitigation measures taken up by the government at the state or central level may be widely promoted through educational and information programmmes.

b. Environmental Health subject may be introduced in medical education syllabus in India.

c. Mass transportation should be strengthened in Delhi. Introduction of metro railway service is a right step in the right direction. On the other hand, people should be advised to maintain their vehicles properly. At the same time they should be encouraged to walk or use bicycles for traveling short distances, and to share vehicles for long distances.

d. In many cases pollution victims are not beneficiary of the polluting facility, such as a pedestrian is affected by the exhaust fumes of a moving car. 'The polluter should pay' philosophy should be introduced in India as well.

e. Consumer awareness should be promoted about air quality benefits by opting to eco-friendly products such as use of water-based acrylic paints that contains less VOCs.

f. From the economic point of view, the total cost of control measures should be measured in terms of total benefits to the society. The monetary benefits of reducing illness and premature mortality associated with a small change in air pollution exposure is important to estimate the value of unit reduction in each pollutant that can serve as an input for a cost-benefit analysis of air pollution mitigation programs.

g. In many cities hot spots are highly influenced by surrounding buildings. Furthermore, due to lack of space, many new developments of infrastructure and housing need to be realized in very small areas, with high risk of creating new hot spots of air pollution. An improvement in our understanding regarding the effect of buildings on hot spot concentrations is needed.

# 7.2 REFERENCES


- Central Pollution Control Board Website Website: www.cpcb.nic.in
- National Air Quality Index by CPCB
- Centre for Science & Environment website: www.cseindia.org
- The Energy and Resources Institute (TERI) website: www.teriin.org
- SAFAR-India website:www.safar.tropmet.res.in
- Johnson, R. A., and Wichern, D. W. (1990) Applied Multivariate Statistical Analysis, Prentice Hall.
- Anderson, T. W. (1984) An Introduction to Multivariate Statistical Analysis, John Wiley.
- Chatfield, C. (1995). Analysis of Time Series: An Introduction. Fifth Edition. Chapman and Hall/CRC. BOCA Raton London New York Washington, D.C.
- Spyros Makridakis, Steven C Wheelwright, Rob J. Hyundman ,Forecasting Methods and Applications ,John Wiley & Sons ,Inc