

ANALYSING AND FORECASTING TIME SERIES DATA

Submitted in partial fulfillment of the requirements for the award of

Master of Science Degree in Mathematics

By

SARIKA G (41630004)



DEPARTMENT OF MATHEMATICS

SCHOOL OF SCIENCE AND

HUMANITIES

**SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY**

(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by

NAAC I 12B Status by UGC I

Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI

SALAI, CHENNAI - 600 119

APRIL – 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC I 12B Status by UGC I Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF MATHEMATICS

BONAFIDE CERTIFICATE

This is to certify that this project report is the bonafide work of **SARIKA G 41630004** who carried out the project entitled "**ANALYSING AND FORECASTING TIME SERIES DATA**" under our supervision from **JANUARY 2023** to **APRIL 2023**.

Internal Guide

Dr. M. NIRMALA

Head of the Department

Dr. M. NIRMALA, M.Sc., M.Phil., Ph.D.,

Head of the Department

Dept. of Mathematics

Sathyabama Institute of Science and Technology

(Deemed to be University), Chennai-600 119.

Submitted for Viva voce Examination held on

11/5/2023

Internal Examiner

External Examiner

DECLARATION

I **SARIKA G 41630004** hereby declare that the project report entitled **“ANALYSING AND FORECASTING TIME SERIES DATA”** done by me under the guidance of **Dr. M. NIRMALA** (Internal) at **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** submitted in partial fulfillment of the requirements for the award of Master of Science Degree in Mathematics

DATE: 26-04-2023

Sarika G

PLACE: Chennai

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of SATHYABAMA for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to Dr. M. Nirmala, Head of the Department, Department of Mathematics for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide Dr. M. Nirmala for her valuable guidance, suggestion and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching staff and non- teaching staff members of the department of Mathematics who are helpful in many ways for the completion of the project.

ABSTRACT:

A prime issue faced in India is the quality of air. Air pollution is the greatest environmental threat to public health globally and accounts for an estimated 7 million premature deaths every year. Chennai is one of the fast-progressing metropolitan city which is battling against air pollution. Ambient air pollutants such as PM_{10} , $PM_{2.5}$, O_3 , SO_2 and NO_2 are well known for their toxic nature and ability to cause many respiratory disorders like bronchitis, emphysema and asthma. For maintaining a sustainable development and ecological balance, it is vital to optimise ambient air quality. In this study, the Air Quality Index (AQI) is analysed based on the database monitored at different monitoring centres in Chennai. For this purpose, a data set consisting of daily Ambient Air Quality data of T-Nagar, Nungambakkam and Adyar with various air pollutants for commercial, residential and traffic area were collected from Tamil Nadu Pollution Control Board. This project work aims to understand air pollution aspects in T-Nagar, Nungambakkam and Adyar by identifying an appropriate time series model and the result were analysed.

KEYWORDS:

- ❖ Air Pollution
- ❖ PM_{10} , $PM_{2.5}$, O_3 , SO_2 and NO_2
- ❖ Air Quality Index (AQI)
- ❖ Time Series Model

CONTENTS

CHAPTER – 1	PAGE NO
1.1 INTRODUCTION	1
1.1.1 COMPONENTS OF TIME SERIES	1
1.2 TIME SERIES ANALYSIS	4
1.3 AIR POLLUTION	4
1.4 TYPES OF AIR POLLUTANTS	4
1.4.1 PRIMARY POLLUTANTS	5
1.4.2 SECONDARY POLLUTANTS	5
1.5 IMPACTS ON HUMAN HEALTH	5
1.6 VARIOUS PARAMETERS OF AIR	6
1.7 AIR QUALITY INDEX	7
CHAPTER – 2	
2.1 AIM OF THE STUDY	11
2.2 OBJECTIVE OF THE PROJECT	11
2.3 SCOPE OF THE PROJECT	11
2.4 STUDY AREA	11
2.5 AIR POLLUTION IN CHENNAI	13
2.6 AREAS TAKEN UNDER STUDY	13
2.7 PARAMETERS TAKEN FOR STUDY	13
CHAPTER – 3	
3.1 LITERATURE SURVEY	15
CHAPTER – 4	
4.1 DESCRIPTIVE STATISTICS	19
4.1.1 PURPOSE OF DESCRIPTIVE STATISTICS	19
4.1.2 TYPES OF DESCRIPTIVE STATISTICS	20
4.1.3 UNIVARIATE DESCRIPTIVE STATISTICS	20
4.1.4 BIVARIATE DESCRIPTIVE STATISTICS	21

4.2	METHODOLOGY FOR FORECASTING	21
4.2.1	QUALITATIVE FORECASTING METHODS	21
4.2.2	QUANTITATIVE FORECASTING METHODS	22
4.2.3	MOVING AVERAGE	22
4.3	ERROR MEASURE	23
4.3.1	MEAN ABSOLUTE PERCENTAGE ERROR(MAPE)	23
4.3.2	MEAN ABSOLUTE DEVIATION (MAD)	23
4.3.3	MEAN SQUARED DEVIATION (MSD)	24
CHAPTER – 5		
5.1	RESULT AND DISCUSSION	26
5.2	ANALYSIS DONE WITH PARAMETER	26
CHAPTER – 6		
6.1	CONCLUSION	42
6.2	FUTURE WORK	42
REFERENCES		43

LIST OF FIGURES

Fig. 1.1.	Showing Components of Time Series	2
Fig. 1.2.	Picture view of the health impacts of air pollution	6
Fig. 1.3.	AQI Calculation Formula	7
Fig. 2.1.	Map of Chennai, India	12
Fig. 4.1.	An outline flowchart for the many types of descriptive statistics	20
Fig 5.1.	Show how the amount of particulate Matter (PM10) varies every week.	26
Fig 5.2.	Show how the amount of particulate Matter (PM2.5) varies every week.	27
Fig. 5.3.	Show how the amount of Sulphur Di-Oxide (SO ₂) varies every week.	28
Fig. 5.4.	Show how the amount of Nitrogen Di-Oxide (NO ₂) varies every week.	29
Fig. 5.5.	Show how the amount of AQI varies every week	30
Fig. 5.6.	Graph's of T-Nagar AQI with different k values and best fit is highlighted.	33
Fig. 5.7.	Graph's of Nungambakkam AQI with different k values and best fit is highlighted.	36
Fig. 5.8.	Graph's of Adyar AQI with different k values and best fit is highlighted.	39

LIST OF TABLES

Table 1.1.	Example of dataset for their respective fields	1
Table 1.2.	AQI category and pollutants	8
Table 1.3.	AQI category and meaning	9
Table 5.1.	Descriptive data on measured air quality for PM ₁₀	26
Table 5.2.	Descriptive data on measured air quality for PM _{2.5}	27
Table 5.3.	Descriptive data on measured air quality for SO ₂	28
Table 5.4.	Descriptive data on measured air quality for SO ₂	29
Table 5.5.	Descriptive data on measured air quality for AQI	30
Table 5.6.	Shows the Moving Average for T- Nagar -AQI	31
Table 5.7.	Shows the Moving Average for Nungambakkam -AQI	34
Table 5.8.	Shows the Moving Average for Adyar -AQI	37
Table 5.9.	Best Error Measure of 3 different area of Chennai AQI	40

CHAPTER -1

INTRODUCTION:

1.1 TIME SERIES:

Time series is a collection of data at different points in a regular interval of time. Daily weather reports, yearly rainfall in Tamil Nadu, monthly sea surface temperature are some examples of time series. Time series analysis convey how data changes over a particular period of time whereas time series forecasting can recognize the direction in which the data is changing. In other words, time series analysis is all about evolving models to gain an understanding of the data to understand the underlying causes. Forecasting then takes the next step of predicting extrapolation of what might happen in the future. There are many real time applications of time series. Some of them are, economic forecasting, sales forecasting, budgetary analysis, stock market analysis, census analysis etc..

Table 1.1 Example of dataset for their respective fields

FIELD	TOPIC	DATA
Medicine	Blood pressure tracking, weighttracking, heart rate monitoring, cholesterol measurements.	MRI scanning and behavioral test dataset.
Epidemiology	Disease rates, mosquito population. Mortality rates.	U.S. cancer incidence rates from the center for disease control
Physical Sciences	Global temperatures, monthly spot observations, pollutionlevels.	Global air pollution fromour world in data
Social sciences	Birth rates, population, migration data, political indicators.	Population without citizenship from Eurostat.
Economics	Gross Domestic Product (GDP),Consumer Price Index (CPI)	India GDP from the federal reserve economic data.

1.1.1 COMPONENTS OF TIME SERIES:

The several reasons which influence the values of an observation in a timeseries. They are the components of time series and the components are,

1. Trend – T
2. Seasonal variation – S
3. Cyclical Variation – C
4. Irregular Variation – I

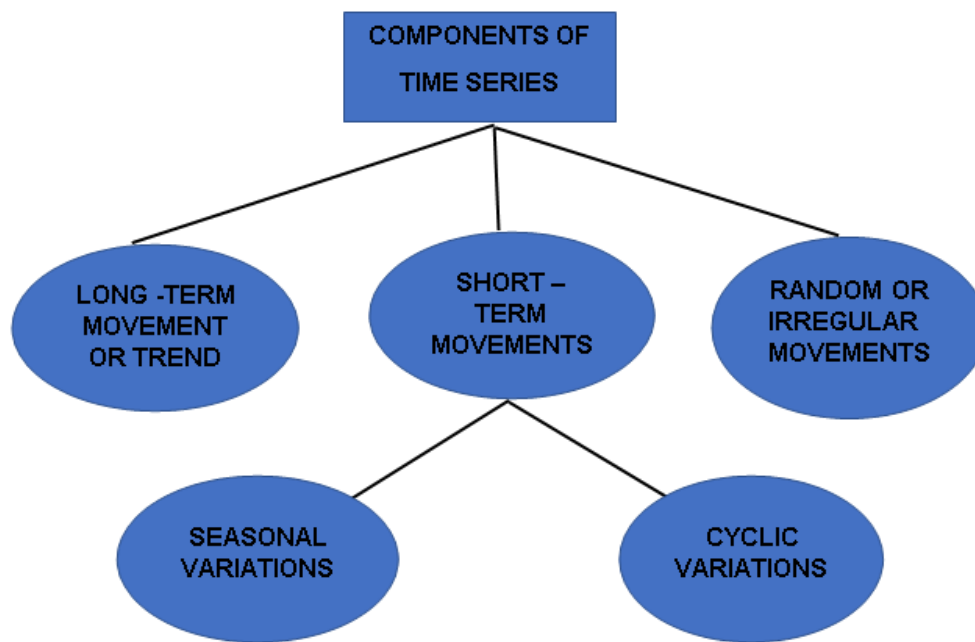


Fig. 1.1. Showing Components of Time Series

TREND - T:

Trend exhibit the general movement of the data to increase or decrease, during a longer period of time. It doesn't necessary that the increase or decrease intrend lies is in the same direction within the given period of time. Trend is normally the outcome of long-term factors such as changes in population, production of goods, demographics, technology or consumer preferences.

SEASONAL VARIATION - S:

The seasonal component is the systematic pattern of variability within particular period of time, such as a year. The variability does not always denote theseasons of the year (i.e., winter, spring, summer, fall). It can also be within-week orwithin-day seasonal behavior.

CYCLICAL VARIATION - C:

Any ordered pattern of series of values above or below the trend line remaining more than one year can be attributed to the cyclical component. Cyclicalvariation is due to multiyear cyclical movements in the economy.

IRREGULAR VARIATION – I:

Irregular variations are purely random or irregular and short in duration. Irregular variation occurs due to phenomenon of unforeseen events like wars, earthquakes or floods.

1.2 TIME SERIES ANALYSIS:

Time series analysis is a specific way of analysing a sequence of data points collected over an interval of time.

Example:

- ❖ Analysing the maximum Heart rate from the heart patients.
- ❖ The daily maximum temperature of a place recorded by the Metrological department over a number of days.
- ❖ Birth and Death rate Analysis.

1.3 AIR POLLUTION:

Air pollution refers to the release of pollutants into the air—pollutants which are detrimental to human health and the planet as a whole. According to the World Health Organization (WHO), each year air pollution is responsible for nearly seven million deaths around the globe. Nine out of ten human beings currently breathe air that exceeds the WHO's guideline limits for pollutants, with those living in low- and middle-income countries suffering the most. Air pollution is hard to escape, no matter how rich an area are you live in. It is all around us. Therefore, air pollution means the undesirable presence of impurities or the abnormal rise in the proportion of some constituents of the atmosphere.

1.4 TYPES OF AIR POLLUTANTS:

There are two basic types of pollutants in the air. They are known as

- ❖ Primary Pollutants
- ❖ Secondary Pollutants

1.4.1 PRIMARY POLLUTANTS:

Pollutants that are formed and emitted directly from particular sources.

Example:

- ❖ Particulate Matter
- ❖ Carbon Monoxide
- ❖ Nitrogen Oxide
- ❖ Sulfur Oxide
- ❖ Radioactive Compounds

1.4.2 SECONDARY POLLUTANTS:

A secondary pollutant is not directly emitted as such, but forms when other pollutants (primary pollutants) react in the atmosphere.

Example:

- ❖ Ozone, which is formed when hydrocarbons (HC) and nitrogen oxides (NO_x) combine in the presence of sunlight.
- ❖ Acid rain, which is formed when sulfur dioxide or nitrogen oxides react with water.

1.5 IMPACTS ON HUMAN HEALTH:

Air pollution has adverse effects on human health.

- ❖ Breathing polluted air puts you at higher risk of Asthma, Wheezing etc.
- ❖ High ozone levels can reduce your lung capacity (the amount of air your lungs can hold) and make it uncomfortable to breathe.
- ❖ More people with lung conditions are admitted to hospital when ozone levels are high, and there's a greater risk of getting pneumonia and bronchitis
- ❖ Pollutants such as fine particles from candles or fireplaces (or from the outdoors) also affect our health.

- ❖ In addition, long-term exposure to air pollution can cause cancer and damage to the immune, neurological, reproductive, and respiratory systems. In extreme cases, it can even cause death.

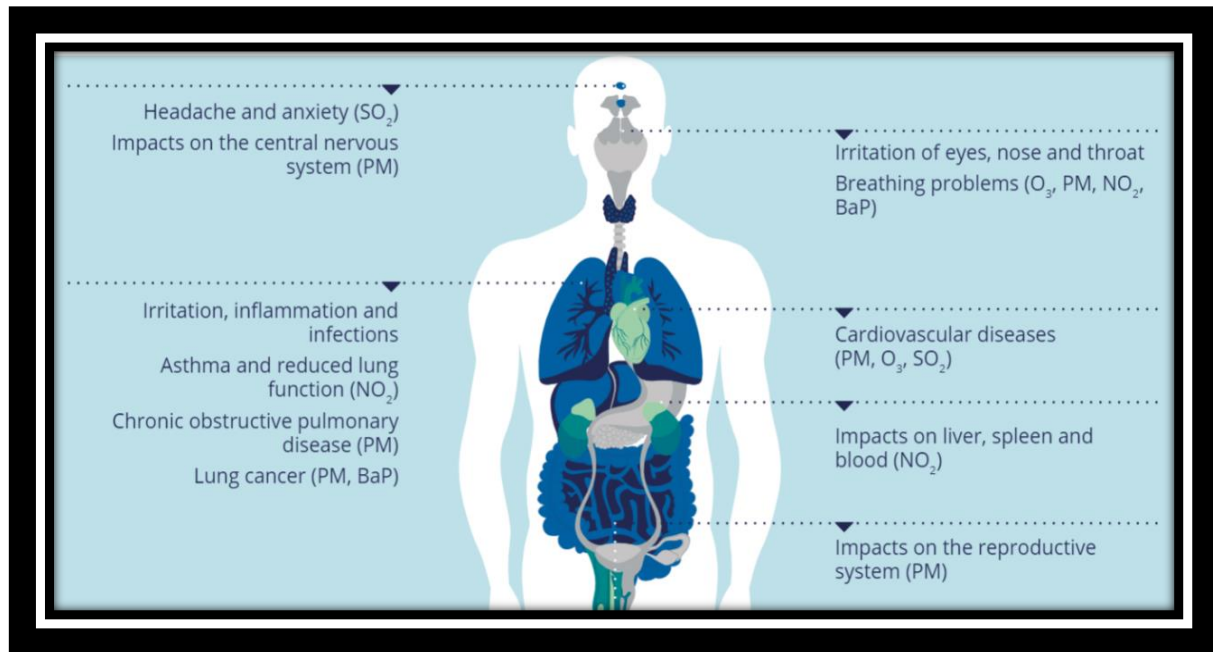


Fig. 1.2. Picture view of the health impacts of air pollution

1.6 VARIOUS PARAMETERS OF AIR:

The following act as major parameters in deriving the AQI of an area

- ❖ Particulate Matter (PM_{10})
- ❖ Particulate Matter ($\text{PM}_{2.5}$)
- ❖ Ozone (O_3)
- ❖ Sulphur di-oxide (SO_2)
- ❖ Nitrogen di-oxide (NO_2)
- ❖ Carbon monoxide (CO)
- ❖ Lead (Pb)
- ❖ Ammonia (NH_3)

1.7 AIR QUALITY INDEX:

An **air quality index (AQI)** is used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become. AQI information is obtained by averaging readings from an air quality sensor, which can increase due to vehicle traffic, forest fires, or anything that can increase air pollution. Pollutants tested include ozone, nitrogen dioxide, sulphur dioxide, among others. The results are grouped into ranges, and each range is assigned a descriptor, a color code and a standardized public health advisory.

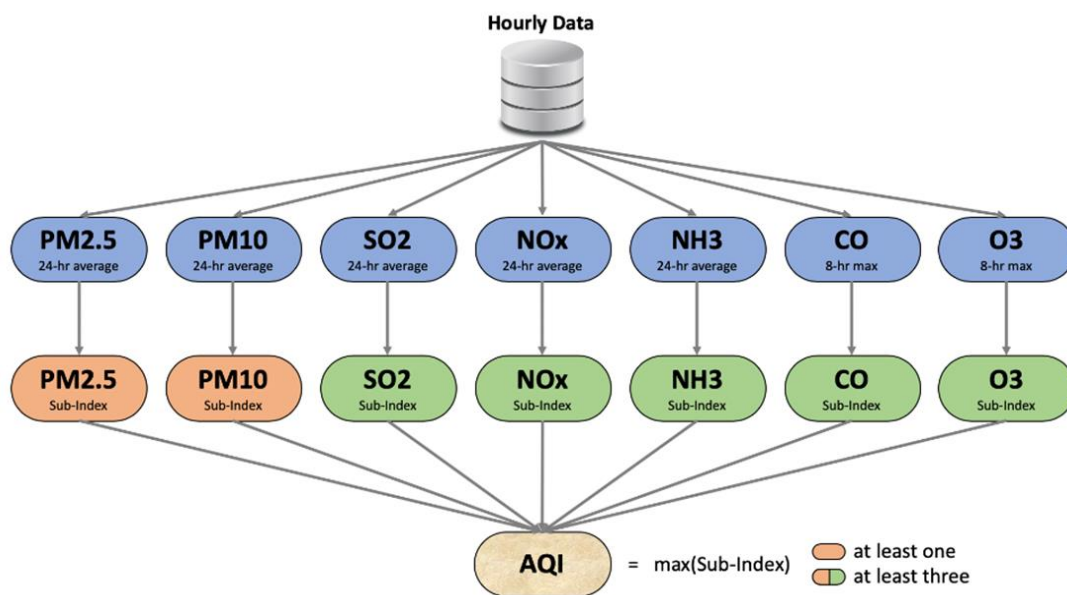


Fig. 1.3. AQI Calculation Formula

- ❖ The AQI calculation uses 7 measures: **PM2.5, PM10, SO2, NOx, NH3, CO and O3**.
- ❖ For **PM2.5, PM10, SO2, NOx and NH3** the average value in last 24-hrs is used with the condition of having at least 16 values.
- ❖ For **CO and O3** the maximum value in last 8-hrs is used.

- ❖ Each measure is converted into a Sub-Index based on pre-defined groups.
- ❖ Sometimes measures are not available due to lack of measuring or lack of required data points.
- ❖ Final AQI is the maximum Sub-Index with the condition that at least one of PM_{2.5} and PM₁₀ should be available and at least three out of the seven should be available.

Table 1.2. AQI category and pollutants

AQI Category (Range)	PM ₁₀ 24-hr	PM _{2.5} 24-hr	NO ₂ 24-hr	O ₃ 8-hr	CO 8-hr (mg/m ³)	SO ₂ 24-hr	NH ₃ 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.5 -1.0
Moderately polluted (101-200)	101-250	61-90	81-180	101-168	2.1- 10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748*	17-34	801-1600	1200-1800	3.1-3.5
Severe(401-500)	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+

Table 1.3. AQI category and meaning

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health warnings of emergency conditions. The entire population is more likely to be affected.
Hazardous	301 to 500	Health alert: everyone may experience more serious health effects.

CHAPTER -2

2.1 OBJECTIVE OF THE PROJECT:

- ❖ Analysing the secondary time series data to understand the underlying behaviour of the dataset.
- ❖ Forecasting the Time series data using advanced statistical techniques.

2.2 SCOPE OF THE PROJECT:

To achieve the objective of this project work, collected daily data of Air Quality Index (AQI) of Chennai from **Tamil Nadu Pollution Control Board** for a period from 1st January 2022 to 4th February 2023 of three different areas in Chennai, having a total of 2000 data with 5 different parameters. In this study, Air Pollution is chosen since the health effects of air pollution are serious. According to World Health Organization, one third of deaths from stroke, lung cancer and heart disease are due to air pollution. This is having an equivalent effect to that of smoking tobacco. Therefore, AQI is identified in this project work, and actions can be made to lower AQI.

2.3 AIM OF THE STUDY:

This study aims at forecasting Air Quality Index (AQI) using statistical methods and to find the most appropriate forecasting technique for the collected secondary time series data.

2.4 STUDY AREA:

CHENNAI

- ❖ Chennai (formerly known as Madras) is the capital city of the Indian state of Tamil Nadu.
- ❖ According to the 2011 Indian census, Chennai is the sixth-most populous city in India. Chennai Population in 2023 is estimated to be 11.8 Million (1.18 Crore)
- ❖ Almost 10% of all of the people in the state live in Chennai. The city is the fourth largest city of India.
- ❖ There are several manufacturing companies situated in Chennai. Most of the manufacturing companies choose Chennai as their land for the headquarters and several branches, as it is the most industrialised and finest city in India.
- ❖ There were around 6,016,717 registered vehicles across the city of Chennai in India at the end of the year 2020.
- ❖ Chennai ranks fifth in carbon emissions from the transport sector among 54 South Asian cities. The city's emission levels in transport is much higher than Kolkata and Delhi.

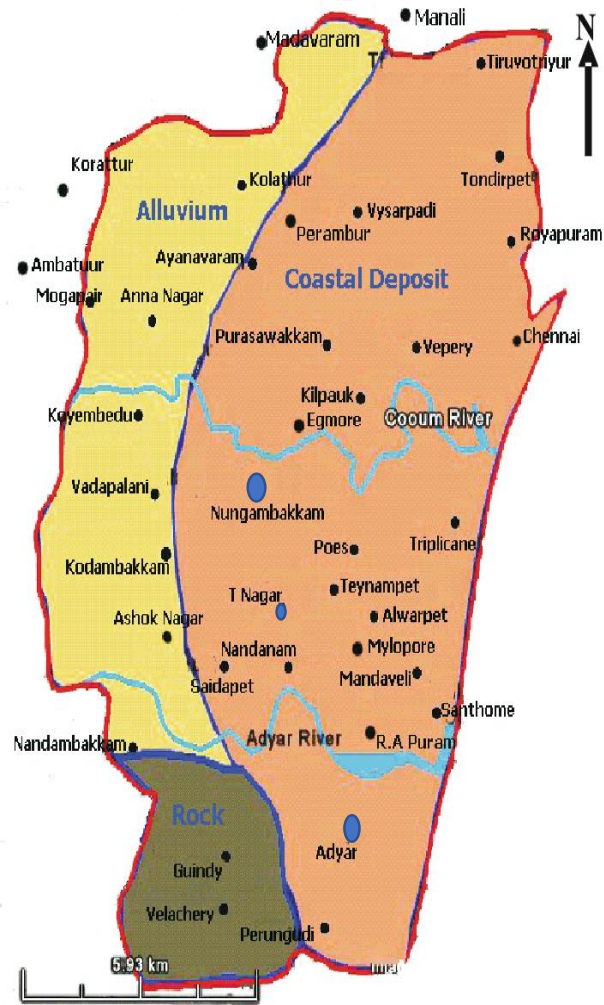


Fig. 2.1. Map of Chennai, India

2.5 AIR POLLUTION IN CHENNAI:

An analysis of air quality data of Chennai by Greenpeace India has shown that the average pollution level in the city was much higher than the WHO air quality guidelines. The levels of PM₁₀, which is particulate matter less than 10-micron size in the air we breathe, have gone up by three to four times in Chennai. The major sources of air pollution at Chennai city are road dust, vehicular emission, construction activities, Industries etc. A report published in the Times of India (TOI) revealed that Greenpeace studies show that Chennai is one of the most polluted cities in India.

2.6 AREAS TAKEN UNDER STUDY:

- ❖ T-Nagar (Commercial)
- ❖ Nungambakkam (Traffic area)
- ❖ Adyar (Residential)

2.7 PARAMETERS TAKEN FOR STUDY:

- ❖ Particulate Matter (PM₁₀)
- ❖ Particulate Matter (PM_{2.5})
- ❖ Sulphur Di-oxide (SO₂)
- ❖ Nitrogen Di-oxide (NO₂)
- ❖ AQI

CHAPTER-3

3.1 LITERATURE SURVEY:

- ✚ Elia Georgiana Dragomir [3] used a traditional data mining method known as K-nearest neighbor which is used to categorize new instances and determine the distance of a new instance using a variety of mathematical distance metrics. It is also used to forecast the value of the Air Quality Index, which is used to classify the amount of pollution and alert the public to potential episodes of pollution.

- ✚ Naveen V et.al [7] analyzed the air quality index using the time series approach and the approach used is based on Box-Jenkins method which is referred as ARIMA. On the basis of a database generated from readings taken at several monitoring stations around the Thiruvananthapuram District in Kerala, India, the study examined and predicted various trends in ambient air quality in terms of the Air Quality Index (AQI). Due to its abundance in the environment, Respirable Suspended Particulate Matter (RSPM) was the pollutant that caused the AQI in all of these locations, according to statistics on air quality from the Kerala State Pollution Control Board (KSPCB).

- ✚ Imran Nadeem et.al [6] adopts a univariate linear stochastic ARMA/ARIMA modelling approach for forecasting. In this study, nine univariate linear stochastic models are used to forecast RSPM, SO₂, and NO₂ concentrations for the three most polluted sites in Chennai city using an ARMA/ARIMA modelling approach. To determine the extent of pollution, comparisons of actual and permitted air pollutant concentrations were made.

- ✚ Nidhi Sharma et.al [8] investigates the trends of various air contaminants using descriptive and predictive analyses. Due to the city of Delhi's rapid urbanization and expansion, air quality has been a significant issue for a very long time. This has caused the city's air pollution to increase at an

alarming rate. Various air pollutants such as sulphur dioxide (SO₂), nitrogen dioxide (NO₂), suspended particulate matter (PM), ozone (O₃), carbon monoxide (CO), and benzene have tendencies that can be studied and predicted using descriptive and predictive analysis. According to forecasts, NO₂ will likely increase and SO₂ will rise by 1.24 µg/m³, and these changes will improve our knowledge of future air quality in the Delhi region.

✚ Evgeniy Marinov et.al [5] effectively used the ARIMA approach to forecast air pollution while taking into account various granularities.. On both the environment and human health, air pollution has a serious effect. A reliable strategy for enforcing regulations and taking preventive action is time series analysis. The air pollution trends at five air quality sites in Sofia, Bulgaria, are examined in this study. Models for statistical analysis are developed using the AutoRegressive Integrated Moving Average (ARIMA) method to predict the concentrations of contaminants. The technique also enables us to study the correlation between levels of a particular pollutant measured in several air quality stations and determine whether the pollutants' levels surpass those set by the World Health Organization (WHO).

✚ Shrivallabha S et.al [9] analysed data on air pollutants using the ARIMA model. Inert carbonaceous cores with numerous layers of absorbed nanocrystals, including metals, organic contaminants, acid salts, and biological components, make up particulate matter (PM). Total suspended particulates (TSP) and total suspended superphosphate matter (TSP) are two of its three categories. From April 2018 to November 2018 in Bengaluru City, Karnataka, India, the future concentration of air pollution was analysed and forecast using the ARIMA model of time series analysis. The Ljung-Box test was performed to test the model, and residual analysis

was employed for cross-validation.

- ✚ Alya Atoui et.al [1] applies the Nave, TBATS, exponential smoothing, and SARIMA models to forecast univariate time series. With the help of the Nave, Exponential Smoothing, TBATS, and Seasonal Autoregressive Integrated Moving Average (SARIMA) models, this study seeks to assess the atmospheric air pollution in Zahleh, Bekaa. As a result of the findings, communities like Zahleh may use the SARIMA model to predict the Air Quality Index (AQI) with the greatest accuracy, allowing them to gauge the level of air pollution and avoid negative health effects.

CHAPTER-4

4.1 DESCRIPTIVE STATISTICS:

Descriptive statistics describe, show, and summarize the basic features of a dataset found in a given study, presented in a summary that describes the data sample and its measurements. It helps analysts to understand the data better. Analysts often use charts and graphs to present descriptive statistics.

EXAMPLE:

A student's grade point average (GPA) is the best example. A GPA gathers the data points created through a large selection of grades, classes, and exams then average them together and presents a general idea of the student's mean academic performance. Note that the GPA doesn't predict future performance or present any conclusions. Instead, it provides a straightforward summary of students' academic success based on values pulled from data.

4.1.1 PURPOSE OF DESCRIPTIVE STATISTICS:

Descriptive statistics can be useful for two things: 1) providing basic information about variables in a dataset and 2) highlighting potential relationships between variables. Graphical/Pictorial Methods are measures of the three most common descriptive statistics that can be displayed graphically or pictorially. It is used to summarise data. Descriptive statistics only make statements about the data set used to calculate them; they never go beyond your data.

4.1.2 TYPES OF DESCRIPTIVE STATISTICS:

There are three primary types of descriptive statistics:

- ❖ **Distribution** – This type focuses on the frequency of each value in a data set.
- ❖ **Central tendency** – This one focuses on the averages of the values.
- ❖ **Variability/dispersion** – This one focuses on how the values spread out.

In univariate analysis, you can use these three types of descriptive statistics to assess one variable at a time. However, in bivariate or multivariate analysis, you can use them to compare two or more variables.

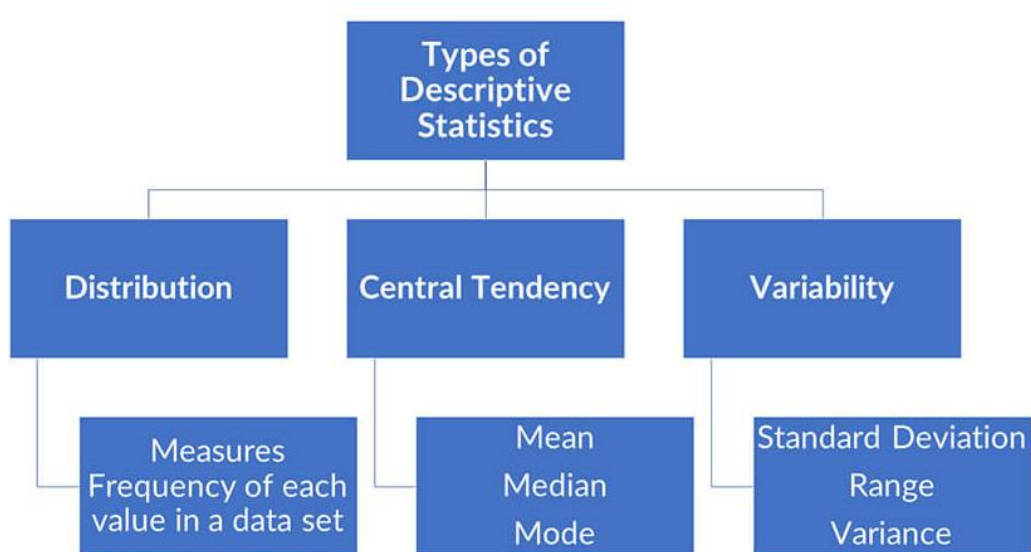


Fig.4.1 An outline flowchart for the many types of descriptive statistics

4.1.3 UNIVARIATE DESCRIPTIVE STATISTICS:

Univariate descriptive statistics focus on only one variable at a time. It is important to examine data from each variable separately using multiple measures of distribution, central tendency and spread.

4.1.4 BIVARIATE DESCRIPTIVE STATISTICS:

When using bivariate descriptive statistics, two variables are simultaneously analysed (compared) to see if there is a connection between them. In other words, bivariate analysis examines any concurrent relationship between two variables or attributes. In order to determine whether there are any differences between the two variables and the potential causes of these differences, this study examines the relationship between the two variables as well as the depth of this relationship.

Multivariate analysis is the same as bivariate analysis but with more than two variables.

4.2 METHODOLOGY FOR FORECASTING:

Forecasting is a method of forecasting the future using the outcomes of past data. In order to anticipate future events, a thorough analysis of past and present trends or events is required. It makes use of statistical methods and tools. As a result, statistical analysis is another name for it. In other terms, we could say that it serves as a planning tool that helps organisations prepare for potential future unpredictability.

The two main categories of forecasting techniques are qualitative and quantitative.

4.2.1 QUALITATIVE FORECASTING METHODS:

- ❖ Delphi Technique
- ❖ Market Research
- ❖ Panel Consensus
- ❖ Historical Analysis

4.2.2 QUANTITATIVE FORECASTING METHODS:

- ❖ Moving Average
- ❖ Regression Analysis
- ❖ Exponential Smoothing
- ❖ Adaptive Smoothing

In time series analysis there are many methodologies that can be used for forecasting. In this project work, **Moving Average** method is used.

4.2.3 MOVING AVERAGE:

Moving average is one of the best approachable techniques for pre-treating of timeseries data. Unweighted moving average is the most common moving average method. It is used to filter random white noise from the data and smooths it by replacing each data point with the average of k neighboring points, where k is called the order or length of moving average. In general, higher the length of moving average, greater the smoothing effect is. For given N data points and T, the length of the moving average, MA(T), the forecast F_{T+1} is given by

$$F_{T+1} = \sum_{i=1}^T \frac{x_i}{T} \text{ ----- (1)}$$

$$F_{T+2} = \sum_{i=2}^{T+1} \frac{x_i}{T} \text{ ----- (2)}$$

Combing the above two equations, we get

$$F_{T+2} = F_{T+1} + \frac{x_{T+1} - x_1}{T} \text{ ----- (3)}$$

From the above equation, we can see that each new forecast, F_{T+1} is an adjustment of the immediate preceding forecast.

4.3 ERROR MEASURE:

Error is a difference between actual (or observed) value and forecasted value. In this research work, Mean Absolute Percentage Error (MAPE), Mean absolute deviation (MAD), Mean squared deviation (MSD) are used to forecast the error. Smaller value usually indicate better forecasted value.

4.3.1 MEAN ABSOLUTE PERCENTAGE ERROR(MAPE):

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula:

$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right| \text{-----} (1)$$

Here,

M – Mean Absolute Percentage Error

n – Number elements in the data set

A_t – Observed value

F_t – Forecasted value

4.3.2 Mean absolute deviation (MAD):

Expresses accuracy in the same units as the data, which helps conceptualize the amount of error. Outliers have less of an effect on MAD than on MSD. It is calculated by the formula

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - m(X)| \text{-----}(2)$$

Here,

n - number of data values

x_i - data values in the set

$m(X)$ - Average value of the data set

4.3.3 Mean squared deviation (MSD):

A commonly-used measure of accuracy of fitted time series values.

Outliers have a greater effect on MSD than on MAD. The Formula is

$$MSD = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{y}_i)^2 \text{-----} (3)$$

Here,

n – number of data points

Y_i – observed values

$\hat{Y}_i$ – predicted values

CHAPTER-5

5.1 RESULT AND DISCUSSION:

After applying the methodologies for forecasting the Air Quality Index (AQI) data, the results which are obtained during this experimental work has been discussed in this chapter.

5.2 ANALYSIS DONE WITH PARAMETER:

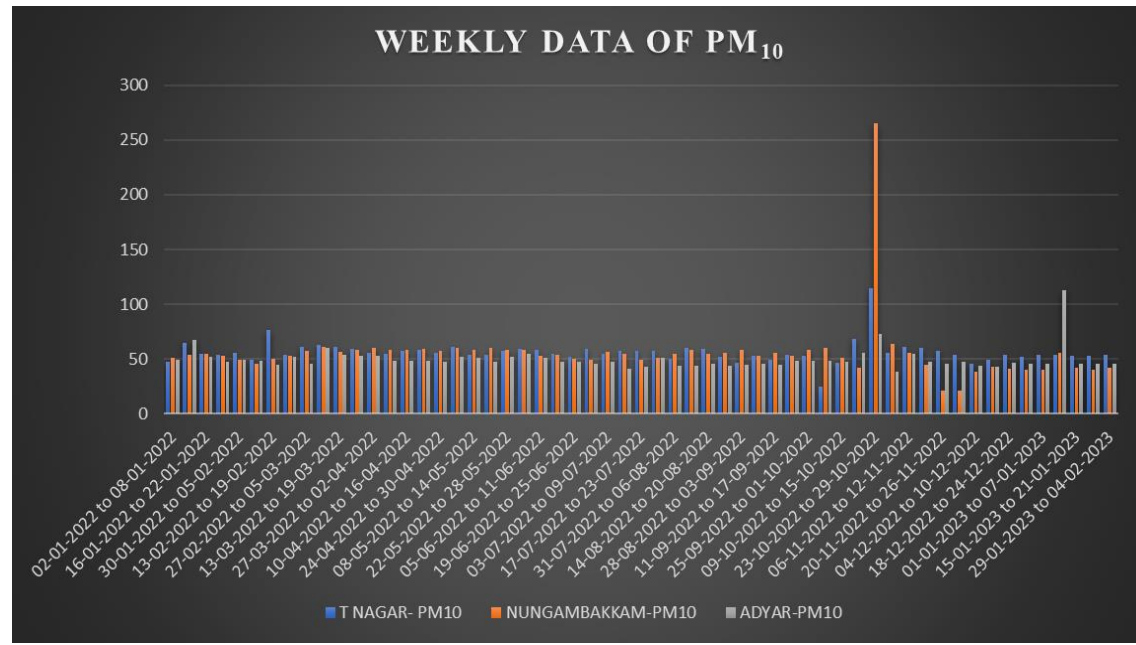


Fig. 5.1. Show how the amount of particulate Matter (PM₁₀) varies every week.

Table 5.1. Descriptive data on measured air quality for PM₁₀

	T-NAGAR- PM₁₀	NUNGAMBAKKAM- PM₁₀	ADYAR- PM₁₀
AVERAGE	55.1258	51.6786	48.6210
MAX	68.6667	63.5	72.5
MIN	24.75	20.9167	38
OUTLIERS	114.6667	265.3333	112.5

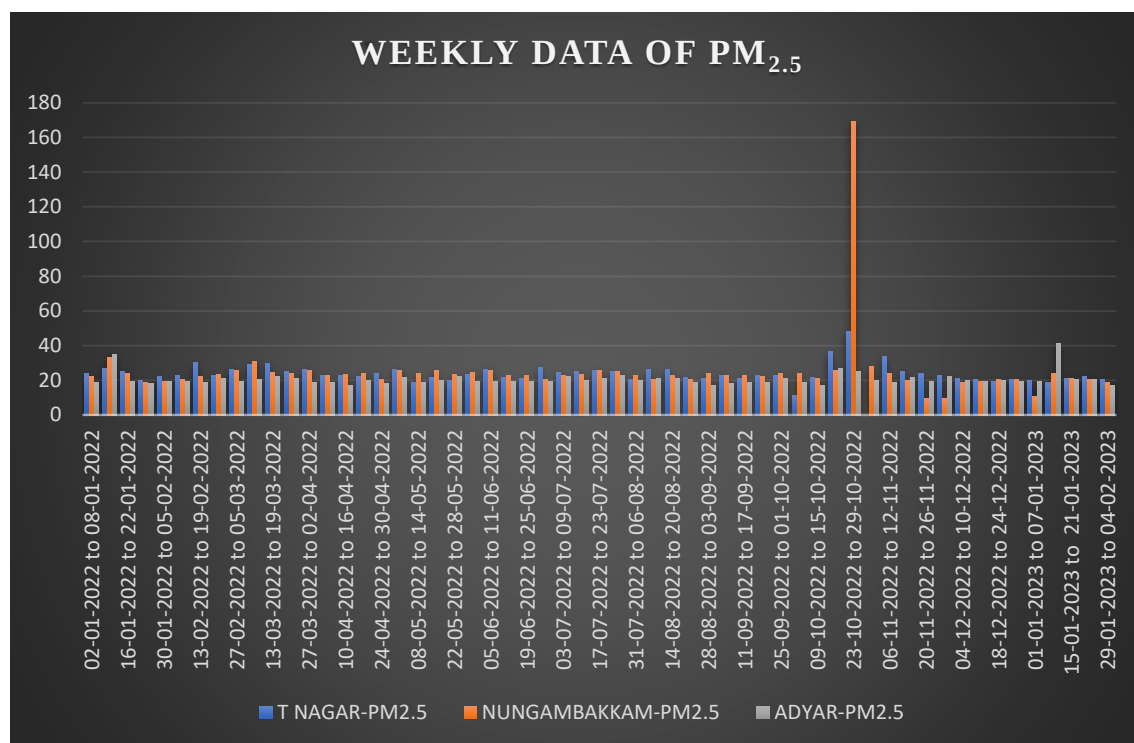


Fig. 5.2. Show how the amount of particulate Matter (PM_{2.5}) varies every week.

Table 5.2. Descriptive data on measured air quality for PM_{2.5}

	T-NAGAR- PM_{2.5}	NUNGAMBAKKAM- PM_{2.5}	ADYAR- PM_{2.5}
AVERAGE	23.4515	22.4196	20.2262
MAX	36.3333	33	35
MIN	11	9.75	17
OUTLIERS	---	169.333	---

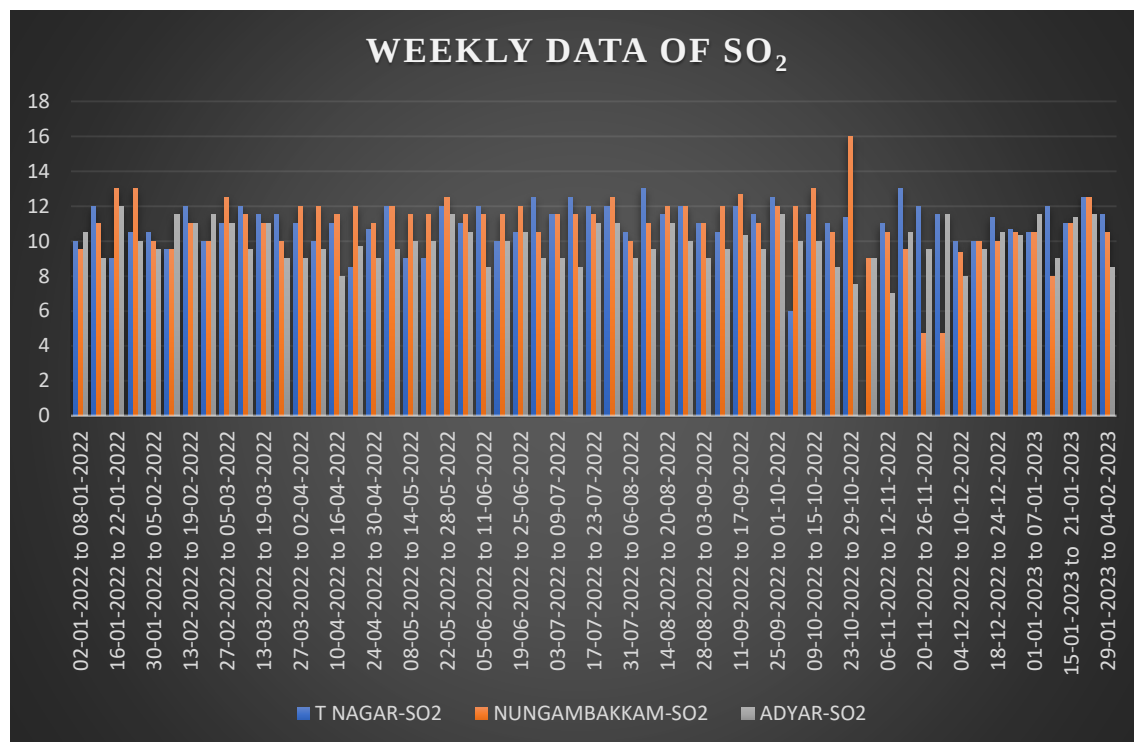


Fig. 5.3. Show how the amount of Sulphur Di-Oxide (SO₂) varies every week.

Table 5.3. Descriptive data on measured air quality for SO₂

	T-NAGAR-SO₂	NUNGAMBAKKAM-SO₂	ADYAR-SO₂
AVERAGE	11.03557	11.0161	9.8713
MAX	13	13	12
MIN	6	4.7083	7
OUTLIERS	---	16	---

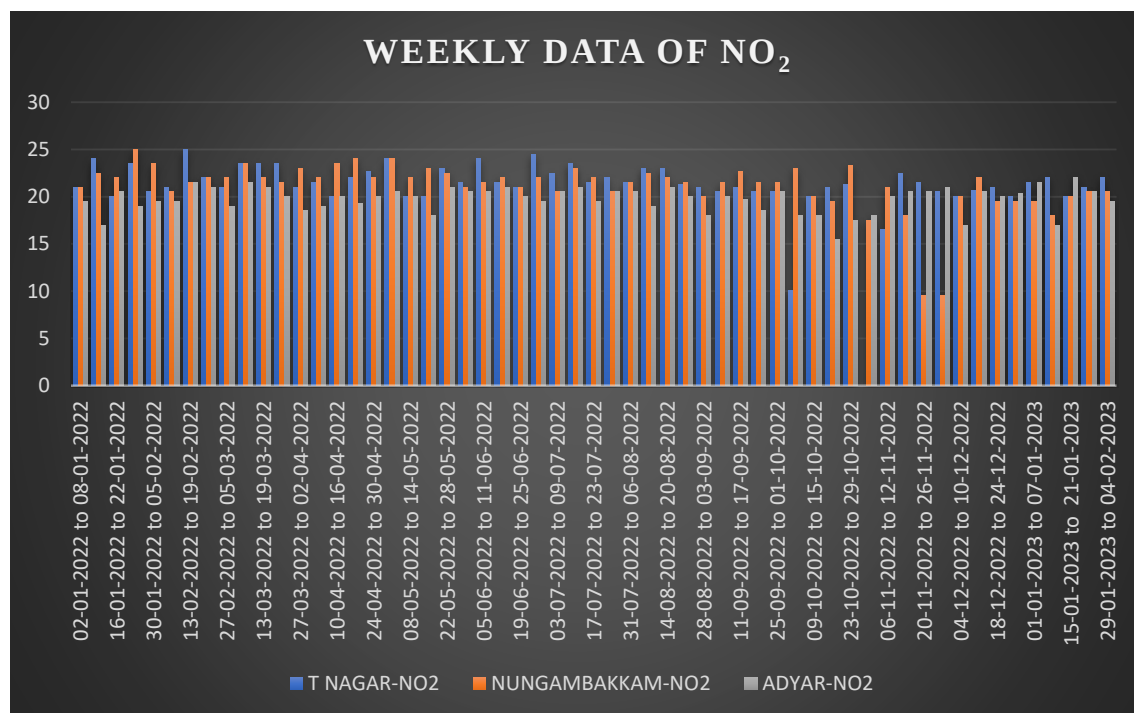


Fig. 5.4. Show how the amount of Nitrogen Di-Oxide (NO₂) varies every week.

Table 5.4. Descriptive data on measured air quality for NO₂

	T-NAGAR- NO₂	NUNGAMBAKKAM- NO₂	ADYAR- NO₂
AVERAGE	21.4040	21.0877	19.6988
MAX	25	25	22
MIN	10.125	9.5	15.5
OUTLIERS	---	---	---

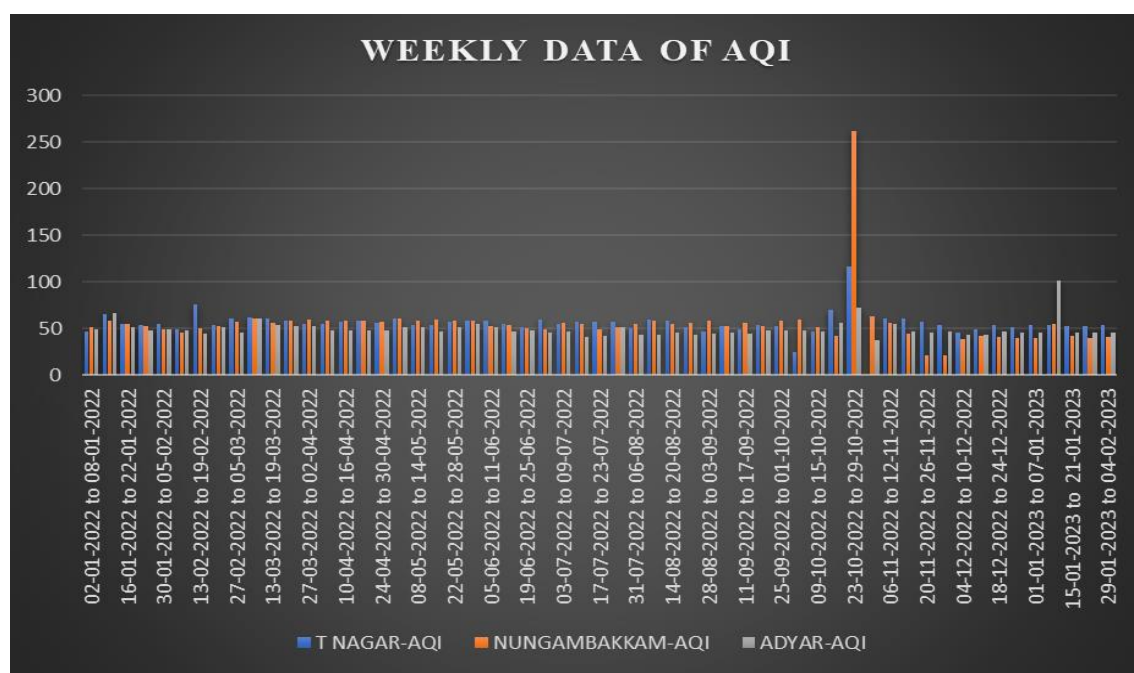


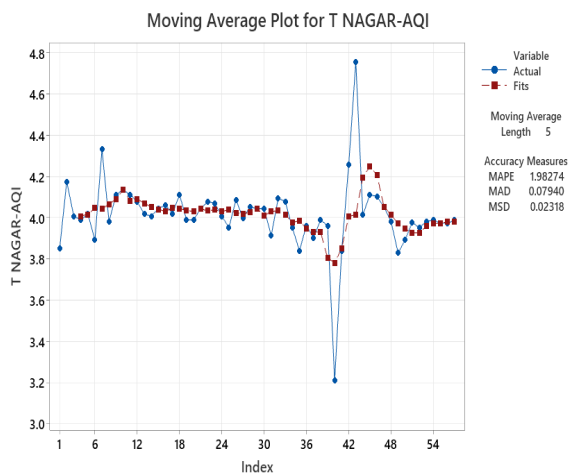
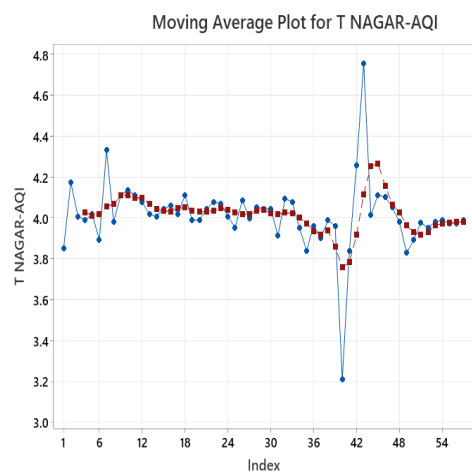
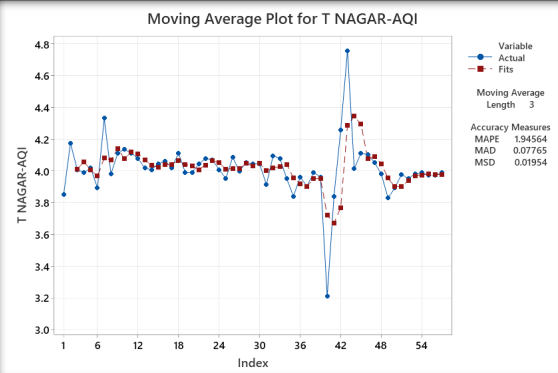
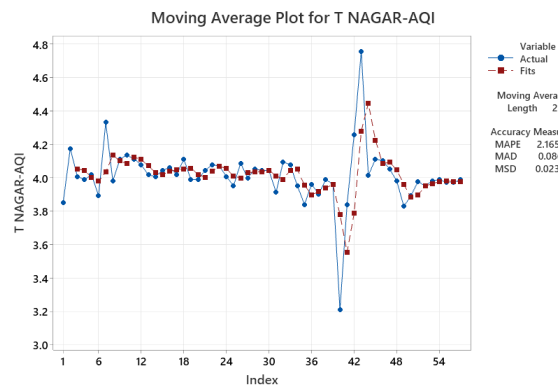
Fig. 5.5. Show how the amount of AQI varies every week.

Table 5.5. Descriptive data on measured air quality for AQI

	T-NAGAR- AQI	NUNGAMBAKKAM- AQI	ADYAR- AQI
AVERAGE	55.1621	51.7857	48.6279
MAX	76	63.5	72.5
MIN	24.75	20.9167	38
OUTLIERS	116.333	262.333	102

Table 5.6. shows the Moving Average for T-Nagar -AQI

MOVING AVERAGE LENGTH	ACCURACY MEASURES		
	MAPE	MAD	MSD
2	2.16579	0.08611	0.02367
3	1.94564	0.07765	0.01954
4	1.95072	0.07800	0.02080
5	1.98274	0.07940	0.02318
6	2.07654	0.08279	0.02447
7	2.14381	0.08523	0.02575
8	2.14311	0.08491	0.02703
9	2.19162	0.08667	0.02810
10	2.18081	0.086312	0.02857



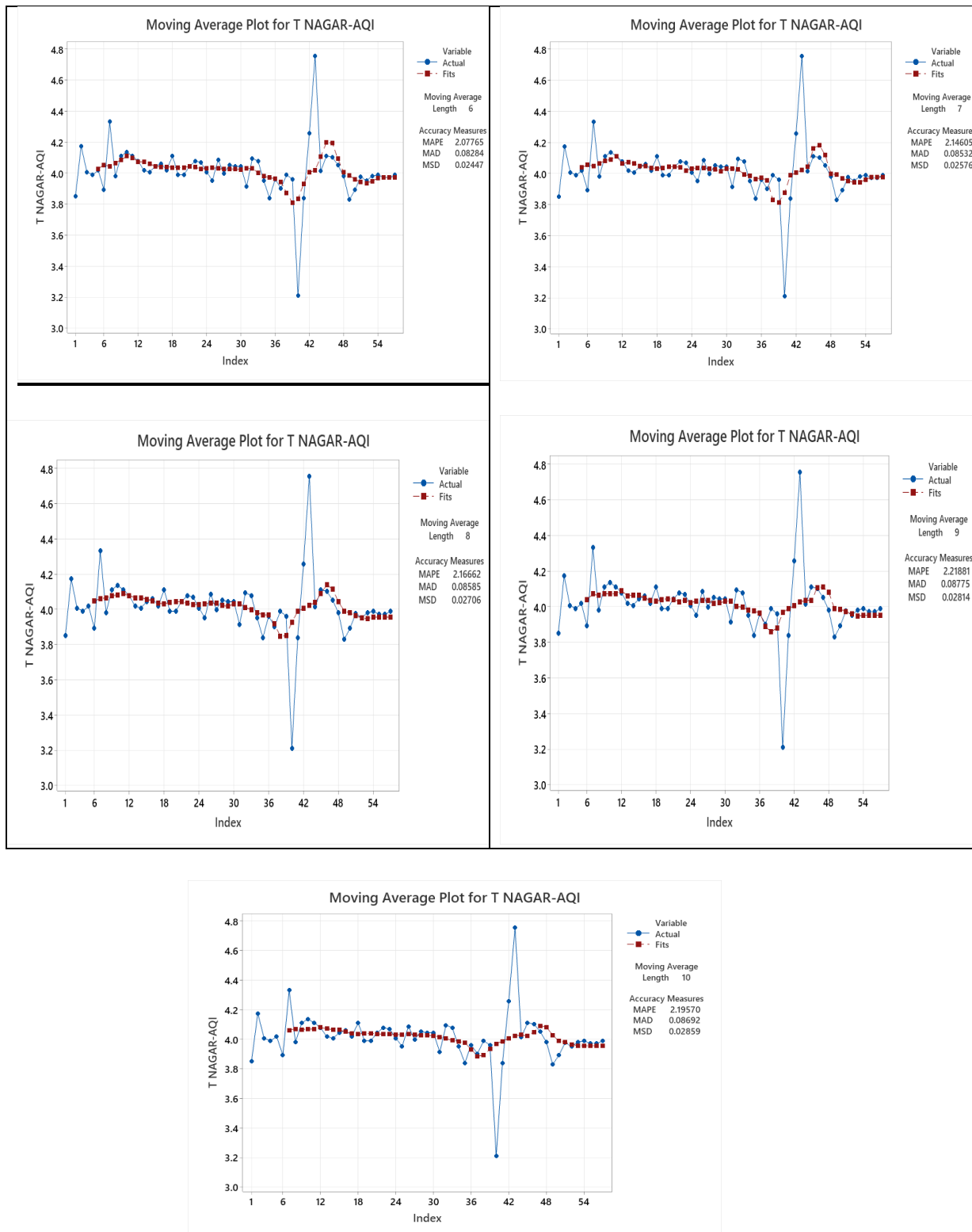
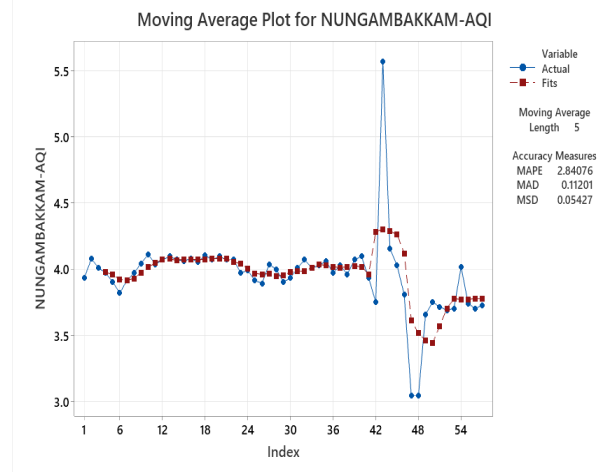
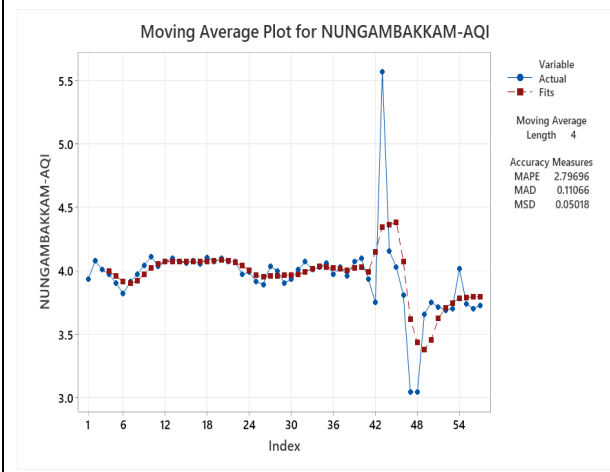
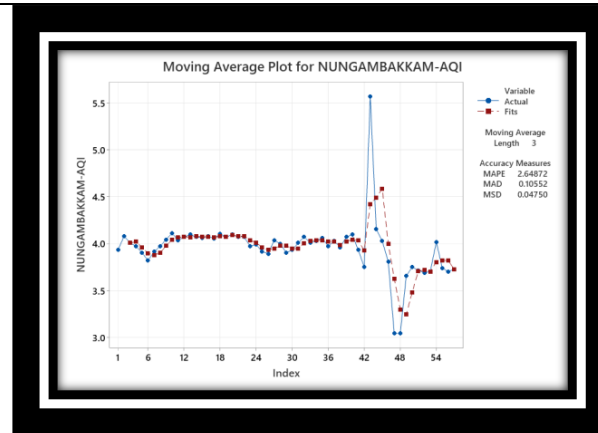
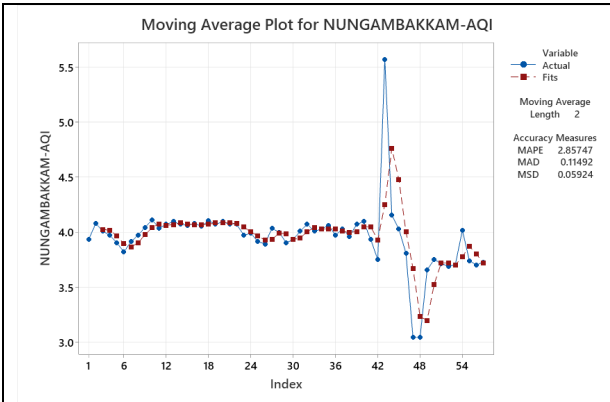


Fig. 5.6. Graph's of T-Nagar AQI with different k values and best fit is highlighted.

The above Table 5.6. and Fig. 5.6. denotes the AQI of T-Nagar measured by using the moving average method with different k values, and the best fit is found by using error measurement. It is clearly seen that length 3 is the minimum when compared with all other lengths. So, the best fit among the moving average is considered as the moving average of **length 3**.

Table 5.7. Shows the Moving Average for Nungambakkam -AQI

MOVING AVERAGE LENGTH	ACCURACY MEASURES		
	MAPE	MAD	MSD
2	2.85747	0.11492	0.05924
3	2.64872	0.10552	0.04750
4	2.79696	0.11066	0.05018
5	2.84076	0.11201	0.05427
6	2.91360	0.11395	0.05989
7	2.96243	0.11519	0.06557
8	2.97723	0.11559	0.07016
9	3.10216	0.12047	0.07414
10	3.23895	0.12618	0.07927



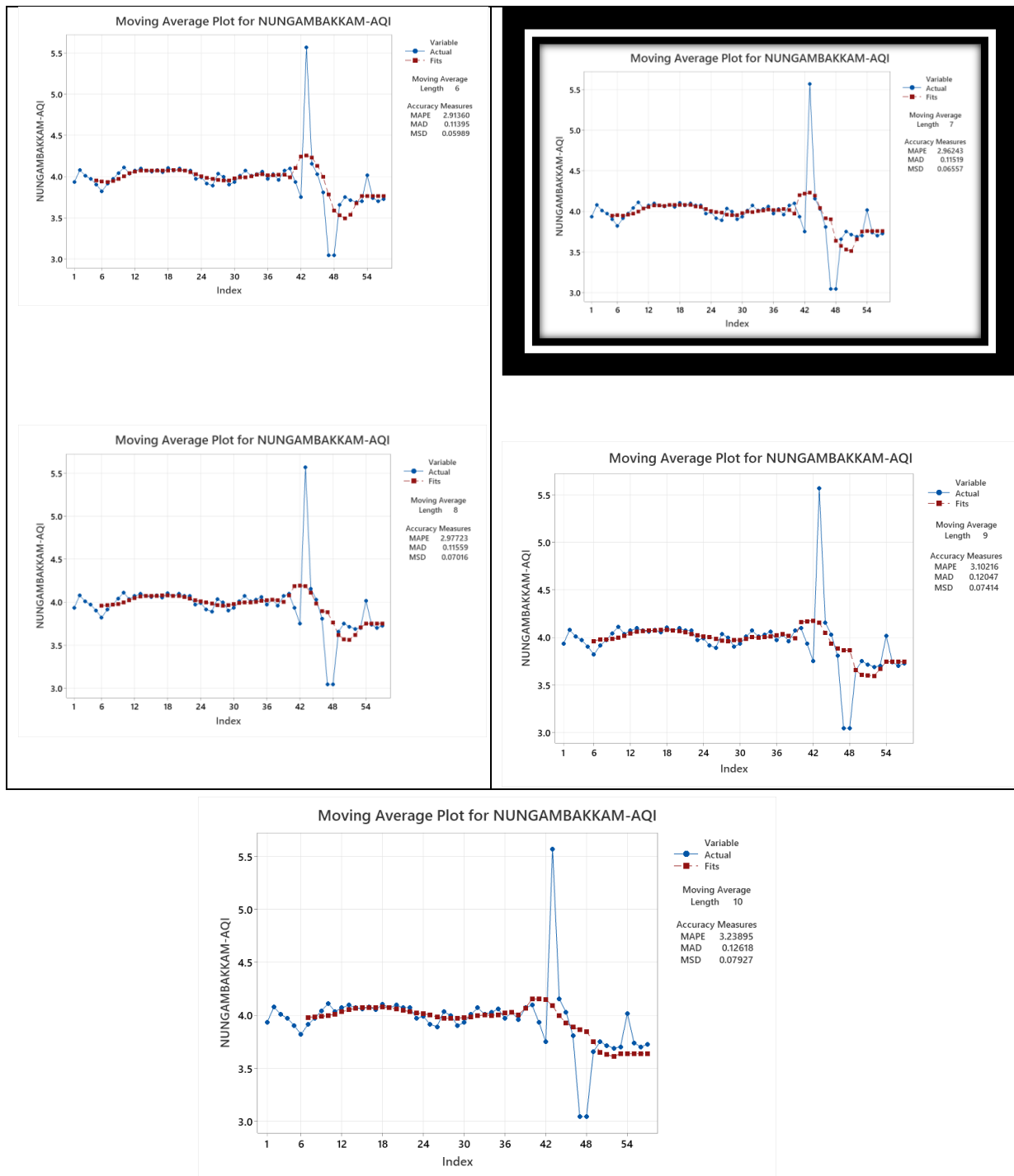
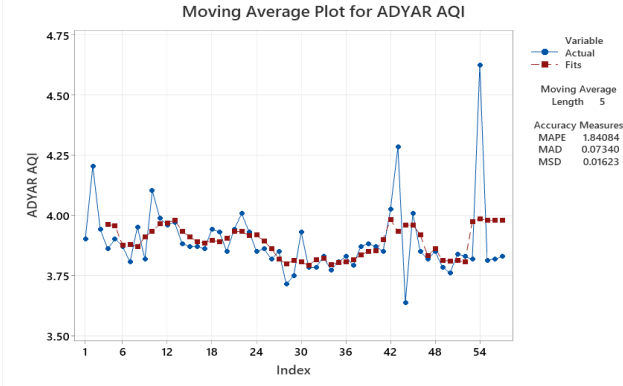
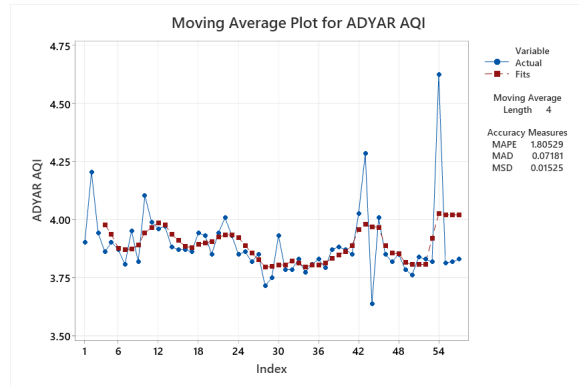
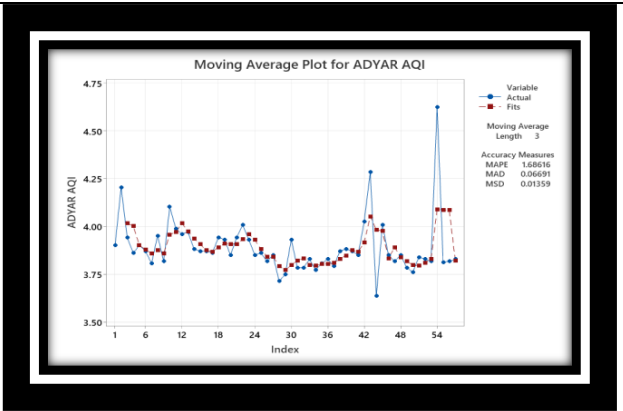
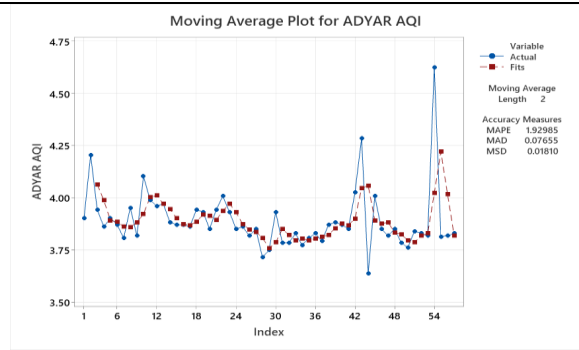


Fig. 5.7. Graph's of Nungambakkam AQI with different k values and best fit is highlighted.

The above Table 5.7. and Fig. 5.7. denotes the AQI of Nungambakkam measured by using the moving average method with different k values, and the best fit is found by using error measurement. It is clearly seen that length 3 is the minimum when compared with all other lengths. So, the best fit among the moving average is considered as the moving average of **length 3**.

Table 5.8. Shows the Moving Average for Adyar -AQI

MOVING AVERAGE LENGTH	ACCURACY MEASURES		
	MAPE	MAD	MSD
2	1.92985	0.07655	0.01810
3	1.68616	0.06691	0.01359
4	1.80529	0.07181	0.01525
5	1.84084	0.07340	0.01623
6	1.90574	0.07609	0.01680
7	2.00182	0.07992	0.01755
8	1.96599	0.07869	0.01788
9	1.98119	0.07935	0.01825
10	2.00126	0.08022	0.01902



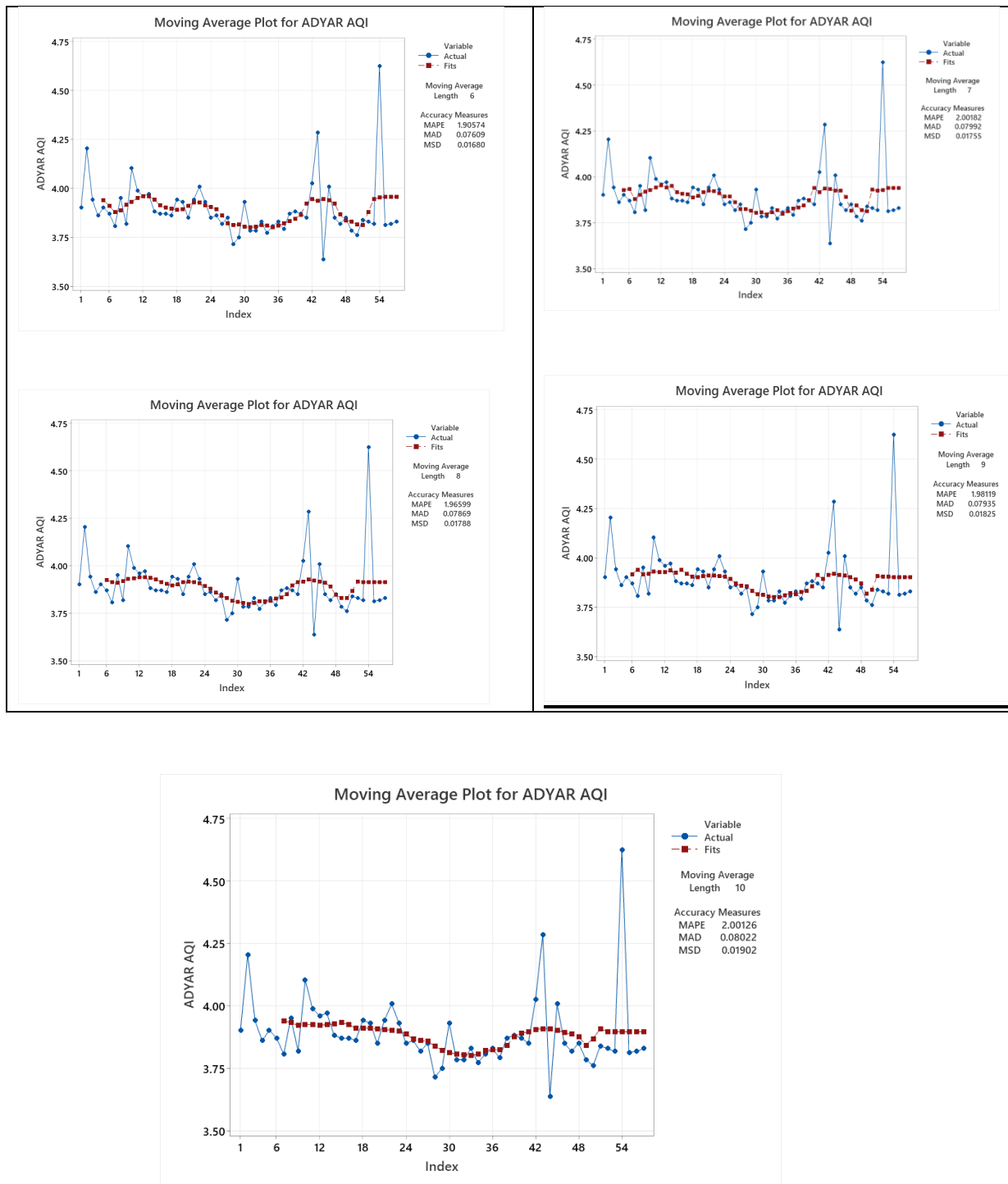


Fig. 5.8. Graph's of Adyar AQI with different k values and best fit is highlighted.

The above Table 5.8. and Fig. 5.8. denotes the AQI of Adyar measured by using the moving average method with different k values, and the best fit is found by using error measurement. It is clearly seen that length 3 is the minimum when compared with all other lengths. So, the best fit among the moving average is considered as the moving average of **length 3**.

Table 5.9. Best Error Measure of 3 different area of Chennai AQI

AREA	MOVING AVERAGE LENGTH	ACCURACY MEASURES		
		MAPE	MAD	MSD
T-Nagar	3	1.94564	0.07765	0.01954
Nungambakkam	3	2.64872	0.10552	0.04750
Adyar	3	1.68616	0.06691	0.01359

The above Table 5.9. illustrates the best error measure (MAPE,MAD,MSD) of 3 different area of Chennai AQI measured by using moving average of **length 3**.

CHAPTER-6

6.1 CONCLUSION:

This study considered daily data of Air Quality Index(AQI) of Chennai for the period of 1st Jan 2022 to 4th Feb 2023 to compare the forecast techniques and to find the most suitable forecasting model. The air quality index (AQI) calculated using the moving average method can provide a useful representation of air quality trends over time. Here, the accuracy of the AQI calculated using the moving average method will depend on the length of the moving average window and the frequency of the data points. Choosing an appropriate window length and data frequency can help to ensure that the AQI accurately reflects the current air quality conditions. Overall, the moving average method can be a useful tool for calculating the AQI and tracking air quality trends over time.

6.2 FUTURE WORK:

In future this work can be extended by comparing the results obtained using the moving average method with other methods for calculating the AQI, such as the maximum concentration approach. This could help to identify any differences or discrepancies in the AQI values obtained using different methods, and could provide insight into the strengths and limitations of each approach. Overall, continued research into the best methods for calculating the AQI can help to improve our understanding of air quality trends and support efforts to protect public health and the environment.

REFERENCES:

- [1] Atoui, A., Slim, K., Andaloussi, S. A., Moilleron, R., & Khraibani, Z. (2022). Time Series Analysis and Forecasting of the Air Quality Index of Atmospheric Air Pollutants in Zahleh, Lebanon. *Atmospheric and Climate Sciences*, 12(4), 728-749.
- [2] Chris Chatfield, The Analysis of Time Series: An Introduction, 6th edition. (2003).
- [3] Dragomir, E. G. (2010). Air quality index prediction using K-nearest neighbor technique. *Bulletin of PG University of Ploiesti, Series Mathematics, Informatics, Physics, LXII*, 1(2010), 103-108.
- [4] George E P Box, Gwilym M Jenkins, Gregory C Reinsel. *Time Series Analysis: Forecasting and control*, 5th edition. (2015)
- [5] Marinov, E., Petrova-Antonova, D., & Malinov, S. (2022). Time Series Forecasting of Air Quality: A Case Study of Sofia City. *Atmosphere*, 13(5), 788.
- [6] Nadeem, I., Ilyas, A. M., & Uduman, P. S. (2020). Analyzing and forecasting ambient air quality of Chennai city in India. *Geography, Environment, Sustainability*, 13(3), 13-21.
- [7] Naveen, V., & Anu, N. (2017). Time series analysis to forecast air quality indices in Thiruvananthapuram District, Kerala, India. *International Journal of Engineering Research and Application*, 7(6), 66-84.
- [8] Sharma, N., Taneja, S., Sagar, V., & Bhatt, A. (2018). Forecasting air pollution load in Delhi using data analysis tools. *Procedia computer science*, 132, 1077-1085.
- [9] Shrivallabha S., Kumaresh P. Nelavigi, (2019). Time Series Analysis of Atmospheric Particulate Matter of Bengaluru City. *International Journal of Scientific Research in Mathematical and Statistical Sciences*, 6(5), 83-85.