**AIR POLLUTION ANALYSIS USING ARIMA MODEL**

**A MINOR PROJECT REPORT**

***Submitted by***

**KESHAV ADKAR [RA1911033010117]**

**SATVIK SRIVASTAVA [RA1911033010155]**

*Under the guidance of*

# Dr. G. Sivashankar

(Associate Professor)

***in partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE & ENGINEERING**

of

**FACULTY OF ENGINEERING AND TECHNOLOGY**



S.R.M.Nagar, Kattankulathur, Chengalpattu District

**NOVEMBER 2022**

# SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section3 of UGC Act,1956)

## BONAFIDE CERTIFICATE

Certified that 18CSP107L minor project report [18CSP108L internship report] titled “AIR POLLUTION ANALYSIS USING ARIMA MODEL” is the bonafide work of “**SATVIK SRIVASTAVA [RA1911033010155], KESHAV ADKAR [RA1911033010117]”** who carried out the minor project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**SIGNATURE**

Dr. G. Shivasankar

**GUIDE**

Associate Professor

SRM Institute of Science and

Technology

**SIGNATURE**

Dr. Annie Uthra

**HEAD OF THE DEPARTMENT**

Professor

Dept.of Computing Technologies

Signature of the

Panel Head

Dr. G. Senthilkumar

Associate Professor

SRM Institute of Science and Technology

## ABSTRACT

Air quality is a significant issue that has long had an impact on ecosystems, plants, animals, and human health. Building, demolition, burning of plastics, and vehicular pollution are the main causes of bad air quality. A model must be created in order to periodically analyse the trends in air quality on a monthly and even seasonal basis. Time series models for analysing future air quality are discussed in this work. These models are used to project and forecast India's monthly air quality. The study analyses the air quality using time series and projects future values using and assist in enhancing or at the very least limiting the deteriorating air quality using the ARIMA model. The ARIMA model, also known as the autoregressive integrated moving average, may analyse and represent both stationary and non-stationary time series. Data on air pollution are regularly analysed using time series analysis. The ARIMA model produces satisfying and accurate results. It will help in the information gathering process so that quick action can be made to watch over and manage the situation before it becomes worse. Performance can be assessed by computing mean square error, mean absolute error, and root mean square error.

Descriptive analysis has been used to study the trends of various air pollutants like Sulphur dioxide (SO2), Nitrogen dioxide (NO2), suspended particulate matter (PM), Ozone (O3) Carbon monoxide (CO), Benzene, Toluene and Xylene (BTX) at daily levels of various stations across multiple cities in India.

**ACKNOWLEDGMENT**

We express our humble gratitude to **Dr C.Muthumizhchelvan**, Vice-Chancellor , SRM IST , for the facilities extended for the project work and his continued support.

We extend our sincere thanks to our Head of Department, **Dr Annie Uthra** Professor , Department of Computational Intelligence, SRMIST, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our Panel Head, Faculty Advisor and Project Supervisor **Dr G. Senthil Kumar**, Associate Professor , Department of Computational Intelligence, SRMIT, for his inputs during the project reviews and support for leading and helping us to complete our project. We also thank him for providing us with an opportunity to pursue my project under his mentorship. He provided us with the freedom and support to explore the research topics of my interest. His passion for solving problems and making a difference and making a difference in the world has always been inspiring.

We sincerely thank the CINTEL Department staff and students , SRM Institute of Science and Technology, for their help during the project.

Dr G. Shivasankar

Dr G. Senthilkumar

## TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **TITLE** | **PAGE No.** |
| 1 | ABSTRACT | 2 |
| 2 | ACKNOWLEDGEMENT |  |
| 3 | TABLE OF CONTENTS | 3 |
| 4 | LIST OF FIGURES | 5 |
| 5 | LIST OF TABLES | 6 |
| 6 | LIST OF ABBREVIATIONS | 7 |
| **1** | **INTRODUCTION**  1.1 Brief Description about  Project | 8 |
| **2** | **PROPOSED METHODOLOGY** |  |
| **3** | **CODING**  3.1 Code Snippets  3.2 Data Visualization |  |

|  |  |  |
| --- | --- | --- |
| **4** | **CONCLUSION** |  |
| **5** | **REFERENCES** |  |

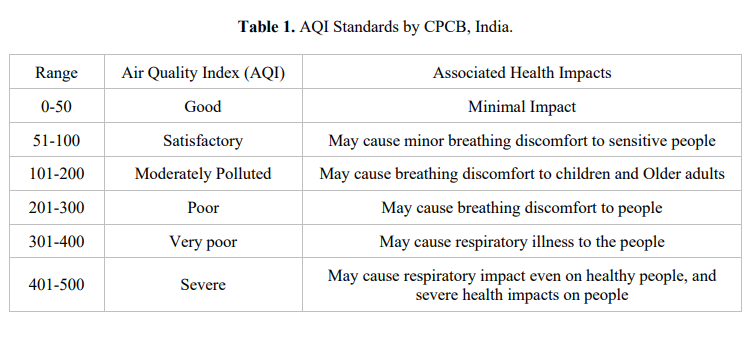
**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **S NO.** | **IMAGE TITLE** | **PAGE NUMBER** |
| 1 | Table 1 AQI Standards | 7 |
| 2 | UML Diagram | 8 |
| 3 | Block Diagram | 9 |
| 4 | Table 2 | 15 |
| 5 | Convolution Demonstration #1 | 16 |
| 6 | Convolution Demonstration #2 | 16 |
| 7 | Convolution Demonstration #3 | 17 |
| 8 | Pooling Demonstration | 18 |
| 9 | Fully Connected Layer | 19 |

**CHAPTER 1**

## INTRODUCTION

* Different types of pollution exist in the universe. One among them is the Air pollution. It affects the health conditions of the human beings, plants, animals and also affects the entire ecosystem.
* Air quality can be defined as the state of air around us. The release of gases from the vehicles, wastes released from the industries pose a great threat for the health factors and also contributes much for the pollution of air. As a result, the quality of air deteriorates much in the cities than in the rural areas .
* Because of the adverse effects of air pollution throughout the world this topic is of prime concern for most of the researchers Increasing and higher concentrations of air pollution can be life intimidating and thereby air pollution became one of the major issue in the world.

****

**CHAPTER 2**

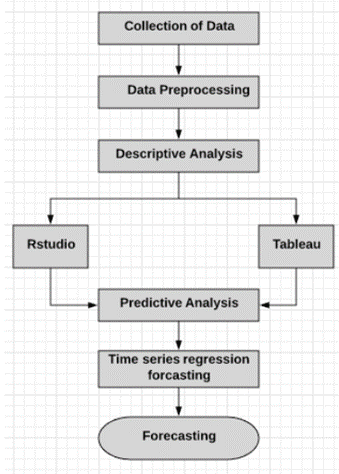
## Proposed Methodology

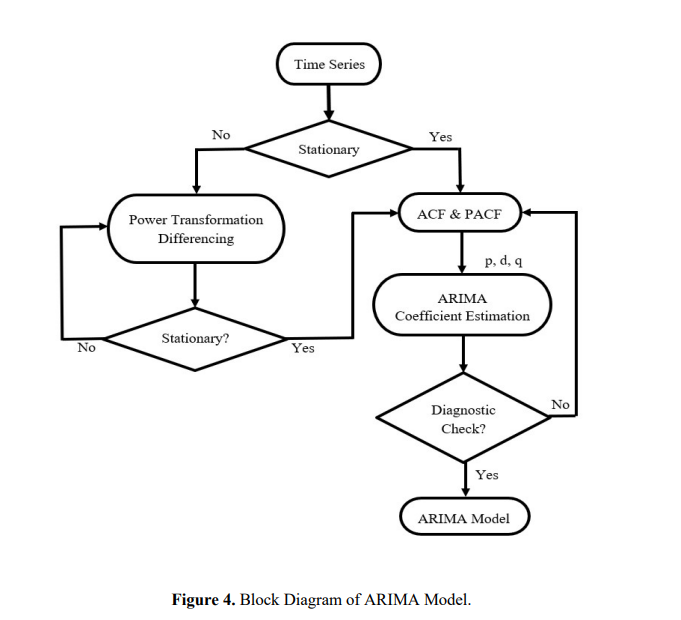
**1. Activity Diagram:**

An activity diagram can be used to show the steps involved in a business or software process. One or more people, pieces of machinery, or pieces of software could carry these out. Activity diagrams are used to document the system's activities, including use cases, business processes, and the actual implementation of those activities.

The following processes are best described with activity diagrams:

* Case studies and the instructions they provide
* processes in business between users and systems,
* The allowable sequence of component interactions in analysis
* software programmes





**SCOPE OF WORK**

MATERIAL USED

1. Arrangement of an Index

There are two main steps in the creation of AQI:

1.a. Creating a subindex for each pollutant

To obtain the AQI, subindices are added up.

In order to create subindices (I1, I2..., In) for n pollutant variables (X1, X2.., Xn), subindex functions that depend on air quality standards are used. Equation (1) states the representation mathematically as II! = ff(XX! ), where ii = 1, 2,..., nn (1)

The subindex also reflects the close connection between pollutant concentrations and health impacts.

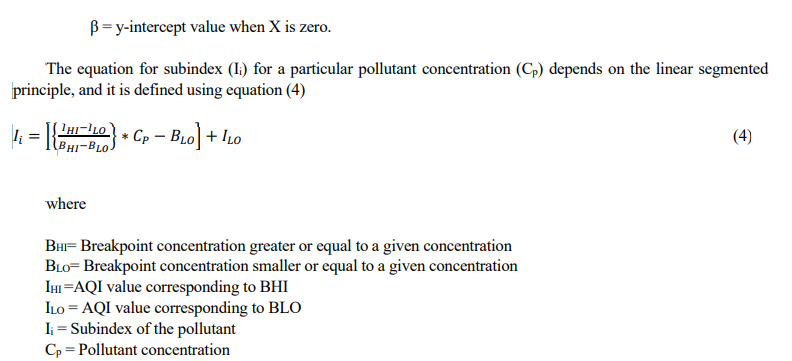
In part 2.2 of this work, the functional link between the subindex value (Ii) and the pollutant variables (Xi) is carefully elucidated. The overall Index (I), also known as the AQI, is then obtained by adding the subindices Ii together using the function (FF) in equation (2).

I = F(I1,I2,…..In)

2. Calculation of Subindices

The subindex function indicates that subindex Ii and pollutant variable Xi have a close association. It aims to demonstrate the connection between environmental effects and the concentration of particular contaminants. Additionally, it can be segmented linear, linear or nonlinear. The correlation between the corresponding subindex Ii and the pollutant concentration Xi is shown in equation .

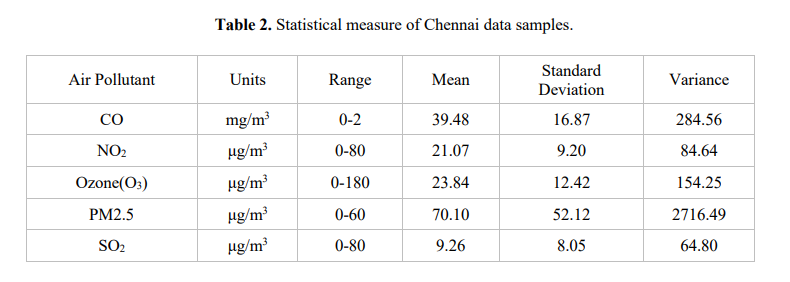


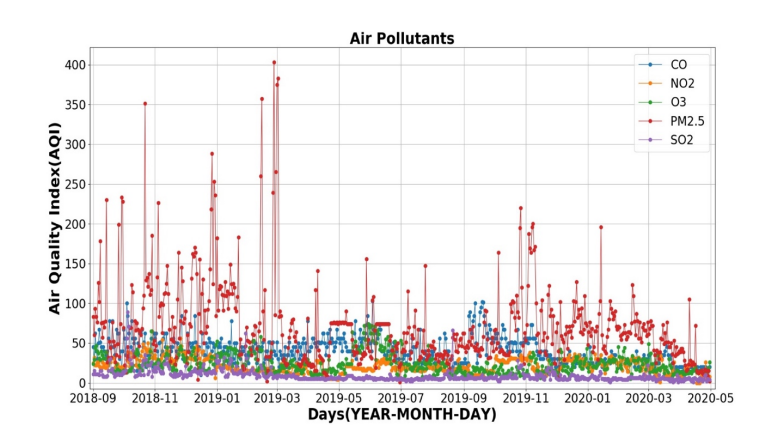


1. Data Preparation

The monitoring and forecasting of air pollution in this study is done using data from Chennai, India.

One of South India's most significant centres for culture, commerce, and education is Chennai. With a population of roughly 10 million, it is one of the most populated cities in India. In and around Chennai, the main air pollutants are NO2, CO, O3, PM2.5, PM10, and SO2. Table 2 lists the air contaminants that were employed in this experiment. Figure 1 depicts the air pollution plot. The information was gathered between September 2018 and April 2020 from the CPCB website.

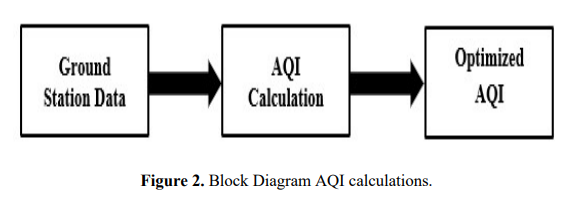




The data set has 607 values total, which are divided into 122 (20%) values for testing prescient models and 486 (80%) values for training prescient models. The suggested strategy examines the data using the Multilinear Regression (MLR) model linearity exists between the optimum AQI and the air contaminants. With the use of the MLR model, AQI predictions are made. For AQI forecasting, the time series analysis-ARIMA model is employed. The error-indexes Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and cross-validation are used to evaluate the performance of each model test.

1. Data Pre-processing

According to CPCB (CPCB, 2019), India, each air pollutant has its own units and subindices. Preprocessing is required before using data collected from various ground stations. So, prior to using machine learning (ML) techniques for the prediction, the raw data was filtered. Table 2 lists the air pollution data gathered from CPCB and their standard ranges. The CPCB website's criteria were used to convert the specific air contaminants to their appropriate AQI. There are 607 air pollutant values in the data set, which takes one reading per day into account. 8% of the data set's values were NANs. The missing values in the data set are imputed using the statistical measure of mean value.



Time series are used in this study to analyse API performance. The Box-Jenkins model serves as the foundation for this study's time series technique. using the ARIMA Box-Jenkins technique (Autoregressive Integrated Moving Average) The model's effectiveness in the forecasting industry has led many academics to still employ it in a variety of fields of research today (Wang and Lu, 2006; Ibrahim et al., 2009; Kumar and Jain, 2010).

The optimum pollution forecasting model was determined in the current investigation using the Box-Jenkins model. For time series forecasting employing auto-regression (AR), differencing, and moving averages, the Box-Jenkins model is used mathematically (MA). To forecast time series data for a single variable, use the model. It first tests for stationery and seasonality before determining the AR and MA parameter. It uses the differencing method to transform non-stationary and produces the ARIMA (Auto Regressive Integrated Moving Average) model. The standard abbreviation for an ARIMA model is ARIMA(p,d,q), where p is the total number of auto-regressive components, d is the order of differencing, and q is the total number of moving average terms.

This model can be used with non-stationary data. If non-stationary data are present, it can be determined using a variety of statistical techniques. Differencing is one method. Differentiation is employed to stabilise the data series before the ARIMA model is applied. After determining the AR and MA parameters using differencing techniques, it makes use of a particular model. Two distinct models based on seasonal effects are the ARIMA and SARIMA models. Apply seasonal ARIMA to forecast values for noteworthy holidays.

**2 Box Jenkins Modelling**

The Box-Jenkins is called after its discoverers, Box and Jenkins (1976). Due to the usage of linear models, this method may produce both stationary and non-stationary timeseries. Researchers typically employ this strategy to forecast univariate time series data. Forecasting involves the use of Box-Jenkins methods, which include Moving Average (MA), Integrated (I), and Autoregressive (AR) models (Cryer and Chan, 2008).

When creating a model using the Box-Jenkins methodology, four steps must be taken into consideration: tentative identification, parameter estimation, diagnostic evaluation, and prediction. This step is essential for selecting the best ARIMA model for time series data (Hanke and Wichern, 2008)

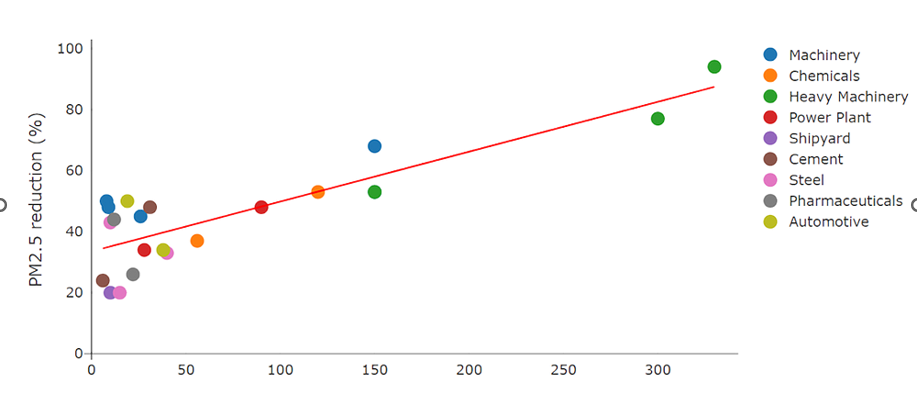
Autoregressive (AR), Moving Average (MA) and Autoregressive Integrated Moving Average (ARIMA) Models: Autoregressive (AR) model is suitable for stationary time series data patterns. A pth-order of autoregressive or AR (p) model can be written in the form

Y t = m0 + m1Yt-1 + m2Yt-2 +……+ mpYt-p + t εt

The current value of the series yt is a linear combination of the p most recent values of itself. The coefficient m0 is related to the constant level of series.

For AR models, forecast depend on observed values in previous time periods. Meanwhile, the dependent variable yt of Moving Average (MA) depends on previous values of the errors rather than on the variable itself. MA models provide forecasts of yt based on linear combination of a finite number of past errors. The errors involved in this linear combination move forward as well.A moving average with qth-order or MA (q) model takes the form

Y = µ + εt – θ1εt-1 – θ2εt-2 -……- θqεt-q



Moving average and autoregressive terms are combined to create an autoregressive moving average model (ARMA). The notation for this model is ARMA (p, q), where p denotes the order of the autoregressive part and q denotes the order of the moving average part.

The ARMA (p, q) is in the form below

Y t = m0 + m1Yt-1 + m2Yt-2 +……+ mpYt-p + εt – θ1εt-1 – θ2εt-2 -……- θqεt-q

ARMA models can be used to explain a wide range of stationary time series behaviour. Because ARMA combines AR and MA, forecasting depends on both the present and historical values of the response Y as well as the present and historical values of the residuals.

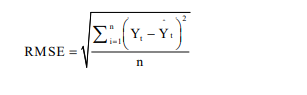
Since the data did not move around a fixed level or mean, the set of data could represent a non-stationary time series data pattern. By taking the difference, the data can be made stationary in one way. Consequently, it is said that the series of data supplied generally as yt after difference follows an integrated autoregressive moving average model, or ARIMA (p, d, q). Normally, the difference would be one or two (d2) for practical purposes.

**SARIMA:** The Box-Jenkins modelling and forecasting approach provides an advantage when analysing seasonal time series data. The model is known as a seasonal ARIMA model or SARIMA model in this situation since it has seasonal components.

The model can be abbreviated as SARIMA (p, d, q) (P, D, Q)S where the lowercase for non-seasonal part meanwhile the uppercase for seasonal part.

**Measure of Accuracy:**

For identification of the models performance, the criteria chosen are the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), the Mean Square Error (MSE) and the Root Mean Square Error (RMSE). Given as:



Where

Yt = The actual value at time t

Y^ = The fitted value at time t

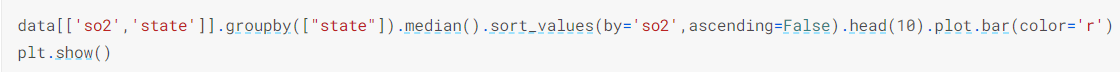
n = The number of observations

The optimal model to utilise for predicting is the one with the lowest MAE, MAPE, MSE, and RMSE values.

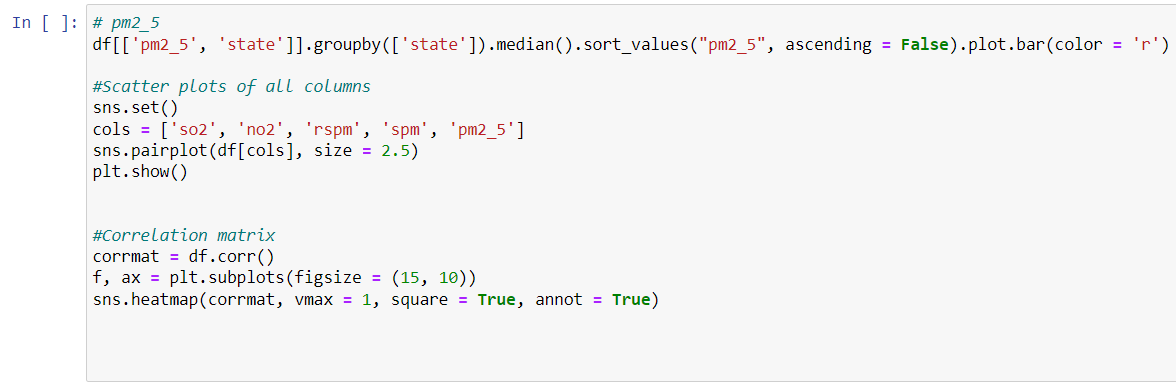
**CHAPTER 3**

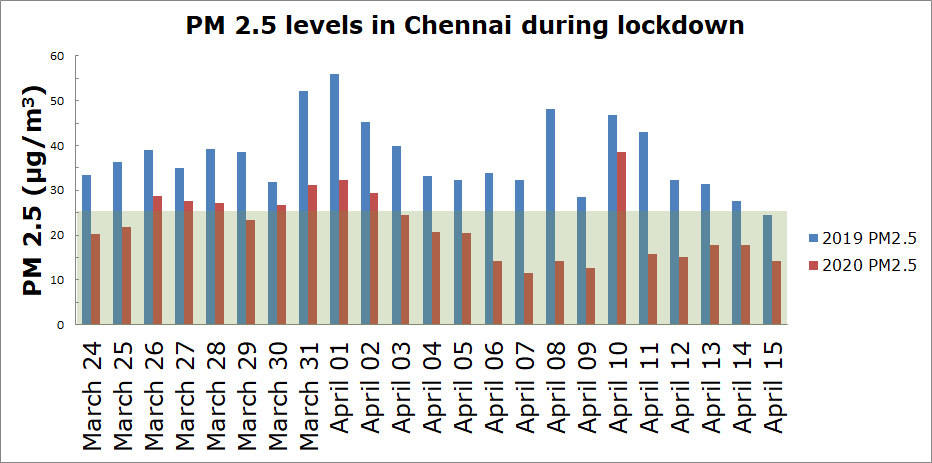
## CODING



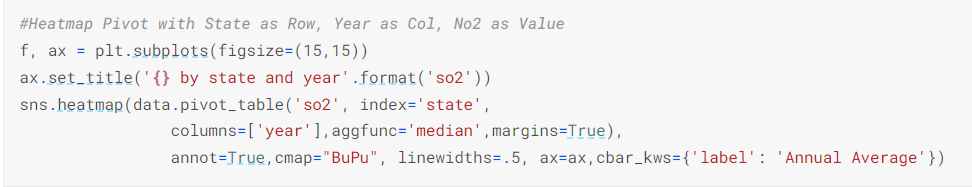




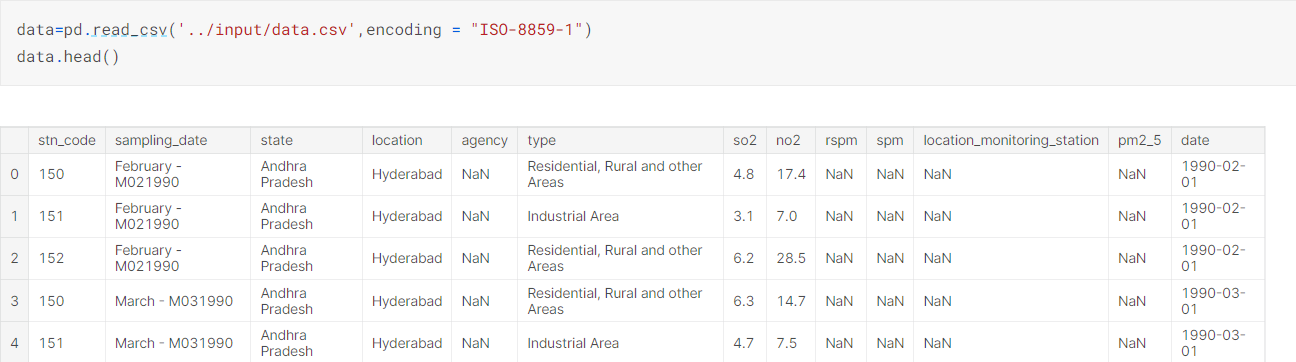




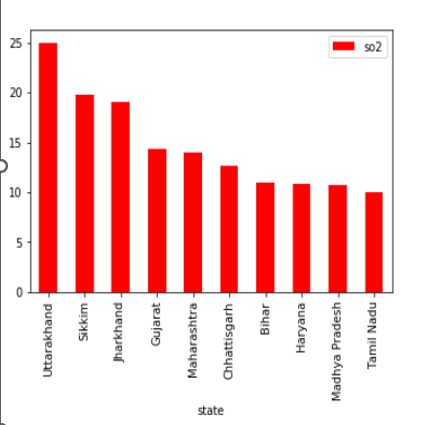


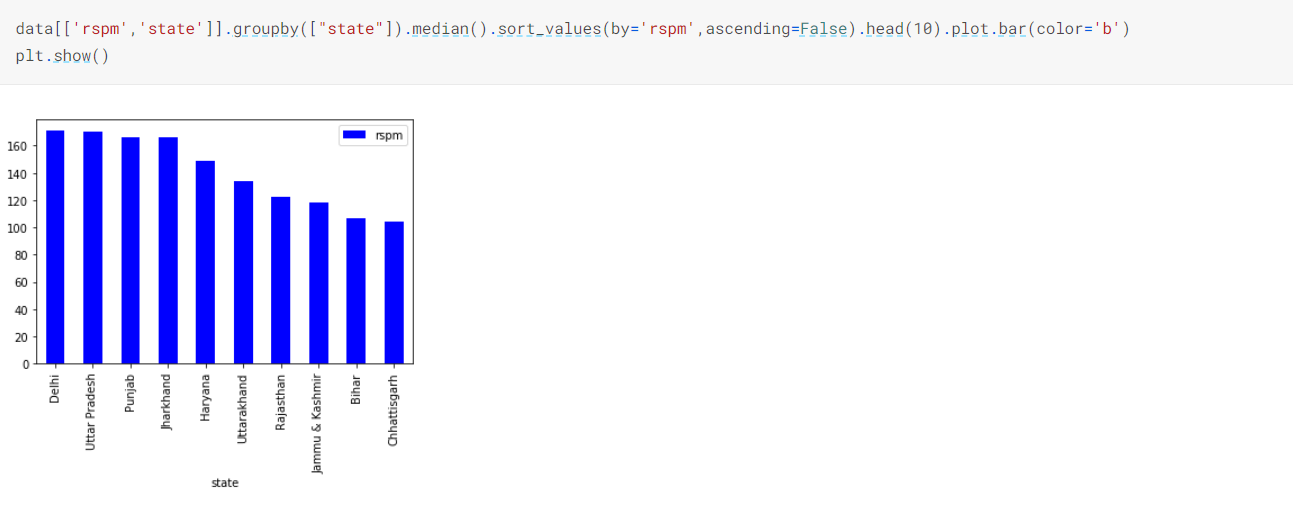


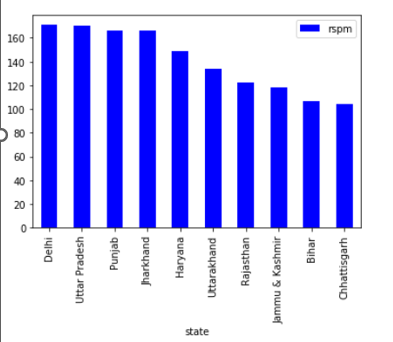
GRAPH :

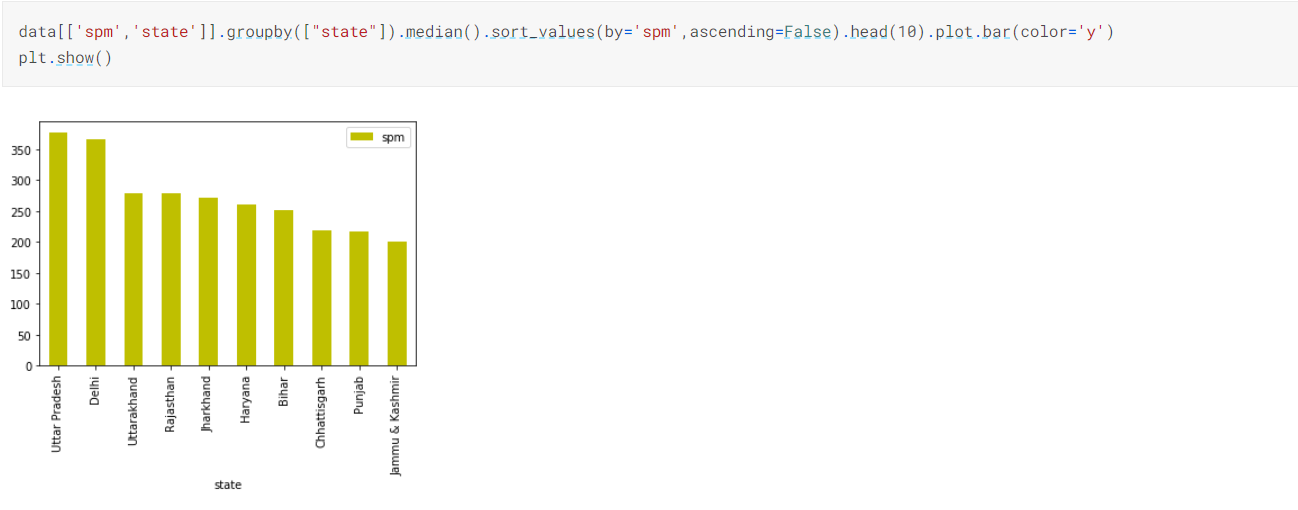


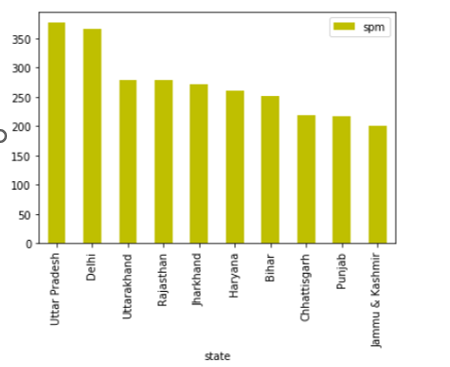


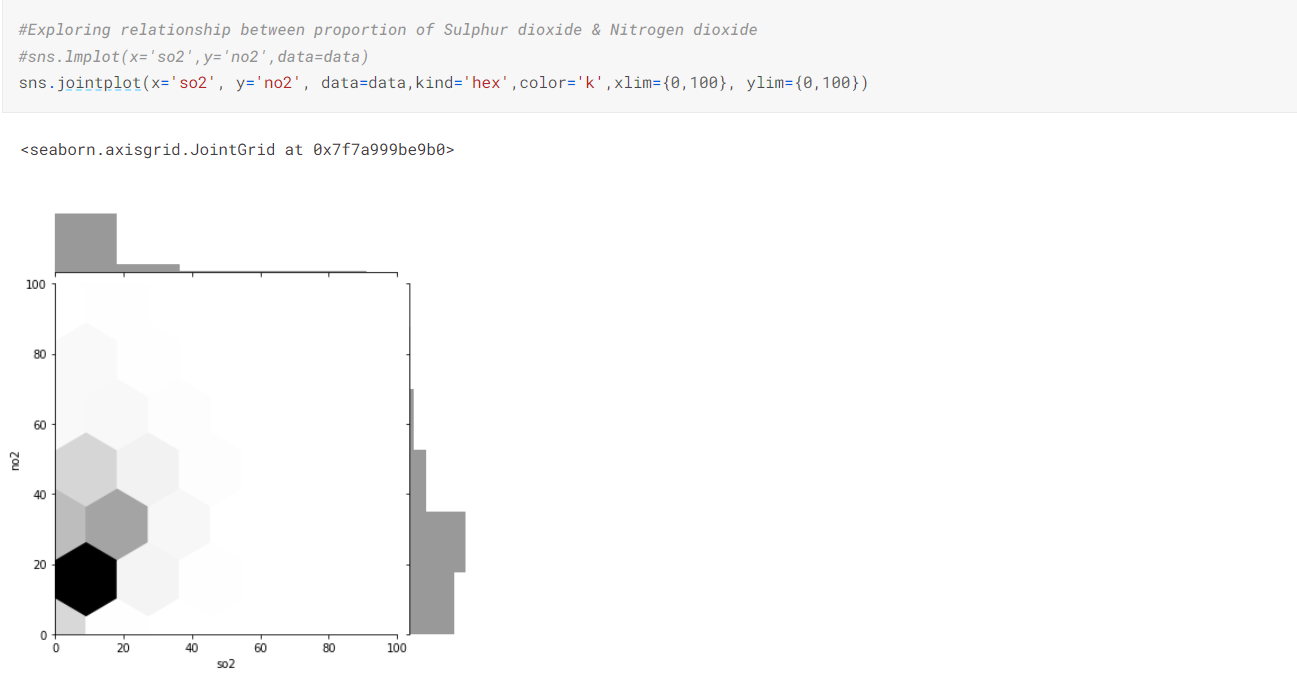


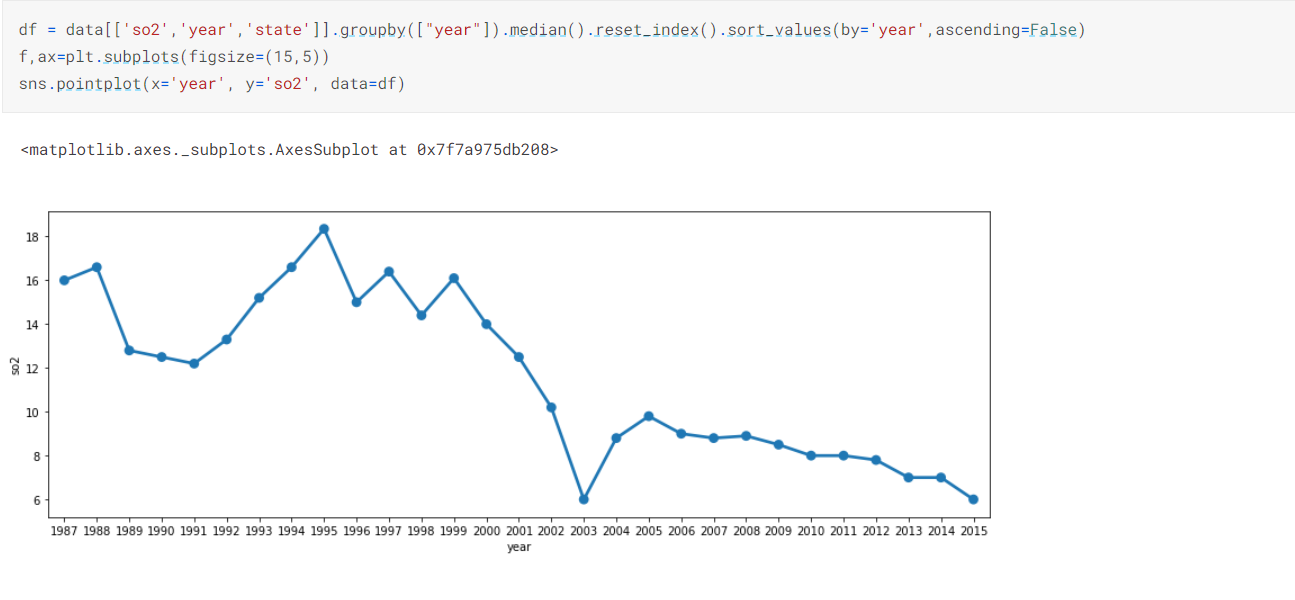


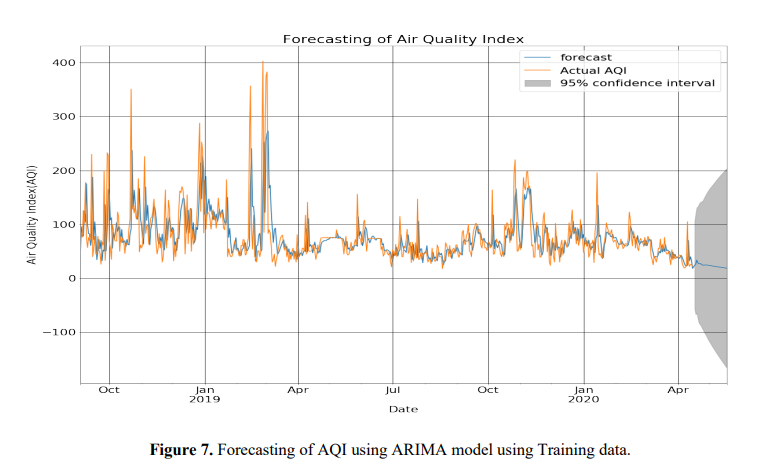














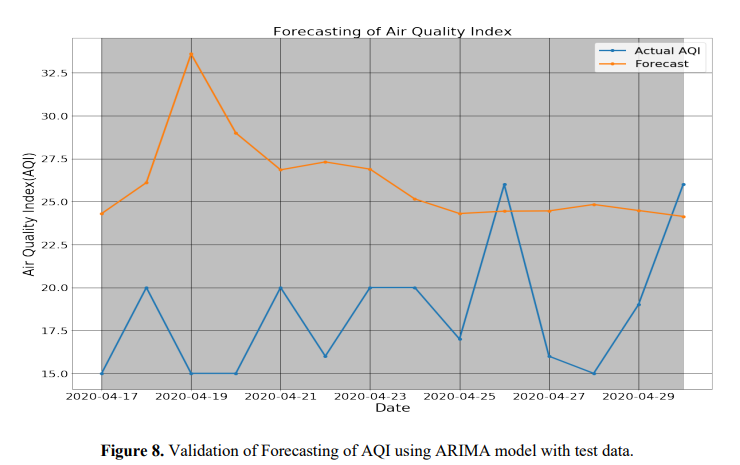
**CHAPTER 4**

## CONCLUSION

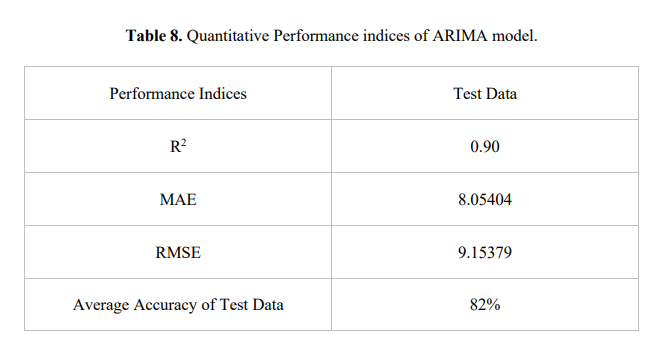
The forecasting and prediction of AQI protects people, the environment, and both from harmful health issues brought on by low AQI. Data on air pollution in the Chennai region is gathered between September 2018 and March 2020 in order to evaluate and validate the created machine learning models. The MLR model is utilised in this study to forecast the AQI.

Following the discovery of the association between air pollutants, the MLR model was trained using the training data set and verified using the test data that had not yet been observed. For both training and test data, it was determined that performance metrics including ask-fold cross-validation R2, MAE, and RMSE were sufficient. Second, the forecasting of AQI is done using the ARIMA model, which has been developed. First, training data are used to train the ARIMA model, and test data are used to evaluate it.

It was determined that the performance indices for the ARIMA model using test data were satisfactory. For AQI prediction, both linear and nonlinear machine learning techniques were examined. Dataset is linear in nature. Therefore, the linear ML method offered the best fit. The dataset is linear and not complicated, however, therefore nonlinear ML approaches provide a poor fit. Due to their complexity, neural networks are difficult to handle. Additionally, it is not appropriate for short-term real-time data changes. The technique should be quick to display the results of forecasting and easy to execute on a hardware platform with little computational overhead. Using data up to April 2020, the ARIMA model projected future values for the following 15 days within a 95% confidence interval. As a result, it is advised to use the suggested methodology to estimate AQI, and AQI forecasting can eventually be implemented in real time.



ACCURACY OF ARIMA MODEL :



**CHAPTER 5**

## REFERENCES

1. Curtis, L.; Rea, W.; Smith-Willis, P.; Fenyves, E.; Pan, Y. Adverse health effects of outdoor air pollutants. Environ. Int. 2006, 32, 815–830.
2. Mayer, H. Air pollution in cities. Atmos. Environ. 1999, 33, 4029–4037.
3. Samet, J.M.; Zeger, S.L.; Dominici, F.; Curriero, F.; Coursac, I.; Dockery, D.W.; Schwartz, J.; Zanobetti, A. The national morbidity, mortality, and air pollution study. Part II: Morbidity and mortality from air pollution in the United States. Res. Rep. Health Eff. Inst. 2000, 94, 5–79.
4. Dockery, D.W.; Schwartz, J.; Spengler, J.D. Air pollution and daily mortality: Associations with particulates and acid aerosols. Environ. Res. 1992, 59, 362–373.
5. Schwartz, J.; Dockery, D.W. Increased mortality in Philadelphia associated with daily air pollution concentrations. Am. Rev. Respir. Dis. 1992, 145, 600–604.
6. American Lung Association. State of the Air Report; ALA: New York, NY, USA, 2007; pp. 19–27.
7. Hinds, W.C. Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles; John Wiley & Sons: Hoboken, NJ, USA, 2012.
8. Soukup, J.M.; Becker, S. Human alveolar macrophage responses to air pollution particulates are associated with insoluble components of coarse material, including particulate endotoxin. Toxicol. Appl. Pharmacol. 2001, 171, 20–26.
9. Environmental Protection Agency (EPA). CFR Parts 50, 51, 52, 53, and 58-National Ambient Air Quality Standards for Particulate Matter: Final Rule. Fed. Regist. 2013, 78, 3086–3286.
10. Schwartz, J. Short term fluctuations in air pollution and hospital admissions of the elderly for respiratory disease. Thorax 1995, 50, 531–538.