DEVELOPMENT OF TIME SERIES MODEL FOR FORECASTING OF AEROSOL OPTICAL DEPTH (AOD) OVER MINING REGIONS

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

BACHELOR OF TECHNOLOGY

IN

MINING ENGINEERING

SUBMITTED BY

Rohit Kumar Singh

113MN0492

UNDER THE GUIDANCE OF

Dr. AMIT KUMAR GORAI

DEPARTMENT OF MINING ENGINEEIRNG

NATIONAL INSTITUTE OF TECHNOLOGY

ROURKELA-769008

NATIONAL INSTITUTE OF TECHNOLOGY-ROURKELA

CERTIFICATE

This is to certify that the project report titled, "DEVELOPMENT OF TIME SERIES MODEL FOR FORECASTING OF AEROSOL OPTICAL DEPTH (AOD) OVER MINING REGIONS" submitted by Rohit Kumar Singh (Roll No. 113MN0492) in partial fulfilments for the requirements for the award of the degree Bachelor of Technology in Mining Engineering during Session 2013-2017 at National Institute of Technology, Rourkela and is an authentic work carried out by him under my supervision and guidance.

(AMIT KUMAR GORAI)

Department of Mining Engineering

National Institute of Technology, Rourkela

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude and respect to my supervisor Dr. Amit Kumar Gorai for his excellent guidance, suggestions, and support and I consider myself extremely lucky to be able to work under the guidance of such a dynamic personality. I would like to render heartiest thanks to my friend whosoever helping nature and suggestion has helped us to complete this present work. I would like to thank all faculty members and staff of the Department of Mining Engineering, N.I.T. Rourkela for their extreme help throughout course. An assemblage of this nature could never have been attempted without reference to and inspiration from the works of others whose details are mentioned in the reference section. We acknowledge our indebtedness to all of them.

Date: Rohit Kumar Singh

Abstract

Air pollution is always a concern of modern world, and in order to keep air pollution level under check it will be helpful if we could understand the nature of air pollution, pollutants and how does different air pollutants interact. Different causal or deterministic relationship behind different factors of air pollution like the relationship between AOD values and dust particles concentration in the atmosphere etc. can be understood by air pollution modelling. In air pollution modelling, we use mathematical theory to understand or predict air pollutants concentrations and the way they behave in the atmosphere. Also, as computational power and amount of collected data is increasing our ability to forecast and understand the nature of air pollution more accurately has enhanced. So, there has been considerable research going on for air pollution modelling methods.

In the present study, time series models were developed for forecasting of AOD level using various techniques like Auto Regressive Integrated Moving Average (ARIMA), Auto Regressive Moving Average (ARMA), and Markov Chain Monte Carlo (MCMC) method. A specific time series modelling method is always suitable for a specific condition like ARIMA is used for non-stationary time series data, ARMA is suitable for stationary time series data, and MCMC method is suitable for any stochastic process. The model diagnostic was performed using Ljung-Box for validating the models. The study used time series data of Aerosol Optical Depth (AOD) observed over Talcher and Godavari Coal Field. The monthly time series data of AOD at 550 nm and 1-degree spatial resolution during January 2000 to December 2016 were downloaded from both the locations from the GIOVANNI web sources. The time series models were developed using the above monthly average data to forecast the monthly AOD levels for the year 2017-2019.

Keywords: Time Series, ARIMA, ARMA, MCMC, Aerosol Optical Depth

List of Figures

List of Tables

Contents

1. Introduction

Mining and Manufacturing are the core industries in India that play a significant role in economic development of the country. However, it degrades the environment. Most of the mining activities contribute directly or indirectly to air pollution. Major sources of air pollution in the mining regions usually come from drilling, unloading, blasting, coal handling plants, losses from exposed over burden dumps, road transport, overburden loading and unloading and workshops [1]. Air pollutants deteriorate air quality and ultimately affect vegetation, people and wild life in and nearby mining regions [2].

The impacts of poor air quality on human wellbeing and the earth can, thusly, have negative monetary effects too. Real expenses, for instance, for hospitalization and therapeutic treatment, unexpected losses, and lost work days. Harm to soils, vegetation, and conduits may diminish the profitability of our agribusiness and ranger service businesses. In urban regions, air contamination can be exorbitant when, for instance, transport is upset (because of substantial scale occasions like volcanic ejections), or consumed structures should be repaired.

Many long terms and short terms studies have been conducted centered on ambient air quality in a mining area in India on SPM and RSPM at Jharia coalfield [3], published physiochemical parameters and proximate analysis of coal dust collected from road surface [4]. Above studies have focused primarily on particulate matters (SPM, RSPM), SO_2 and NO_x. Since aerosol optical depth (AOD) has significant importance [5] in air pollution as it measures atmospheric turbidity caused by aerosols. In this project monthly long term (Jan 2000- Dec 2016) AOD forecasting using time series analysis methods over Godavari and Talcher coal mines in India have been conducted.

Here, ARIMA, ARMA, MCMC, methods have been accommodated to predict aerosol optical depth over two coal mining area in India. The time series analysis is a handy tool in decision-making, planning and operating of atmospheric variations.

1.1 Time Series Models and Analysis

Time Series Analysis consists of methods for analyzing time series data in order to extract meaningful statistics and other characteristics of data [6]. It accounts for the fact data points collected over time may have an internal structure like auto-correlation, trend or seasonal variation that should be accounted for [7]. A time series is a series of data points indexed according to time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time [8].

In its broadest shape, time series analysis is about construing what has happened to a progression of information focuses in the past and endeavoring to anticipate what will transpire what's to come. Be that as it may, we will adopt a quantitative, measurable strategy to time series, by expecting that our time series are acknowledged of successions of arbitrary factors [9]. That is, we will accept that there is some fundamental creating process for our time series in light of at least one factual dispersions from which these factors are drawn.

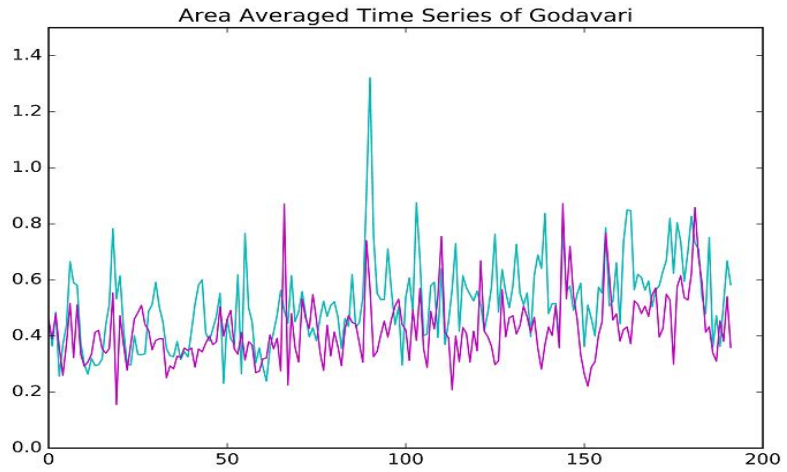

Figure 1.1: Time Series Plot of AOD_{550} in Talcher (blue) and Godavari (purple)

1.1.1 Characteristics of a Time Series Data

The following are the characteristics of a time series data.

Trend: A trend is a predictable directional development in a period arrangement. These trends will either be deterministic or stochastic. The previous enables us to give a hidden justification to the trend, while the last is an irregular component of an arrangement that we will be probably not going to clarify.

Seasonal Variation: Many time series data contain seasonal variations. This is particularly true in series representing air pollution or climate levels. In air pollution, we often see seasonal variation in pollution level.

Serial Dependence: A standout amongst the most vital attributes of time series, especially air pollution series, is that of serial connection. This happens when time series perceptions that are near to one another in time have a tendency to be connected. Instability grouping is one part of the serial connection that is especially imperative in quantitative exchanging [10].

Time Series analysis methodology can be broadly classified into two classes: frequency domain methods and time-domain methods. In frequency domain methods, one analyzes mathematical functions with respect to frequency instead of time and in time domain methods, time series is analyzed with respect to time. For atmospheric data, time domain methods are preferred, and frequency domain methods are preferred in electronics and statistics. The frequency domain methods are Fourier Series Analysis for repetitive signals and oscillating systems, Fourier transform for non-repetitive signals, transients, etc. and the time domain methods are Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), etc. This project has been focused on time domain methodologies. There are many other time domain methods too, but the present study only used ARIMA, ARMA and MCMC method for time series model development.

1.1.2 Stationarity and Non-Stationarity

A stationary time series is a stochastic process whose joint probability distribution does not change when shifted in time [11]. Consequently, parameters like mean and variance also do not change over time. The most common cause of violation of stationarity are trends in mean, which can be either due to the presence of a unit root or of a deterministic trend [12]. Stationarity is an important property of stochastic processes in order to implement many time series methods. Stationarity can be of strict or weak in nature. In strict stationarity probabilistic behavior of every collection of values is identical to the time shifted set [13]. Strict stationarity is too strict for most applications hence the study used conditions such as time independent mean value function and co-variance function for weak stationarity. In the case of non-stationary processes, stationarity can be achieved by one or more differencing of the series. In the present study, the stationarity of the time series data was examined using augmented Dicker-Fuller Test (ADF).

1.1.3 Time Series Models

1.1.3.1 Additive Time Series Model

In additive models, the assumption is that the components of time series affect additively. This can be represented as

$$\text{Data} = \text{Trend} + \text{Cyclical} + \text{Seasonal Effect} + \text{Residual}$$

By default, an additive model holds the assumption that the difference between seasonal values is approximately the same in each year or the amplitude of seasonal factor is same each year. The additive model also assumes that the residuals are random components and that adds on to the other components in the same way at all parts of the series.

1.1.3.2 Multiplicative Time Series Models

In many time series involving quantities, the absolute difference in the values are of less interest and importance than the percentage changes. For example, in seasonal data, it might be more useful to model that the AOD value in winter month is the same proportion higher than the value in summer month in each year, rather than the assumption that their difference is constant. Assuming that the seasonal and other effects act proportionately on the series is equivalent to multiplicative model. A multiplicative time series model can be represented as:

Data = Seasonal Effect X Trend X Cyclical X Residual

A multiplicative model is equally easy to fit data as an additive model. The trick to fitting a multiplicative model is to take logarithms of both sides of models as given below.

Log(Data) = Log (Seasonal Effect X Trend X Cyclical X Residual)

This can also be represented as:

Log(Data) = Log(Seasonal Effect) + Log(Trend) + Log(Cyclical) + Log(Residual)

After taking logarithms the four components of the time series again act additively. Though AOD_{550} values are quantities, we have assumed the series to be additive in nature.

1.2 Objectives

Our aim is to develop a time series forecasting model for forecasting of AOD values over mining region. The specific objectives are:

- Development of time series model for forecasting of AOD value over Godavari and Talcher coal mining regions.
- Study of the comparative performances of different time series model.

1.3 Thesis Outline

This thesis contains six chapters each consisting of different sections.

Chapter 1 demonstrates brief introduction of air quality, the theory of different time series analysis models, and objectives of the work.

Chapter 2 discusses different properties of time series, stationarity of time series, checking and removing stationarity, different components of time series, and additive & multiplicative time series.

Chapter 3 explains about the materials and methods of the proposed study.

Chapter 4 demonstrates the detailed methodology of model development and diagnosis.

Chapter 5 discusses the summary and Conclusions of the study.

Chapter 2

Literature Review

2.1 Introduction

Satellite data with high spatial and temporal diversity can cover larger and any region on earth which gives us the flexibility to use the larger geographic area as per our interest. Also, it is easy to select any area for collecting data since only changing the coordinates desired variable values can be easily accessed. The vast amount of available data stored by satellites presents us with the opportunity to perform any mathematical modelling to simulate any process, find out underlying characteristics present in data, to analyze and forecast data for air pollution [5]. In the past many studies have been conducted to understand, analyze and forecast air pollution level at different geographic locations. Most of those studies have been based on machine learning algorithms or neural networks or a combination of machine learning algorithms and neural networks. Time Series forecasting techniques have also been used in the past to forecast atmospheric changes and air pollution. Any variable that has the possibility to be collected over a time period could be analyzed or predicted by time series forecasting, for such methods time series analysis is preferred [6]. For better results always time series combined with some other machine learning or artificial neural network techniques are used. Even in time series, there are many different methodologies to handle data of different nature. For non-stationary data, we have methods like Autoregressive Moving Average (ARMA), Moving Average, for non-seasonal methods we have Autoregressive Integrated Moving Average (ARIMA), for seasonal methods, we have Holt-Winters Exponential Smoothing, Seasonal Autoregressive Integrated Moving Average (SARIMA), for multivariate analysis there is Vector Autoregressive (VAR)

method, for non-linear analysis there are different Autoregressive Conditional Heteroskedasticity (ARCH) methods.

In a study in 2008 [6] SARIMA and VAR were used to model the times series of monthly maximum 1-hour CO concentration in California South Coast Area. The dependence of current month air pollution level on previous month's air pollution was shown by SARIMA model. The connection between CO concentrations meteorological variables like solar radiation, precipitation was studied by VAR model. For the study data was used from open source websites and sources like California Irrigation Management Information System (CIMIS).

In an, another recent study in 2014 [5] Aerosol Optical Depth over 11 coal mine areas was studies using time series analysis method ARIMA and regression analysis. AOD is the measure of turbidity of atmosphere and an indicator of air pollution. In the study $AOD_{550\ nm}$ data was used from the year 2000-2012 over 11 different areas in India. This data was collected from Giovanni website. A thorough assessment of the air quality was performed, and different statistics like root mean square error, R^2 values, Bayesian Information Criterion (BIC) and Ljung-Box test was used for fitting best model and model diagnosis. Also, $AOD_{550\ nm}$ values were predicted for the year 2013-2015 with error margins [13].

Time series analysis has also been performed to figure out the relationship between air pollution and daily mortality. Poison Regression was used to find out the relationship between different air pollutants like SO_2, NO_x cause-specific deaths [7]. Significant correlation was fund between different air pollutants and daily mortality. It was found in the study that for every 100 micrograms/m^3 increase in SO_2 respiratory deaths increased by 4.21%, coronary heart disease death 10.68%, Cardiovascular death by 3.97% and chronic obstructive pulmonary death by 19.22%.

Different methodologies that have been used in the project to perform time series analysis have been discussed briefly in next sections.

2.2 Autoregressive Moving Average (ARMA)

ARMA refers to Autoregressive Moving Average. ARMA is the combination of Autoregressive (AR) and Moving average (MA) parts. It is used on weakly stationary stochastic processes [8]. Stationarity shall be explained later.
This method was first described in 1951 and can be mathematically written as:

$$\boldsymbol{X_t = c + \in_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \emptyset_i \in_{t-i}} \qquad (2.1)$$

where p and q are the order of AR and MA part respectively, $\boldsymbol{\varphi_i}$ are the parameters of AR part, $\emptyset_\mathbf{i}$ are the parameters of MA part and $\in_{\boldsymbol{t-i}}$ are white noise error terms.
Here it has been assumed that $\mathbf{X_t}$ depends linearly on its past values and error terms. If this is not the case model shall be specifically called nonlinear autoregressive moving average (NARMA) [8]. This methodology can be applied to finding out the correct order of AR i.e. p and correct order of MA part i.e. q. In order to find out p and q, we plot auto-correlation function (ACF) and partial auto-correlation function (PACF) against different time lags for the give time series. The table 2.1 below shows the behavior of the ACF and PACF for ARMA models.

Table 2.1: Behaviour of ACF and PACF for ARMA Models

	AR (p)	**MA (q)**	**ARMA (p,q)**
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

We can easily implement ARMA in python and R using “ARIMA” function in “statsmodels” module and “arima” function in base package respectively.

2.3 Autoregressive Integrated Moving Average (ARIMA)

ARIMA refers to Autoregressive Integrated Moving Average. It is the generalization of an Autoregressive Moving Average (ARMA) model. This method is applied where

there is inherent non-stationarity in the data, where an initial differencing step is applied one or more time to remove the non-stationarity. This method can be mathematically written as:

$$(\mathbf{1} - \sum_{i=1}^{p} \boldsymbol{\varphi_i} \boldsymbol{L^i})(\mathbf{1} - \boldsymbol{L})^d \boldsymbol{X_t} = \left(\mathbf{1} + \sum_{i=1}^{q} \emptyset_i \boldsymbol{L^i}\right) \in_t \qquad (2.2)$$

where L is the lag operator, d is the degree of differencing, p and q are the order of AR and MA part respectively, $\boldsymbol{\varphi_i}$ are the parameters of AR part, $\emptyset_\mathbf{i}$ are the parameters of MA part and $\in_t$ are white noise error terms. Here also, the secret of applying this methodology lies in finding out the correct order of AR i.e. p, correct degree of differencing, d and correct order of MA part i.e. q. In order to find out p and q we plot auto-correlation function (ACF) and partial auto-correlation function (PACF) against different time lags for the give time series [14]. Also, Akaike Information Criterion (AIC) is used to identify order. It is written as:

$$\boldsymbol{AIC} = \mathbf{2}(\boldsymbol{q} + \boldsymbol{p} + \boldsymbol{k} + \mathbf{1}) - \mathbf{2}\boldsymbol{log}(\boldsymbol{L}) \qquad (2.3)$$

where L is the likelihood of the data, p and q are the orders of AR and MA part and k is the number of parameters in the model being fitted to the data. The ACF and PACF table for ARIMA model is same as the ARMA model since in ARIMA only difference is instead of applying ARMA on actual data points we apply ARMA on differenced data points. We can easily implement ARIMA in python and R using "ARIMA" function in "statsmodels" module and "arima" function in base package respectively.

2.4 Markov Chain Monte Carlo (MCMC)

The theory of MCMC is just like the theory of Ordinary Markov Chain (OMC), except that stochastic dependence in the Markov chain changes the standard error [11]. OMC is the special case of MCMC in which X_1, X_2 …. are independent and identically distributed, in which case the Markov Chain is stationary and reversible. We start as in OMC with an expectation that we cannot do other than by Monte Carlo. To begin the

discussion, suppose that X_1, X_2, ... is a stationary Markov chain having initial distribution the same as the distribution of X [15].

We never use stationary Markov chains in MCMC, because if we could simulate X_1 so that it has the invariant distribution, then we could also simulate X_2, X_3, ... in the same way and do OMC (Ordinary Markov Chain).

It is a theorem, however, that, under a condition that is easier to verify than the CLT, if the CLT [12] holds for one initial distribution and transition probability, then it holds for all initial distributions and that same transition probability, and the asymptotic variance is the same for all initial distributions. Although the theoretical asymptotic variance formula contains variances and co-variances for the stationary Markov chain, it also gives the asymptotic variance for nonstationary Markov chains having the same transition probability distribution (but different initial distributions). In practice, this does not matter, because we can never calculation exactly except in toy problems and must estimate it from our simulations.

Markov Chain Monte Carlo process is very math intensive, but when applied on time series data or any other stochastic process it gives better results. For its implementation in python, one can use latest “fbprophet” module published by Facebook and for R one can use “glm” function in “mcmc” library.

Chapter 3

Materials and Methods

3.1 Data

For any forecasting or modelling method, it is important to have access to consistent, sufficient (in amount or duration in case of time series) and reliable data in order to make better modelling and more accurate forecast. Satellites collect data on a regular basis, so they solve the problem of consistency and sufficiency of data. Also for most of the climatic and atmospheric study satellite data are used since values of a specific variable over a wide region of earth can be easily gathered using satellites. Hence, we have used satellite data since no mine in India collects AOD data on a regular basis and even if we get AOD data, it is difficult to get for a decade or so. All the methods mentioned here were tested using the time series data of AOD over Godavari and Talcher region, Odisha, India. The AOD data obtained from the Giovanni web source for the Deep blue sensor at 550 nm. The area average values of the Godavari region (active mining region) between 23˚ 45' and 23˚ 50' North and 85˚ 30' and 86˚ 03' East, for land (corrected) for each month during 2001 to 2016 was derived.
Similarly, the area average values of the Talcher region between 20˚ 31' and 21˚ 40' North and 84˚ 15' and 85˚ 23' East, for land (corrected) for each month during 2001 to 2016 was derived.

3.2 Methodology

The detailed methodology adopted to achieve the objective of the project are:

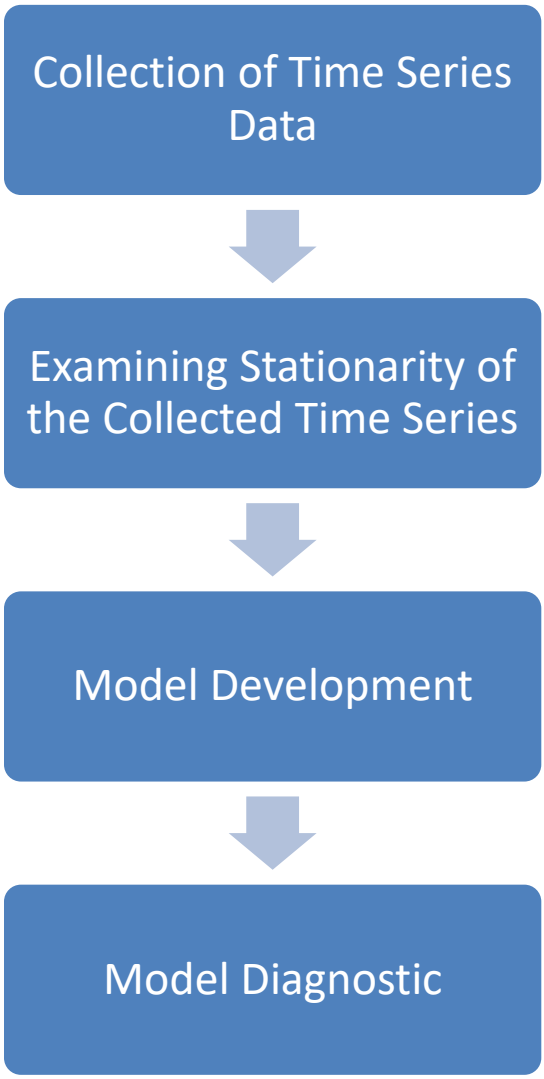

Fig 3.1: Flow Chart of the time series modelling process adopted in the project

3.2.1 Collection of Time Series Data

As mentioned above in section 3.1 we have used satellite data for Aerosol Optical Depth 550 nm from Giovanni website by entering coordinates of the two mining fields in which we are interested Godavari and Talcher namely. Data has been collected from 2001 to 2016.

3.2.2 Examine the Stationarity of the Collected Time Series

Stationarity is one of the most significant properties of a time series that helps in determining the suitable time series analysis methods. There are many intuitive ways and statistical tests to find out whether the series is stationary or not. One of the most common intuitive methods is an interpretation of a decomposed time series plot. The presence of increasing or decreasing trend in the trend part of the decomposed plot implies non-stationarity in data.

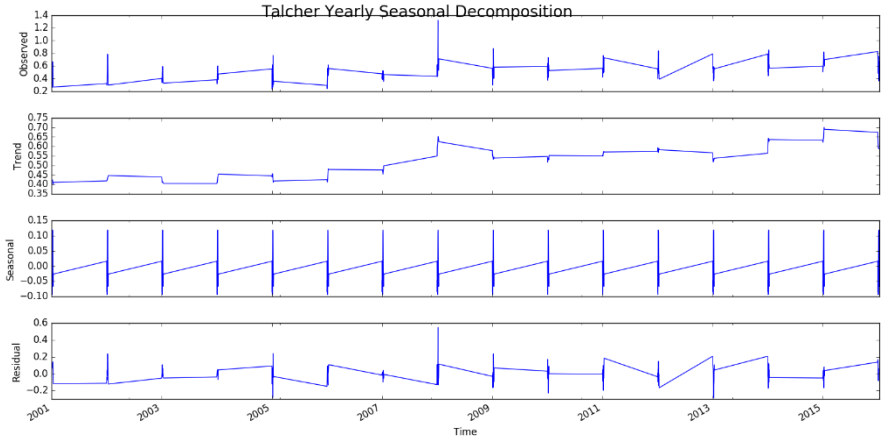

Fig 3.2: Decomposition of Talcher series into Trend, Seasonal and Residual Components

In the above figure 3.2 one can see the decomposition plot of Talcher Series has increasing trend hence one can easily say that Talcher series is non-stationary.

Similarly, a popular statistical test used to verify stationarity is Augmented Dickey-Fuller test. This test tests the null hypothesis of a unit root present in a time series sample. The alternative hypothesis is different depending on which version of the test is used but is usually stationary or trend-stationary. It is an augmented version of Dickey-Fuller test for a larger and more complicated set of time series models. The augmented Dickey-Fuller test statistic, used in the test is a negative number. The more negative number it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence. The intuition behind the test is that if the series is integrated then lagged values of the series **(y_{t-1})** will provide no relevant information in predicting the change in **y_t** besides the one obtained in the lagged changes **(y_{t-k}).** In the case, $\Upsilon = 0$ and the null hypothesis is not rejected.

3.2.3 Model Development

Once the stationarity of the data is checked next step is to select the appropriate model on the basis of stationarity or non-stationarity of the data. In this project ARMA, ARIMA and MCMC methods have been used. ARMA is used for stationary data, ARIMA is used for non-stationary data, and MCMC can be applied to any kind of stochastic process. After selection of the method, one needs to follow up requisite steps as per the method. For example, in ARMA one needs to plot correlation plots i.e. Auto-Correlation and Partial Auto-Correlation plots for the data. And using both the plots orders for autoregressive and moving average part is found out and ARMA model is fit with the obtained orders. In next step, subsequent orders which are close to previously obtained order are assumed and ARMA model has fit the assumed orders. Then one calculates the errors of the residuals of the model fit for all the assumed AR and MA orders. Different measures of error that have been used in this project are the root mean square error (RMSE), R Squared and Akaike Information Criterion (AIC). Thus model for ARMA methodology is developed.

Similarly, for ARIMA process before starting all the steps of ARMA one performs differencing on the time series. In differencing all the data points in a time series is subtracted from their immediate past data point. One performs differencing until stationarity of the differenced series is achieved. And one shall check the stationarity

using augmented dickey-fuller test. Once stationarity is achieved then further steps to be followed in ARIMA are just like ARMA.

For MCMC methodology one can skip the stationarity test as MCMC can be applied to any stochastic process. MCMC is a highly math intensive method, and it uses the spread of the data, the trend in the data, cyclic nature of the data to fit the model as well using the same inherent characteristics to predict the future values.

3.2.4 Model Diagnostic

In Model diagnostic, it is important that basic hypothesis made for the residuals are true. It is the model validation step in any modelling process. Here validation of the model is done using different statistical tests and other intuitive methods. In this project, Ljung-Box statistical test has been used to test the correlation of the residuals and the presence of white noise like behaviour in the residuals. Ljung-Box test is the standard residual analysis test used in time series analysis. This test has no of lags as the parameter, and this parameter depends on the seasonality present in series. If a series is found stationary, then a minimum of the 2x seasonal period (e.g. 4 for data showing quarterly seasonal behavior) or length of the series/5 is used as the no of lags, and if the series is non-stationary, then a minimum of 10 or length of series/5 is used as no of lags. Ljung-box test gives characteristic as a result which should be compared with Chi-distribution. In Ljung-Box test if (Ljung-box Test Static) $Q > \chi^2$ (Chi-square distribution) at the h-k degree of freedom where h is the total no of lags and k is a total number of parameters at significance level α then we can say that the model lack fits at that significance level. Correlation plots of residuals could also be used to get an intuition of validation of the fitted model. If one finds significant auto-correlation and partial auto-correlation in the data, then it suggests model lack fit.

Chapter 4

Model Development and Diagnostic

4.1 Autoregressive Moving Average (ARMA) Model

Before developing an ARMA model, it is essential to examine the stationarity of the data. The stationary data can only be used in ARMA model. The stationarity was examined by Augmented Dickey-Fuller test for time series AOD data of both the regions (Godavari and Talcher). The results of augmented Dickey-Fuller test results for Godavari and Talcher region are shown in Table 4.1 and Table 4.2 respectively.

Table 4.1: Results of augmented Dickey Fuller test for time series data of Godavari

Test Static	-6.712688e+00
P value	3.636860e-09
#Lags used	1.00
No of observations used	1.900000e+02
Critical Value (1%)	-3.465244e+00
Critical Value (5%)	-2.876875e+00
Critical Value (10%)	-2.574945e+00

It can be inferred from the results that the test statistic is less than all the critical values. This indicates that the data is stationary. The test statistic is lower than critical value at 1%, and this tells that the time series data at Godavari region is stationary with 99% confidence.

Table 4.2 Results of augmented Dickey Fuller test for time series data of Talcher

Test Static	-1.784760
P value	0.388054
#Lags used	12.00
No of observations used	179.00
Critical Value (1%)	-3.467420
Critical Value (5%)	-2.877826
Critical Value (10%)	-2.575452

It can be inferred from the results that the test statistic is higher than all the critical values. This indicates that the data is non-stationary. Thus, ARMA model cannot apply directly to the time series data of Talcher region.

Thus, the time series model using ARMA was developed for Godavari region. The auto-correlation (ACF) and partial auto-correlation (PACF) for the time series data of Godavari region are plotted to find out the orders of autoregressive and moving average.

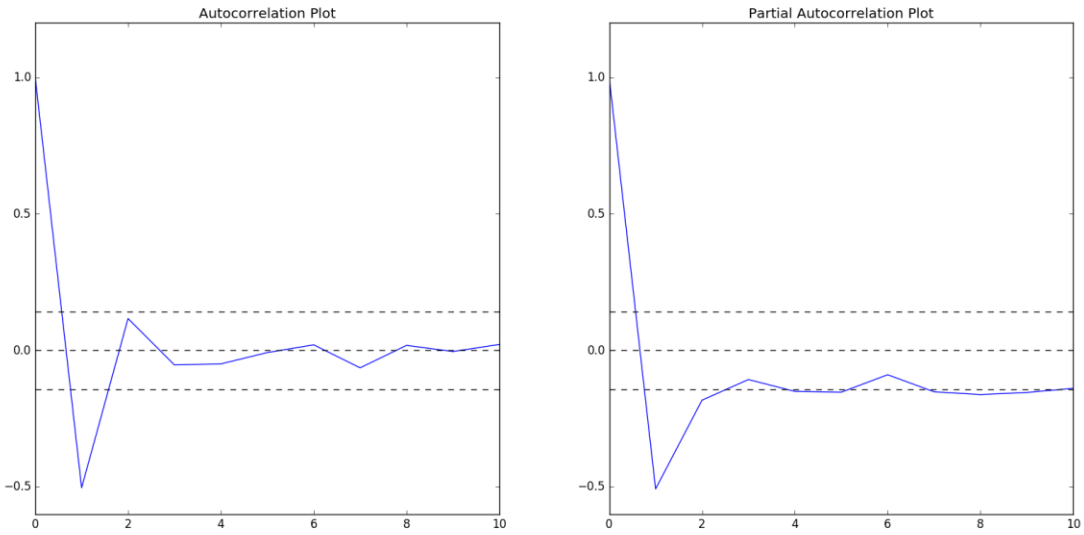

Fig 4.1: ACF and PACF plots of time series data of Godavari region

It was reported in Table 2.1 that both ACF and PACF should tail off and this can be observed in Fig. 4.1. The lag value after which ACF and PACF plots tail off would be the order of the parameters. Here it appears that both ACF and PACF is tailing off after lag 1. Hence, ARMA (1,1) model have been used to fit the data.

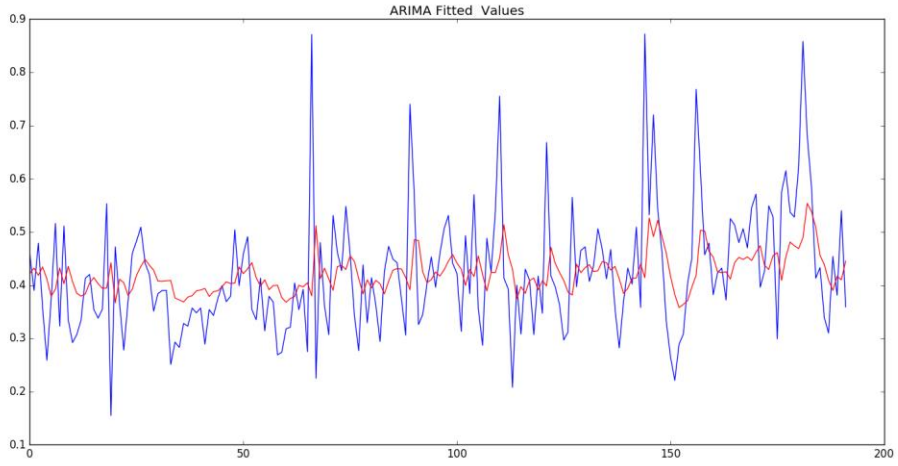

Fig 4.2 Godavari Series (blue) and ARMA (1,1) (red) model fit for Godavari Series

But, the correlation plots tell that other nearby models can be preferred and thus need to be examined. Hence, ARMA (2,2), ARMA (3,3), ARMA (4,4) models were also developed and examined for its suitability. The model performances were examined using the three parameters viz. R^2, Root mean Square Error (RMSE) and AIC values. The model performance parameters for different models are reported in Table 4.3.

Table 4.3 RMSE, R Squared and AIC for different ARMA Models

ARMA (p,q) model	**RMSE**	**R Squared**	**AIC**
ARMA (1,1)	0.112094	0.773047	-287.379
ARMA (2,2)	0.11081	0.77563108	-287.7214
ARMA (3,3)	0.10854	0.7802384	-291.135
ARMA (4,4)	0.10861	0.7800956	-286.79807

For a best fit model, the RMSE and AIC values should be low and R^2 value should be high. The results shown in Table 4.3 indicates that ARMA (3,3) has least RMSE and AIC values and highest R^2 value. Therefore, ARMA (3,3) model can be considered as a best fit model for forecasting of AOD values.

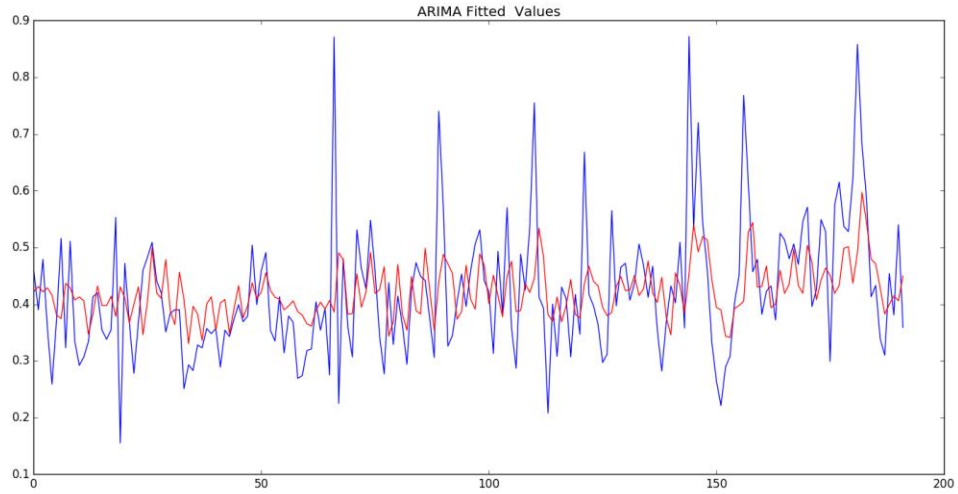

Fig 4.3: Observed (blue) and predicted AOD (red) values using ARMA (3,3) model in Godavari region

The above result indicates the fact that sometimes it is difficult to find out correct parameters only from correlation plots. So one of the best methods to find the best fit model is to simulate more than one ARMA model for identifying the best model. The model with optimum values of the diagnostic parameter can be selected as the best fit. The residual ACF and PACF plots (shown in Fig. 4.4.) was also analysed for different lag values.

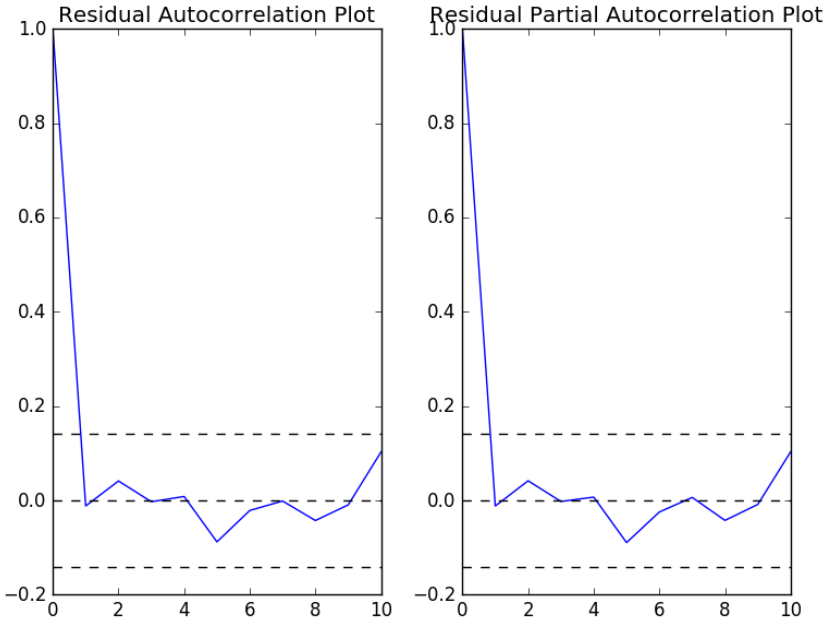

Fig 4.4: Residual Correlation plots for Godavari ARMA (3,3) fitted model

The residual correlation plots of ARMA (3,3) model indicate that the correlation values are within significance limit for all the lags. Thus, it can be inferred that there is no correlation between residuals for ARMA (3,3) model fitted to time series data in Godavari region.
The Ljung-Box result for time series data in Godavari region of ARMA (3,3) model is shown in Table 4.4. The results indicate that model is fitted at 0.1 significance level.

Table 4.4 Ljung-box test result for Godavari ARMA (3,3)

Lags	**p Value of ARMA (3, 3) model**	**Test Statistics ARMA (3, 3) model**
1	0.868799	0.027286
2	0.835341	0.359831
3	0.948099	0.361384
4	0.984455	0.375219
5	0.860487	1.917146
6	0.918921	2.008439
7	0.959353	2.008781
8	0.967199	2.377481
9	0.983554	2.395862
10	0.918822	4.555625

4.2 Model Development using Autoregressive Integrated Moving Average (ARIMA)

ARIMA is the extended version of ARMA and used with non-stationary time series data. In this method, first, non-stationarity has been removed by differencing of the series and then applied ARMA process. In ARIMA, 'I' stand for integrated that gives us the order of differencing of the series. Earlier we have seen that the time series data is non-stationary (Table 4.2) and thus it can be removed by differencing. The first difference is obtained by subtracting each current value in the series with its immediate past value.

After first differencing, the time series data of Talcher region has become stationary as shown in Table 4.5. Since the test static value is less than the critical values.

Table 4.5: Result of Augmented Dickey Fuller Test for first differenced time series data of Talcher region

Test Static	-5.295418
P value	0.0000
#Lags used	15.00
No of observations used	175.00
Critical Value (1%)	-3.468280
Critical Value (5%)	-2.878202
Critical Value (10%)	-2.575653

Now after differencing the test statistic is less than the any of the critical values hence the series becomes stationary. Then ARMA process on differenced series have been applied, and correlation plots (ACF and PACF) for the differenced series are plotted (shown in Fig. 4.5).

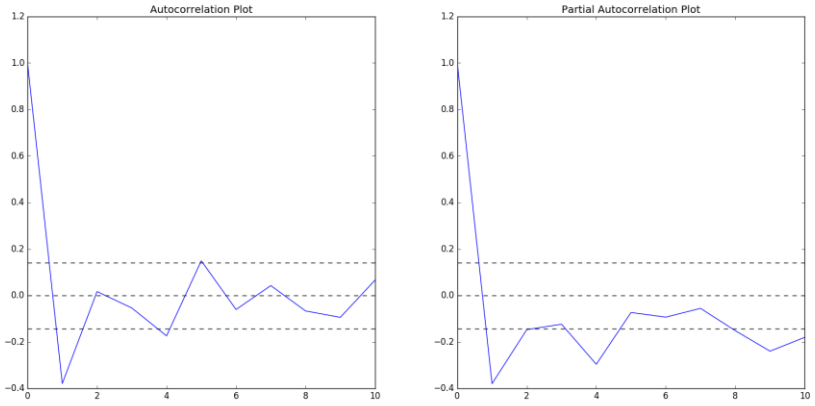

Fig 4.5 ACF and PACF plots for first differenced time series data in Talcher region

It is known that both ACF and PACF should tail off for selecting the order parameters of a time series model. The lag value after which ACF and PACF plots tail off would be the order of the parameters. Here, it seems that ACF is tailing off after 4 lags hence q = 4 and PACF is tailing off after 8 lags and

hence p = 8. One important thing is that the parameter for moving average part was found out by ACF plot and for the Autoregressive part is found out by PACF plot. But as in ARMA case, nearby models have also been checked and then choose the best fit accordingly. Hence, in this case also, ARIMA (0,1,1), ARIMA (0,1,2), ARIMA (0,1,3), ARIMA (1,1,1), ARIMA (1,1,4), ARIMA (4,1,4), ARIMA (4,1,1) and ARIMA (8,1,1) were developed for identifying the best fit model. The model performance parameters (RMSE, AIC, and R^2) for different cases were determined and reported in Table 4.6. For a best fit model, the RMSE and AIC values should be low and R^2 value should be high. The results shown in Table 4.6 indicates that ARIMA (0,1,3) and ARIMA (1,1,4) perform equally good and thus can be chosen either one as best fitted model.

For best fit model, low RMSE and AIC values and high R Squared values are preferred. For the above case R Squared value should not be considered because it uses mean of the series and since, the series is not stationary, using one mean for the series is not correct. In time series data of Talcher region, the R^2 value cannot be chosen as model performance parameters as in the case of Godavari data. This is because the R^2 value for a non-stationary series does not make sense. Thus, RMSE and AIC values were used to find the best fit model. The observed versus predicted values of ARIMA (0,1,3) and ARIMA (1,1,4) model are shown in Fig 4.6 and Fig 4.7 respectively. Since the time series data in Talcher region has been differenced and thus before plotting differenced values have to be scaled back so that the observed and predicted values could be compared.

Table 4.6: RMSE, R Squared, and AIC for different ARIMA values

ARIMA (p,d,q)	**RMSE**	**R Squared**	**AIC**
(0,1,1)	0.66327	-0.327	-209.1023
(0,1,2)	0.5478	-0.09	-225.8954
(0,1,3)	0.45067	0.0980	-231.098
(1,1,1)	0.4760565	0.04725755	-229.869
(1,1,4)	0.44960	0.1001857	-229.178
(4,1,4)	0.547122	-0.094968	-231.0526
(4,1,1)	0.47232410	0.054727	-229.7787
(8,1,1)	0.608150	-0.217104	-229.3233

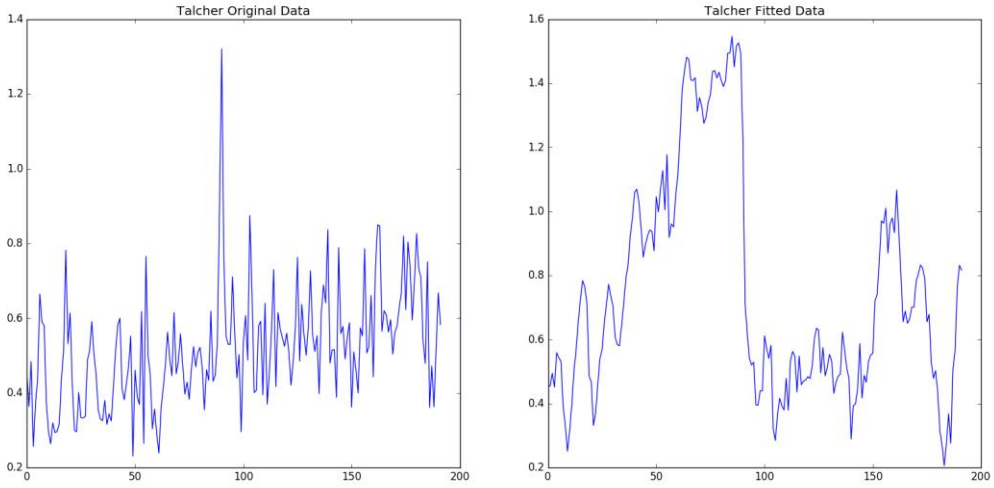

Fig 4.6 (Left) Observed and (Right) Predicted values of time series data in Talcher region using ARIMA (0,1,3) model

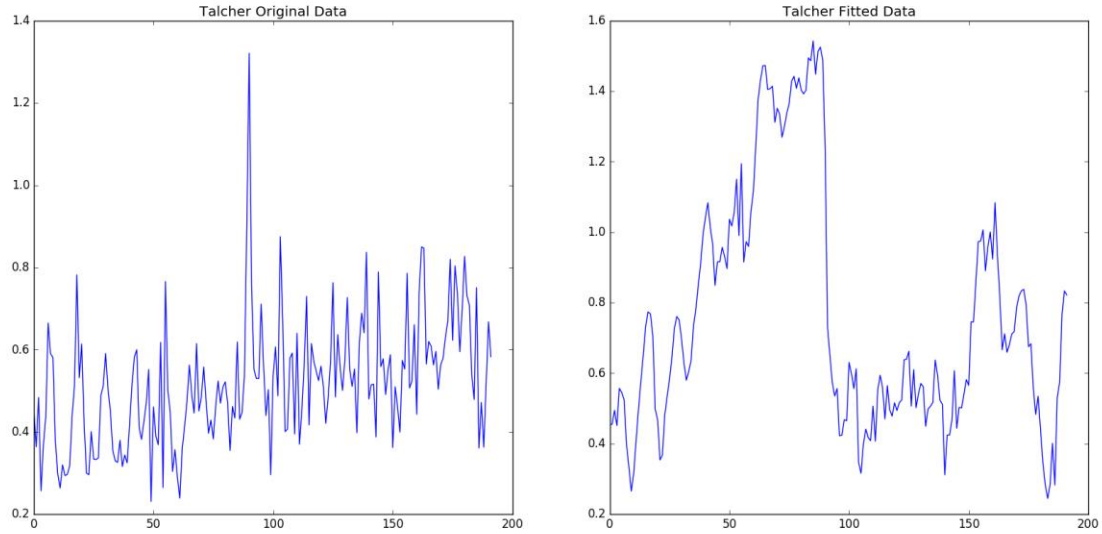

Fig 4.7: Fig 4.6 (Left) Observed and (Right) Predicted values of time series data in Talcher region using ARIMA (1,1,4) model

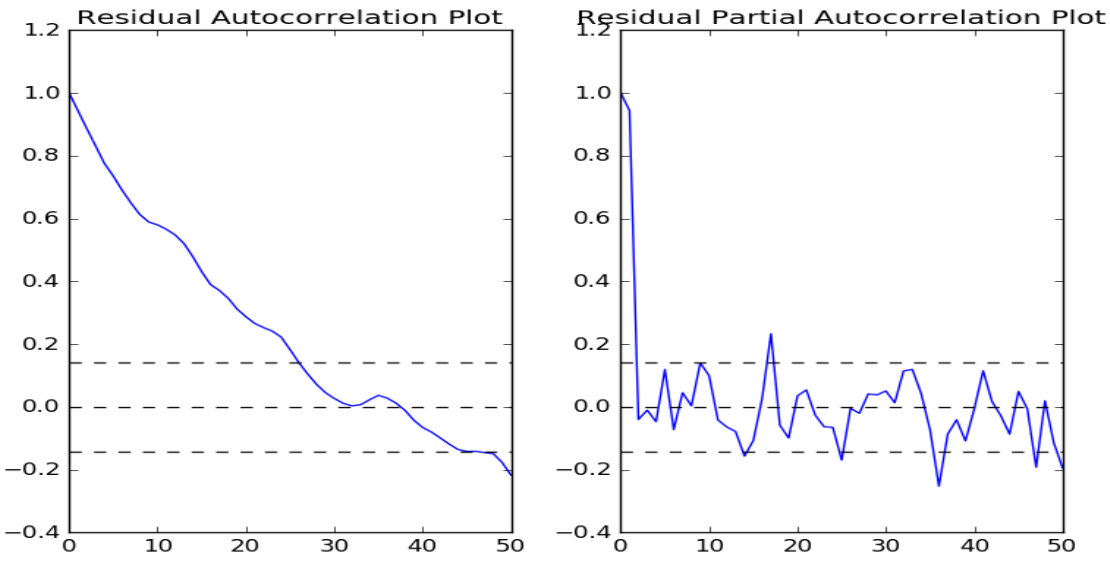

Fig 4.8: Residual Correlation Plots of ARIMA (0,1,3) model in Talcher region

The residual correlation plots of ARIMA (0,1,3) model indicate that the correlation values are linearly decreasing and have significant auto-correlation for many lag values. In Ljung-Box test, if $Q > \chi^2$ at the h-k degree of freedom where k is a total number of parameters at significance level α then we can say that the model lack fits at that significance level. The Ljung-Box result for time series data in Godavari region of ARIMA (0,1,3) model and ARIMA (1,1,4) model is shown in Table 4.7.

Table 4.7: Ljung-Box test for Talcher ARIMA (0,1,3) and ARIMA (1,1,4)

Lags	**p Values of ARIMA(0,1,3) model**	**Test Statistics ARIMA(0,1,3) model**	**p Values of ARIMA(1,1,4) model**	**Test Statistics ARIMA (1,1,4) model**
1	1.2E-39	173.6235	1.71E-39	172.909
2	6.84E-72	327.728	1.61E-71	326.0154
3	3.2E-100	463.8544	1.4E-99	460.9215
4	7.7E-125	582.9288	3.7E-124	579.7886
5	5.3E-147	690.6059	3.6E-146	686.7263
6	1.6E-166	785.9768	1.1E-165	782.0803
7	9.1E-184	870.9189	8.3E-183	866.4754
8	4.2E-199	946.9387	4.2E-198	942.2901

4.3 Model development using Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) process can be used for any stochastic process stationary as well as non-stationary. Hence, the present study applied MCMC on time series data of both the regions (Godavari and Talcher). MCMC is a very math intensive and complex process in comparison to that of the ARMA process. In this case, we have used this method to predict future AOD values for the year 2017-2019. It can be easily implemented in python using 'Prophet()' in 'fbprophet' module. This module also considers the uncertainty level of the fit and to forecast values the number of time periods with their frequencies.

Fig 4.9 and 4.10 shows the fitted and forecasted values of the time series data in Talcher and Godavari region respectively. Fig 4.11 and Fig 4.12 shows the weekly and yearly AOD values as well as the trend of the time series data Godavari and Talcher region respectively. It can also be observed that a sudden rise in AOD_{550} values during January to March in Godavari region and a sudden drop in February in Talcher region. Another interesting pattern is the high AOD_{550} values during the weekend in Talcher coalfield area.

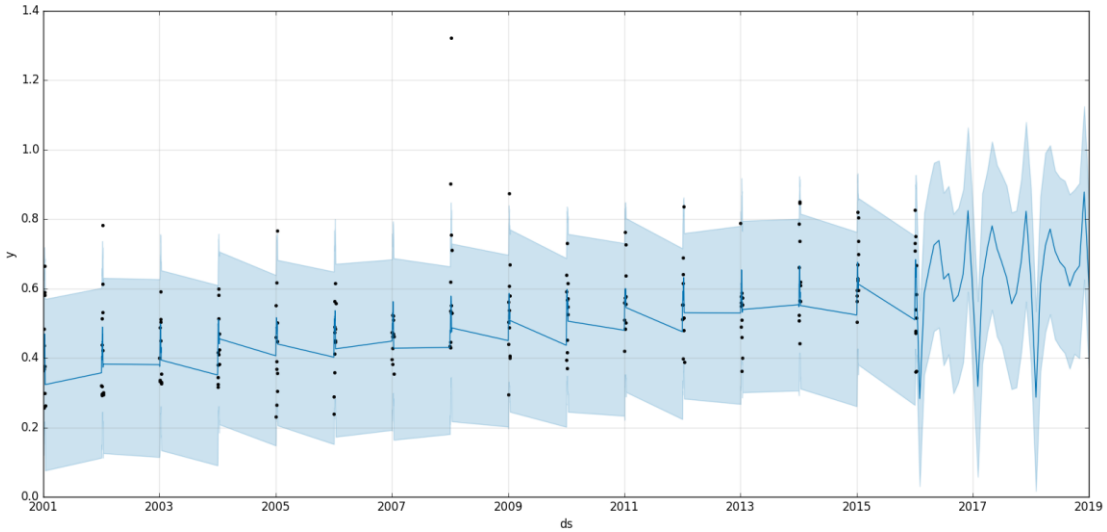

Fig 4.9: MCMC model fit plot with future predicted values in Talcher region. Dots represent observed values; blue curve represents forecasted values and the uncertainty intervals of our forecast is shown by shaded region.

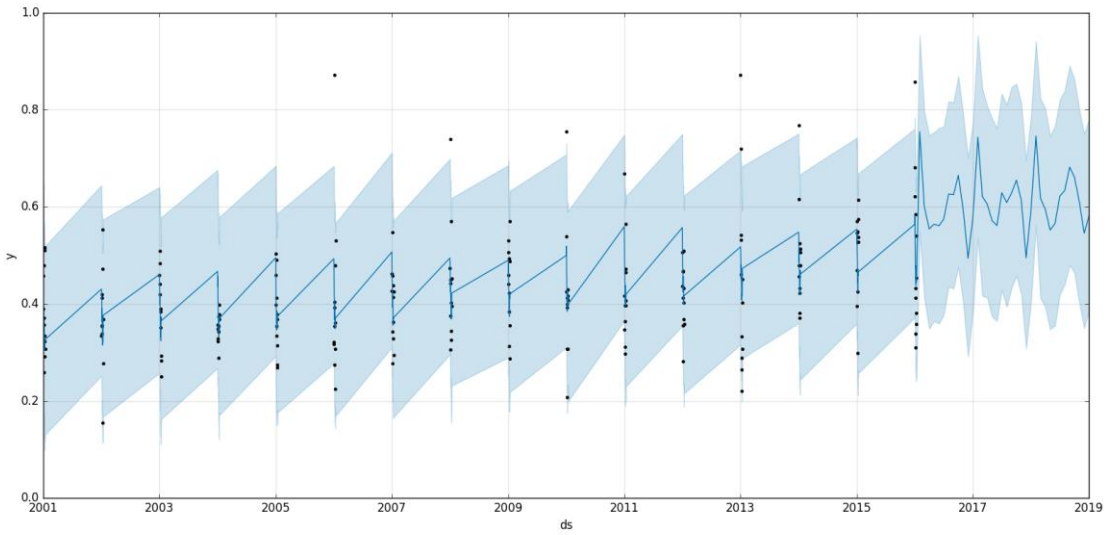

Fig 4.10: MCMC model fit plot with future predicted values of time series data in Godavari region. Dots represent observed values, blue curve represent forecasted values, and the uncertainty intervals of our forecast is shown by shaded region

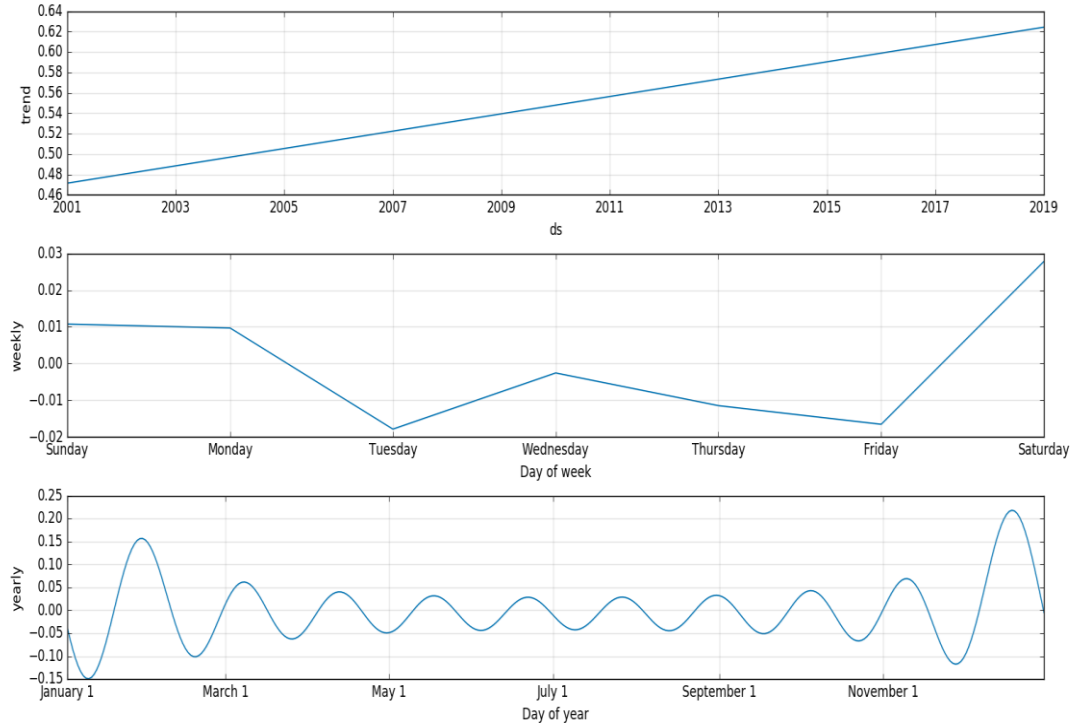

Fig 4.11: Annual, Monthly, and daily AOD predicted values of MCMC model in Godavari region

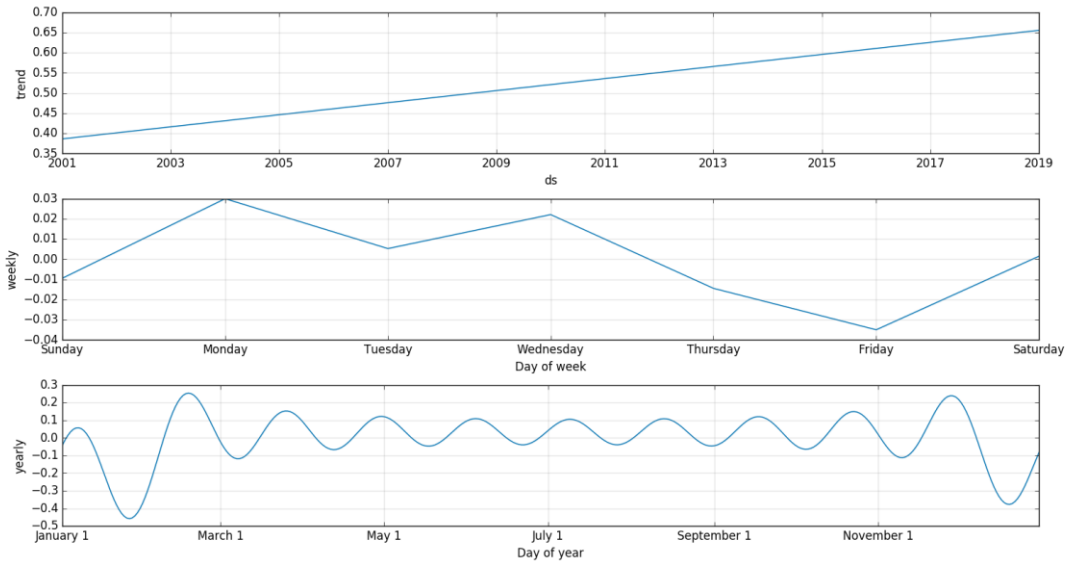

Fig 4.12: Annual, Monthly, and daily AOD predicted values of MCMC model in Talcher region

Table 4.8: Ljung-Box result for MCMC fitted Godavari data

Lags	P value of MCMC model	Test Statistics of MCMC model
1	0.485987	0.485398
2	0.051533	5.931051
3	0.098565	6.284381
4	0.115973	7.40498
5	0.168638	7.782481
6	0.064561	11.886
7	0.056921	13.69289
8	0.059584	14.97753
9	0.083026	15.29971
10	0.108503	15.70147

The Ljung-Box test results for MCMC model are shown in Table 4.9 and Table 4.10 respectively for Godavari and Talcher region. The Ljung-Box test results suggest that it is a poor fit than ARMA (3,3) model in Godavari region. In contrast to this, the Ljung-Box test results suggest that MCMC model is better than that of the ARIMA (0,1,3) and ARIMA (1,1,4) model designed for time series data in Talcher region.

Table 4.9: Ljung-Box result for MCMC fit of Talcher data

Lags	P value of MCMC model	Test Statistics of MCMC model
1	0.000131	14.62335
2	0.000173	17.32983
3	0.00048	17.81467
4	0.000587	19.64285
5	0.000602	21.6813
6	0.001292	21.84416
7	0.00258	21.96074
8	0.000842	26.55833

Chapter 5

Conclusion and Future Work

5.1 Conclusions and Future Work

The present study attempts to develop different time series model for forecasting of AOD over Godavari and Talcher coal mining region. Three types of time series models (ARIMA, ARMA, and MCMC) were used to fit the time series AOD data over the two regions. The best fit models were selected based on the three model diagnostic parameters (RMSE, R squared, and AIC). The Ljung-Box test was used to validate the fit of the model. The best fit model for Godavari region was observed to ARMA (3,3), and that of Talcher region was MCMC. The MCMC models were fitted with 95% confidence interval and have chosen as the best fit.

Multivariate time series analysis could be performed in future to establish a correlation between different air pollutants like AOD and particulate matter. Also, further work can be done to improve the accuracy of this work for non-stationary data. Also, Seasonal Autoregressive Moving Average could be used.

References

[1] CMRI (1998) Determination of emission factor for various opencast mining activities report GAP/9/EMG/MOEF/97. Central Mining Research Institute Environmental Management Group, Dhanbad.

[2] Xio Han Cai, (2008) Time Series Analysis of Air Pollution CO in California Coast Area, with Seasonal ARIMA and VAR Model, International Journal of Forecasting, 23:40-45

[3] Ghosh MK, Majee SR (2007) Characteristics of hazardous airborne dust around an Indian Coal Mining area. Environ Monit Assess 130:17-25.

[4] Chaulya SK, Kumar AK, Tripathi M, Singh N, Mishra RS, Bandyopadhyay LK (2012) Assessment of coal mine road dust properties for controlling air pollution, Mining and Science Technology, 22:25-35

[5] Soni K, Parmar SK, Kapoor S (2014) Time Series Model Prediction and trend variability of aerosol optical depth over coal mines in India. Enviro Sci Pollut Res 10:2-14.

[6] Wikipedia "Time Series – Wikipedia the free encyclopedia," 2017, [online accessed; 9-May 2017]. [Online]. Available: https://en.wikipedia.org/wiki/Time_series.

[7] NIST "Introduction to Time Series," 2013, [Online accessed; 9-May 2017]. [Online]. Available: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm.

[8] Wikipedia "Autoregressive-moving-average model – Wikipedia the free encyclopedia," 2017, [Online accessed; 9-May 2017]. [Online]. Available: https://en.wikipedia.org/wiki/Autoregressive-moving-average_model.

[9] Wikipedia "Autoregressive integrated moving average – Wikipedia the free encyclopedia," 2017, [Online accessed; 9-May 2017]. [Online]. Available: https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average.

[10] Methodology of the monthly index of services. Annex B: The Holt-Winters forecasting method, 2017, [Online accessed; 9-May 2017]. [Online]. Available: https:// annexbtheholtwinterforecast.pdf

[11] Geyer C.J, Introduction to Markov Chain Monte Carlo, [Online accessed; 9-May 2017]. Available: http://www.mcmchandbook.net/HandbookChapter1.pdf

[12] Wikipedia "Stationary Process – Wikipedia the free encyclopedia," 2017, [Online accessed; 9-May 2017]. [Online]. Available: https://en.wikipedia.org/wiki/Stationary_process.

[13] Shumway R.H et al., (2006) Time Series Analysis and Its Applications with R examples, Second Edition. Springer New York, ch- 1-4.

[14] Chang G. et al. (2003) Time Series Analysis on the relationship between air pollution and daily mortality in Beijing. Chinese Science Bulletin, 23: 24-27

[15] Gooijer J.G.D. et al. (2006) 25 years of time series forecasting. International Journal of Forecasting, 22:20-25