A

Dissertation On (Major Project-II)

"To predict accurate Silver MCX trend in Indian Stock Market using Box-Jenkins method"

Submitted in Partial Fulfillment of the Requirement
For the Award of Degree of

Master of Technology

In

Software Technology

By

Akul Taneja
University Roll No. 2K15/SWT/504

Under the Esteemed Guidance of

Mr. Vinod Kumar
Associate Professor, Department of Computer Science & Engineering

2015-2019(Jan)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
DELHI – 110042, INDIA

STUDENT UNDERTAKING

Delhi Technological University
(Government of Delhi NCR)
Bawana Road, Delhi- 110042

This is to certify that the thesis entitled **"To predict accurate Silver MCX trend in Indian Stock Market using Box-Jenkins method"** done by me for the Major project-II for the achievement of **Master of Technology** Degree in **Software Technology** in the **Department of Computer Science & Engineering**, Delhi Technological University, Delhi is an authentic work carried out by me under the guidance of Associate Prof. Vinod Kumar.

Signature:
Student Name
Akul Taneja
2K15/SWT/504

Above Statement given by Student is Correct.

Project Guide:
Mr. Vinod Kumar
Associate Professor,
Department of Computer Science & Engineering, DTU

ACKNOWLEDGEMENT

I would like to express sincere thanks and respect towards my guide **Mr. Vinod Kumar, Associate Professor, Department of Computer Science & Engineering, Delhi Technological University Delhi.**

I consider myself very fortunate to get the opportunity for work with her and for the guidance I have received from her, while working on this project. Without her support and timely guidance, the completion of the project would have seemed a far. Special thanks for not only providing me necessary project information but also teaching the proper style and techniques of documentation and presentation.

AKUL TANEJA
M.Tech (Software Technology)
2K15/SWT/504

ABSTRACT

In today's world we are surrounded by loads of data. Data in form of text, images, videos, locations etc. We are moving into a new era that is termed as the Data Era. Data is going to play a vital part by changing the way people interact with their surroundings on daily basis. Experts says Data will lead to the fourth industrial revolution. There are around 2.5 quintillion bytes of data that is being gathered currently each day. This a lot of data and the potential of benefits from this data is unimaginable. Social media platforms like Facebook, Twitter, and Google are one of the major contributors of data in towards world.

Today we are well equipped with the technology to deal with this amount of data and use the same for our advantage. Technologies like Machine Learning, Deep Learning, Artificial Intelligence, Block Chain, and 5G provide us with a great opportunity to use in almost each and every field. One of the advantages that data could provide is in prediction of future demands of products, stock market prices, investment plans and market requirements. This could be beneficial to both product/ service providers and the users who avail them.

Many big organizations have started using data for their advantage. With the growing competition and demand for better products and services, companies are seeking helping of Machine Learning algorithms to learn the shopping patterns of the customers, get more insight into the department of their company lacking behind and many more such information.

As we know today a lot of people are interested in buying stocks and commodities from the market due the promising returns these tools present to the customer. So the price of such tools changes day to day. Every day the stock and commodity have a different price tag attached to it. So there is a series of values associated with the stock or commodity over a period of time. Such process can be termed as Time-Series process in the language of Data Science.

Here we are taking up the task of finding the patterns in Time-Series processes so that the future value of such process can be predicted with utmost accuracy. To find the patterns and predict the

future values will be using the Box-Jenkins approach. This approach has proved to be very accurate for finding patterns and predicting the future values of the Time-Series processes. The language used to apply the Box-Jenkins approach here is python. Python provides various libraries to efficiently work with various machine learning algorithms. The few libraries used here are pandas, numpy, and matplotlib.

TABLE OF CONTENTS

CHAPTER 6:

CONCLUSION

Chapter 1: Introduction

1.1 PROBLEM STATEMENT

In the era of the data both companies and individuals are aiming to use data to predict future requirements or prices of shares and various commodities in the stock market. This trend can be seen globally where companies and individuals are trying to outsmart each other by buying and selling of shares and commodities are right time to maximize their profits.

With the power of data and field of data science it is very much now possible to predict future of the targeted variable with help of dependent variables. Though we are surrounded with data, it still requires a lot of effort and time in identifying the right kind of data. Then processing of data so it could be used to successfully predict the future value of the targeted variable.

So, hereby we trying to accurately predict the trend of Silver MXC in the Indian stock market with the help of using Box-Jenkins approach for time-series process.

1.2 WHAT ARE THE STEPS INVOLVED IN PREDICTIVE ANALYSIS

The lifecycle of such prediction involves many steps. First being having the clarity of the problem statement in hand and identification of the solution to the problem. Once we are clear with the above two points we will be in a position to select our target variable. Second step one of the major steps that is identification of the dependent variables which will help to predict the target variable with the required accuracy. Accuracy to be achieved is also dependent upon the target variable selected. For example to predict the loan defaulters an accuracy of 90% is also considered as a good percentage, but if we talk about predication of some disease or medical case an accuracy of 90% is not acceptable. In such cases accuracy needs to be approaching 100%. Similarly for various fields the accuracy varies as per the stakes associated with the targeted variable.

Third steps involves data preparation also known as pre-processing of the data set. This step is very important in successful building of prediction model and almost takes 80% of total time required to build a prediction model. Data can be gathered from various kinds of sources such as text documents, databases, files in different formats, images and videos supporting all kinds of formatting, spreadsheets and etc. There we can be having both kind of data that is structured or unstructured data. After the data has been successfully gathered as per the problem statement in hand, next is to explore the data carefully. One should go through the data in hand and identify the type of data, types of variable, range of variables, identify if the variable are categorical or

discrete in nature, format of the variables, length of the variables and other such properties of the variables.

Once the analyst has understanding of the data in hand he/she can go on with the next step of data validation.

Data validation, data can have various problems or errors that need to be dealt with before the data can be used as a training and test data set. Problems likes missing values, outliers, incorrect variable type, correlated variables, redundant variables and many more need to be taken care of. There are a lot techniques to handle the above cases. There are few modelling techniques which are self-capable to handle the above mentioned cases, but other require separate effort to take care of above situations. Dimension reduction technique also an important technique which helps in reducing the total number of variables in the process of building model. Once the data is validated thoroughly that is now the data understanding the problems with the data in hand are identified, analyst can start with data cleaning process.

Data cleaning step is about correcting the errors and problems identified in data exploration and validation steps. It is after this step data is ready to be used to train the required predication model. In this step problems such as missing values, outliers are to be dealt with. There are no shortcuts that can be used in this step, each case to be dealt separately keeping in mind the problem statement and type of data in hand.

1.3 THESIS MOTIVATION AND GOAL

Currently there is a lot of information and data associated with the different commodities available in the market. Data such as the daily prices of the commodities, various factors that affect the price of these commodities can be easily found and used over the internet.

Also with the introduction of machine learning algorithms and increasing efficiency and accuracy of these algorithms, it makes easier to find patterns in the prices of these commodities and predict their trends.

With more and more people getting involved in buying and selling of such commodities, it makes sense to develop such systems which could help user to predict the trend of various commodities so that he/she could wisely make decision of buying or selling the commodities.

So it totally makes sense to develop such a system which can first find the patterns in the price of a commodity. Next step is to predicting the future trend of the commodity. Final step would to accurately predicting the value of a commodity in the near future.

1.4 THESIS ORGANIZATION

The thesis is classified into six different chapters.

Chapter 1 defines the problem statement for the thesis, which is accurate trend prediction of the Sliver MXC price in Indian Stock Market using Box-Jenkins approach. This technique has shown accurate results in case of predictive analysis of the time-series process.

Chapter 2 describes the related work done in the areas of predictive analysis of the time-series process using various techniques. This thesis deals predicting the trend of Silver MXC price using the Box-Jenkins approach.

Chapter 3 explains the research topics used in detail. In our research details, we explain the terms, which are being used in the thesis like Box-Jenkins, Deep ,AR process, MA process, ARMA process, ARIMA process, Time-series process, ACF plots, PACF plots, IACF plots, Data modelling, Data Validation, Data Cleaning, Python, Numpy, Pandas, Matplotlib and etc.

Chapter 4 is explaining the proposed approach for the solution. This section of the thesis covers the various steps involved in the accurate prediction of the Silver MCX price in Indian Stock Market using Box-Jenkins approach. The process of data gathering is explained, that is how data is gathered for predictive analysis of the process in hand. The next section explains the importance and steps of the pre-processing of the data so that the data can be used in model building, which in turn will predict the trend of the process in hand. Then followed by the section which checks whether the time-series process is stationary or not. If not the process needs to be made stationary using various techniques. These techniques are also explained in the section. Next explains how to identify the process type of the given time-series process. Finally the chapter explains once the process is identified how to find the results and compare the result values with the original values.

Chapter 5 illustrates the step by step results of the complete approach starting from processing of the data to the result comparison of the predicted values with the original values. Results are shown in the form of graphs and plots for better visualization. Finally the implications of the results are discussed in this chapter.

Chapter 6 is the conclusion of the thesis. It describes the benefits of the Box-Jenkins approach for predictive analysis of the time-series process. Discuss about the accuracy of the approach. Also the future work which can be done on the current work.

Chapter 2: Related Work

[1] Demand forecasting for available seats in airlines is important to maximize the expected revenue by setting the appropriate fare levels for those seats.

[2] Airline passenger forecasting using neural networks and Box–Jenkins.

[3] ARIMA Implementation to Predict the Amount of Antiseptic Medicine Usage in Veterinary Hospital.

[4] Prediction of Rupiah against US Dollar by Using ARIMA.

[5] Short-term Traffic Flow Prediction Using a Methodology Based on ARIMA and RBF-ANN.

[6] Forecasting Method of Aero-Material Consumption Rate Based on Seasonal ARIMA Model.

[7] Forecasting of Raw Material Needed for Plastic Products Based in Income Data Using ARIMA Method approach.

[8] Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network.

Chapter 3: Research Background

3.1 Time Series Process

Time-Series process is a process which has particulars varying over a period of time. Time period over which the process varies is also fixed. This means we have an equal time interval between varying data of the Time Series Process. Data like the number of flight booked for an airline in a month or number of loans applied to a particular bank in a month, all examples like these come under the category of the Time Series process. If these companies are successfully able to predict the future values of number of flight booking or number of loans that will be applied in the upcoming time interval, this predication if done successfully can be very helpful for the companies in allocating funds and resources and efficiently meet the requirements.

3.2 Time Series Analysis

Time Series process undergo Time Series analysis so that using the previous data future prediction can be made and useful patterns can be drawn from the historical data that could be beneficial in future dealings. Time Series involves three major steps; Descriptive analysis, modelling and forecasting the future values. First step that is the descriptive analysis is all about understanding the properties of the Time Series process in hand. It is about looking for trends, seasonality and the behavior of the series. Based on the first step, the second step of modelling is performed. Identification of the correct technique to be used to prepare the model is finalized. Once the model is fully prepared using the training data, the next steps comes into the picture that is forecasting the values using the prepared model in second step. Using the test data set forecasting power or accuracy of the model is tested. If the model does not meet the required threshold of accuracy the model is rebuilt. Hence it is an iterative process until the desired accuracy is achieved.

3.3 Box-Jenkins Approach

One of the most popular techniques for Time-Series analysis is Box-Jenkins approach. Box-Jenkins approach publicized the ARIMA (auto-regressive integrated moving average) technique of modelling a time-series. Box-Jenkins approach did significant advancements in the ARIMA

method by simplifying the application of the method. There are many fields where this approached has proved to be very useful in accurate prediction of the time series process. Many work has already been done using this approach and has shown very positive results. One needs to have good understanding of AR, MA, ARMA and ARIMA processes for implementation of Box-Jenkins approach to do predictive analysis of Time Series processes.

3.4 AR Process

AR process is the process in which the previous values of the time series process have major effect on the current values of the process. AR (p) is the symbol used to represent the AR processes. P denotes the number of previous values on which the current value depends.

Let G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}, G_{t-5}, G_{t-6}, G_{t-7}, G_{t-8}…… G_{t-k}, be a given values of a time series process. If the series is represented by symbol AR (1), then current value of G that is J_t will be given as follows.

$$G_t = a_1 * J_{t-1} + \epsilon_t,$$

where a_1 denotes total or quantified impact on G_t and ϵ_t denotes error at time t also sometimes known as white noise. Similarly we can see if process is represented by symbol AR (3) then the current value of G is given as

$$G_t = a_1 * G_{t-1} + a_2 * G_{t-2} + a_3 * G_{t-3} + \epsilon_t,$$

3.5 MA Process

MA process is the process in which the previous values do not put any consequence on the next set of values of the process, rather current values depend upon the error or noise of the previous values.MA (p) is the symbol used to represent the MA processes. P denotes the number of previous values on which the current value depends.

Let G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}, G_{t-5}, G_{t-6}, G_{t-7}, G_{t-8}…… G_{t-k}, be a given values of a time series process. If the series is represented by symbol MA(1) , then current value of G that is G_t will be given as follows.

$$G_t - \mu = b_1 * \epsilon_{t-1} + \epsilon_t,$$

where b_1 denotes total or quantified impact on ϵ_{t-1}, ϵ_t denotes error at time t also sometimes known as white noise and μ denotes mean of the series. Similarly we can see if process is represented by symbol MA (2) then the current value of J is given as

$$G_t - \mu = b_1 * \epsilon_{t-1} + b_2 * \epsilon_{t-2} + \epsilon_t,$$

In terms of effect on series AR component which shows a long term pattern while MA component which shows a short term pattern.

3.6 ARMA Process

ARMA process depicting behavior of both AR process and MA process, hence called ARMA process. So there is both long term effect on current values of previous values as well as short term effect of noise of previous values. ARMA (p, q) denotes an ARMA process where p depicts order for AR process and q denotes order for MA process.

Let G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}, G_{t-5}, G_{t-6}, G_{t-7}, G_{t-8}…… G_{t-k}, be a given values of a time-series process. If the series is represented by symbol ARMA (1, 1), then current value of G that is G_t will be given as follows.

$$G_t = a_1 * G_{t-1} + \epsilon_t + b_1 * \epsilon_{t-1},$$

where a1 denotes total or quantified impact on G_t, ϵt denotes error at time t and b1 denotes total or quantified impact on ϵt-1. Similarly we can see if process is represented by symbol ARMA (2, 1) then the current value of $G_t$is given as

$$J_t = a_1 * J_{t-1} + a_2 * J_{t-2} + \epsilon_t + b_1 * \epsilon_{t-1},$$

3.7 Working of Box-Jenkins Approach

Now with understanding of the AR, MA, and ARMA process we are in position to apply Box-Jenkins method to time series prediction. One of the most important condition required to apply Box-Jenkins approach on any time series process is to make sure the time series is stationary. If the time series is not stationary it becomes necessary to make the series stationary so that Box-Jenkins approach could be applied on the series.

A time-series can be surely termed as stationary if we there are no patterns or seasonality trends identified in the series. In statistical terms we can say there is no systematized change in mean or variance of the series. In practical problems time series process are mostly non-stationary. Hence the first step is to make the series stationary. Stationarity of the series can be tested both visual and statistical methods. Statistical techniques are preferred over the visual techniques in practical problems.

Plotting Rolling statistics is one of the visual techniques used to check whether the process under investigation is stationary or not. A graph is plotted to check if the moving variance or moving average varies with respect to time. Moving variance/average means taking cumulative values for a defined period of time. Dickey-Fuller test is a statistical test used to check for stationarity of the series. In this technique null hypotheses is that the series is not stationary. After performing the DF test on the series the P-value is noted down. If the P-value comes out to be lesser than 5% then null hypotheses can be rejected that means the series is stationary, otherwise if the P-value is greater than 5% the null hypotheses is accepted and series is considered as non-stationary.

Now comes the next step, if the series is predicted as a non-stationary, series need to be made stationary as it is pre-condition of applying Box-Jenkins approach on time series process. By making series stationary means removing trends and seasonality from the series. Few techniques which can be used to make the series stationary are Aggregation, Smoothing and Polynomial fitting.

Moving Average, in this approach we club particular number of values together to reduce the effect of trends and seasonality in the time series. Suppose if we have monthly data of a certain process we can combine the values of a time period of say twelve months by taking average of those values to make the series stationary. There is better version of this technique known as weighted moving average which gives more weightage to the close by values as they depict better change in the current values.

Differencing, another technique widely used to remove trends from the time series process. In this particular technique the difference is taken at a particular value with the previous value. If we the difference is taken with just previous value it is known as first order differencing.

Decomposition, this techniques handles both trends and seasonality separately. It removes both trends and seasonality and return the stationary time series.

Once the time-series is made stationary, comes the important step to recognize the type of time series process. This step is important as correct identification of process determines the accuracy of the model prepared for prediction of values of the time series process. One need to identify out of AR, MA, and ARMA process which category the times series process belongs. Once the process has been successfully identified, next is to predict the order the process finalized.

Identification of the process type cannot done by visually analyzing the graph of time series process. Certain metric are required to correctly identify the process type. PACF and ACF functions and plots are used to identify to which process the time series truly belong.

For identification of the AR process ACF function and its plot is used. The rule to be followed is to check the ACF plot, if the plot dies down to zero or shows a tendency of reducing to zero then we can safely say that the time series process belongs to AR process. Now to identify the order of AR process, PACF plot need to be checked. The point where the PACF plot cuts off indicated the order of the AR process.

For identification of the MA process PACF function and its plot is used. The rule to be followed is to check the PACF plot, if the plot dies down to zero or shows a tendency of reducing to zero

then we can safely say that the time series process belongs to MA process. Sometimes PACF plot does not clearly show the trends, hence IACF plot is used in that case. IACF plot is nothing but just inverse of ACF plot. Now to identify the order of MA process, ACF plot need to be checked. The point where the ACF plot cuts off indicated the order of the MA process.

For identification of the ARMA process all the three plots ACF, PACF, IACF need to be considered. If all the three plots damp down to zero, it can be easily concluded that the process in hand in ARMA type of process. Identification of ARMA process is easy but to find the order of ARMA process further techniques are required. Techniques like SCAN and ESACF are used to determine the order of the ARMA process. Both techniques will provide various options for the order of the ARMA process, the least complicated option is selected as the order of the ARMA process.

Once the type of process and the order of the process are identified successfully, next step is to estimate the parameters of the equation formed as per the process and its order. Optimization techniques are utilized to identify the parameters of the equation in hand.

After all the above steps of process identification, order estimation and parameter estimation, next step is to forecast the values of time series by using the equation formed.

Finally having the values forecasted for the time series process it is very important to the accuracy of the model developed. There are various techniques to test the accuracy of the model using the predicted values and the estimated values. Few of the techniques used are Mean absolute deviation, Mean absolute percentage error, and Mean square error. It always a good practice to show the error rates along with the predicted values of the time series.

Chapter 4: Proposed Approach

The very first step involved in building a times series prediction model of a time series process is to gather the relevant data. It is correctness of the data gathered on which the accuracy of the model built depends. Following is the figure depicting the initial structure of the data gathered for building a model to predict the trend of the Silver MCX at India Stock Market. Also give are the data types of the various columns in the data gathered.

	Date	Price	Open	High	Low	Vol.	Change %
0	Dec 17	39,237	37,724	39,313	36,672	219.24K	4.54%
1	Nov 17	37,532	38,814	40,208	37,422	307.73K	-3.31%
2	Oct 17	38,818	39,314	40,632	38,688	257.26K	-1.62%
3	Sep 17	39,457	39,780	41,927	39,309	291.78K	-0.86%
4	Aug 17	39,798	38,672	40,275	36,935	440.39K	2.98%

```
 DATA TYPES :
Date         object
Price        object
Open         object
High         object
Low          object
Vol.         object
Change %     object
dtype: object
```

Fig 4.1

As we can see in the above data we have a separate index, so the first step to convert the above data into time series process is to make the ‘Date’ column as index also convert format of the date.

```
In [11]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
              Price    Open    High     Low     Vol. Change %
Date
2017-12-01  39,237  37,724  39,313  36,672  219.24K    4.54%
2017-11-01  37,532  38,814  40,208  37,422  307.73K   -3.31%
2017-10-01  38,818  39,314  40,632  38,688  257.26K   -1.62%
2017-09-01  39,457  39,780  41,927  39,309  291.78K   -0.86%
2017-08-01  39,798  38,672  40,275  36,935  440.39K    2.98%
Price       object
Open        object
High        object
Low         object
Vol.        object
Change %    object
dtype: object
```

Fig 4.2

In the above figure we can see that indexing is now done on column ‘Date’ also the format of the date has been changed.

Now we need to focus on the column ‘Price’ and also need to remove delimiters and null values from the data.

```
In [12]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
Date
2017-12-01    39237
2017-11-01    37532
2017-10-01    38818
2017-09-01    39457
2017-08-01    39798
Name: Price, dtype: int64
```

Fig 4.3

Now we have the finally processed the raw data into a times series process indexed on the date column in the required format. We can start with our process of building the prediction model for time-series analysis using Box-Jenkins approach.

Below is the graph depicting the complete time series process.

Fig 4.4

Next step is to check if the given time series process is stationary or not. We will be using here rolling statistics and dicker fuller test for stationary check. Below are the results for the test performed in the initial time series process.


```
Test Stats                      -1.198737
p-value                          0.674139
#lags used                       0.000000
No. of observations used       159.000000
Cirtical value (1%)             -3.472161
Cirtical value (5%)             -2.879895
Cirtical value (10%)            -2.576557
dtype: float64
```

Fig 4.5

Here we can see the ‘p-value’ is greater than 5%, it is around 67%. Hence we can easily be said that the series under investigation is non-stationary.

So now we need to use different techniques to make the series stationary. Now we try to make series stationary by applying function of log to the time-series process. After applying the log function to the time series, the graph plotted is as follows.

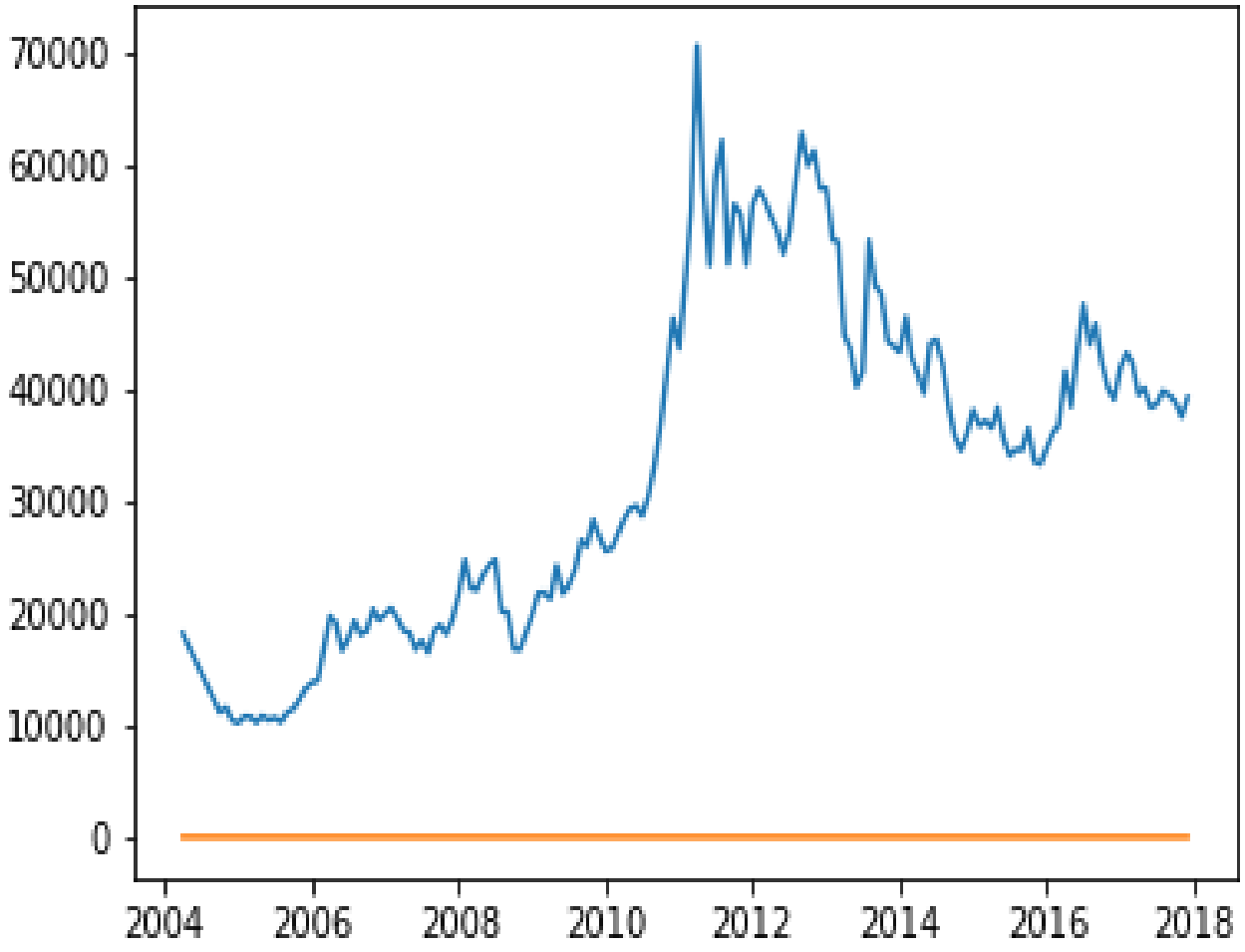
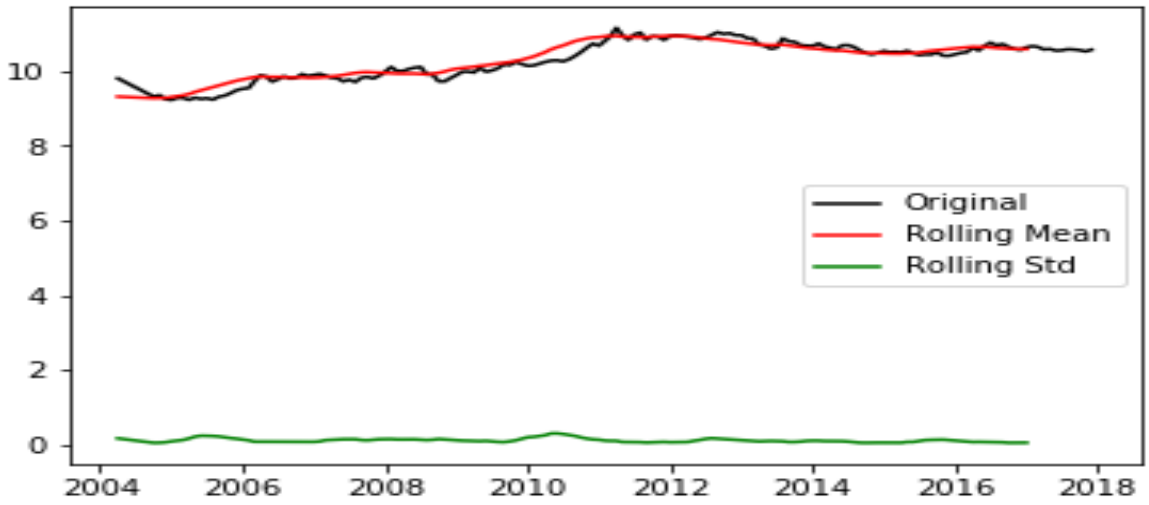


Fig 4.6

Now we test the above log value of time series for stationarity.




```
Test Stats                      -0.932920
p-value                          0.776938
#lags used                       0.000000
No. of observations used       159.000000
Cirtical value (1%)             -3.472161
Cirtical value (5%)             -2.879895
Cirtical value (10%)            -2.576557
dtype: float64
```

Fig 4.7

Here we can see that the ‘p-value’ is still not less than 5%, hence time series process requires further processing to make the series stationary.

Now we try to use next techniques were weights are assigned to the previous values with decay factor. Let us see the results of the same on time series process.

```
In [30]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
Date
2017-12-01    10.577375
2017-11-01    10.554521
2017-10-01    10.558796
2017-09-01    10.565372
2017-08-01    10.571234
Name: Price, dtype: float64
```

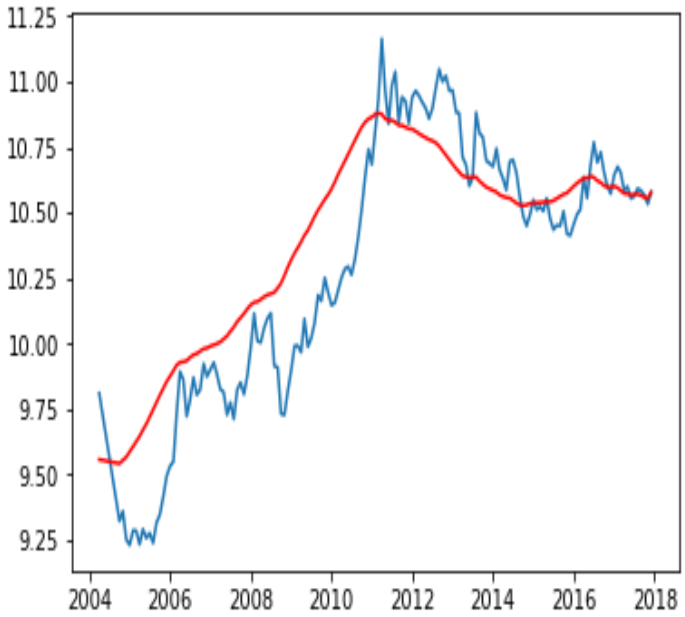
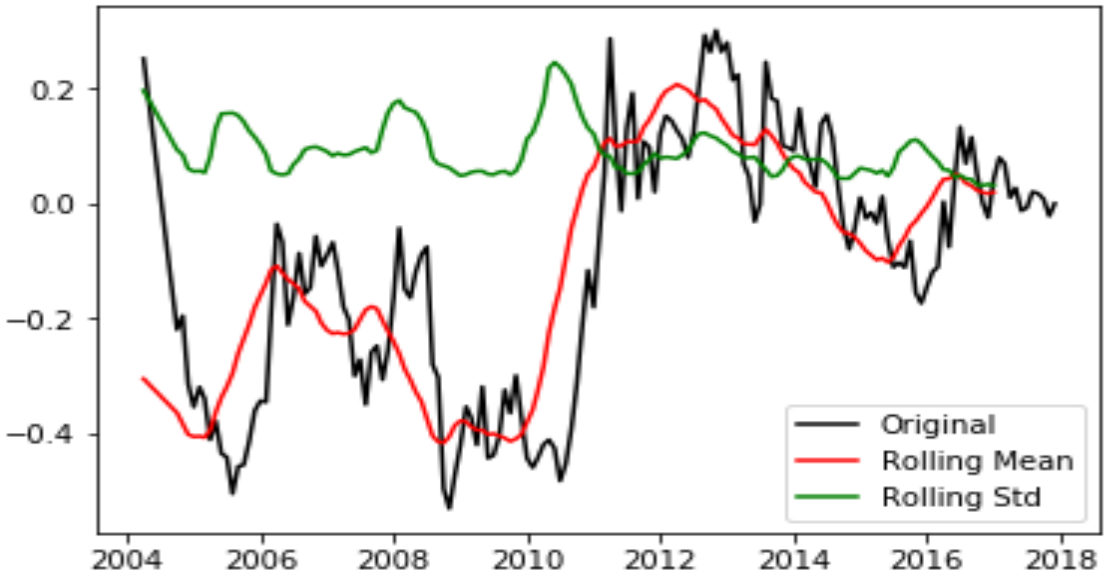


Fig 4.8

Now run the stationary test of the weighted time series process.




```
Test Stats                      -2.189452
p-value                          0.210096
#lags used                       0.000000
No. of observations used       159.000000
Cirtical value (1%)             -3.472161
Cirtical value (5%)             -2.879895
Cirtical value (10%)            -2.576557
dtype: float64
```

Fig 4.9

Here we can see the ‘p-value’ value has dropped considerably but still is above 5% threshold.

Now we use the most common technique known as differencing which takes care of both seasonality and trends. Let us see the plot of the times series after differencing.

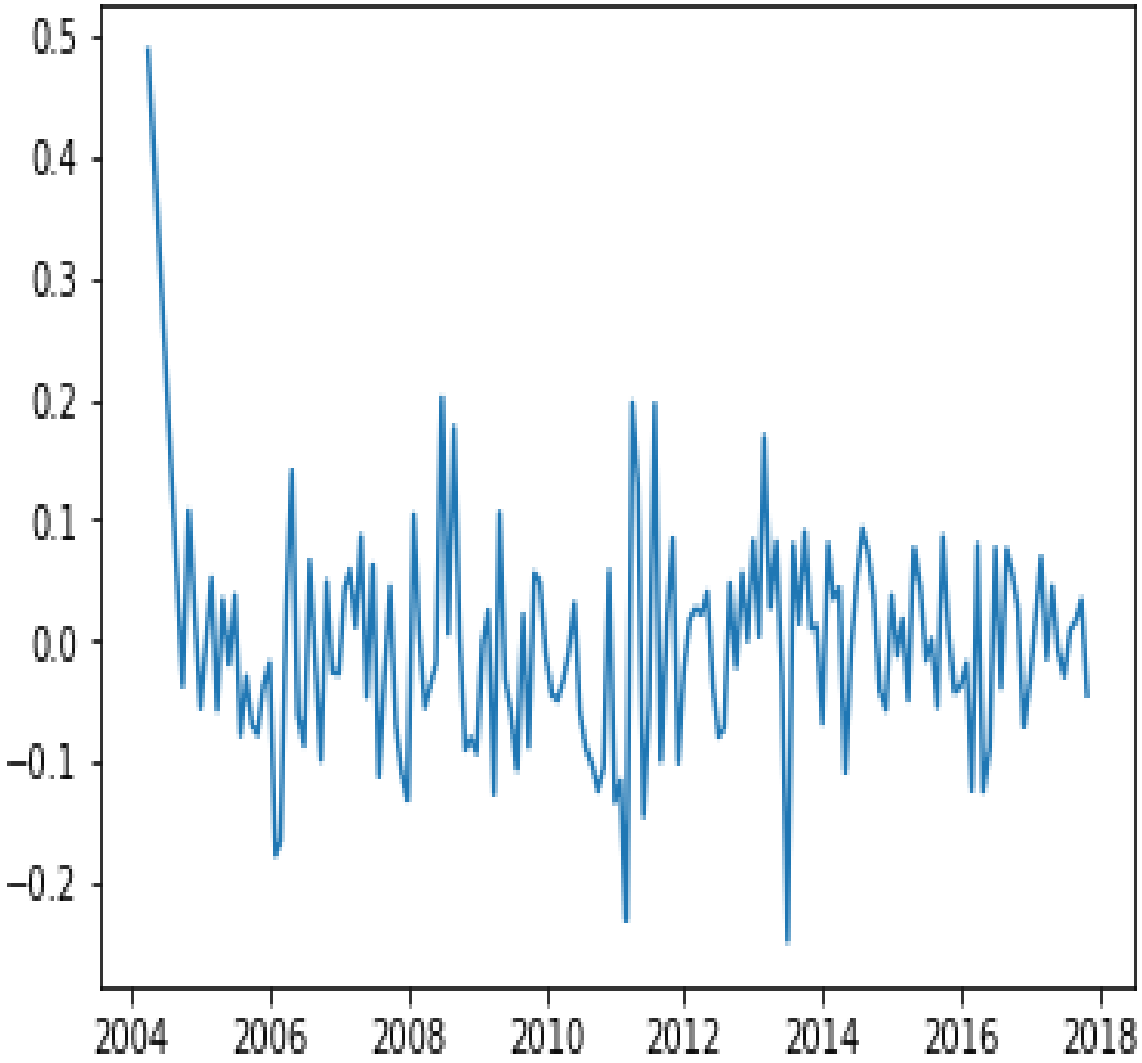
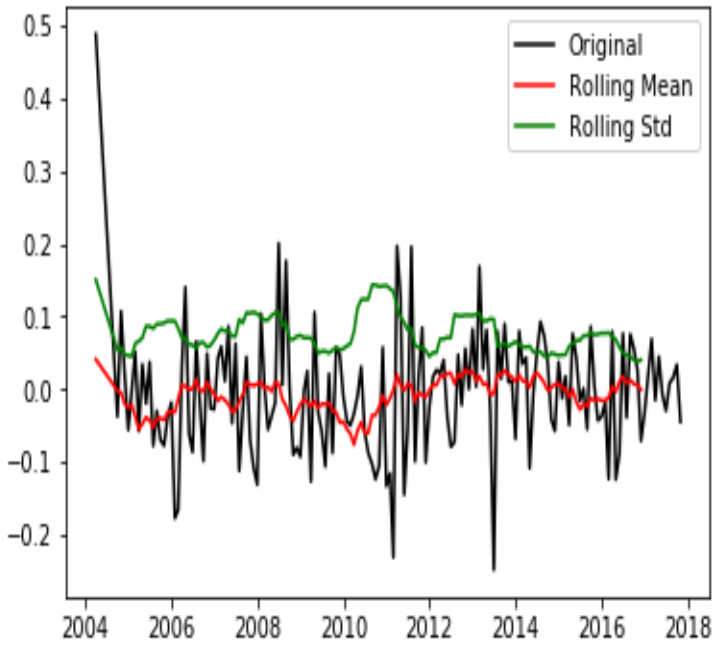


Fig 4.10

Now we perform stationarity test on the above time series process.

```
In [36]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```




```
Test Stats                      -1.139461e+01
p-value                          7.914642e-21
#lags used                       0.000000e+00
No. of observations used         1.580000e+02
Cirtical value (1%)             -3.472431e+00
Cirtical value (5%)             -2.880013e+00
Cirtical value (10%)            -2.576619e+00
dtype: float64
```

Fig 4.11

Here we can clearly see that the ‘p-value’ is lower than the 5%, hence we have made the time series process stationary using differencing technique.

Now once the series has been made stationary we can start with process identification, whether time series process belongs to AR, MA, or ARIMA process. For that purpose we need to plot two graphs, ACF and PACF graphs.

ACF GRAPH

```
In [37]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```

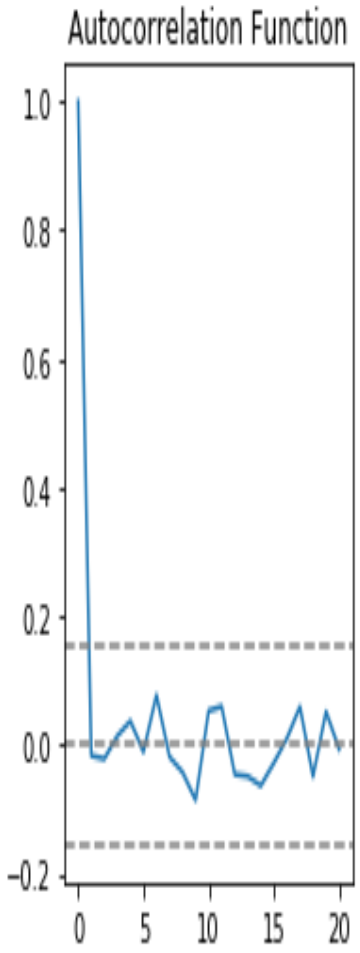


Fig 4.12

PACF GRAPH

```
In [39]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```

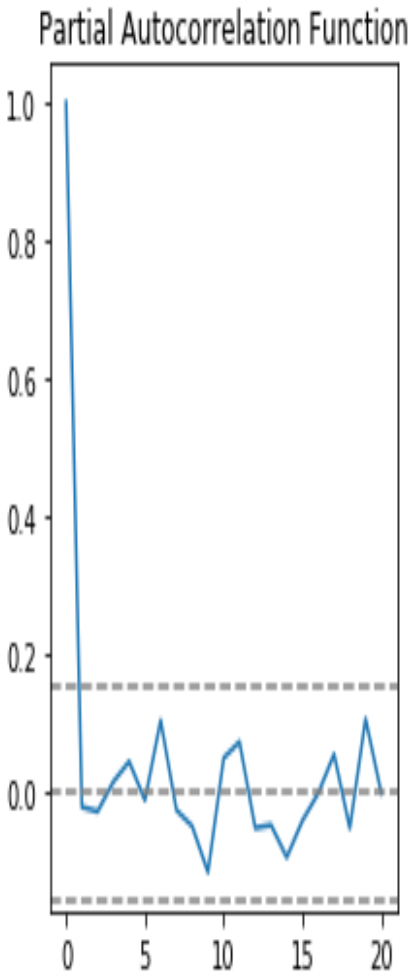


Fig 4.13

Now using the above graph we will prepare all the three models that is AR, MA, ARIMA process, the one with lowest residual factor will be selected.

AR MODEL

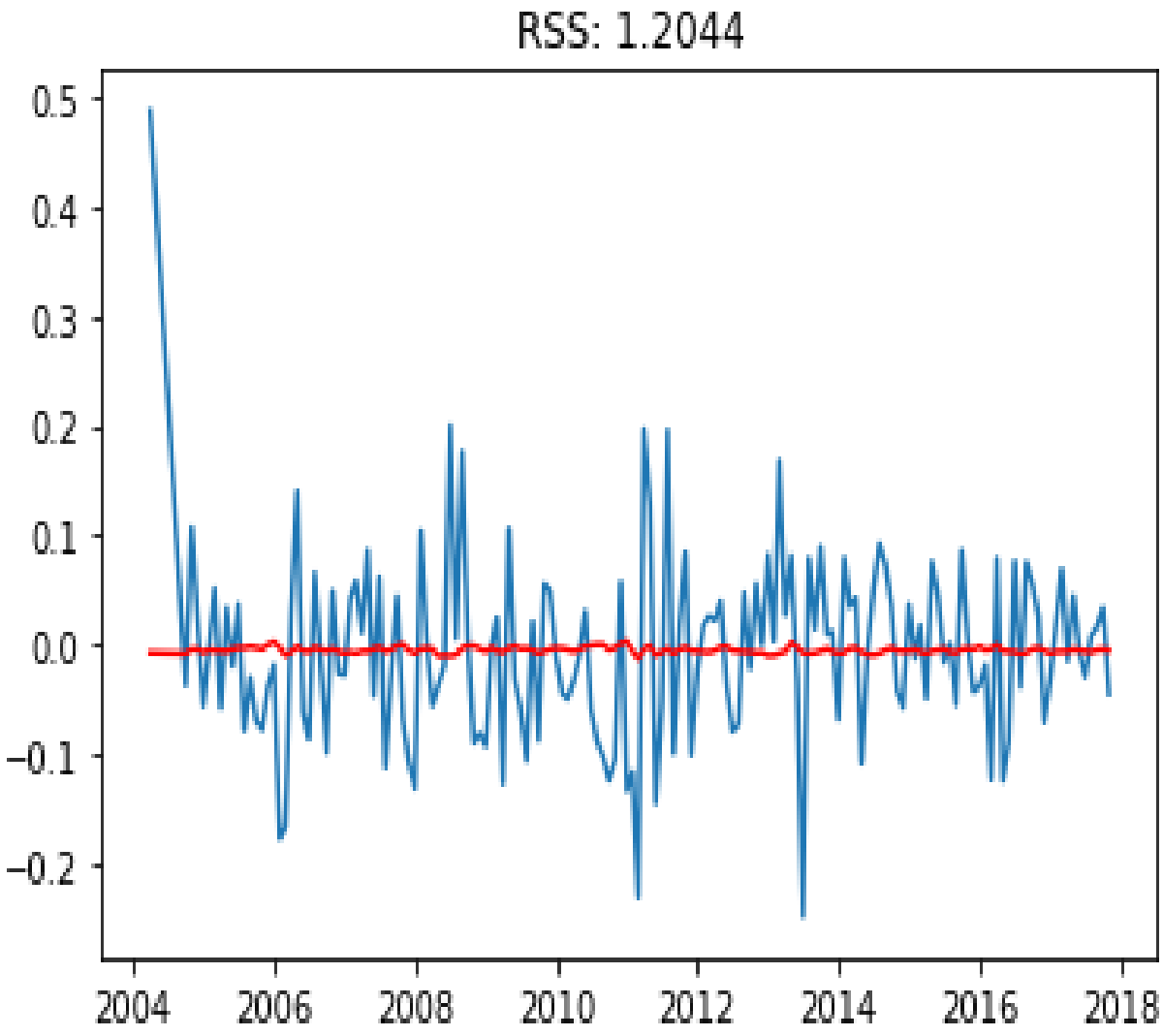


Fig 4.14

MA MODEL

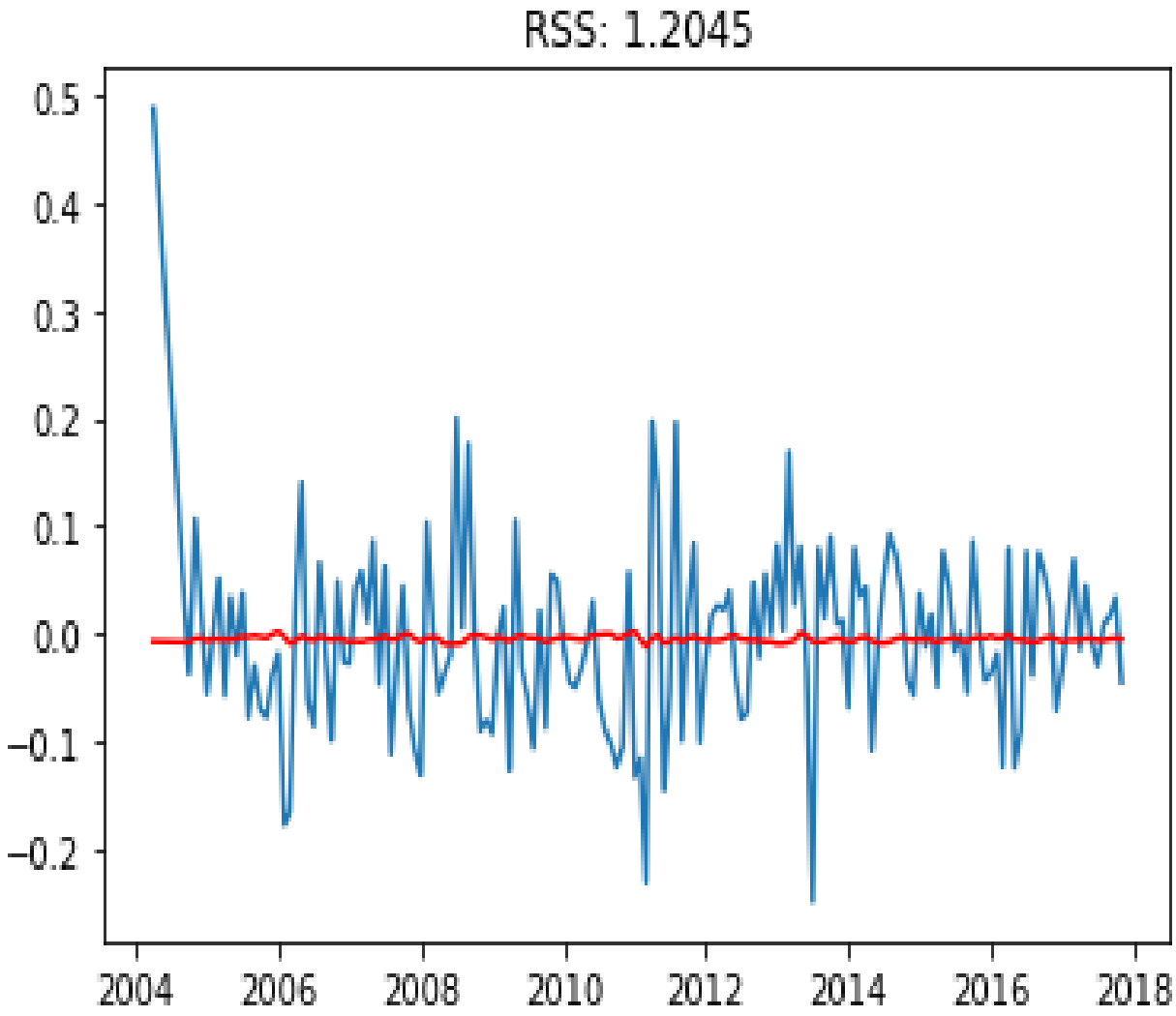


Fig 4.15

ARIMA MODEL

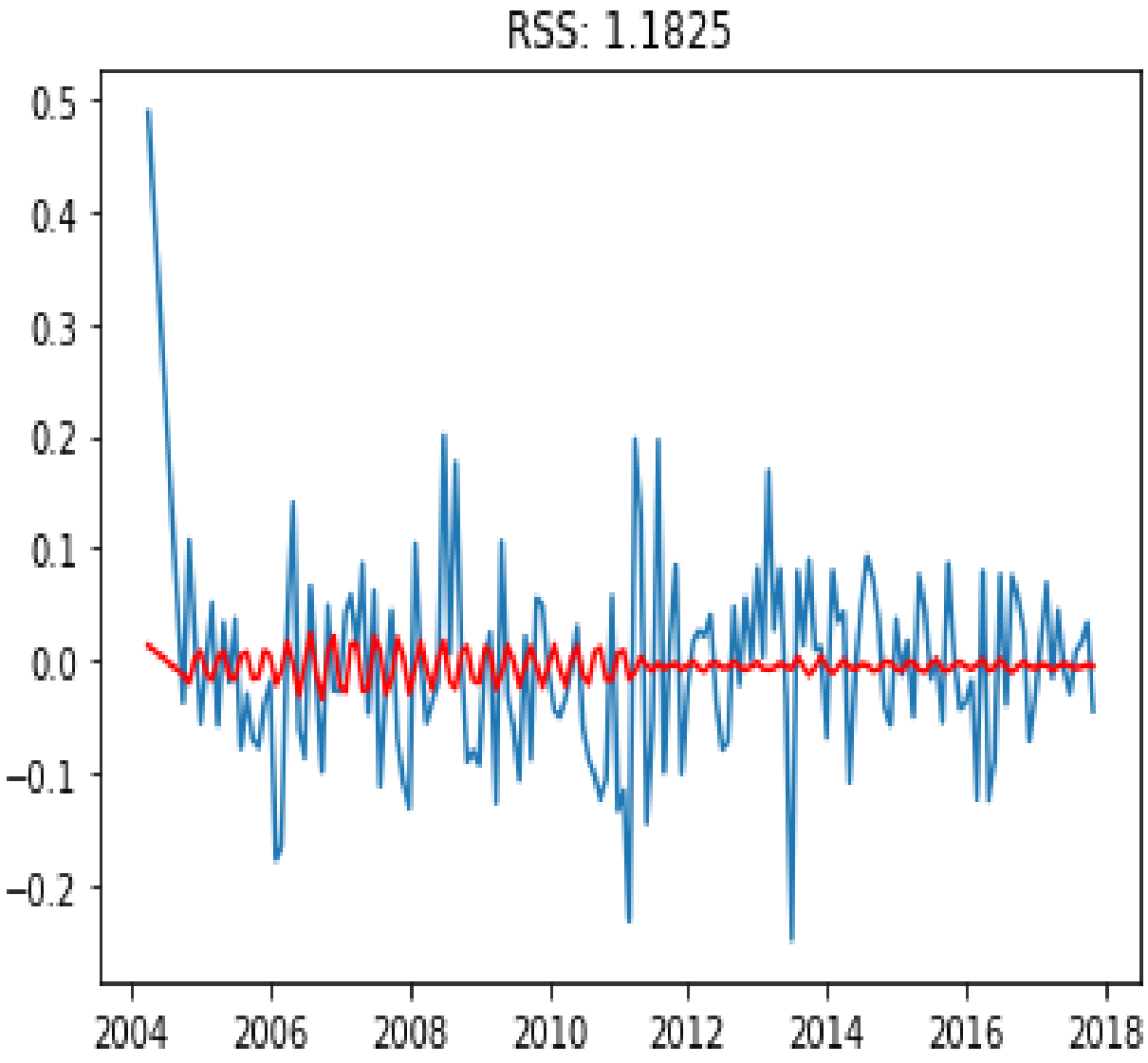


Fig 4.16

We can see RSS value for AR and MA is almost same, whereas combined model of ARIMA has a lower RSS value , so we move forward with the ARIMA model.

Now we have selected our final model we can now compare the results of the output of our model with the initial data we had gathered.

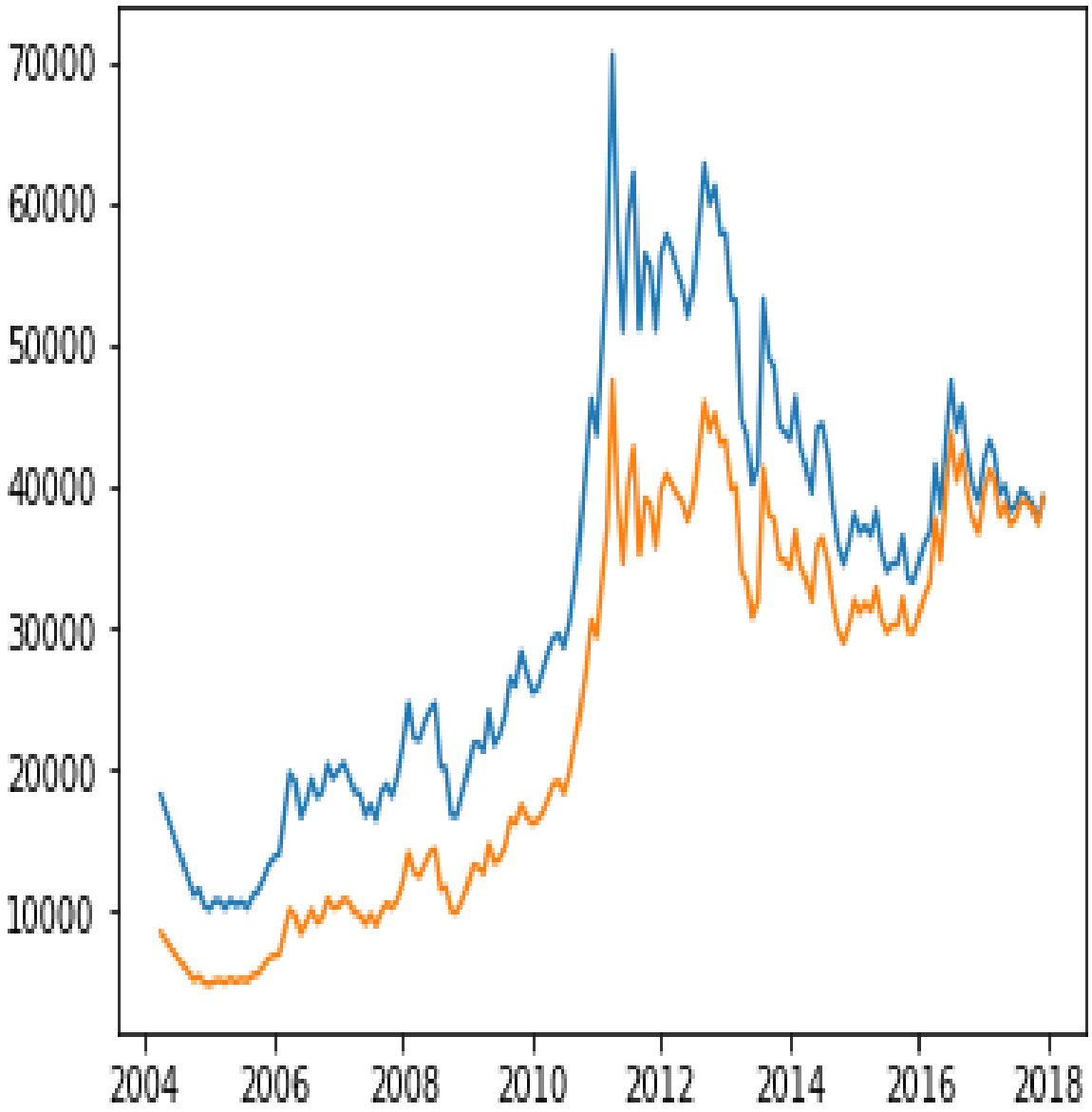


Fig 4.17

Chapter 5: Results

5.1 Data Assembly and Preprocessing Details

The raw data that is the price of Silver commodity was gathered as an initial step. This raw data was required to study the trend of the price of Silver in Indian Stock market for last few years. Therefore daily price of Silver MCX at Indian Stock market was fetched for last 14 years that is from 2004 to 2018. This volume of data helped to train the prediction model and achieve desirable results for the project of predicting trend of price of Silver MCX in Indian Stock Market. Also the data fetched was in raw form so it required a lot of understanding and pre-processing. Firstly the format of the data was important to be understood so that the data could be converted into the desired format. The data fetched was in excel sheet and the data had various fields associated with it. Then it required to check for data types of each field in the data. The data had different indexing from a time-series process, hence data needed to be converted into time series process by making indexing on date column of the data.

Then data was searched for various delimiters such as commas, dollar signs, full stops and etc. These delimiters were then removed from the data to help to use the data in training the prediction model in hand. Also fields in data which were not required were removed from the data by performing selective query on the entire data.

Once the raw data was converted into a time-series process successfully, next step was to check for stationarity of the time-series process. The time-series process in hand was not found stationary. Hence it was important to make the time-series process stationary first before the building of model could be started with. 'P-value' test was performed on the initial time-series process to confirm that the process in hand is not stationary. The 'p-value' obtained was higher than 5 % hence it was confirmed the process is not stationary. Different techniques were used to convert the time-series process stationary thereafter.

Firstly log function was used to make the time-series process stationary. After applying log function still the time-series process was not found stationary as the 'p-value' was still above 5%. Next approached applied to make the time-series process stationary was to apply a decay factor or weights to the foregoing values in the time-series process. Then stationary test was performed again, 'p-values decreased significantly but still were above 5%. Finally the most effective technique of making the time-series process stationary was applied that is differencing was applied to remove trends and seasonality from the time-series process. The stationarity test performed after the differencing technique showed that the time–series process had been made stationary as the 'p-value' obtained was less than 5%.

5.2 Process Identification

Once the raw data was successfully processed and time-series process obtained then was made stationary. Then came the step to identify the process type of the time-series process to be used to predict the trend of Silver MCX price at Indian Stock Market. There are three different types of projects Auto-Regressive process, Moving-Average process and Auto-Regressive Integrated Moving-Average process. To identify the type of process to which the time-series process in-hand belongs required to plot two graphs. The two graphs plotted were Auto-correlation Function graph and Partial Auto-correlation Function graph. For identifying the time-series process as AR, MA or ARIMA process graphs were plotted and corresponding residual factor was noted down.

The one with the lowest residual factor was selected. The residual factor values obtained for AR process was 1.2044, for MA process was 1.2045 and for ARIMA process was 1.1825. Hence the time-series process was categorized as ARIMA process.

RESIDUAL FACTOR WITH ARIMA PROCESS

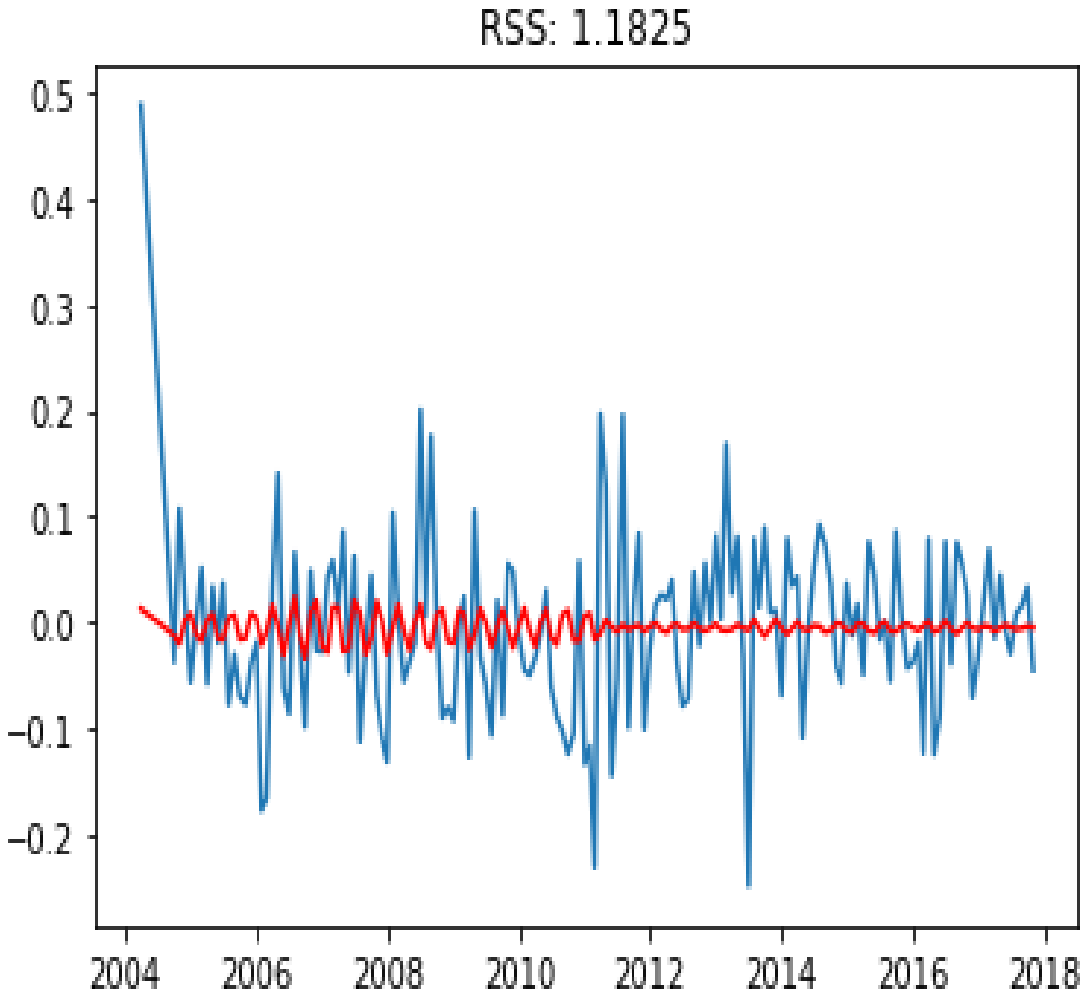


Fig 5.1

5.3 Final Results

After the process was identified as Auto-Regressive Integrated Moving-Average process, then the model was run and the following was the output obtained with comparison to the original data.

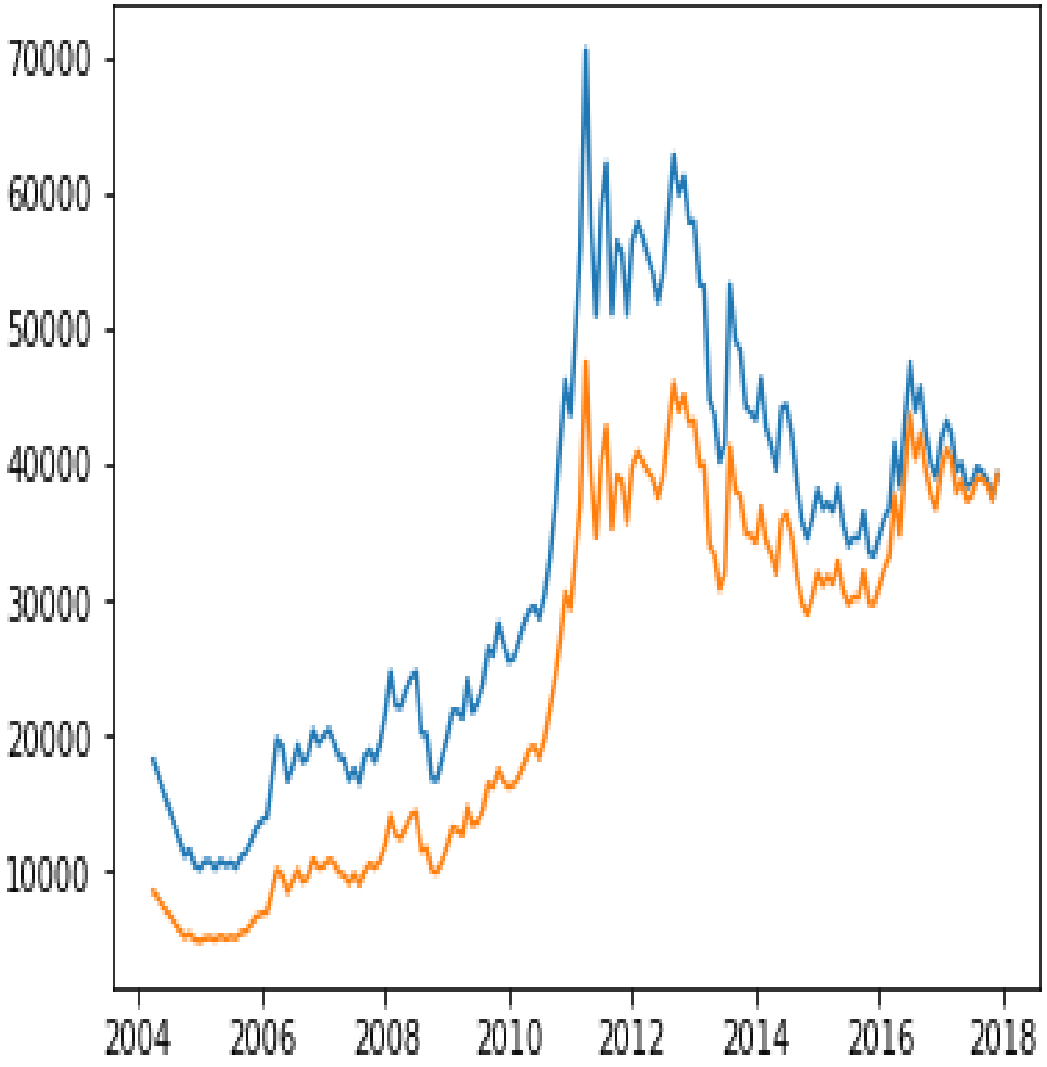


Fig 5.2

The original data is represented in the above graph with blue line whereas the predicted values of Silver MCX price by the prediction model are represented by orange line in the graph. It can be seen that predicted values by the prediction model based on Box-Jenkins approach successfully predict the trend of the values when compared to the trend of the original values. It can be noted that the trend is similar of prediction model when compared to the original data. But the predicted values differ from the original values.

Chapter 6: Conclusion

6.1 Conclusive Results

Box-Jenkins method is one the powerful techniques used in Time series analysis of the time series process. Using this approach the trend for the value of Silver MCX at Indian stock market was predicted with very good accuracy. Project helped to understand the importance of gathering the right data for building a successful project. Data processing is also a very important part of time series analysis, as various problems of formatting and null values was dealt with during the project.

Times series need to be stationary, this being one of the important pre-condition of time series analysis. Different techniques of making the times series stationary we used, Differencing came out to be the most effective technique handling both trend and seasonality at the same time. Model was prepared using all three available methods of AR, MA, and ARIMA model. ARIMA model proved to be the best fit with the least residual factor.

6.2 Future Work

The future task would be to determine accurate values of Silver MCX price at Indian Stock Market. For this it would require to understand all the factors on which the Silver MCX price depends upon. So the prediction model will be based on multiple variables.

REFERENCES

[1] Airline passenger forecasting using neural networks and Box–Jenkins, S.M.T. Fatemi Ghomi and K. Forghani Department of Industrial Engineering Amirkabir University of Technology Tehran, Iran.

[2] Time series forecasting using improved ARIMA, Soheila Mehrmolaei Computer Engineering Qazvin Branch, Islamic Azad University Qazvin, Iran

[3] ARIMA Implementation to Predict the Amount of Antiseptic Medicine Usage in Veterinary Hospital, Hans Pratyaksa, Adhistya Erna Permanasari, Silmi Fauziati, Ida Fitriana,Department of Electrical Engineering and Information Technology, Department of Pharmacology Universitas Gadjah Mada ,Indonesia

[4] Prediction of Rupiah Against US Dollar by Using ARIMA, Adiba Qonita, Annas Gading Pertiwi, Triyanna Widiyaningtyas,Electrical Engineering Department,Universitas Negeri Malang,Malang, Indonesia

[5] Short-term Traffic Flow Prediction Using a Methodology Based on ARIMA and RBF-ANN,Kui-lin Li,Chun-jie Zhai,Jian-min Xu,School of Automatic Science and Engineering South China University of Technology Guangzhou,China

[6] Forecasting Method of Aero-Material Consumption Rate Based on Seasonal ARIMA Model, Yanming Yang, Chenyu Liu, Feng Guo, Qingdao Campus, Naval Aeronautical University, Qingdao 266041, China

[7] Forecasting of Raw Material Needed for Plastic Products Based in Income Data Using ARIMA Method, Baihaqi Siregar, Erna Budhiarti Nababan, Alexander, Yap, Ulfi Andayani, Department of Information Technology, University of Sumatera Utara, Medan, Indonesia

[8] Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network, Yulin Du, School of Management ,Fudan University ,Shanghai,China