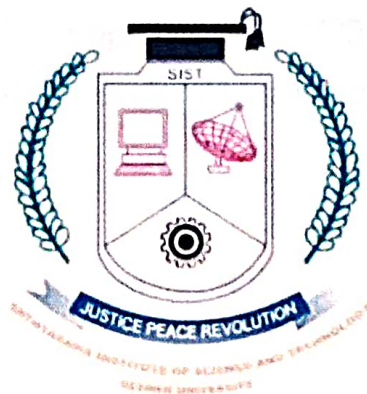


FORECASTING OF CRUDE OIL RATE USING ARIMA MODEL

Submitted in partial fulfillment of the requirements for the
award of Master of Science Degree in Mathematics.
(M.Sc. Mathematics)

By

MADHUMITHA T (40630005)



**DEPARTMENT OF MATHEMATICS
SCHOOL OF SCIENCE AND HUMANITIES**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

MAY - 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

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

www.sathyabama.ac.in

DEPARTMENT OF MATHEMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **MADHUMITHA T (40630005)** who carried out the project entitled "**FORECASTING OF CRUDE OIL RATES USING ARIMA MODEL**" under our supervision from **November 2021** to **May 2022**.

Dr . E. PRIYADARSHINI

Internal Guide

(Name in capital letters
with signature)

External Guide(if Applicable)

(Name in capital letters
with signature)

Head of the Department

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

Head of the Department

Dept. of Mathematics

(Name in Capital letters with Seal & Signature)

Sathyabama Institute of Science and Technology
(Deemed to be University), Chennai-600 119.

Submitted for Viva voce Examination held on _____

R. Manoharan

Internal Examiner

(Name in capital letters
with signature)

(Dr. R. Manoharan)

External Examiner

(Name in capital letters
with signature)

(S. Ganesan)

DECLARATION

I Madhumitha T (40630005) hereby declare that the Project Report entitled "FORECASTING OF CRUDE OIL RATE USING ARIMA MODEL" was done by me under the guidance of Dr. E. Priyadarshini (Internal) and _____ (External) at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Master of science degree in Mathematics.

DATE : 12.05.2022

PLACE: CHENNAI

Madhumitha.T
SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

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

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

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

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

ABSTRACT

Time series forecasts are the most important quantitative models with significant interest in the literature. The main objective of this paper is to help decision-makers to reduce the risks of crude oil and improve the accuracy of crude oil price forecasts. The average crude oil rate from May 2021 to December 2021 were analyzed by Box – Jenkins model. Autocorrelation and Partial Autocorrelation functions were calculated for the data and appropriate Box-Jenkins Autoregressive Integrated Moving Average model was fitted. Validity of the model was tested using standard statistical error measures like MSE, RMSE and MAPE.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	V
	LIST OF TABLES	Viii
	LIST OF FIGURES	IX
	LIST OF SYMBOLS AND ABBREVIATIONS	X
1	INTRODUCTION TO FORECASTING OF ARIMA MODEL	2
	1.1 Introduction	2
	1.2 Crude oil characteristics	3
	1.3 Application of Crude oil usages in various purposes	4
	1.3.1 Plastics	4
	1.3.2 Clothing Materials	4
	1.3.3 Food Production	5
	1.3.4 Transportation	5
	1.3.5 Home Products	5
	1.3.6 Crude oil Production	6
2	LITERATURE SURVEY	9
3	RESEARCH METHODOLOGY	13
	3.1 Fitting Versus Forecasting	13
	3.2 The Ljung-Box Test	14

	3.3 Kolmogorov-Smirnov Test for Normality	14
	3.4 Normal Q-Q Plot	15
	3.5 Akaike and Schwarz Bayesian Information Criteria	15
4	ARIMA METHODOLOGY	17
	4.1 Forecasting with ARIMA Model	17
	4.2 ARIMA(AUTO-REGRESSIVE MOVING AVERAGE)MODEL	17
	4.2.1 Box- Jenkins Model Identification	18
	4.2.2 Model Parameter Estimation	20
	4.2.3 Diagnostic Checking for Model Forecasting	20
	4.2.4 Forecasting	21
	4.2.5 ARIMA Model Assumptions	22
	4.2.6 Auto Correlation Function	23
5	ANALYSIS OF CRUDE OIL RATES USING ARIMA	25
	5.1 Data Analysis	25
	5.1.1 Model Identification	25
	5.1.2 Model Estimation and Verification	33
	5.2 Forecasting with ARIMA Model	33
6	CONCLUSION	37
	REFERENCES	38

LIST OF TABLES

TABLE NO	TITLE	PAGE NO
4.1	IDENTIFICATION RULES FOR ARIMA ORDERS	21
5.1	MODEL FIT OF ARIMA (0,1,1)	26
5.2	MODEL FIT OF ARIMA (0,1,2)	27
5.3	MODEL FIT OF ARIMA (1,1,0)	28
5.4	MODEL FIT OF ARIMA (1,1,1))	29
5.5	MODEL FIT OF ARIMA (2,1,0)	30
5.6	MODEL FIT OF ARIMA (2,1,1)	31
5.7	MODEL FIT OF ARIMA (2,1,2)	32
5.8	ANALYSIS OF ARIMA MODELS	33
5.9	ERROR MEASURES OF THE FITTED ARIMA MODEL	34

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
4.1	Flow chart for ARIMA methodology	19
5.1	Observed versus fit of crude oil rates using ARIMA model (0,1,1)	26
5.2	Observed versus fit of crude oil rates using ARIMA model(0,1,2)	27
5.3	Observed versus fit of crude oil rates using ARIMA model(1,1,0)	28
5.4	Observed versus fit of crude oil rates using ARIMA model(1,1,1)	29
5.5	Observed versus fit of crude oil rates using ARIMA model(2,1,0)	30
5.6	Observed versus fit of crude oil rates using ARIMA model(2,1,1)	31
5.7	Observed versus fit of crude oil rates using ARIMA model(2,1,2)	32
5.8	ACF and PACF of residuals of fitted ARIMA model	34
5.9	Forecast of crude oil rate	35

LIST OF SYMBOLS AND ABBREVIATION

INR	- Indian Rupee
ACF	- Auto Correlation Function
PACF	- Partial Auto Correlation Function
ARIMA	- Auto Regressive Integrated Moving Average
BIC	- Bayesian Information Criterion
DF	- Degrees of Freedom
MAPE	- Mean Absolute Percentage Error
MaxAE	- Maximum Absolute Error
MaxAPE	- Maximum Absolute Percentage Error
MAE	- Mean Absolute Error
MSE	- Mean Squared Error
RMSE	- Root Mean Square Error
SPSS	- Statistical Package for the Social Sciences
SE	- Standard Error

CHAPTER 1

INTRODUCTION TO FORECASTING

1.1 INTRODUCTION

Crude oil, liquid petroleum that is found accumulated in various porous rock formations in Earth's crust and is extracted for burning as fuel or for processing into chemical products. Crude oil is a naturally occurring petroleum product composed of hydrocarbon deposits and other organic materials. A type of fossil fuel, crude oil is refined to produce usable products including gasoline, diesel, and various other forms of petrochemicals. Crude oil is typically obtained through drilling, where it is usually found alongside other resources, such as natural gas (which is lighter and therefore sits above the crude oil). Although it is often called "black gold," crude oil has a range of viscosity and can vary in colour from black to yellow depending on its hydrocarbon composition. Distillation, the process by which oil is heated and separated into different components, is the first stage in refining. Although fossil fuels like coal have been harvested for centuries, crude oil was first discovered and developed during the Industrial Revolution, and its industrial uses were developed in the 19th century. Newly invented machines revolutionized the way we do work, and they depended on these resources to run. Today, the world's economy is largely dependent on fossil fuels such as crude oil, and the demand for these resources often sparks political unrest, as a small number of countries control the largest reservoirs. Like any industry, supply and demand heavily affect the prices and profitability of crude oil. The United States, Saudi Arabia, and Russia are the leading producers of oil in the world.

1.2 CHARACTERISTICS

Crude oil is a mixture of comparatively volatile liquid hydrocarbons (compounds composed mainly of hydrogen and carbon), though it also contains some nitrogen, sulfur, and oxygen. Those elements form a large variety of complex molecular structures, some of which cannot be readily identified. Regardless of variations, however, almost all crude oil ranges from 82 to 87 percent carbon by weight and 12 to 15 percent hydrogen by weight.

Crude oils are customarily characterized by the type of hydrocarbon compound that is most prevalent in them: paraffins, naphthenes, and aromatics. Paraffins are the most common hydrocarbons in crude oil; certain liquid paraffins are the major constituents of gasoline (petrol) and are therefore highly valued. Naphthenes are an important part of all liquid refinery products, but they also form some of the heavy asphalt like residues of refinery processes. Aromatics generally constitute only a small percentage of most crudes. The most common aromatic in crude oil is benzene, a popular building block in the petrochemical industry. Because crude oil is a mixture of such widely varying constituents and proportions, its physical properties also vary widely. In appearance, for instance, it ranges from colourless to black. Possibly the most important physical property is specific gravity (i.e., the ratio of the weight of equal volumes of a crude oil and pure water at standard conditions). In laboratory measurement of specific gravity, it is customary to assign pure water a measurement of 1; substances lighter than water, such as crude oil, would receive measurements less than 1. The petroleum industry, however, uses the American Petroleum Institute (API) gravity scale, in which pure water has been arbitrarily assigned an API gravity of 10°. Liquids lighter than water, such as oil, have API gravities numerically greater than 10. On the basis of their API gravity, crude oils can be classified as heavy, medium, and light as follows:

Crude oil also is categorized as “sweet” or “sour” depending on the level of sulfur, which occurs either as elemental sulfur or in compounds such as hydrogen sulfide.

Sweet crudes have sulfur contents of 0.5 percent or less by weight, and sour crudes have sulfur contents of 1 percent or more by weight. Generally, the heavier the crude oil, the greater its sulfur content. Excess sulfur is removed from crude oil during refining, because sulfur oxides released into the atmosphere during combustion of oil are a major pollutant.

1.3 APPLICATION OF CRUDE OIL USAGES IN VARIOUS PURPOSES

The most common use of crude oil is to create fuels for vehicles of all sizes, as well as for generators, heaters, and other equipment. Through the process of fractional distillation and cracking, this resource supports production in many industries. Even things like food need crude oil for their plastic packaging and processing. Here are some common uses for crude oil that go beyond fuel you may not have realized.

1.3.1 *PLASTICS*

As noted above, plastic requires crude oil for its production. Since there are so many products that use plastic, this resource is in high demand. You're probably reading this article on a device that uses plastic in its construction. You may be surrounded by plastic products in your home or at work. It's hard to get away from this versatile product.

1.3.2 *CLOTHING MATERIALS*

Petroleum is essential for the creation of clothing. It helps to make clothing that is colourful and non-flammable. Polyester, rayon, nylon, and some fake furs all use crude oil.

1.3.3 FOOD PRODUCTION

You may want to cringe thinking of crude oil being used in the production of food items, but it is a necessary piece of the process. Some fertilizers used to grow your food need this product. Transportation is a more “tame” use for crude oil in the industry, with commercial vehicles delivering food all over the country. The packaging that your food comes in is also made using plastic.

1.3.4 TRANSPORTATION

Your car is another item that uses crude oil in the manufacturing process. Your car’s bumper, interior, and even engine block components are made with high-performance plastics. These are used because they are lightweight and can help improve the fuel economy of the car. They can also help to enhance the safety of the vehicle.

1.3.5 HOME PRODUCTS

Believe it or not, if you’re sitting on your couch or other furniture with a cushion, you’re probably sitting on a product created using crude oil. Polyurethane foam is a petroleum-based material that is often used for couch cushions. It’s durable, so it can take a beating from years of people sitting and laying on it while being lightweight. Also, your flooring could be another hidden crude oil product in your home. Both linoleum and carpet are made using petroleum. Many of the appliances that are in your kitchen require crude oil as well. Your refrigerator is a prime example. It contains foam insulation, door liners, and molded interior panels for shelving that are all made with crude oil. One product, insulation, that you rely on to keep your home warm in the winter and stay cool in the summer is made using crude oil. Insulation helps to keep your home’s temperature regulated.

1.3.6 CRUDE OIL PRODUCTION

Crude oil occurs underground, at various pressures depending on depth. It can contain considerable natural gas, kept in solution by the pressure. In addition, water often flows into an oil well along with liquid crude and gas. All these fluids are collected by surface equipment for separation. Clean crude oil is sent to storage at near atmospheric pressure, usually aboveground in cylindrical steel tanks that may be as large as 30 metres (100 feet) in diameter and 10 metres (33 feet) tall. Often crude oil must be transported from widely distributed production sites to treatment plants and refineries. Overland movement is largely through pipelines. Crude from more isolated wells is collected in tank trucks and taken to pipeline terminals; there is also some transport in specially constructed railroad cars. Overseas transport is conducted in specially designed tanker ships. Tanker capacities vary from less than 100,000 barrels to more than 3,000,000 barrels.

The primary destination of crude oil is a refinery. There any combination of three basic functions is carried out:

(1) Separating the many types of hydrocarbon present in crude oils into fractions of more closely related properties, (2) chemically converting the separated hydrocarbons into more desirable reaction products, and (3) purifying the products of unwanted elements and compounds. The main process for separating the hydrocarbon components of crude oil is fractional distillation. Crude oil fractions separated by distillation are passed on for subsequent processing into numerous products, ranging from gasoline and diesel fuel to heating oil to asphalt.

The proportions of products that may be obtained by distillation of five typical crude oils, ranging from heavy Venezuelan Boscan to the light Bass Strait oil produced in Australia. Given the pattern of modern demand (which tends to be highest for transportation fuels such as gasoline), the market value of a crude oil generally rises with increasing yields of light products.

In the United States, the conventional practice for the petroleum industry is to measure capacity by volume and to use the English system of measurement. For this reason, crude oil in the United States is measured in barrels, each barrel containing 42 gallons of oil. Most other areas of the world define capacity by the weight of materials processed and record measurements in metric units; therefore, crude oil outside the United States is usually measured in metric tons. A barrel of API 30° light oil would weigh about 139 kg (306 pounds). Conversely, a metric ton of API 30° light oil would be equal to approximately 252 imperial gallons, or about 7.2 U.S. barrels.

CHAPTER 2

LITERATURE SURVEY

Liu (1991) used Box-Jenkins technique to study the dynamic relationships between US crude oil prices, gasoline prices. Kumar (1992) analyzed time series models to investigate and compare the forecast accuracy of future prices of crude oil. Chin *et al.* (2005) examined the predictive content of energy futures by examining the relationship between spot and futures prices for energy commodities. Sadorsky (2006) found that the out-of-sample forecasts of a single equation GARCH model are best for those of Vector Auto-regression. Agnolucci (2009) pointed out diverse kinds of GARCH models and mentioned instability models to predict daily WTI future price instability. Marzo and Zagaglia (2010) utilized numerous GARCH models to predict the instability of daily futures prices of crude oil traded on NYMEX. Hou and Saudy (2012) analyzed an alternative technique involving nonparametric method to model and forecast oil price return volatility. Ahmed and Shabri (2013) tested fitted GARCH model to crude oil spot prices. This was done in order to illustrate the advantages of nonlinear models. Morana (2001) displayed how the oil price allocation can be predicted by using the GARCH properties of oil price changes over short-term horizons. Sadorsky (1999) exposed that oil price instability alarms have dissymmetrical effects on the financial system.

Khashman and Nwulu (2011a) exhibited an intelligent system that forecasts the crude oil price. Xiao-Lin and Hai-Wei (2012) accede three basic kernel functions of SVM (Support vector machine) to build the prediction model of the crude oil price. Khashman and Nwulu (2011b) inquired and compared the applying of a back propagation neural network and an SVM to the task of forecasting oil price. Radchenko (2005) enlarges the Pindyck study with performing a long term forecast of energy prices with 127 years data that covers the same period with Pindyck, and examines forecasting ability of the model. Lanza *et al.* (2005) calculated the relationship among heavy crude oil prices and product prices. The authors execute a comparison among ten heavy crude oil price series and fourteen petroleum product price series in Europe and America. Wang *et al.* (2005) accomplished ARIMA approach to model the linear component of crude oil price time series, they use monthly WTI crude oil data from January 1970 to December 2003. Xie *et al.* (2006) predicted WTI crude oil prices with applying ARIMA method.

Fernandez (2010) showed an out-of-sample forecasting for short and long term horizons with using ARIMA model, the author employs daily natural gas and Dubai crude oil prices from 1994 to 2005. Bopp and Lady (1991) inquired the impact of lagged future and lagged spot oil prices on future spot prices. Green and Mork (1991) investigated efficiency and unbiasedness among official oil prices and ex-post spot prices. Sami (1992) examined WTI crude oil future prices as function of WTI spot prices and interest rate. Morana (2001) used a semi parametric approach proposed by Barone-Adesi et al. (1998) to forecast the volatility of Brent crude oil price, which is based on the relationship among spot and future prices. Cortazar and and Schwartz (2002) showed a financial analysis on oil prices with extension of two-factor model of Schwartz (1997) to a three-factor model and examined the relationship between spot and future prices.

Fong and See (2002) applied a Markov regime switching (MRS) model to explain the volatility of oil prices. Chernenko et al. (2004) investigated the efficiency and unbiasedness of broad sort of forward and future rates that among them we only consider crude oil and natural gas price futures. Abosedra (2005) employed a simple univariate model to examine unbiasedness and efficiency of crude oil spot and future prices. Yousefi et al. (2005) analyzed wavelet methodology in providing out-of-sample forecasted for oil prices and forecasted time horizons, and then investigated the efficiency of future contracts on spot prices. Coppola (2008) explored in long run relationship between spot and future oil prices; using weekly WTI spot and future prices traded on NYMEX.

Alizade et al. (2008) are the first to apply the markov regime switching approach to estimate the time varying minimum variance hedge ratio (Hung et al. 2011) by introducing Markov regime switching error correction model with GARCH error structure. Murat and Tokat (2009) examined the relationships between crude oil price with crack spread futures, and future prices. Kulkarni and Haidar (2009) found that applying 3 days simple moving average to the original data, then transforming it into relative change is the best method among the others tested. Silva et al. (2010) investigated the performance of hidden Markov model (HMM) to forecast the medium term future crude oil price movements. Dritsaki (2018) found that a certain

configuration of a hybrid ARIMA-GARCH model was the most suitable for forecasting the returns of oil prices over a short term. Funk (2018) tested many forecasting models and combinations of models to determine what combination of individual models is beneficial for the forecasting performance.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 FITTING VERSUS FORECASTING

If Y_i is the actual value, $\hat{Y}_i$ is the fitted or forecast value, the error is defined as

$$e_i = y_i - \hat{Y}_i$$

For a time series with n terms, the following statistics are used to measure model accuracy

$$(1) \text{ Mean Absolute Error (MAE)} = \frac{1}{n} \sum_{i=1}^n |e_i|$$

$$(2) \text{ Mean Absolute Percentage Error (MAPE)} = \frac{1}{n} \sum_{i=1}^n \left| \frac{e_i}{y_i} \right|$$

$$(3) \text{ Mean Squared Error (MSE)} = \frac{1}{n} \sum_{i=1}^n e_i^2$$

$$(4) \text{ Root Mean Square Error (RMSE)} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n}}$$

$\overline{R^2}$ is a general statistic used to judge the relative accuracy of any model. $\overline{R^2}$ is most often seen as a number between 0 and 1.0, used to describe how well a regression line fits a set of data. A $\overline{R^2}$ near 1.0 indicates that a regression line fits the data well, while a $\overline{R^2}$ closer to 0 indicates a regression line does not fit the data very well. It provides a measure of how well future outcomes are likely to be predicted by the model.

3.2 THE LJUNG-BOX TEST

The Ljung–Box test is a type of statistical test of whether any of a group of autocorrelations of a time series is different from zero. Instead of testing randomness at each distinct lag, it tests the overall randomness based on a number of lags, and is therefore a test. The Ljung–Box test can be defined as follows:

H₀: The data are independently distributed

H_a: The data are not independently distributed.

The test Statistic is:

$$Q = n(n+2) \sum_{k=1}^h \frac{\hat{\rho}_k^2}{n-k}$$

Where n is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k , and h is the number of lags being tested. For significance level α , the critical region for rejection of the hypothesis of randomness is $Q > \chi_{1-\alpha, h}^2$

Where $\chi_{1-\alpha, h}^2$ is the α -quantile of the chi-squared distribution with h degrees of freedom (Ljung and Box 1979).

3.3 KOLMOGOROV-SMIRNOV TEST FOR NORMALITY

If the sample size is larger than 50, we use the Kolmogorov-Smirnov test. If the sample size were 50 or less, we would use the Shapiro-Wilk statistic instead. The null hypothesis for the test of normality states that the actual distribution of the variable is equal to the expected distribution, i.e., the variable is normally distributed. Since the probability associated with the test of normality is < 0.001 is less than or equal to the level of significance (0.01), the null hypothesis is rejected.

3.4 NORMAL Q-Q PLOT

In order to determine normality graphically we can use the output of a normality distributed then the data points will be close to the diagonal line. If the data points stray from the line in an obvious non-linear fashion then the data are not normally distributed.

3.5 AKAIKE AND SCHWARZ BAYESIAN INFORMATION CRITERIA

Schwarz (1978) developed the Bayesian Information Criterion(BIC) and Akaike (1978) developed the Information Criterion (AIC) , for selecting models that trade off model complexity and the error in fitting so as to achieve the most accurate out of sample forecasts. They suggest selecting models that have the lowest AIC or BIC respectively.

Akaike proposed the equation,

$$AIC = n \text{ Log (SSE) } + 2k$$

Where k = number of parameters that are fitted in the model and the Log is natural logarithm.

Schwarz proposed the following criteria,

$$BIC = n \text{ Log (SSE) } + k \text{ Log (n)}$$

In the next chapter, ARIMA methodology will be discussed in detail.

CHAPTER 4

ARIMA METHODOLOGY

4.1 FORECASTING WITH ARIMA MODEL

George Box and Gwilym Jenkins (1976) had studied Auto Regressive Integrated Moving Average (ARIMA) model extensively and their names are associated with ARIMA processes for time series analysis, forecasting and control. ARIMA model building requires a large set of time series data and technical expertise on the part of the forecaster.

4.2 ARIMA (AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE) MODELING

ARIMA processes have been a popular method of forecasting because they have well developed mathematical structure from which it is possible to calculate various model features such as prediction intervals. Autoregressive Integrated Moving Average (ARIMA) is a generalization of Autoregressive moving average (ARMA) in Economics, Statistics and mainly in Time series analysis. In general, it is either used for better understanding of data or to forecast. ARIMA can be subdivided into three major steps. Evolving variable of interest is reverted on its own prior value this is designated as the first part 'AR'. The other part 'MA' stipulates that the regression error is in fact a linear combination of terms of error that occurred in the process. The last part 'I' designates that the difference between data values and the previous values have replaced the data values.

The Auto-Regressive Integrated Moving Average model is denoted by ARIMA (p,d,q) where p denotes the order of the auto regressive part, d denotes the amount of differencing and q indicates the order of moving average. An ARIMA (p,d,q) model is given by the equation

$$Y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \dots - \theta_q \epsilon_{t-q}$$

The basis of the approach of Box and Jenkins consists of four phases consisting of identification, estimation, diagnostic testing and application.

4.2.1 Box-Jenkins Model Identification

The information is inspected to check for the foremost fitting course of ARIMA forms through selecting the order of the successive and seasonal differencing required to create arrangement stationery, as well as specifying the arrangement of the standard and regular ARIMA demonstrate fundamental to satisfactorily speaks to the time arrangement demonstrate. The Autocorrelation Function (ACF) and the Fractional Autocorrelation work (PACF) are the foremost vital components of time arrangement investigation and estimating. The ACF measures the sum of linear reliance between perceptions in a time arrangement that is isolated by a slack k . The PACF plot helps to decide how numerous backward terms are vital to uncovering one or more of the following characteristics: time slacks where high correlations show up, the regularity of the arrangement, slant either within the mean level or within the fluctuation of the arrangement.

First determine whether the series is stationary by considering the Auto Correlation Function (ACF) graph. The time series can be considered non-stationary if the ACF of the time series cuts off or dies down slowly. The series can be made stationary by differencing (d) until the graph of ACF either cuts off or dies down fairly quickly. The first thing to note is that most time series are non-stationary, so that the characteristics of the underlying stochastic process change over time. The Auto-Regressive (AR) and Moving Average (MA) aspects of an ARIMA model refer only to a stationary time series. For an ARIMA (p, d, q) process, the series becomes stationary after differencing in d times. This procedure is applied to the available data, and the best ARIMA fit is constructed and then they are used for forecasting the future prices. To find provisional orders of AR and MA models, Box and Jenkins suggest examining:

- i. The sample autocorrelation function (ACF, correlogram)
- ii. The sample partial autocorrelation function (PACF)

If data is AR (p) then

- i. ACF will decline steadily, or follow a damped cycle
- ii. PACF will cut off suddenly after p lags

If data is MA (q) then

- i. ACF will cut off suddenly after q lags
- ii. PACF will decline steadily, or follow a damped cycle.

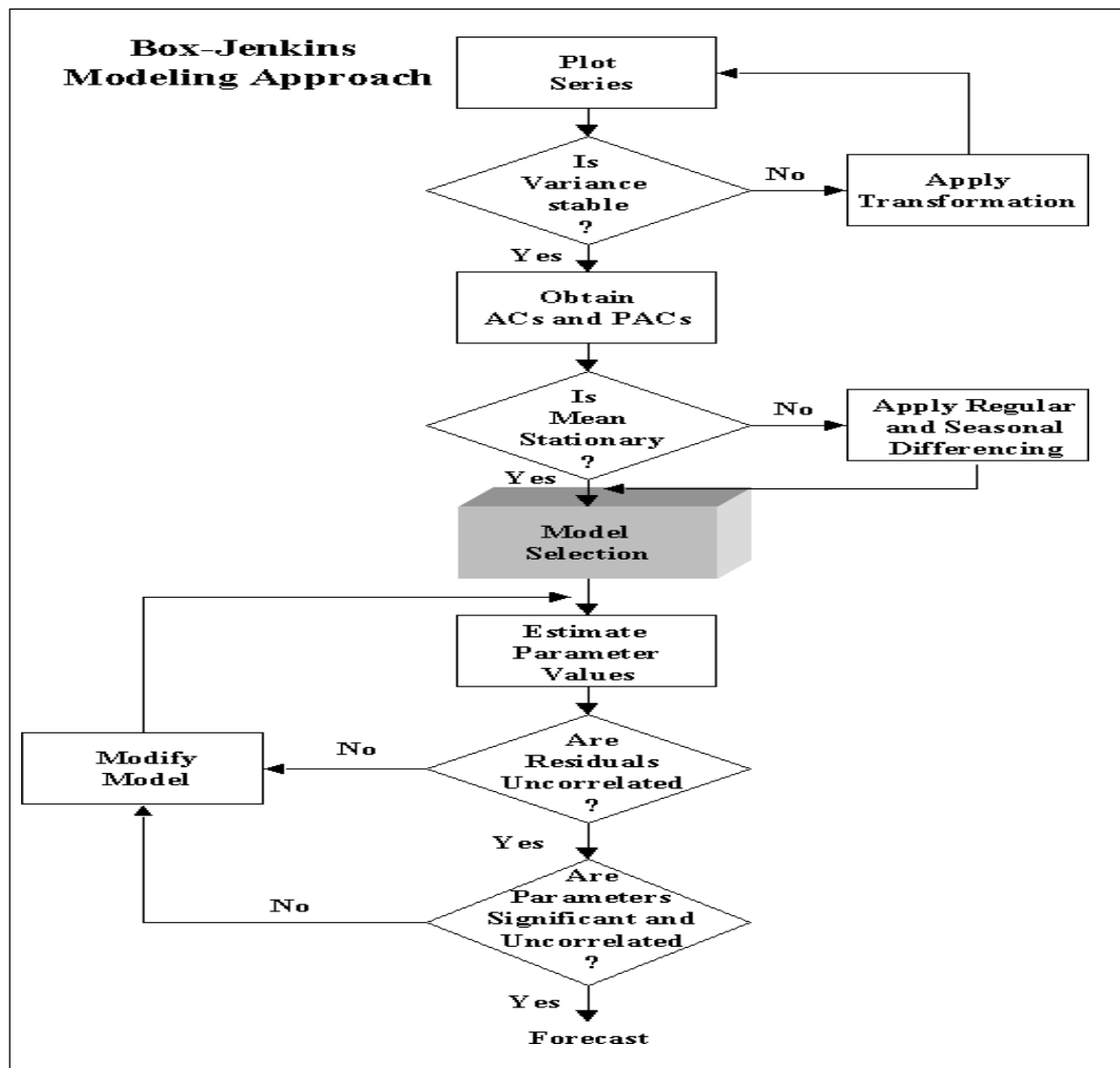


FIG 4.1 Flow chart of ARIMA Methodology

4.2.2 Model Parameter Estimation

The Box and Jenkins model incorporates autoregressive and moving normal parameters as well as differencing within the model. The three types of parameters within the process are the autoregressive parameters (p), the number of differencing (d), and moving average parameter (q). Box –Jenkins demonstrates are summarized as ARIMA (p,d,q). For the case, if the model is depicted as ARIMA (1,1,0) implies that it contains 1 autoregressive (p) parameter and 0 moving average (q) parameter for the time arrangement information after it was differenced once to accomplish stationarity.

The next step is to estimate the parameters for a tentative model. One or more models are tentatively chosen that seem to provide statistically adequate representation of the available data then we attempt to obtain precise estimates of the model by least squares as advocated by Box & Jenkins. Different models can be obtained for various combinations of AR and MA individually and collectively. For the models obtained, perform diagnostics tests using (a) Residual ACF (b) Box pierce Chi-square tests

4.2.3 Diagnostic Checking for Model Forecasting

In analyzing the checking step, the residuals from the fitted demonstrate should be inspected against accuracy of the model. This was done by relationship examination through the residual ACF plots and the goodness-of-fit test by implies of BIC and white noise utilizing the Ljung-Box Q measurement. On the off chance that the residuals are related, at that point the show should be refined as in step one over. Something else, the autocorrelations are white commotion and the show is adequate to speak to our time arrangement. Consequently, the demonstration created consolidates the fundamental measurable properties of the time arrangement information into its parameters. This three-step demonstrate building handle is regularly rehashed several times until a satisfactory model is at long last chosen. The ultimate chosen demonstrate can at that point be utilized for forecasting.

To transform the time series into a stationary one, the data is checked for the correct ARIMA process by choosing the order of the differencing so that the model represents an appropriate time series model. The most necessary elements of time series analysis are the auto correlation and partial autocorrelation functions. The amount of dependence between the observations of time series separated using a lag are measured by the ACF and time lags where high correlations appear, seasonality of the series, trend either in the mean level or in the variance of the series and the number of auto regressive terms necessary to reveal one or more of the following characteristics are determined by the PACF.

4.2.4 Forecasting

ARIMA models are developed basically to forecast the corresponding variable. There are two kinds of forecasts-sample forecasts and post sample forecasts the former are used to develop confidence in the model and the latter to generate genuine forecasts for future planning. ARIMA model can be used to yield both these kinds of forecasts.

TABLE 4.1 Identification Rules for ARIMA orders

ARIMA(p,d,q)	ACF	PACF
(0 ,0 ,0)	No significant lags	No significant lags
(0, 1 ,0)	Linear decline at lag 1 with many significant lags	Single significant peak at lag 1
(1, 0 ,0)	Exponential decline with first two or more lags	Single significant peak at lag 1

$\Phi_1 > 0$	Significant	
$(1, 0, 0)$ $\Phi_1 < 0$	Alternating exponential decline with a negative ACF (1)	Single significant negative peak at lag 1
$(0, 0, 1)$ $\theta_1 > 0$	Single significant negative peak at lag 1.	Exponential decline of negative values with first 2 or 3 lags significant
$(0, 0, 1)$ $\theta_1 < 0$	Single significant positive peak at lag 1.	Alternating exponential decline starting with positive.

4.2.5 ARIMA Model Assumptions

A time series is a linear function of past values and random shocks (i.e. error terms).

$$Y_t = f(Y_{t-k}, e_{t-k}) + e_t, k > 0,$$

$$Y_t = \text{Pattern} + e_t$$

The purpose of time series analysis is to extract the right pattern from the time series so that those are distributed as white noise (a normally and independently distributed variable –NID –having no pattern, a zero mean and an error variance lower than the variance of Y_t). The process of model fitting for the Crude oil price was done by using a computer software using SPSS 18 version.

4.2.6 Auto Correlation Function

An important guide to the properties of a time series is provided by a series of quantities called sample autocorrelation coefficient or a serial correlation coefficient, which measures the correlation between observations at different distances apart. These coefficients often provide insight into the probability model which generated the data. The sample autocorrelation coefficient is similar to the ordinary correlation coefficient between two variable (x) and (y), except that it is applied to a single time series to see if successive observation is correlated. Auto Correlation Function refers to the ways the observation in a Time Series are related to each other and is measured by the simple Correlation between Current Observation (Y_t) and observation from P Periods before the current one (Y_{t-p}) that auto correlation at lag P is the Correlation between the pair (Y_t, Y_{t-p}). In the next chapter, the analysis of crude oil rate was carried out using ARIMA modeling.

CHAPTER 5

ANALYSIS OF CRUDE OIL RATES USING ARIMA

5.1 DATA ANALYSIS

The data for analysis is taken from the average monthly Crude oil rates (in Indian Rupees) for a period of Eight months from May 2021 to December 2021 obtained from West Texas Intermediate (WTI).

5.1.1 *Model Identification*

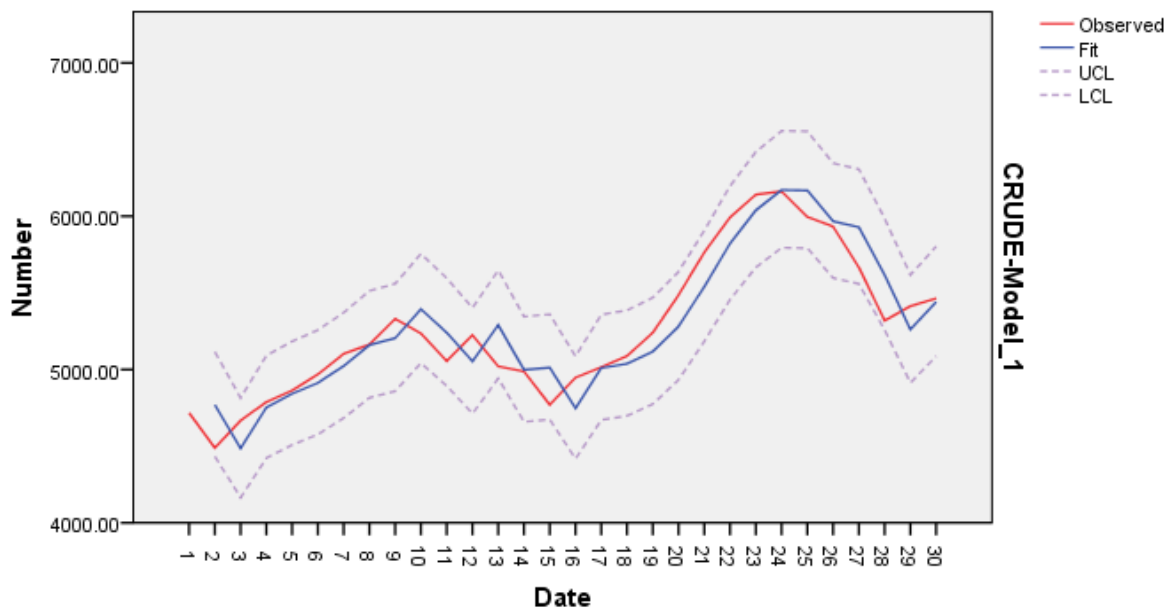
ARIMA model is estimated only after transforming the variable under forecasting into a time series. The stationary series is the one whose values vary over time only around a constant mean and constant variance. The most common method is to check stationarity through examining the time plot of the data. Non-stationarity in mean is corrected through appropriate differencing of the data. In this case difference of order 1 was sufficient to achieve stationarity in mean.

The next step is to identify the values of p and q . For this, the autocorrelation and partial autocorrelation coefficients of various orders are computed. The ACF and PACF (FIG 5.8) shows that the order of p and q can at most be seven tentative ARIMA models were examined and the model which has minimum BIC (Bayesian Information Criterion) was chosen.

i. ARIMA MODEL (0,1,1)

Table 5.1 MODEL FIT OF ARIMA (0,1,1)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary	.058	.	.058	.058	.058	.058	.058	.058	.058	.058	.058
R-squared	.866	.	.866	.866	.866	.866	.866	.866	.866	.866	.866
RMSE	171.108	.	171.108	171.108	171.108	171.108	171.108	171.108	171.108	171.108	171.108
MAPE	2.545	.	2.545	2.545	2.545	2.545	2.545	2.545	2.545	2.545	2.545
MaxAPE	6.271	.	6.271	6.271	6.271	6.271	6.271	6.271	6.271	6.271	6.271
MAE	132.972	.	132.972	132.972	132.972	132.972	132.972	132.972	132.972	132.972	132.972
MaxAE	297.216	.	297.216	297.216	297.216	297.216	297.216	297.216	297.216	297.216	297.216
Normalized	10.633	.	10.633	10.633	10.633	10.633	10.633	10.633	10.633	10.633	10.633
BIC											



**FIG 5.1 OBSERVED VERSUS FIT OF CRUDE OIL RATES USING ARIMA
MODEL (0,1,1)**

ii. ARIMA MODEL (0,1,2)

Table 5.2 MODEL FIT OF ARIMA (0,1,2)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.093	.	.093	.093	.093	.093	.093	.093	.093	.093	.093
R-squared	.872	.	.872	.872	.872	.872	.872	.872	.872	.872	.872
RMSE	170.792	.	170.792	170.792	170.792	170.792	170.792	170.792	170.792	170.792	170.792
MAPE	2.590	.	2.590	2.590	2.590	2.590	2.590	2.590	2.590	2.590	2.590
MaxAPE	6.030	.	6.030	6.030	6.030	6.030	6.030	6.030	6.030	6.030	6.030
MAE	134.928	.	134.928	134.928	134.928	134.928	134.928	134.928	134.928	134.928	134.928
MaxAE	308.213	.	308.213	308.213	308.213	308.213	308.213	308.213	308.213	308.213	308.213
Normalized BIC	10.745	.	10.745	10.745	10.745	10.745	10.745	10.745	10.745	10.745	10.745

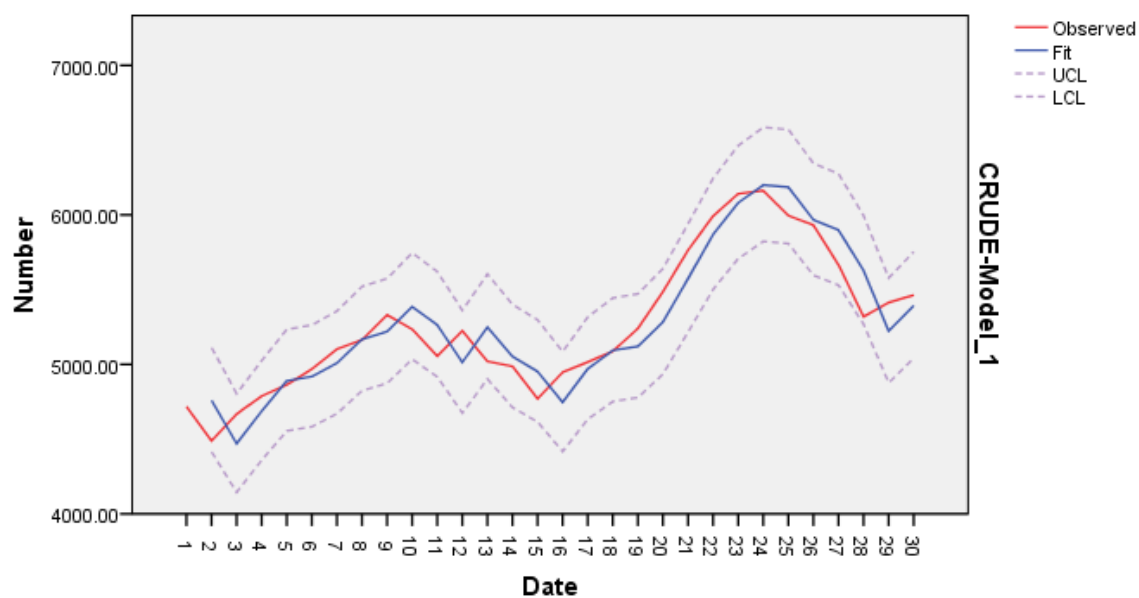


FIG 5.2 OBSERVED VERSUS FIT OF CRUDE OIL RATES USING ARIMA MODEL (0,1,2)

iii. ARIMA MODEL (1,1,0)

Table 5.3 MODEL FIT OF ARIMA (1,1,0)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.073	.	.073	.073	.073	.073	.073	.073	.073	.073	.073
R-squared	.869	.	.869	.869	.869	.869	.869	.869	.869	.869	.869
RMSE	169.259	.	169.259	169.259	169.259	169.259	169.259	169.259	169.259	169.259	169.259
MAPE	2.513	.	2.513	2.513	2.513	2.513	2.513	2.513	2.513	2.513	2.513
MaxAPE	6.125	.	6.125	6.125	6.125	6.125	6.125	6.125	6.125	6.125	6.125
MAE	131.012	.	131.012	131.012	131.012	131.012	131.012	131.012	131.012	131.012	131.012
MaxAE	281.072	.	281.072	281.072	281.072	281.072	281.072	281.072	281.072	281.072	281.072
Normalized BIC	10.611	.	10.611	10.611	10.611	10.611	10.611	10.611	10.611	10.611	10.611

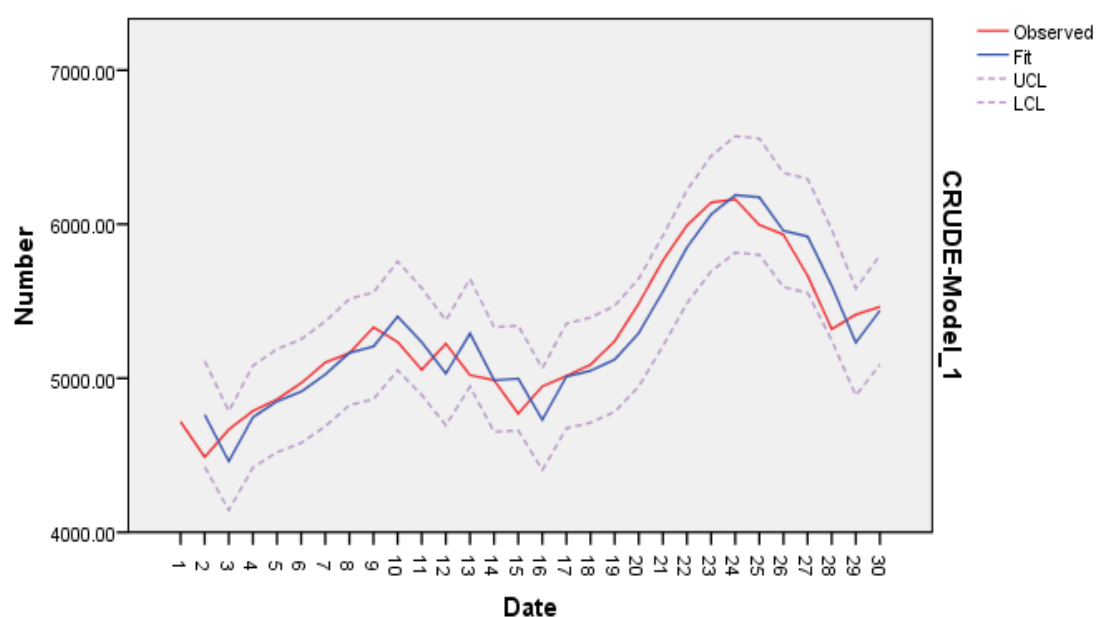


FIG 5.3 OBSERVED VERSUS FIT OF CRUDE OIL RATES USING ARIMA MODEL (1,1,0)

iv. **ARIMA MODEL (1,1,1)**

Table 5.4 MODEL FIT OF ARIMA (1,1,1)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary	.085	.	.085	.085	.085	.085	.085	.085	.085	.085	.085
R-squared	.871	.	.871	.871	.871	.871	.871	.871	.871	.871	.871
RMSE	171.393	.	171.393	171.393	171.393	171.393	171.393	171.393	171.393	171.393	171.393
MAPE	2.544	.	2.544	2.544	2.544	2.544	2.544	2.544	2.544	2.544	2.544
MaxAPE	6.034	.	6.034	6.034	6.034	6.034	6.034	6.034	6.034	6.034	6.034
MAE	132.616	.	132.616	132.616	132.616	132.616	132.616	132.616	132.616	132.616	132.616
MaxAE	278.380	.	278.380	278.380	278.380	278.380	278.380	278.380	278.380	278.380	278.380
Normalized BIC	10.752	.	10.752	10.752	10.752	10.752	10.752	10.752	10.752	10.752	10.752

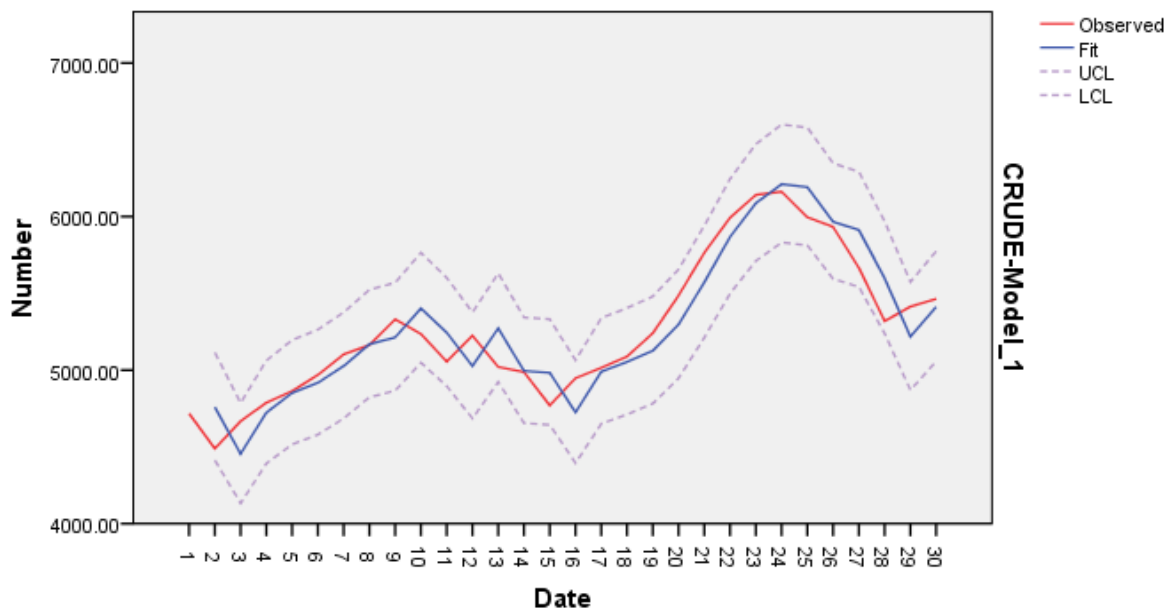


FIG 5.4 OBSERVED VERSUS FIT OF CRUDE OIL RATES USING ARIMA MODEL (1,1,1)

v. ARIMA MODEL (2,1,0)

Table 5.5 MODEL FIT OF ARIMA (2,1,0)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.096	.	.096	.096	.096	.096	.096	.096	.096	.096	.096
R-squared	.872	.	.872	.872	.872	.872	.872	.872	.872	.872	.872
RMSE	170.431	.	170.431	170.431	170.431	170.431	170.431	170.431	170.431	170.431	170.431
MAPE	2.573	.	2.573	2.573	2.573	2.573	2.573	2.573	2.573	2.573	2.573
MaxAPE	5.963	.	5.963	5.963	5.963	5.963	5.963	5.963	5.963	5.963	5.963
MAE	134.133	.	134.133	134.133	134.133	134.133	134.133	134.133	134.133	134.133	134.133
MaxAE	288.770	.	288.770	288.770	288.770	288.770	288.770	288.770	288.770	288.770	288.770
Normalized BIC	10.741	.	10.741	10.741	10.741	10.741	10.741	10.741	10.741	10.741	10.741

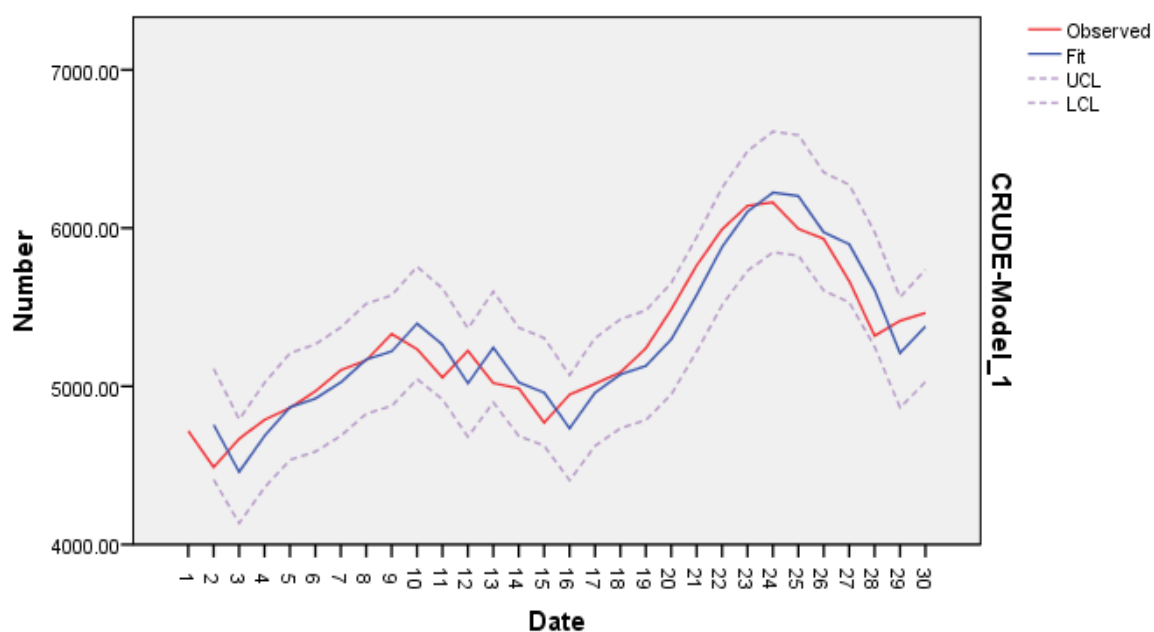


FIG 5.5 OBSERVED VERSUS FIT OF CRUDE OIL RATES USING ARIMA MODEL (2,1,0)

vi. ARIMA MODEL (2,1,1)

Table 5.6 MODEL FIT OF ARIMA (2,1,1)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary	.097	.	.097	.097	.097	.097	.097	.097	.097	.097	.097
R-squared	.873	.	.873	.873	.873	.873	.873	.873	.873	.873	.873
RMSE	173.896	.	173.896	173.896	173.896	173.896	173.896	173.896	173.896	173.896	173.896
MAPE	2.574	.	2.574	2.574	2.574	2.574	2.574	2.574	2.574	2.574	2.574
MaxAPE	5.957	.	5.957	5.957	5.957	5.957	5.957	5.957	5.957	5.957	5.957
MAE	134.127	.	134.127	134.127	134.127	134.127	134.127	134.127	134.127	134.127	134.127
MaxAE	290.176	.	290.176	290.176	290.176	290.176	290.176	290.176	290.176	290.176	290.176
Normalized BIC	10.897	.	10.897	10.897	10.897	10.897	10.897	10.897	10.897	10.897	10.897

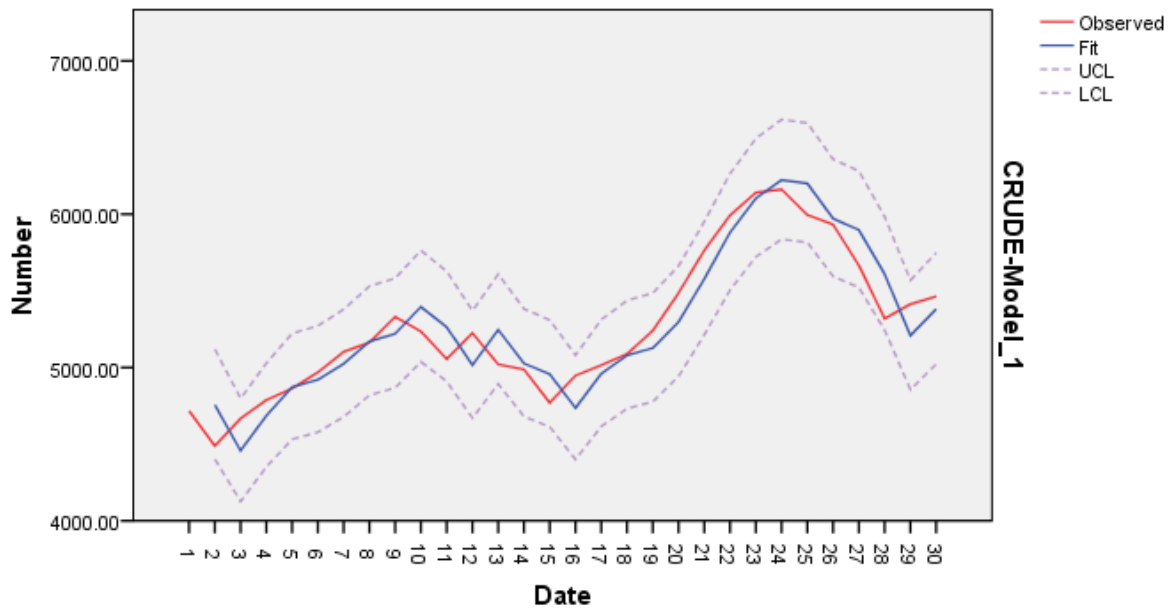


FIG 5.6 OBSERVED VERSUS FIT OF CRUDE OIL RATES USING ARIMA MODEL (2,1,1)

vii. ARIMA MODEL (2,1,2)

Table 5.7 MODEL FIT OF ARIMA (2,1,2)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.244	.	.244	.244	.244	.244	.244	.244	.244	.244	.244
R-squared	.895	.	.895	.895	.895	.895	.895	.895	.895	.895	.895
RMSE	160.978	.	160.978	160.978	160.978	160.978	160.978	160.978	160.978	160.978	160.978
MAPE	2.268	.	2.268	2.268	2.268	2.268	2.268	2.268	2.268	2.268	2.268
MaxAPE	5.771	.	5.771	5.771	5.771	5.771	5.771	5.771	5.771	5.771	5.771
MAE	117.277	.	117.277	117.277	117.277	117.277	117.277	117.277	117.277	117.277	117.277
MaxAE	269.408	.	269.408	269.408	269.408	269.408	269.408	269.408	269.408	269.408	269.408
Normalized BIC	10.859	.	10.859	10.859	10.859	10.859	10.859	10.859	10.859	10.859	10.859

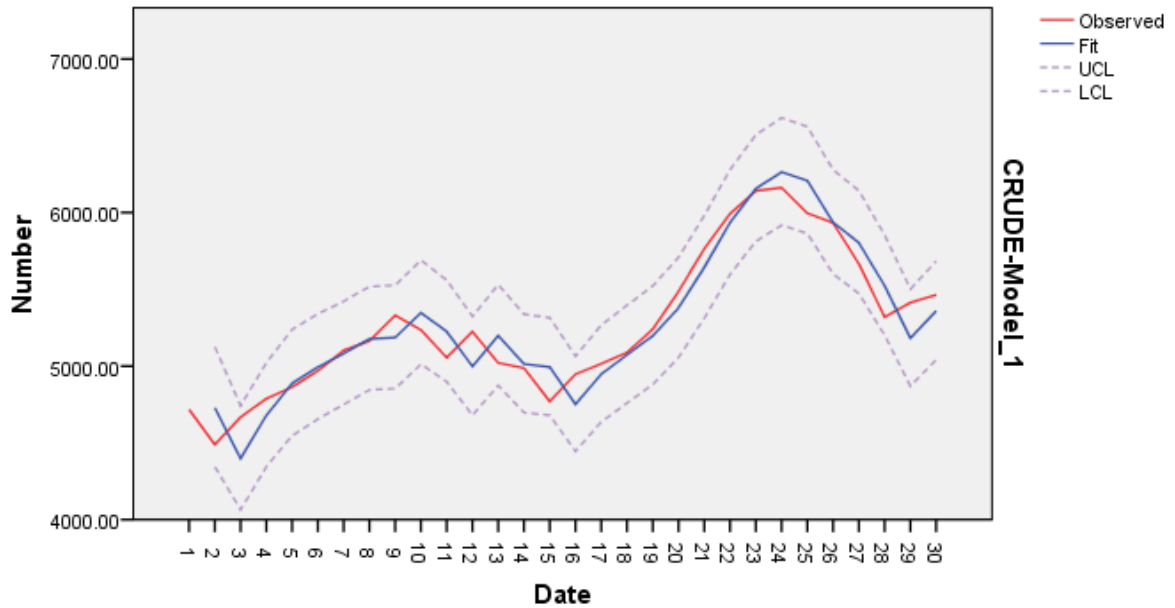


FIG 5.7 OBSERVED VERSUS FIT OF CRUDE OIL RATES USING ARIMA MODEL (2,1,2)

5.1.2 Model Estimation and Verification

Model parameters were estimated using SPSS package. To improve in the chosen ARIMA the residuals of the model is check to see if there is any systematic pattern. The autocorrelations and partial autocorrelation are examined for residuals of different order. For this purpose, the various correlations upto 17 lags were computed and the same along with their significance which is tested by Box-Ljung test are provided in (Table 5.9). As the results indicate, none of these correlations is significantly different from zero at a reasonable level. This proves that the selected ARIMA model is an appropriate model. The ACF and PACF of the residuals (FIG 5.8) also indicate 'good fit' of the model (Table 5.8).

Table 5.8 ANALYSIS OF ARIMA MODELS

ARIMA MODEL (P, D, Q)	MAPE VALUE	NORMALIZED BIC VALUE	MAE VALUE	R- SQUARED VALUE	RMSE VALUE
(0,1,1)	2.545	10.633	132.972	.866	171.108
(0,1,2)	2.590	10.745	134.928	.872	170.792
(1,1,0)	2.513	10.611	131.012	.869	169.259
(1,1,1)	2.544	10.752	132.616	.871	171.393
(2,1,0)	2.573	10.741	134.133	.872	170.431
(2,1,1)	2.574	10.897	134.127	.873	173.896
(2,1,2)	2.268	10.859	117.277	.895	160.978

Crude oil rates has been forecasted using all the models of ARIMA and ARIMA(1,1,0) has been selected as best Normalized BIC model which provided least value when compared with other models.

5.2 FORECASTING WITH ARIMA MODEL

Based on ARIMA Model: The process of model fitting for the crude oil price was done by using a computer software using SPSS 18. The following table indicates the estimated model parameters of ARIMA (1, 1, 0) model.

Table 5.9 ERROR MEASURES OF THE FITTED ARIMA MODEL

Model	Number of Predictors	Model Fit statistics								Ljung-Box Q(18)		
		Stationary R-squared	R-squared	RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Statistics	DF	Sig.
Crude oil Rate	1	.073	.869	169.259	2.513	131.012	6.125	281.072	10.611	18.887	17	.335

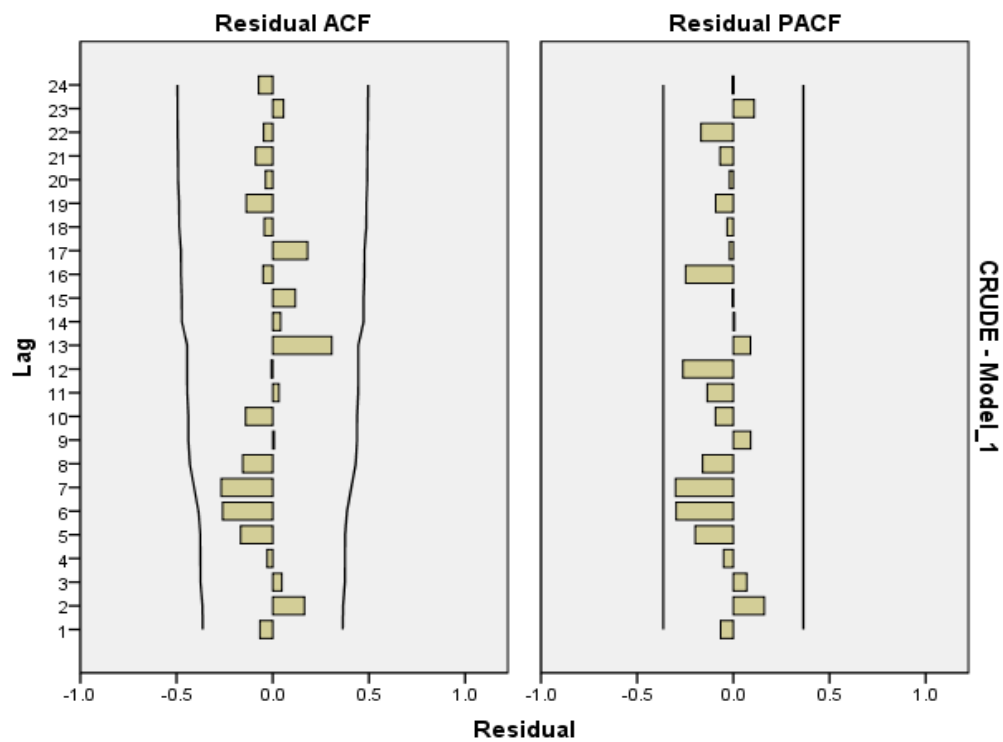


FIG 5.8 ACF AND PACF RESIDUALS OF FITTED ARIMA MODEL

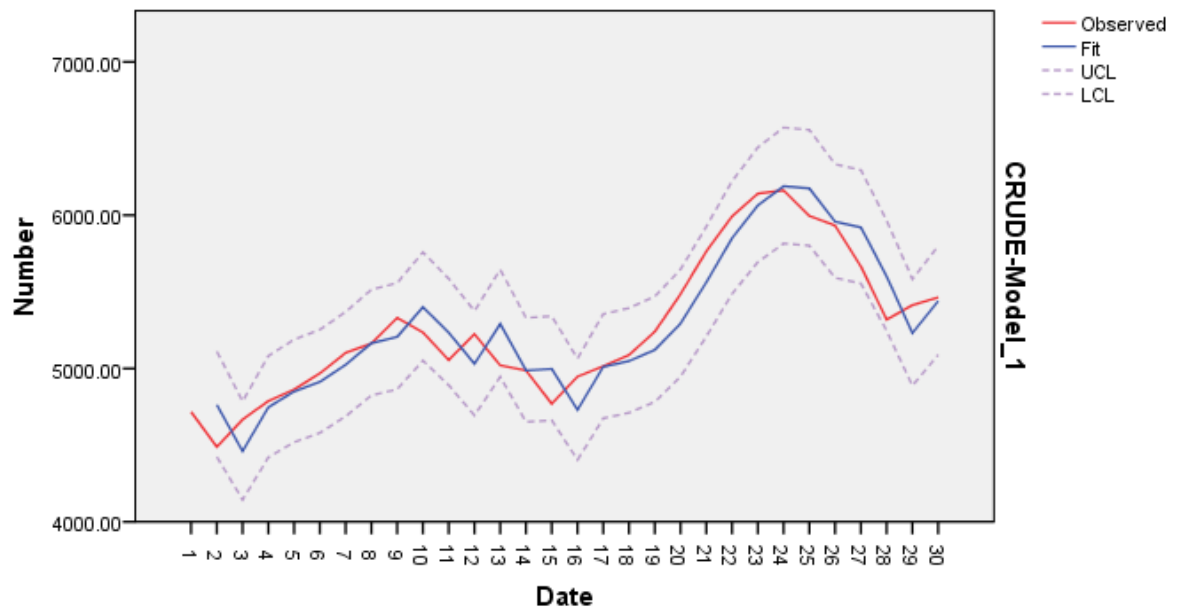


FIG 5.9 FORECAST OF CRUDE OIL RATES

Important measures of the sample period forecasts accuracy were computed to judge the forecasting ability of the fitted ARIMA model. The Mean Absolute Percentage Error (MAPE) for Crude oil rates was found out to be 2.513. This measure indicates that the forecasting inaccuracy is very low (Table 5.9). The observed fit and forecast graph for crude oil rate from (2nd MAY to 26th DEC) is shown in (FIG 5.9).

CHAPTER 6

CONCLUSION AND SUMMARY

In this study, 190 data set was taken for the time period 2nd MAY to 26th DEC and the data was analyzed with the help of Auto Regressive Moving Average. To find the exact ARIMA model seven tentative models were analyzed using SPSS 18 and their model fit values were found out. Out of these, ARIMA (1, 1, 0) was found to be best fit for forecasting of crude oil rates based on the Statistical standard error measures like Root Mean Square Error (RMSE) value, Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE) and Bayesian Information Criterion (BIC) value. MAE (Mean Absolute Error) was found to be 131.012, RMSE (Root Mean Square Error) was found to be 169.259 and MAPE (Mean Absolute Percentage Error) as 2.513 . This indicates that the forecasting error is very low and so this model can be taken as the best fit for the given time series. Also this study enables us to understand how time series model help us to provide accurate result for decision making in present and future market scenario. Its strength lies in the fact that the method is suitable for any time series with any pattern of change and it does not require the forecaster to choose a prior the value of any parameter. Its limitations include its requirement of a long time series.

REFERENCES

1. Agnolucci, P, Volatility in crude oil futures: A comparison of the predictive ability of GARCH and implied volatility models(2009). *Energy Econ.*, 31: 316-321. DOI: 10.1016/j.eneco.2008.11.001.
2. Ahmed, R.A. and A.B. Shabri, Fitting GARCH models to crude oil spot price data(2013). *Life Sci. J.*, 10:654-661.
3. Bollerslev.T , Generalized autoregressive conditional heteroskedasticity (1986). *J. Econometr.*, 31:307-321. DOI: 10.1016/0304-4076(86)90063-1.
4. Box, G.E.P. and G. Jenkins, *Time Series Analysis: Forecasting and Control*.1st Edn., Holden Day, San Francisco, ISBN-10: 0816211043, (1976) pp: 575.
5. Burges, C.J.C., A tutorial on support vector machines for pattern recognition. *Data Mining Knowl(1998) Discovery*,2:121-167.DOI:10.1023/A:1009715923555.
6. Chinn, M.D., M.R. LeBlanc and O. Coibion, *The Predictive Content of Energy Futures: An Update on Petroleum, Natural Gas, Heating Oil and Gasoline*. 1st Edn., National Bureau of Economic Research,(2005) pp: 17.
7. Engle, R.F., Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation(1982). *Econometrica*, 50: 987-1007.
8. Hou, A. and S. Suardi, A nonparametric GARCH model of crude oil price return volatility (2012). *Energy Econ.*, 34:618-626. DOI: 10.1016/j.eneco.2011.08.004.
9. Huang, W., Y. Nakamoria and S. Wang, Forecasting stock market movement direction with support vector machine (2005) . *Comput. Operat. Res.*, 32: 2513-2522. DOI: 10.1016/j.cor.2004.03.016.

10. Khashman, A. and N.I. Nwulu, Intelligent prediction of crude oil price using support vector machines. Proceedings of the IEEE 9th International Symposium on Applied Machine Intelligence and Informatics, Jan. 27-29, IEEE Xplore Press, Smolenice (2011a), pp: 165-169. DOI:10.1109/SAMI.2011.5738868.

11. Khashman, A. and N.I. Nwulu, Support vector machines versus back propagation algorithm for oil price prediction. Proceedings of the 8th International Conference on Advances in Neural Networks, May 29-Jun. 01, Springer Verlag Berlin, Guilin, China (2011b), pp: 530-558. DOI: 10.1007/978-3-642-21111-9_60.

12. Kim, K.J., Financial time series forecasting using support vector machines(2003). Neurocomputing, 55: 307-319. DOI: 10.1016/S0925-2312(03)00372-2.

13. Kumar, M.S., The forecasting accuracy of crude oil futures prices(1992). Int. Monetary Fund, 39: 432-461. DOI: 10.2307/3867065.

14. Laux. P, Vogl.S , Qiu.W , Knoche H.R. and Kunstmann.H , Copula-based statistical refinement of precipitation in RCM simulations over complex terrain(2011). Hydrol. Earth Syst. Sci., 15: 2401-2419.
DOI: 10.5194/hess-15-2401-2011.

15. Liu, L.M., Dynamic relationship analysis of us gasoline and crude oil prices(1991). J. Forecast., 10: 521-547. DOI: 10.1002/for.3980100506.

16. Marimoutou. V, Raggad.B and Trabelsi. A , Extreme value theory and value at risk: Application to oil market(2009). Energy Econ., 31: 519-530. DOI: 10.1016/j.eneco.2009.02.005.

17. Marzo.M and P. Zagaglia, Volatility forecasting for crude oil futures (2010). Applied Econ. Lett., 17:1587-1599. DOI: 10.1080/13504850903084996.

18. Morana. C, A semiparametric approach to short term oil price forecasting (2001). Energy Econ., 23:325-338. DOI:10.1016/S0140-9883(00)00075-X.
19. Muller, K.R., J.A. Smola and B. Scholkopf, Predicting time series with support vector machines(1997). Proceedings of the International Conference on Artificial Neural Networks, Oct. 8-10, Switzerland, pp: 999-1004. DOI: 10.1007/BFb0020283.
20. Sadorsky. P , Oil price shocks and stock market activity(1999). Energy Econ.,21: 449-469. DOI: 10.1016/S0140-9883(99)00020-1.