# Box-Jenkins Modelling and Forecasting of WTI Crude Oil Price

Maitham A. Rodhan[1] , Adnan H. Jaaz²
{Maitham.Rodhan@buog.edu.iq[1], adnan.hadi@buog.edu.iq[2]}

Basra University for Oil and Gas. College of Industrial Management For oil and Gas¹'²

**Abstract.** Oil is such an important commodity that its price and volatility have a significant impact on everyone's lives. Oil has a significant impact on economic growth. Unexpected changes in oil prices have an impact on both supplier and production countries' economic stability. As a result, estimating crude oil prices is critical. Trend, regression, moving average, and other methods are all used to model and forecast time-series data. The ARIMA approach is used to forecast the following year in this research. ARIMA approaches are used to analyse time-series data and are mostly utilised for loading forecasting because of their accuracy, mathematical soundness, and flexibility (due to the inclusion of AR and MA factors over a regression analysis). From March 1990 to March 2021, data on WTI prices was acquired from the US Energy Information Administration (EIA). The data is evaluated for stationarity and autocorrelation before using residuals procedures to choose amongst different ARIMA models for analysis. The outcomes of many models were evaluated, and the ARIMA (1,1,4) model was shown to be the most accurate forecasting model. The tables and figures below provide a full explanation of the above and summarise it.

Keywords: WTI crude oil, ARIMA Model, Stationarity, Oil Price Forecast, JEL Codes: C22, C51, E37, P28, G12

## 1 Introduction

Oil is one of the primary sources of energy consumption. Global consumption of crude oil and other petroleum products in 2014 for example reached more than 34 billion barrels, and its value exceeded $ 3.3 trillion. Therefore, the price of oil affects the world economy. Sudden variations in crude oil prices can have enormous economic implications; falls in crude oil prices hinder economic activity, while price booms produce severe inflation. As a result, energy academics, corporate leaders, and regulators are interested in modelling and forecasting crude oil prices. Both governments and corporations rely on accurate forecasting to make economic strategy decisions, yet it is difficult to foresee because to its chaotic and unpredictable nature.

Because there are so many variables that change over time, it's difficult to model the price of oil. As seen in Figure 1, the price of oil can fluctuate drastically in a short period of time, making it impossible to forecast. Oil has gone through some ups and downs during the last three decades.. The Asian Financial Crisis, combined with Iraq's decision to increase oil production,

IMDC-IST 2021, September 07-09, Sakarya, Turkey

DOI 10.4108/eai.7-9-2021.2314841

caused oil prices to plummet in early 1999. However, the market rapidly adjusted and in November of 2000, it had hit approximately US$ 35. The global economy then began to regain momentum, resulting in a period of bullishness lasting a few years. As a result, oil prices have been rising. The price of oil reached an all-time high of US$ 133.88 in June 2008. In the same year, the United States had a housing bubble, which led to an unprecedented credit crisis. This pushed the prices to drop below $ 40 a barrel within a few months, but the prices improved after that until they exceeded $ 100 in 2014, but it did not last long, as it decreased to the level of $ 30 in 2016 due to the increased supply of crude oil, in 2020, due to the repercussions of Covid-19, prices fell to their lowest level in two decades, reaching $ 16.55 in April. So, all these fluctuations complicate the process of forecasting crude oil prices in general.

There are different types of oil, including West Texas Intermediate (WTI) crude. WTI is very high quality. It is the main benchmark for crude oil in the Americas. In this study, we will choose West Texas Intermediate crude oil to predict its price because it is the main benchmark for crude oil in the US, which is the largest consumer of crude oil in the world, as it consumes more than 20% of the total global consumption of crude oil, as well as the largest producer of crude oil in the world. Finally, we can say that forecasting crude oil prices is a very complicated matter because many factors affect prices. So, we will limit our paper to forecasting for one year to achieve the highest possible accuracy. Medium and long-term forecasts usually lack accuracy based on the origin nature of the global oil market.

The goal of this project is to discover factors that influence oil prices and develop an accurate oil price forecast model.

## 2 Factors Affecting Oil Prices

Crude oil is the most volatile supplies in terms of price due to a variety of factors such as the dollar's exchange rate, the prices of other energy sources such as natural gas, coal, and renewable energy sources of all kinds, supply and demand for oil, crude oil market speculation, and weather factors. As well, an important factor that affects crude oil prices is the geopolitical factor, which is a very important variable. Crude oil production centers, in general, are in areas experiencing continuous political turmoil. And thus, these disturbances affect crude oil supplies and hence prices. So, studying these variables can help us a lot in understanding the mechanism for determining prices for crude oil and thus enables us to find a more accurate predictive model.

### 2.1 Economic Factors

Oil demand and consumption by all sectors rises in response to economic, industrial, and population growth, and falls in response to economic downturns. Supply and demand play a role on oil pricing. When demand increases (or supply decreases), prices should rise, and when demand decreases, prices should fall (or supply increases). The pricing of gas and oil reserves reflect supply and demand pressures from a specific time period. Prices change to maintain future supply levels while taking current production levels into account when demand for crude oil products is high. High prices, on the other hand, encourage production, hence boosting supply and lowering the price of oil and its derivatives[1].

### 2.2 Exchange Rate

Crude oil is exchanged in US dollars all over the world, but petroleum products are purchased in local currencies. As the dollar depreciates against other currencies, countries with non-dollar appreciating currencies profit from cheaper oil, whilst customers in USD-pegged countries pay a greater price for the same barrel of oil. As a result, changes in the value of the US dollar will affect global oil demand. When the US dollar depreciates versus other currencies, the cost of acquiring a dollar decreases. This will increase demand for crude oil in currencies other than the US dollar, driving up prices. As a result, a negative correlation is expected between the US Dollar exchange rate and crude oil price movements[2].

### 2.3 Geopolitical Events

Political upheaval has had a significant impact on oil supply and prices, particularly in oil-producing regions such as the Middle East. Long-standing historic rivalries between governments and tribes, religious differences, and ownership of critical resources such as petroleum could all contribute to such tensions. The highest oil prices in history have been connected to tensions with Iran; Iraq's production has been lowered in recent decades due to protracted war periods. Nigerian oil production is harmed by violence and insurgency activity in the Niger Delta. Nigeria's production output has been lowered due to frequent abductions of foreign employees, pipeline attacks, and sabotage of oil infrastructure. Venezuela's political unrest and nationwide strike harmed oil exports and had a direct influence on the international oil market. Libya experienced the same thing. The price of crude oil is rising as a result of all of this upheaval[3] .

### 2.4 Climatic Factors

Seasonal weather changes affect the oil demand. In the winter, more heating oil is consumed, whereas in the summer, more gasoline is consumed. As a result, the price of oil varies according to the season. Hurricanes, tsunamis, and thunderstorms, especially in major oil-producing areas, can cause physical damage to production facilities and infrastructure, disrupting oil supply and raising prices[4].

American oil production stopped after Hurricane ‘Evan’, which deprived the United States of America of more than 10 million barrels of production in September 2004. In addition to the suspension of production in some fields of Mexico due to the damage caused by this hurricane, which continued to stop production for more than 3-4 months.  And based on the above, this hurricane caused a decrease in oil supplies to the markets. The price of a barrel of oil increased by $13m when Hurricane 'Katrina' hit the southern United States in 2005, affecting 19% of the US oil supply. The flooding of the Mississippi River in May 2011 caused oil prices to fluctuate[5].

### 2.5 Future Market Speculation

Future pricing above spot prices, which lead to anticipation of higher future prices, can persuade oil producers to keep their oil and sell it later for a bigger profit. This might diminish the existing supply of oil and have a significant impact on pricing[6]. Only about 3% of futures transactions result in the buyer of the futures contract getting possession of the commodity being exchanged, hence most futures trading is done by speculators. Furthermore, market sentiment plays an important part in determining oil prices. When speculators and hedge funds buy oil future contracts, for example, the simple concept that oil demand will rise dramatically at some point in the future might lead to a big surge in oil prices now[7].

### 2.6 Technological Development and Renewable Energies

The reliance on crude oil has been reducing as other energy technologies have advanced and their capital costs have continued to fall. To meet diverse areas of the world's energy needs, crude oil and other energy sources such as renewables, natural gas, and hydroelectric power are employed. While crude oil is still the most abundant significant energy source, it is largely utilised to make transportation fuels, whereas renewable energy is primarily used to generate power (electricity). As a result, because crude oil has no direct substitutes, demand for the other energy source does not always increase when the price of one rises[8] .

## 3 ARIMA Model

The ARIMA model combines autoregression (AR) and moving average (MA). AR makes use of the dependent relationship between an observation and a set of lagged observations. The term "integrated" refers to the use of raw observation differencing to keep a time series stable. MA is a model that takes into account the link between an observation and a residual error from a lagged moving average model.

AR (p) stands for autoregressive of order (p), that is:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} \ldots\ldots\ldots \phi_p Y_{t-p} + \varepsilon_t \qquad (1)$$

Where, $Yt$ - dependent variable at time $t$; $\varphi0$, $\varphi1$, $\varphi2$, --- $\varphi p$- the predicted factors, $\varepsilon t$ – the error. an error describing the effects of variables that are not considered in the model. The term MA (q) stands for a q-order moving average model and is represented by:

$$Yt = \mu + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \ldots\ldots + \omega_q \varepsilon_{t-q} \qquad (2)$$

Where, Yt-dependent variable at time t; -constant process average; t-error at time t; 1, 2,---, q-estimated coefficients

The ARMA (p, q) model is a mixture of AR (p) and MA (q) models that is represented by:

$$Y_t=\mu+ \phi_1 Y_{t-1} + +....+\square\phi_p Y_{t-p} - \omega_1 \varepsilon_{t-1}.... - \omega_q \varepsilon_{t-q} + \varepsilon_t \qquad (3)$$

The above-mentioned t is a AWGN. That is, it is a set of variables with a zero mean, a two-variable variance, and a zero correlation across time, for example, E $(\varepsilon_u - \varepsilon_v) = 0$ if $u \neq v$. They are also independent and identically distributed.
To obtain ARIMA, the ARMA model can be used to differentiate the time series for d times to achieve a stationary instance (p,d,q). The standard difference is:

$$\text{regular difference} = (1-B)^d \; X_t \qquad (4)$$

B is for backward operator, while d stands for non-seasonal order of differences.

Although the differencing procedure can be repeated numerous times in theory, only one or two differencing operations are employed in practise[9]. To determine stationarity, utilise the Dickey-Fuller test, whose null hypothesis is that time-series data is nonstationary. The critical p-value was set to 0.05, which indicates that if the test p-value is greater than 0.05, the Dickey-Fuller null hypothesis, claiming that data is stationary, will be accepted.

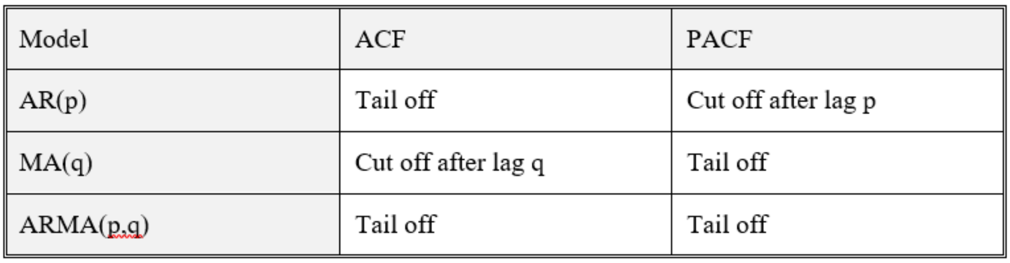
## 4 Box-Jenkins Methodology

Box and Jenkins presented a time series analysis methodology for determining the optimum time function to create future projections. There are four steps to the methodology:

1. Identification.
2. Estimation
3. Diagnostic testing of the indicated model's suitability for simulating an event
4. Model application, such as forecasting.

For time series analysis and forecasting, the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) are the two most important analytical tools (PACF). They look at statistical correlations between observations in a particular data collection. ACF has the advantage of allowing you to measure the level of linear dependence between observations in a time series separated by a lag k. The PACF image shows how many automobiles regressive keywords are needed to uncover one or more time lags with substantial correlations. The behaviour of the various models on the ACF and PACF can be seen as follows:

**Table 1** Details of ACF and PACF Patterns.



| Model | ACF | PACF |
|---|---|---|
| AR(p) | Tail off | Cut off after lag p |
| MA(q) | Cut off after lag q | Tail off |
| ARMA(p,q) | Tail off | Tail off |

There may be more than one option for the final model in some cases. In this instance, we may want to select the most suited model using an alternative strategy based on information criteria. This is accomplished by employing the criteria outlined below.

### 4.1 Akaike Information Criterion AIC

It should be emphasised that for successful models, the model with the fewest variables delivers the best predicting results, for example, for a time series with multiple successful ARIMA models. In this scenario, the classical AR and/or MA should be selected. This is accomplished by selecting the best ARIMA model among successful models using Akaike's Information Criterion (AIC) (Akaike, H. 1974). The AIC value with the least value should be picked.

Akaike's Information Criterion (AIC) may be written as:

$$\mathrm{AIC}(p, q) = \ln(\sigma^{2}) + 2(p + q) / T \qquad (5)$$

Where $\sigma^2$ is the white noise variance's maximum likelihood estimate, sample size is T, and the total number of parameters found in the ARMA (p, q) model is (p + q). The correct model is then found by selecting the set of p and q values that minimise the AIC (p, q). To avoid over parametrization, the term 2 (p+q) T could be used as a penalty term..

### 4.2 Bayesian Information Criterion (BIC) / Schwarz Criterion (SC)

The BIC is given by[9] .

$$\mathrm{BIC}(p, q) = \ln(\sigma^{2}) + \ln(T)(p + q)/T \qquad (6)$$

with $\ln(T) / T > 2 / T$ for all $T \geq 8$

### 4.3 Determination Coefficient $R^2$ (R-squared)

The coefficient of determination shows how near the constructed model is to the optimum constant. If the model incorporates a free term, the determination coefficient can range from 0 to 1. (constant). It can be read as the proportion of the variance of the dependent variable Yt that can be explained using the independent and lag variables in the model, in the form that they exist in the model, in this case. If this is not the case, the coefficient of determination may be negative..

### 4.4 Adjusted Coefficient of Determination $R^2_{adj}$

Because the computed $R^2$ will not decrease as more lags are added to the assessed model, it cannot be used as a good indicator of the model's quality. For additional regressors, a fine is applied when calculating the adjusted determination coefficient (lag variables). As a result, the adjusted determination coefficient values do not surpass those of the standard determination coefficient. R2adj can be decreased by adding more variables to the model.

### 4.5 Standard Regression Error (s.e.regr)

The Standard Regression Error represents the variation of the time series in reference to the generated model. Adjusted R-squared, AIC value, SC value, and S.E. of regression are some of the model selection criteria. For rating and selecting the best model, the AIC and SC criteria are most typically utilised. The higher the coefficient of determination, the smaller the AIC, SC, and residual variance are. The ARMA (p,q) model that corresponds to it is superior.

### 4.6 Residual Analysis

If the Box - Jenkins model is good enough for time series data, we anticipate the residual to be a realisation of white noise. The residual, in other words, must be self-contained and conform to the normal distribution. This is visualised using the time series residual plot, residual correlogram, and Q-Q plot. The Ljung-Box test is also utilised, which is predicated on the autocorrelation being greater than zero. This hypothesis is supported by the Ljung-Box test with test statistic $Q^{10}$:
H0: independently distributed
H1: not independently distributed.

## 5 ARIMA Model

### 5.1 Data Description and Sources

The model is based on the West Texas Intermediate crude oil spot price. Time series data for Monthly WTI prices from Jan-1990 to Mar- 2021is used in this paper. This data was obtained from **the** US Department of Energy. As shown in Fig 1.

### 5.2 Stationarity Test

The PWTI data series ( Jan-1990 to Mar- 2021) is shown in Fig 1.

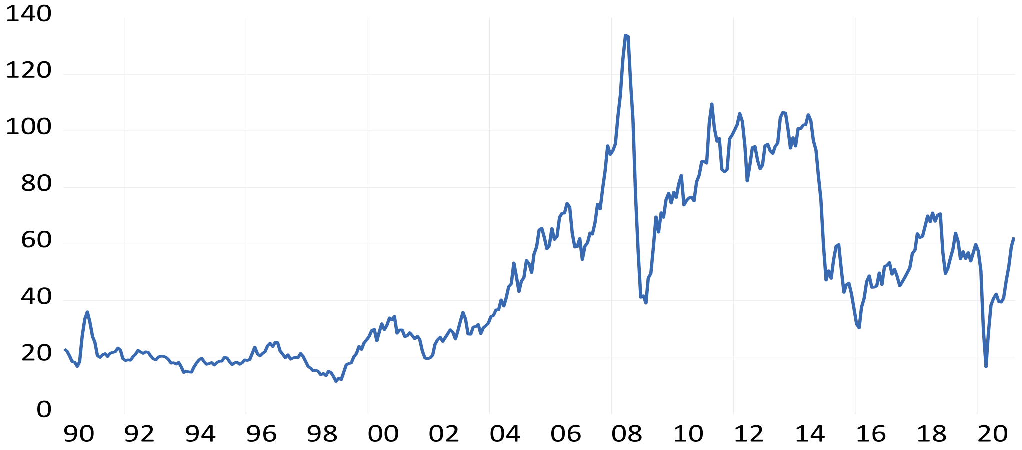


**Fig 1**. The PWTI Data ( Jan-1990 to Mar- 2021).

The outcomes result of the stationarity process is stated in Table 2.

**Table 2.** ADF Test on PWTI.

Null Hypothesis: Unit root assigned to PWTI

Exogenous: Constant

maxlag=16

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Dickey-Fuller | | -2.464102 | 0.1252 |
| Used data: | 1% | -3.447675 | |
| | 5% | -2.869071 | |
| | 10% | -2.570849 | |

From Table 2, the original non-stationary PWTI sequence is illustrated. Figure 2 Oil prices have a large difference when considering the highest and lowest, so by taking the natural logarithm we can reduce the variation of original PWTI data but it is also still nonstationary as demonstrated in Table 3. Therefore, we take the first-order difference, and now it is stationary as represented in Table 4.

**Table 3.** ADF Test on LPWTI.

Null Hypothesis: LPWTI has a unit root

Exogenous: Constant

maxlag=16

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Dickey-Fuller | | -2.101028 | 0.2445 |
| Values that must be tested: | 1% | -3.447675 | |
| | 5% | -2.869071 | |
| | 10% | -2.570849 | |

**Table 4.** ADF Test on D(LPWTI).

Null Hypothesis: D(LPWTI) has a unit root

Exogenous: Constant

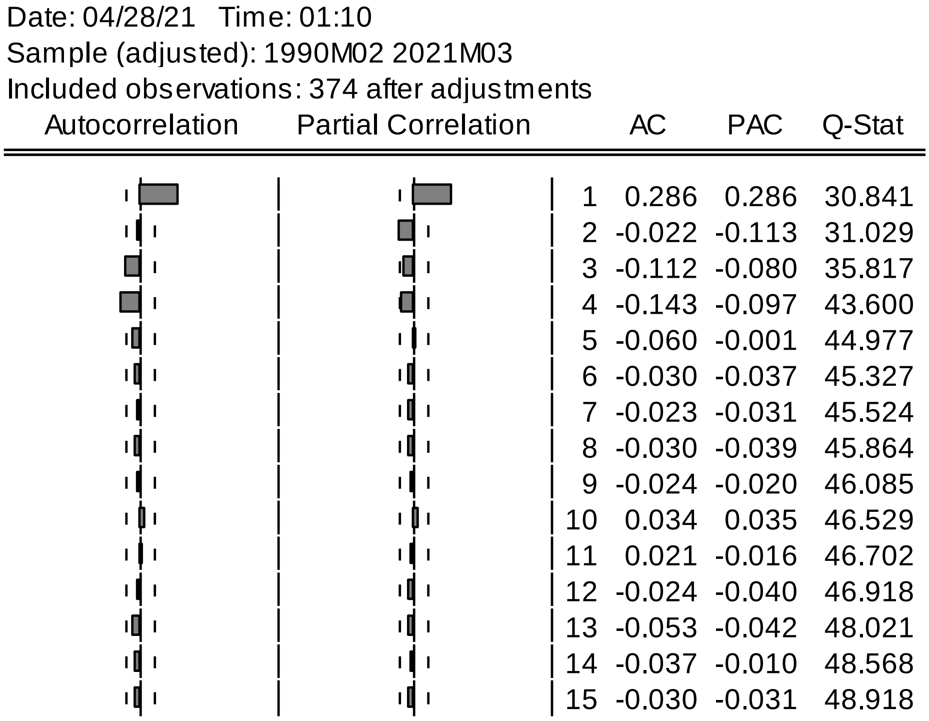
maxlag=16

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Dickey-Fuller | | -14.34564 | 0.0000 |
| Test data: | 1% | -3.447675 | |
| | 5% | -2.869071 | |
| | 10% | -2.570849 | |

### 5.3 Select the Model

The AC and PAC Graphs of the DLPWTI are shown in Table 5.

**Table 5.** AC and PAC Graphs of the DLPWTI.



```
Date: 04/28/21   Time: 01:10
Sample (adjusted): 1990M02 2021M03
Included observations: 374 after adjustments
   Autocorrelation      Partial Correlation        AC      PAC     Q-Stat

                                                1   0.286   0.286   30.841
                                                2  -0.022  -0.113   31.029
                                                3  -0.112  -0.080   35.817
                                                4  -0.143  -0.097   43.600
                                                5  -0.060  -0.001   44.977
                                                6  -0.030  -0.037   45.327
                                                7  -0.023  -0.031   45.524
                                                8  -0.030  -0.039   45.864
                                                9  -0.024  -0.020   46.085
                                               10   0.034   0.035   46.529
                                               11   0.021  -0.016   46.702
                                               12  -0.024  -0.040   46.918
                                               13  -0.053  -0.042   48.021
                                               14  -0.037  -0.010   48.568
                                               15  -0.030  -0.031   48.918
```

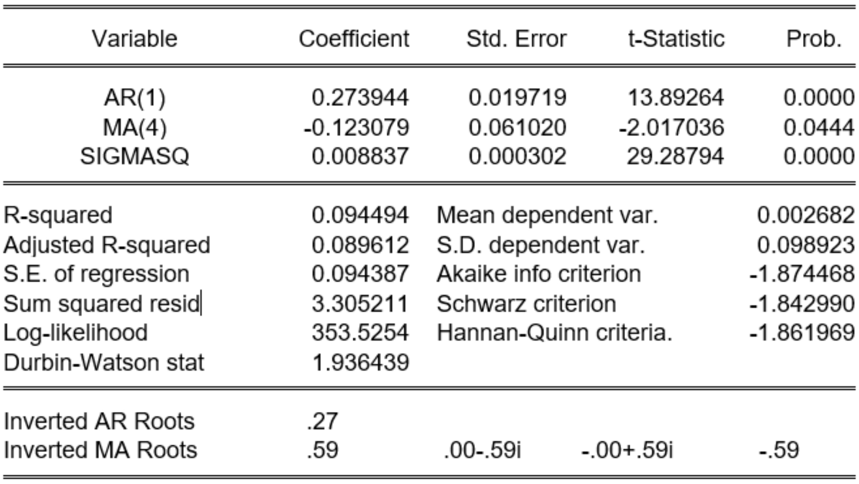
From Table 5. The q can be taken 1 or 3 or 4. And p can be taken 1 or 2. Therefore, we have multiple ARMA models. By following Adj-$R^2$, AIC, SC, and S.E. of relapse It appears that the most accurate model is ARMA (1,4).

## 5.4 Model Estimation

Table 6 summarized the data computation:

**Table 6.** Summarized Results.

```
Dependent Variable: D(LPWTI)
ARMA and OPG - BHHH
Date: 05/07/21   Time: 23:23
No. of data: 1990M02 2021M03
observations: 374
```



| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| AR(1) | 0.273944 | 0.019719 | 13.89264 | 0.0000 |
| MA(4) | -0.123079 | 0.061020 | -2.017036 | 0.0444 |
| SIGMASQ | 0.008837 | 0.000302 | 29.28794 | 0.0000 |
| R-squared | 0.094494 | Mean dependent var. | | 0.002682 |
| Adjusted R-squared | 0.089612 | S.D. dependent var. | | 0.098923 |
| S.E. of regression | 0.094387 | Akaike info criterion | | -1.874468 |
| Sum squared resid | 3.305211 | Schwarz criterion | | -1.842990 |
| Log-likelihood | 353.5254 | Hannan-Quinn criteria. | | -1.861969 |
| Durbin-Watson stat | 1.936439 | | | |
| Inverted AR Roots | .27 | | | |
| Inverted MA Roots | .59 | .00-.59i | -.00+.59i | -.59 |

The final model is ARIMA (1, 1, 4), and specific form of the model is shown in Equation (7)

$$DLPWTI= 0.273226 \quad DLPWTI_{t-1} - 0.1236 \; e_{t-4} \qquad (7)$$

From Fig 2 we can see that the AR roots and all MA roots lie inside the united cycle. So, the ARIMA covariance is stationary and invertible and now we can forecast with this model.

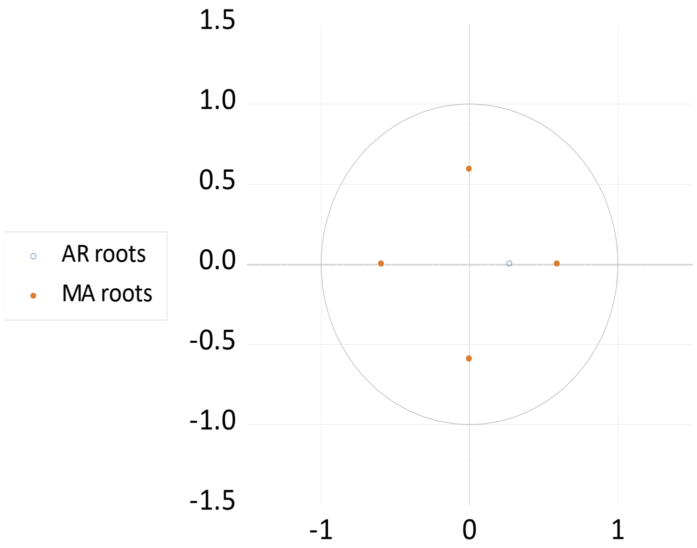


**Fig 2.** AR/MA Polynomials.

## 5.5 Forecasting

The PWTI values are predicted using this model from November 2021 to October 2022. Table 7 summarises the findings.

**Table 7.** PWTI Forecast from November 2021 to October 2022.

| Nov 2021 | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct 2022 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 58.18 | 58.34 | 58.49 | 58.64 | 58.80 | 58.95 | 59.10 | 59.26 | 59.41 | 59.57 | 59.72 | 59.88 |

## 6 Conclusion

The ARIMA model has been used to predict the price of WTI crude oil, and two primary features were studied that affects the Oil prices. The study looked at 375 months of WTI crude oil price data from January 1990 to March 2021. The time series is unit root nonstationary, according to the Dickey-Fuller test. When comparing the greatest and lowest oil prices, we can see that there is a significant difference. We use the log of the original PWTI data to lessen the data's fluctuation and better fit it into the model. We can see from the log that the processed data has a general rising trend, and that the price differences between early and recent years have narrowed substantially, making data analysis and model training easier. First-order differencing was employed to convert the non-stationary time series into a stationary one, allowing the univariate box- Jenkins modelling approach to be used. The time series, ACF, and PACF plots of the first-order difference of the WTI crude oil price data revealed many ARMA (p, q) models with p (1,2) and q (1,3). (1,3,4). The ARIMA (1,1,4) model, which has the minimum AIC, BIC S.E. of regression statistics, emerges as the optimal model for the underlying generating process of WTI crude oil price data and was utilised to create a forecast.

## References

Akaike, H: A new look at the statistical model identification. Transactions on Automatic Control, Vol.19, No.16, (1974), 723., Available on http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/MiscDocs/Akaike_1974.

Clover Global: "Ten Factors that Affect the Price of Oil", Clover Global Solutions, 2012, webpage, *https://c1wsolutions.wordpress.com/2012/04/30/factors-affect-price-of-oil/*. Accessed: May 1, 2021.

Francis, X.: On the power of Dickey-Fuller tests against fractional alternatives," Business Cycles: Durations, Dynamics and Forecasting, Princeton University Press.1999 p.258.

Henry, d- Acquah, G Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in the selection of an asymmetric price relationship, Journal of Development and Agricultural Economics Vol. 2(1), January 2010,p6.

King, K.- Deng, A.- Metz, D: "An Economic Analysis of Oil Price Movements: The Role of Political Events and Economic News, Financial Trading, and Market Fundamentals", *Bates White Economic Consulting*, Washington DC, 2011,p14.

Kwasi,E :Box-Jenkins modelling and forecasting of Brent crude oil price, Munich Personal RePEc Archive,2015,p6.

Le, B. : “Crude Oil’s Impact on Renewable Energy: Energy Alternative or Energy Staple?”,2015, *http://www.altenergymag.com/article/2015/06/crude-oil%E2%80%99s-impact-on-renewable-energy-energy-alternative-or-energy-staple/20384*. Accessed: April 28, 2021.

Lioudis, N. : “The Determinants of Oil Prices”,2021: Investopedia webpage, http://www. investopedia.com/ask/answers/012715/what-causes-oil-prices-fluctuate.asp. Accessed: May 3, 2021.

Olimb, M- Ødegård, : “Understanding the Factors Behind Crude Oil Price Changes A Time-varying Model Approach”*, 2010,Unpublished MSc Thesis Report*, Norwegian University of Science and Technology, p8.

Olimb, M- Ødegård, T.: “Understanding the Factors Behind Crude Oil Price Changes A Time-varying Model Approach”*, Unpublished MSc Thesis Report*, Norwegian University of Science and Technology,2010, p11.

Pankratz, A : Forecasting with Univariate Box-Jenkins Models Concepts and Cases. John Wiley & Sons, Inc. New York, USA,1983, p: 414.

Salas, J- Delleur, J- Yevjevich, V and Lane, Applied : Modeling of Hydrologic Time series. Water Resources Publication, Michigan, USA.1980, P44.

Seth, S.: Do Oil and Natural Gas Prices Rise and Fall Together”,2015, Investopedia webpage, *http://www.investopedia.com/articles/active-trading/032515/do-oil-and-natural-gas-prices-rise-and-fall-together.asp*. Accessed: May 5, 2021.

U.S. EIA,: Cushing, OK WTI Spot Price FOB U.S. Energy Information Administration (EIA)2021.https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=M. accessed 14 April,