Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)

Volume 1 Issue 1 June, 2024, pp 17-26
https://idjams.com/index.php/jms/index

ARIMA Modeling for Monthly West Texas Intermediate Crude Oil Price Forecasting

Agus Sofian Eka Hidayat[1], Jason Marvel Wiharja[1,*]

[1] *Study Program of Actuarial Science, School of Business, President University, Bekasi, 17550, Indonesia*

[*]*Corresponding author: jason.wiharja@student.president.ac.id*

Copyright: ©2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License-(CC-BY-SA) (https://creativecommons.org/licenses/by-sa/4.0/)

Article Info

Article history:
Received: June 2024
Revised: August 2024
Accepted: August 2024
Available online: August 2024

Abstract

Crude oil is a non-renewable natural resource. It is one of the world's most important natural resources since it has many utilities. One of the best parameters for crude oil price is West Texas Intermediate (WTI). It is the main oil benchmark for North America sourced from the United States. It is important to predict WTI price since the price of crude oil will impact the price of other commodities, which also will impact human life. In this work, the Autoregressive Integrated Moving Average (ARIMA) approach is used to predict WTI price for the next 12 months with the monthly data from April 1, 2019, to April 1, 2023, which was collected from investing.com. The results ARIMA (1,2,1) with the best model for forecasting WTI price with accuracy value from error measurements MSE is 38,129117536, MAPE is 6,8128441%, and RMSE is 6,174882619. This forecasting will be useful as a consideration for various stakeholders and the government.

Keywords: ARIMA; Crude Oil; Natural Resources; West Texas Intermediate

I. Introduction

Crude oil is a non-renewable natural resource. It is a mixture of hydrocarbons (composed only of the elements Carbon (C) and Hydrogen (H)) that exists in a liquid phase in natural underground reservoirs and remains liquid at atmospheric pressure after surface separating [1]. Crude oil is a significant resource in the world, it is impacting our lives. When Russia invaded Ukraine the crude oil price increased significantly, causing inflation, and the increasing price of goods for our needs. In other words, the crude oil price impacts other commodities economically it is called the spillover effect (indirect effect).

West Texas Intermediate (WTI) is the main oil benchmark for North America as it is sourced from the United States, primarily from the Permian Basin. It was established in 1983, in that year the oil market was mostly in the fall, with Iraq and Iran at war and OPEC finding its footing. In that, situation visionaries at NYMEX want to establish a transparent market for crude oil so market participants could manage the risk [1]. WTI is known as a light sweet oil because it contains less than 0.50% sulfur (between 0.24% and 0.34%) making it have a low density. Nowadays, WTI is the underlying commodity of the New York Mercantile Exchange (NYMEX which is part of CME Group) which oil futures contract and is considered a high-quality oil that is easily refined [2].

From previous research, some researchers also use the ARIMA approach. In According to Rodhan and Jaaz [3] WTI was predicted using the ARIMA with data from March 1990 to March 2021. The research shows the best ARIMA model is (1,1,4). The other one, Derek [4] uses the ARIMA, GARCH, and APARCH model with data from April 1993 until March 2012 which is divided into 2 periods the first one is April 1993 to March 2022 and the second is April 2003 to March 2012. The result of this

research is that the GARCH and APARCH model is used for the second period and ARIMA is for the first period.

Ramlan [5] uses the ARIMA model to forecast Malaysian goods export for 2021-2022 with the data from January 2020 until May 2021 with monthly frequency. The study shows the best ARIMA model used to predict is ARIMA (2,1,2) with the forecast result being very close to the actual growth rate. Research from Mashitasari [6] uses the ARIMA model to predict coal sales with monthly data from September 1997 until August 2007. The results show that ARIMA (0,1,1) is the best model to predict bricket coal sales. Research from John and Latupeirissa [7] predicts gold prices in 2020 using ARIMA Box-Jenkins with the data collected from 2014-2019 with monthly frequency and the result shows that the best ARIMA model is ARIMA (1,1,1) with a little difference between actual value and forecasted value because in March 2020 Indonesia experienced covid-19 that impacted the price of gold. In this work, the price of WTI will be predicted using the ARIMA modelling with data from April 1, 2019, to April 1, 2023. The difference with the previous work compares the model which model is effective in predicting the movement of one object, in this work, the researcher tries to find how effective the ARIMA model is in predicting the oil index in the middle of international tension.

II. Method

Forecasting is a statistical method to predict the future as accurately as possible. It helps and provides a guide to long-term strategic planning and measurement of the risks that can be prevented such as the scheduling of production, transportation and personnel, financial possibility and many more [5]. One of the forecasting methods is the Box-Jenkins method It refers to a systematic method of identifying, fitting, checking and applying Autoregressive Integrated Moving Average (ARIMA(p, d, q)) models created by 2 mathematicians Box and Jenkins in 1970. This project applies box Jenkins with the ARIMA approach to predict the time series data [6].

A. Autoregressive (AR)

In the autoregression model, the forecast method uses a linear combination of past values of the data hence the term autoregression indicates that it is a regression of the variable against itself. Autoregressive (AR) with order p can be written as:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (1)$$

The equation above is the autoregressive model with lags p. The value of p is determined from the partial autocorrelations (PACF) plot of the differenced time series model. If the partial autocorrelations cut off after a few lags, the last lag with a large value would be the estimated value of p [6].

B. Differencing

Time series with trends or seasonality that are not stationary will affect the value of the time series at different times. One way to make the non-stationary time series stationary is by computing the differences, also known as differencing and denoted as d. It can help to stabilize the mean and variance of time series by reducing the trend and seasonality [7].

C. Moving Average (MA)

Moving average model uses past forecast error in a regression model. The moving average (MA) with order q can be written as:

$$y_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

The value of q is found by the autocorrelations (ACF) plot of the differenced time series model. If the autocorrelations cut off after a few lags, the last lag with a large value would be the estimated value of

q [6].

D. ARIMA (p, d, q)

ARIMA is the abbreviation of Auto-Regressive Integrated Moving Average (ARIMA) with order p (AR), d, q (MA). In other words, the ARIMA model implements the differenced ARMA (Auto Regressive-Moving Average) model. Time series $\{Y_t\}$ is said to follow an Integrated Moving Average if the d^{th} difference $W_t = \nabla^d Y$ is a stationary ARMA process. If { }follows an ARMA (p, q) model, it can be said that $\{Y_t$ }is an ARIMA process with the equation as follows [7]:

$$Y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3)$$

Simplify the Box-Jenkins method with the ARIMA approach can be illustrated by the flow chart below:

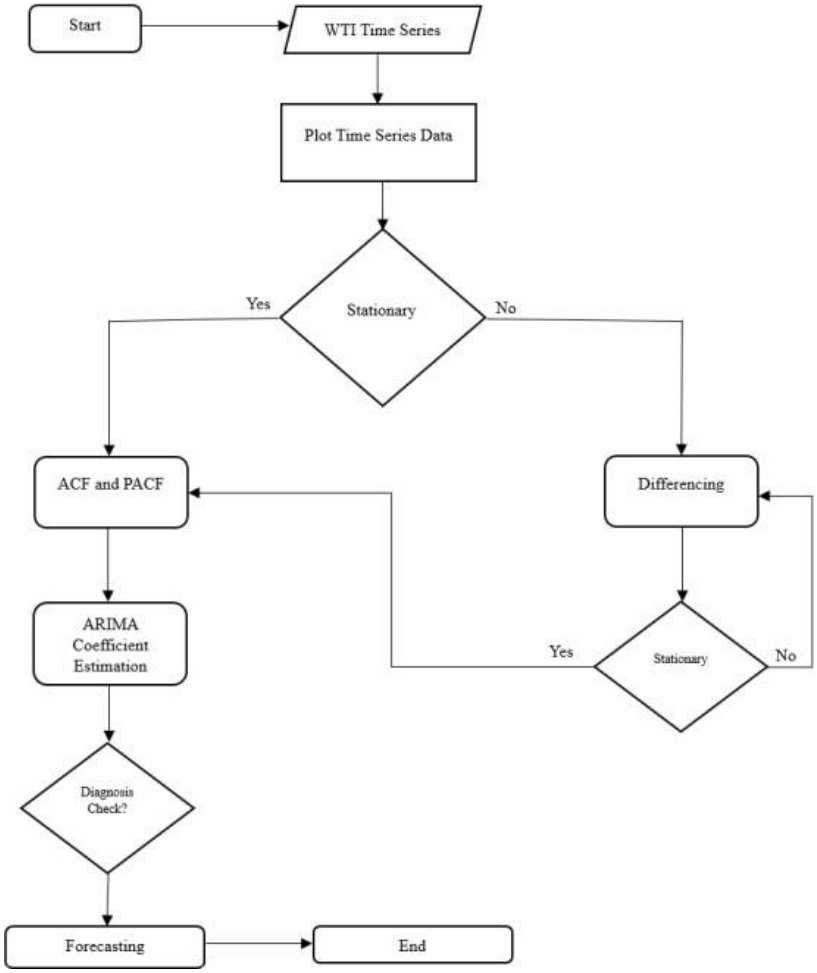


Figure 1. Box-Jenkins flowchart

Figure 1 illustrates the workflow of this research, starting by collecting the WTI data in a time series and then plotting the data. When the data is already stationer in mean and variance checks the ACF and PACF on the contrary if not stationer does the differencing until stationary. After checking the ACF and PACF determine the coefficient of the equation and do a diagnosis check the stationarity using the Ljung Box test and Shapiro test, if the equation fulfils the requirement it is allowed to be used, if not try to find the equation to use and the lastly use the equation to forecast.

III. Result and Discussion

A. Stationary Test

The data used is West Texas Intermediate (WTI) with monthly frequency from April 1, 2019, to April 1, 2023, was collected from Investing.com [8]. The first step is to check the stationarity of the data using the Augmented Dickey-Fuller Test (ADF-Test). A stationary test is a test to check whether the time series has a trend or seasonal component. if the p-value $< 0{,}05$, it indicates that the data is stationary, otherwise if p-value $> 0{,}05$ then the data is not stationary. Augmented Dickey-Fuller Test shows that the p-value is $0{,}5092 > 0.05$ indicating that the data is not stationer. In addition, it can be seen by the plot below that the plot does not perform any trend or seasonal component.

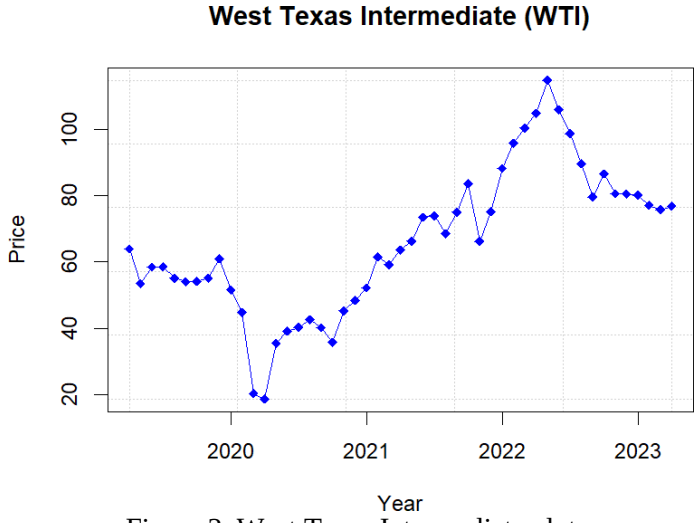


Figure 2. West Texas Intermediate plot

Because the data is not yet stationary, the next step is to perform differencing until the data is stationary and the p-value <0.05 is obtained. After the second differencing, $d = 2$, the ADF-Test give the p-value $0{,}01 < 0{,}05$. It indicates the data has been stationer and it is also proven by the plot.

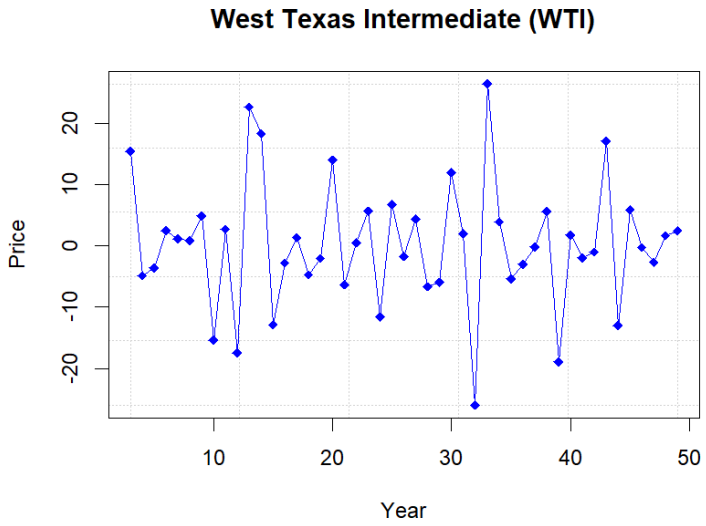


Figure 3. WTI Plot after second differencing

The process continues to determine the best ARIMA model for the data. The ARIMA Model specification is determined using the PACF and ACF. It is known that the ARIMA Model consists of p and q where p is Auto-Regressive (AR) determined by PACF, and q is Moving Average (MA) determined by ACF.

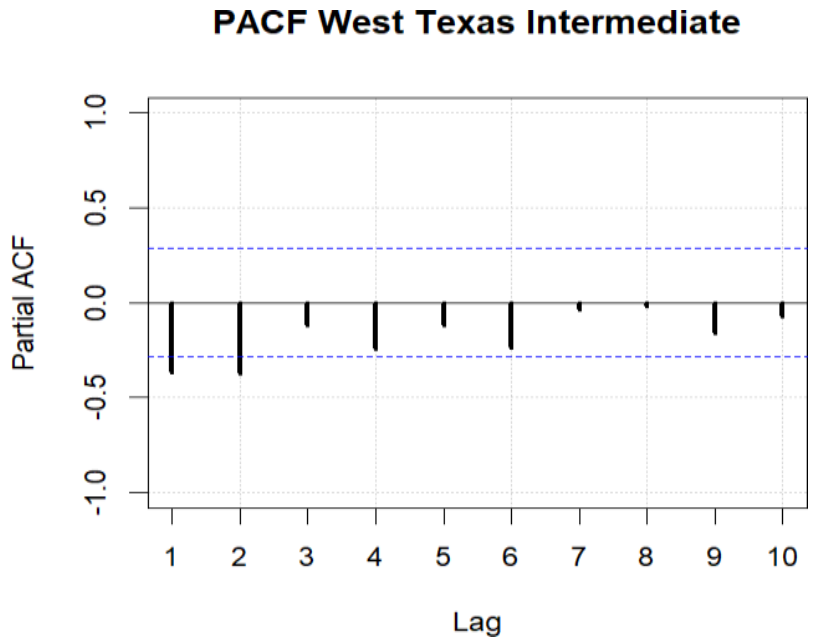


Figure 4. PACF West Texas Intermediate

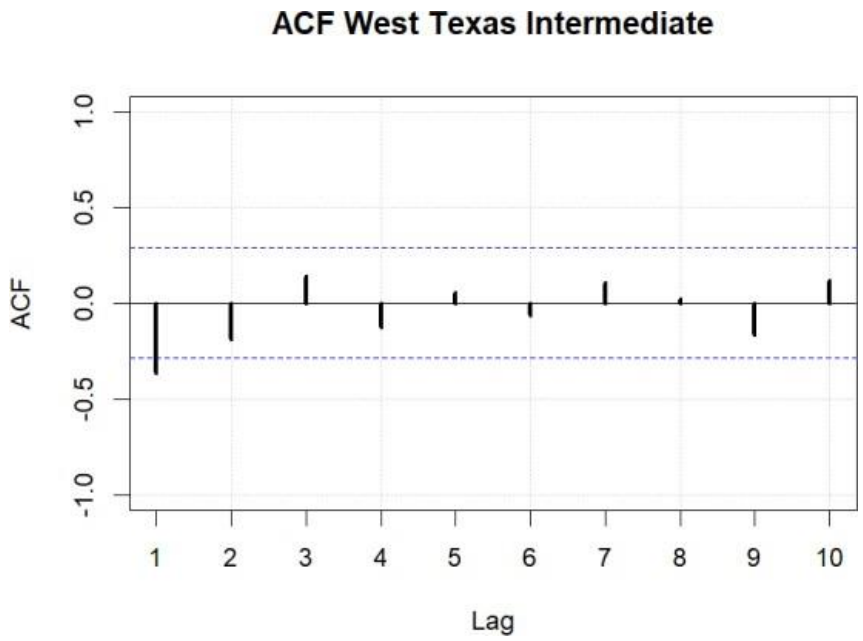


Figure 5. ACF West Texas Intermediate

From Figures 4 and 5, it can be seen that PACF cuts off the lag at 2 and the ACF at 1 respectively. Therefore, it can be determined that the value of p is 2 and the value of q is 1. In addition, from the previous section, 2 times differencing has been done to make the data stationary hence the value of d is

2. Based on these predictions, there are 6 ARIMA Model specifications.

Table 1. ARIMA model spesification

No	ARIMA Model	*p*	*d*	*q*
1	ARIMA (2,2,1)	2	2	1
2	ARIMA (2,2,0)	2	2	0
3	ARIMA (1,2,0)	1	2	0
4	ARIMA (1,2,1)	1	2	1
5	ARIMA (0,2,1)	0	2	1
6	ARIMA (0,2,0)	0	2	0

B. Parameter Estimation

Parameter estimates were obtained by computation using R-Studio including Log-likelihood and also AIC (Akaike Information Criteria).

Table 2. Parameter estimation of ARIMA model

No	ARIMA Model	AR1	AR2	MA1	Log Likelihood	AIC
1	ARIMA (2,2,1)	0,1404	-0,0833	-1	-164,3	336,61
2	ARIMA (2,2,0)	-0,525	-0,3833		-168,8	343,6
3	ARIMA (1,2,0)	-0,379			-172,51	349,03
4	ARIMA (1,2,1)	0,1315		-1	-164,46	334,92
5	ARIMA (0,2,1)			-1	-164,85	333,71
6	ARIMA (0,2,0)				-176,05	354.11

It can be seen from the table above that the smallest AIC is 323,09 and the ARIMA Model is ARIMA (0,2,1). It does not determine that the ARIMA (0,2,1) model is the best model, but it can be one of the indicators.

C. Residual Analysis

The candidate model will be selected by using the Shapiro test and the Ljung-Box test as the requirements. When the p-value is greater than 0,05 for the Shapiro test as well as the Ljung-Box test, the model passes the requirement. On the other hand, if one of which is smaller than 0,05, the candidate cannot pass the requirement as well as if the p-value is smaller than 0,05 for both tests.

Table 3. Residual analysis result

No	ARIMA Model	Ljung-Box	Shapiro Test	Result
1	ARIMA (2,2,1)	0,998	0,1313	Passed
2	ARIMA (2,2,0)	0,7568	0,3323	Passed
3	ARIMA (1,2,0)	0,3305	0,96237	Passed
4	ARIMA (1,2,1)	0,9538	0,1776	Passed
5	ARIMA (0,2,1)	0,3738	0,2113	Passed
6	ARIMA (0,2,0)	0,00769	0,1236	Not Passed

From the result of residual analysis, it can be seen that almost all of the ARIMA model passes the

requirements only the ARIMA (0,2,0). To narrow down the result and choose the best model, this research uses error analysis such as Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE).

Table 4. Error analysis result

No	ARIMA Model	ME	RMSE	MAE	MPE	MAPE	MASE	AIC
1	ARIMA (2,2,1)	1,035841	7,506651	5,760481	1,164626	10,90493	0.9762492	336,61
2	ARIMA (2,2,0)	0,294856	8,556785	6,306299	0,540115	12,79688	1,068751	343,6
3	ARIMA (1,2,0)	0,279887	9,291211	6,65291	0,801041	13,25261	1,127492	349,03
4	ARIMA (1,2,1)	0,982947	7,54569	5,739086	1,125247	10,89567	0,972623	334,92
5	ARIMA (0,2,1)	1,020228	7,587894	5,822466	0,983221	10,90372	0,986754	333,71
6	ARIMA (0,2,0)	0,23346	10,03471	7,106482	0,992173	14,44339	1,204361	354.11

Error analysis shows the smallest MAE, MAPE, and MASE value for the ARIMA Model (1,2,1). In addition, AIC (Akaike Information Criteria) shows the ARIMA Model (1,2,1) as the AIC with the second smallest value hence this forecasting will use ARIMA (1,2,1) as the most optimal model.

D. Forecasting

From the previous section, it is already found that this forecasting will use ARIMA (1,2,1) with the p =1, d = 2, and q = 1. The value of AR 1 (p) is 0,1315 and the value of MA 1(q) is -1. By these data, the ARIMA Model (1,2,1) can be expressed in the form of the equation below:

$$Y_t = 0{,}1315\, Y_{t-1} + e_t - e_{t-1} \tag{4}$$

The next step is plotting the data using ARIMA (1,2,1) which is presented below:

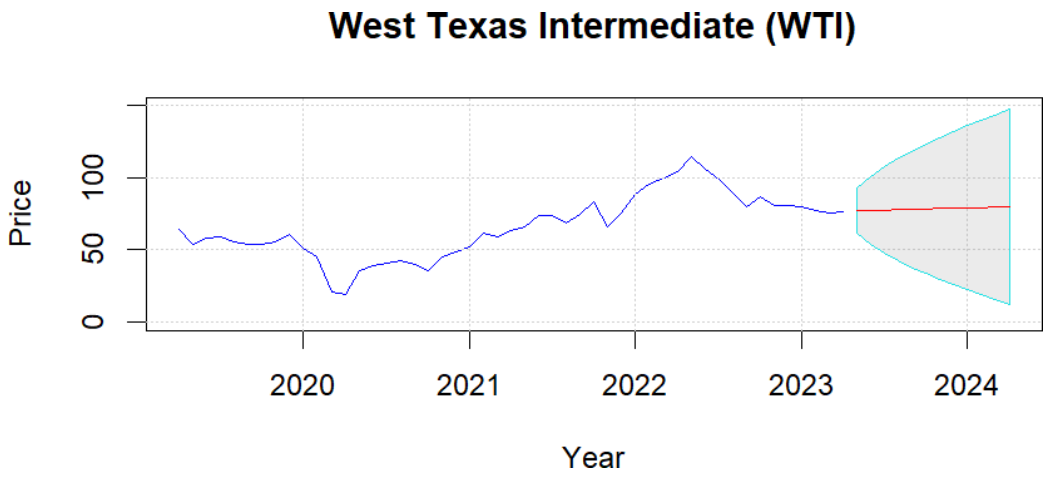


Figure 6. Forecasted value of ARIMA (1,2,1)

In Figure 6, the blue line in the grey area indicates the prediction data, and the grey area indicates the confidence level area of 95 % which means that the true population forecast lies somewhere between lower boundary and upper boundary values. In other words, there is a 95% chance that the actual value falls within this range.

Table 5. Forecasting Value of ARIMA (1,2,1)

Date	Forecast	Actual	Lower Bound	Upper Bound	MSE	MAPE	RMSE
01/05/2023	77,13204786	68,09	61,87591497	92,38818076	81,75862952	13,2795533%	81,75862952
01/06/2023	77,38443642	70,64	54,11616975	100,6527031	45,48742269	9,5476167%	45,48742269
01/07/2023	77,62372129	81,8	48,09939341	107,1480492	17,44130385	5,1054752%	17,44130385
01/08/2023	77,86128322	83,63	42,99429652	112,7282699	33,2780933	6,8979036%	33,2780933
01/09/2023	78,09861861	90,79	38,44499883	117,7522384	161,0711617	13,9788318%	161,0711617
01/10/2023	78,33592421	81,02	34,27175174	122,4000967	7,204262868	3,3128558%	7,204262868
01/11/2023	78,57322589	75,96	30,37143359	126,7750182	6,828949551	3,4402658%	6,828949551
01/12/2023	78,81052706	71,65	26,67904474	130,9420094	51,27314776	9,9937572%	51,27314776
01/01/2024	79,04782816	75,85	23,15071773	134,9449386	10,22610494	4,2159897%	10,22610494
01/02/2024	79,28512925	78,26	19,75531109	138,8149474	1,050889983	1,3099019%	1,050889983
01/03/2024	79,52243034	83,17	16,46984648	142,5750142	13,3047644	4,3856795%	13,3047644
01/04/2024	79,75973143	85,11	13,27684493	146,2426179	28,62537373	6,2862984%	28,62537373
Result					38,12917536	6,8128441%	6,174882619

Table 5 shows the comparison between the predicted value and the forecast value of WTI from May 1, 2023 until April 1, 2024. It indicates the forecasting runs well. The predicted value is not much different from the actual value and does not go outside the below and upper bounds. In addition, there are also error measurements such as MSE is 38,129117536, MAPE is 6,8128441%, and RMSE is 6,174882619.

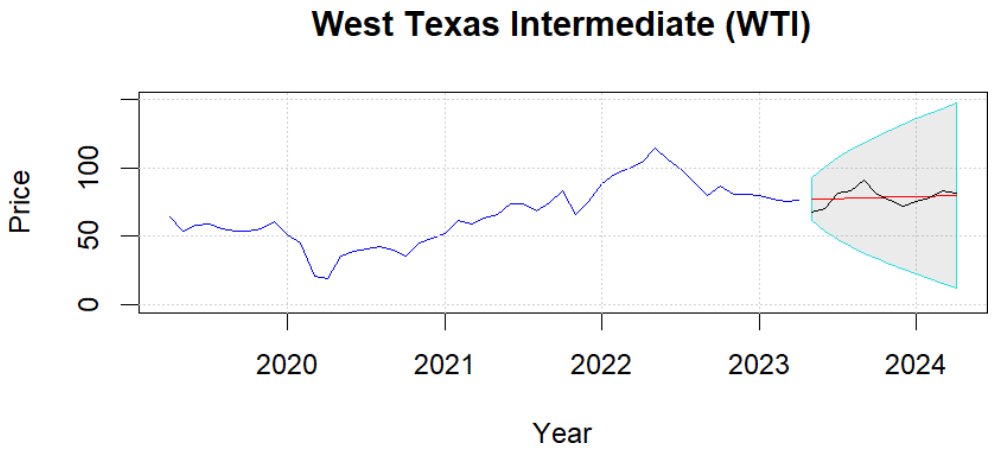


Figure 7. Comparison of the actual and forecasted

Figure 7 shows the comparison between the predicted value and the forecasted value from May 1, 2023 until April 1, 2024. It is shown that ARIMA (1,2,1) only catch the trending trend and does not catch the seasonality trend that is included in crude oil prices but the actual price is still in the range between lower bound and upper bound. This is caused by ARIMA struggles when the historical data has a seasonal pattern since it assumes the historical data is stationary in mean and variance. Therefore, ARIMA (1,2,1) is accurate to capture the trend but not for seasonality.

IV. Conclusion

This research uses data on oil prices in West Texas Intermediate (WTI) starting from April 1, 2019, until April 1, 2023, with monthly frequency obtained from investing.com. The result of the predicted price indicates the oil price continuously increasing from Mei 1, 2023-April 1, 2024 (12 months). The best model used to predict WTI oil prices is ARIMA (1,2,1) with the lowest value of MAE, MAPE and MASE with the following equation:

$$Y_t = 0{,}1315\, Y_{t-1} + e_t - e_{t-1}.$$

The research ran successfully since the predicted value does not go far away from the actual value. In addition, the MAPE value of this forecasting is 6,8128441% since it has a lower value than 10% indicating a small error between the predicted value and the actual value. The comparison plot proves that the ARIMA (1,2,1) successfully captured the trend but without seasonality. For the next research, it is recommended to do the forecasting with other methods such as SARIMA, GARCH, APARCH and even the Neural Network Model (NNM) to get more accurate results that can capture the trend and seasonality.

Acknowledgment

We would like to thank Study Program of Actuarial Science President University for supporting the research, and the anonymous reviewers for providing constructive comments to improve the earlier version of the manuscript.

References

[1] Environmental Protection Agency, "Crude oil." [Online]. Available: https://www3.epa.gov/carbon-footprint-calculator/tool/definitions/crudeoil.html#:~:text=Crude%20oil%20means%20a%20mixture,passing%20through%20surface%20separating%20facilities [Accessed: May 5, 2024].

[2] U.S. Energy Information Administration, "Use of oil." [Online]. Available: https://www.eia.gov/energyexplained/oil-and-petroleum- products/use-of-oil.php#:~:text=We%20use%20petroleum%20products%20to,intermediate%20and%20end%2Duser%20goods. [Accessed: Apr. 22, 2024].

[3] J. Chen, "West Texas Intermediate (WTI): Definition and use as a benchmark," Investopedia. [Online]. Available: https://www.investopedia.com/terms/w/wti.asp. [Accessed: Apr. 22, 2024].

[4] L. Derek, Time Series Modelling of Monthly WTI Crude Oil Returns. Open University, 2013.

[5] M. A. Rodhan and A. H. Jaaz, "Box-Jenkins modelling and forecasting of WTI crude oil price," *Proc. Int. Mult-Discip. Conf*, 2022. DOI: https://eudl.eu/doi/10.4108/eai.7-9-2021.2314841.

[6] M. N. Ramlan, "Evaluating forecast performance of Malaysian goods export for 2021-2022 with Box-Jenkins methodology and Arima model," *Foc. on Res. in Contemp. Econom.*, vo. 2, no. 2, 2021.

[7] D. Mashitasari and Soehardjoepri, "Coal Briquette Sales Forecasting Analysis Using the ARIMA Box-Jenkins Method," *J. Appl. Stat., Mat. Data Scie.*, vol. 1, no. 2., pp 42-51, 2023, DOI: https://doi.org/10.31537/estimator.v1i2.1695

[8] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and practice, 2nd ed. Monash University, 2018.

[9] "The Box-Jenkins method." [Online]. Available: https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/The_Box-Jenkins_Method.pdf. [Accessed: Apr. 17, 2024].

[10] J. D. Cryer and K. S. Chan, Time Series Analysis with Applications in R, 2nd ed. Springer, 2008.

[11] Investing.com. "Crude Oil Futures Price Today (WTI)," [Online]. Available: https://www.investing.com/commodities/crude-oil. [Accessed: Apr. 17, 2024].

[12] GeeksforGeeks, "ARIMA vs SARIMA model," Mar. 26, 2024. [Online]. Available: https://www.geeksforgeeks.org/arima-vs-sarima- model/#arima-vs-sarima-seasonality. [Accessed: May 5, 2024].