Oil Price Forecasting based on Improved SARIMA Model

Yaobin Wang *

Department of Economics, University of Michigan, Ann Arbor, 48109, US

* Corresponding author email: yaobin@umich.edu

Abstract. Oil Price is important to countries and people. As Black Gold, Oil is considered the blood of the industry. People in modern cities are surrounded by crude oil and its derivatives, such as petrol for cars, plastic products for storing food, etc. Thus, the changing price of oil can have a profound effect on people and countries. The prediction of oil prices can lessen the impact. Once people and the government know the future price of oil, then people and government can adjust their behavior. This study leverages the Seasonal Autoregressive Integrated Moving Average (SARIMA) model to find the parameters of estimation and forecasts. The study uses Brent crude oil prices from May 1987 to July 2022. The study result shows that the price of oil is hard to predict based on the Brent crude oil prices. The study also discussed some protentional reasons why the SARIMA model cannot predict the price so accurately even if the method is correct.

Keywords: Oil Price; SARIMA Model; Price Prediction; Crude Oil.

1. Introduction

After years of development, oil as the blood of industry became one of the most critical sources of energy. People and industries all over the world are directly or indirectly consuming oil and its derivatives, so the changing price of oil affects the costs in all production sectors. From the economic perspective, a lot of countries' economy is depending on the trading of oil and its products, in this way, the prediction of the oil price is a prior task. By predicting the oil price, the governments or people can make different police or adjust behavior to eliminate oil price changes' effect on their economy or business. From 2008 -2009, the world is suffering from a finical crisis, the oil price dropped from $133.88 per barrel to $39.09 per barrel. It brings a huge fiscal deficit to those countries which are highly dependent on oil products. From 2019 to 2022, the world is suffering from some events and those events affect the oil price profoundly. According to data from datahub, oil prices have increased from $ 63.96 to $ 111.93 [1]. The price is almost twice that of two years ago. It damaged some countries' economies like German, France, etc which are highly dependent on oil imports. The unstable oil prices bring a lot of trouble to people and the government. People have to spend more money on energy and product, and the government have to spend more money on import and get less money from taxes.

The Time series model has been widely used in science areas, but there is a limited application for forecasting oil prices. The SARIMA model as one of the time series models can describe time series data that have non-stationary behaviors across seasons, and it only requires very few datasets. Also, major statistical software like Stata, SPSS, and SAS can easily implement the SARIMA model [2]. Oil prices are mostly depending on demand and supply. The demand has seasonal trends such as during the winter and summer, the demand for oil will increase and during spring and fall will decrease. So, using the SARIMA model will be suitable for this study.

Most research shows a great prediction of the oil price based on the Brent Oil Price index or WIT by using the SAIMA model. However, using the same data and similar methods, the prediction is not quite like what appeared in their paper. This study shows that based on the historical data of oil prices cannot make a such great prediction. Because too many uncertain elements can affect oil prices and make predicting oil prices so hard. By implement the SARIMA model, it shows a great difference between the real data and predictions, which means the model is not performing well. There are some protentional reasons such as the model is not considered sudden events. Oil prices can be affected by many external factors, such as the finical crisis, covid-19 causing a decrease in oil demand, or the

sudden increase of the US oil production technology causing an increase in oil supply. Those sudden events are not predictable which can make the model inaccurate. Also, the data, itself may have problems too, Machine Learning uses historical data to train the model, but it is hard to ensure that the historical data will have the same pattern or relationship to the current or future data, especially for the oil price. Oil as a product will be affected by inflation, so the price from the previous date will not have the same value as the recent. The price drop or increase caused by a sudden event will not have the same effect nowadays as past. It can be a reason why the prediction is not as accurate as expected.

2. Method

This section's first part is dataset and processing, the second part is the method of the SARIMA model, and the following is how to utilize the SARIMA model to make the prediction. Finally, using the MSE and RMSE to show how the model performs.

2.1 Dataset and Processing

The data is from one of the most important oil price indicators – the Brent Complex. Dated Brent is a barometer for the oil market, and the price in the Brent complex is a reference in physical terms and spot deals daily. Oil companies use the Brent Complex as a reference in their official trading, and governments use it to tax [3]. The data shows the price of oil from 1987 to 2022 by month. There is a total of 442 data points. Before using the dataset, SARIMA modeling needs a stationary dataset because stationarity data is relatively easy to predict: the future statistical properties should be the same as in the past [4]. To check the stationarity, using the null hypothesis test will be good for this study. To find the p-value, python provides a method called “adfuller” to check the p-value. If the p-value is smaller than 0.05, it means the dataset is stationarity. If the p-value is greater than 0.05, the dataset needs to use the difference which is

$$Dif_{\_1} = Y(t) - Y(t-1) \tag{1}$$

to make the dataset become stationary. However, for the SARIMA model, there is always a seasonal pattern, so the difference should consider the pattern, the formula for SARIMA will be

$$Dif_{1} = Y(t) - Y(t-pattern) \tag{2}$$

2.2 SARIMA Model

The SARIMA model is a time series model that has a regular pattern of changes over an S number of periods. SARIMA model, seasonal AR, and MA terms predict using data values and errors at times with lags that are multiples of S [5]. The notation for the SARIM model is

$$ARIMA(p,d,q) \times (P,D,Q)S \tag{3}$$

with non-seasonal AR: p, non-seasonal differencing: d, non-seasonal MA: q, seasonal AR: P, seasonal differencing D, seasonal MA Q, and period S. The formal formula for SARIMA is

$$\Phi(B^s)\emptyset(\mathrm{B})(x_t - \mu) = \Theta(B^s)\theta(\mathrm{B})w_t \tag{4}$$

where the non-seasonal component is:

$$\begin{aligned} &AR\text{: } \emptyset(B) = 1 - \emptyset_1 B - \cdots - \emptyset_p B^p \\ &\mathrm{MA}\text{: } \theta(B) = 1 + \theta_1 \mathrm{B} + \cdots + \theta_{\mathrm{q}} \mathrm{B}^{\mathrm{q}} \end{aligned} \tag{5}$$

The Seasonal component is:

$$\begin{aligned} Seasonal\ AR: \Phi(B^S) &= 1 - \Phi_1 B^S - \cdots - \Phi_P B^{PS} \\ Seasonal\ MA: \Theta(B^S) &= 1 + \Theta_1 B^S + \cdots + \Theta_Q B^{QS} \end{aligned} \tag{6}$$

To find the MA can use an autocorrelation function (ACF) plot of the differenced series: it shows the coefficients of correlation between a time series and lag in a bar chat [6].

$$\hat{Y}_t = \mu + \emptyset_1 Y_{t-1} \tag{7}$$

To find the AR can use partial autocorrelation (PACF) plot of the differenced series: it is similar to ACF, but it shows partial correlation coefficients between the series and lag in a bar chat [6].

$$\hat{Y}_t = \mu + Y_{t-1} - \theta_1 e_{t-1} \tag{8}$$

To estimate the parameter of the SARIMA, there is a way to draw a graph of ACF and PACF. By looking at the graph, there are some points lying in the confidence interval, and those points are potential perimeters for SARIMA. However, there is another way, the study will use called grid search, which is a method to automatically test all combinations of parameters to find the optimal parameters for the model. Also, to check the best model parameters, the study uses Akaike's information criterion (AIC) to compare the quality of a set of statistical models to each other [7]. The combination with the smallest AIC is the best combination for this model to perform.

2.3 Evaluation Matrix

To evaluate the performance of the model, this study will use MSE and RMSE. Mean squared error (MSE) measures the amount of error in statistical models. It evaluates the average squared difference between the observed and predicted values [8]. The formula for MSE is

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \tag{9}$$

The root mean squared error is the root of MSE the formula is:

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}} \tag{10}$$

The smaller value for MSE and RMSE shows the error is small and the performance is good.

3. Result

This section will show the result of using the SARIMA model to predict the oil price. The first part will be the data analysis, and this part will implement the data processing from the Method part. The following part will show the result of SARIMA and the prediction.

3.1 Data Analysis

First, the study will graph the data from the Brent oil price complex, data shows the price of oil from 1987 to 2022 by month.

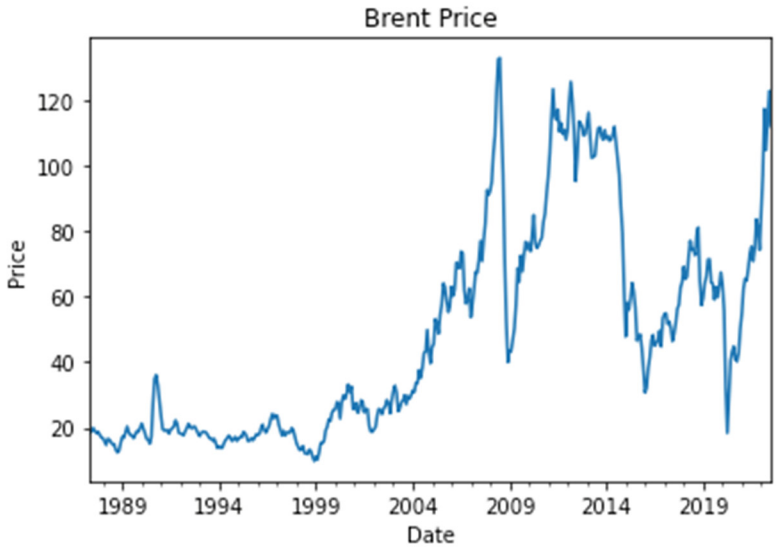

Fig 1. Brent oil price complex

Figuer 1 shows there are a lot of ups and downs, and to decompose the data can be better to see the inside of the data. Figure 2 is the decomposing of the data.

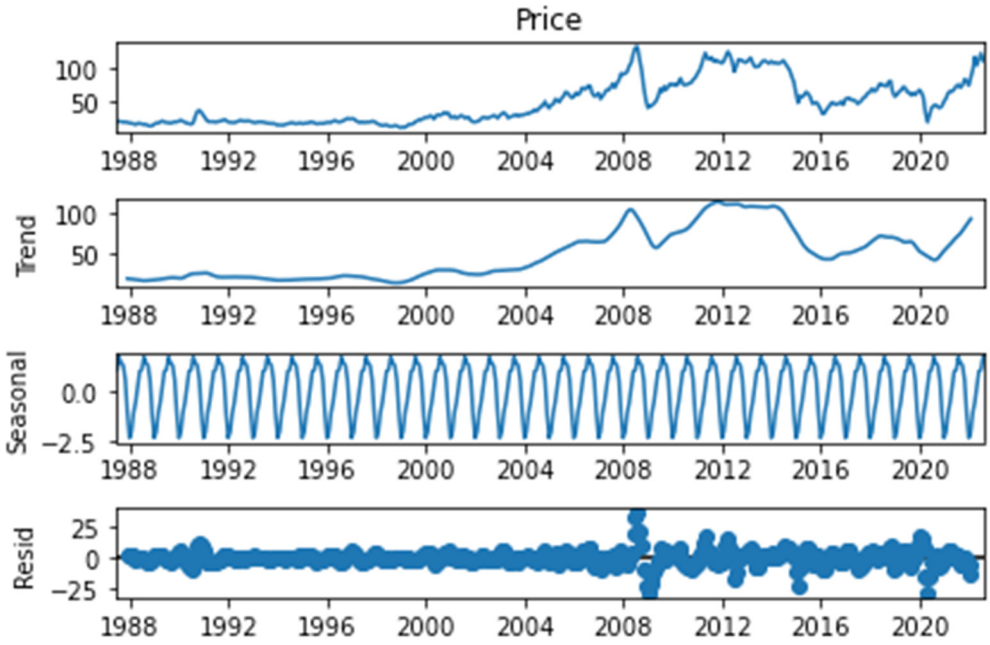

Fig 2. The decomposition of Brent oil price complex

Figure 2 shows the raw data of oil prices has a trend and seasonal pattern, and one of the critical characteristics of utilizing the SARIMA model is to make sure the data is stationary. The decomposition graph shows there is a trend and seasonality which means the data is not stationary. To be more numerical, the study uses adfuller to check if the data is stationary.

The p-value from the adfuller test is 0.315 which is greater than 0. 05 and it means the null hypothesis is rejected and failed to prove this data is stationarity. This shows that the data is not stationarity. The following step is to make the data stationary for future use. This study uses the seasonal difference. From figure 2, the seasonal pattern is about 12. So, the diff_1 function will be

$$Dif_{\_1} = Y(t) - Y(t-12) \tag{11}$$

By using the adfuller to test the dif_1 data, the p-value from adfuller test is 0.000388 which is smaller than 0.05 and this can prove that this data set is stationary and the graph of the diff_1 can also show there is some pattern in the data. This diff_1 will be the data that this study uses to train and predict.

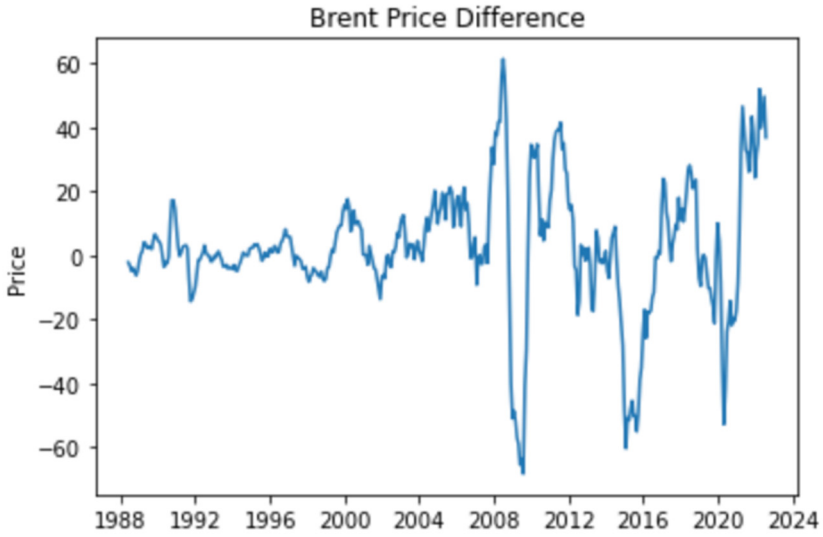

Fig 3. The graph of diff_1

3.2 Results of the SARIMA Model

The SARIMA model requires 7 different perimeters. First, to make finding perimeters easier, use a graph to estimate. For SA and MA using PACF and ACF to do the estimation. Figure 4 shows the result of graphing.

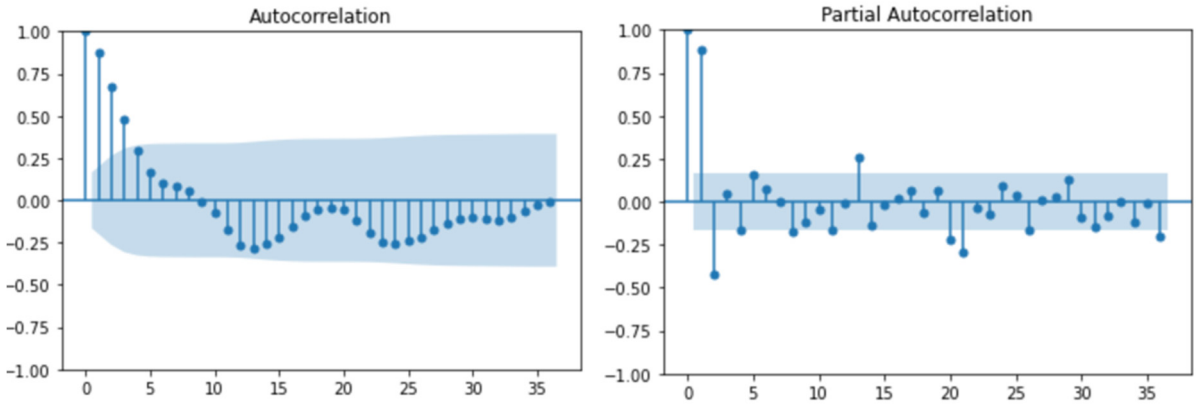

Fig 4. Graph of ACF and PACF for diff_1

However, using a grid search is a better way to find the best combination of 7 different perimeters. And find the smallest AIC to make sure it is the best one.

From table 1, the best combination of the SARIMA model parameters will be (1,1,0) (8,1,1) [12]. Those parameters will be used to build the prediction model. There is a portion of historical data will be used as training data and another portion will be testing data. In this study, the diff_1 from year1988 to 2008 will be the train data and the rest will be testing data.

After deciding all parameters and training data, the next step is to train the data and show the result. Figure 5 is the result.

Table 1. The gid search for the SARIMA model

ARIMA (1,1,0) (2,1,0) [12]	AIC=1729.847
ARIMA (1,1,0) (3,1,0) [12]	AIC=1688.450
ARIMA (1,1,0) (4,1,0) [12]	AIC=1663.793
ARIMA (1,1,0) (5,1,0) [12]	AIC=1647.370
ARIMA (1,1,0) (6,1,0) [12]	AIC=1635.279
ARIMA (1,1,0) (7,1,0) [12]	AIC=1622.443
ARIMA (1,1,0) (8,1,0) [12]	AIC=inf
ARIMA (1,1,0) (7,1,1) [12]	AIC=inf
ARIMA (1,1,0) (6,1,1) [12]	AIC=inf
ARIMA (1,1,0) (8,1,1) [12]	AIC=1590.011
ARIMA (1,1,0) (8,1,2) [12]	AIC=1591.868
ARIMA (1,1,0) (7,1,2) [12]	AIC=1590.468
ARIMA (0,1,0) (8,1,1) [12]	AIC=1631.627
ARIMA (2,1,0) (8,1,1) [12]	AIC=1590.310
ARIMA (1,1,1) (8,1,1) [12]	AIC=1590.730
ARIMA (0,1,1) (8,1,1) [12]	AIC=1600.733
ARIMA (2,1,1) (8,1,1) [12]	AIC=1592.068
ARIMA (1,1,0) (8,1,1) [12]	AIC=1591.998
Best model:	ARIMA (1,1,0) (8,1,1) [12]

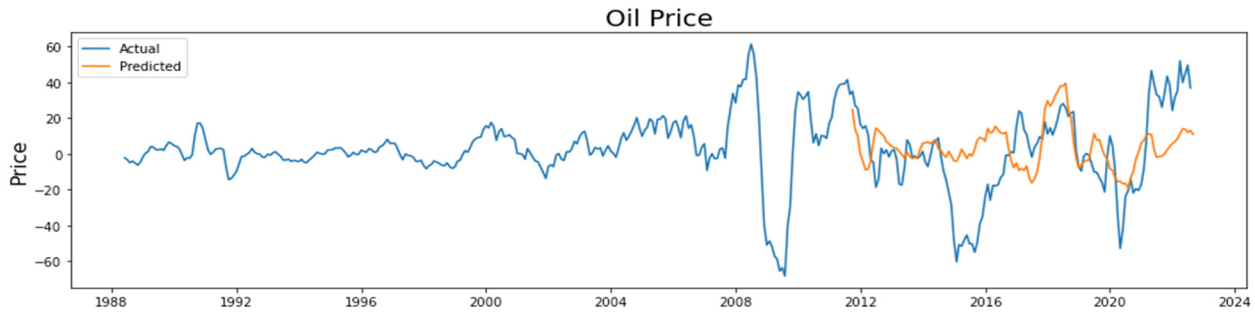

Fig 5. The result for the SARIMA model

3.3 The Evaluation of the SARIMA Model

The graph shows a similar trend to the historical data, but to determine whether a model is well-performed need to use a statistical method such as MSE and RMSE. Table 2 shows the MSE and RMSE between prediction and real data.

Table 2. The MSE and RMSE for the prediction

Mean Squared Error	4856.43
Root Mean Squared Error	69.69

Table 2 shows a large value of SME and RMSE which can mean the prediction is not so accurate and the performance of the model is not well.

4. Discussion

The result shows the prediction model is not performing well. There are some protentional reasons. From the graph in Figure 5, there is a big change during the period of 2008 – 2009. This period is well known as the “Global Financial Crisis.” The price of oil reached a high of $ 133.88 in June 2008

and went down to the lowest of $39.09 in February 2009. During the Financial Crisis, U.S. GDP went down about 0.3% and 2.8% in 2008 and 2009, while the unemployment rate reach about 10% [9]. As many people lose jobs, the demand for production will decrease. The factories will decrease their demand for oil due to the lower production demand. That could be a reason why the oil price drop is unusual, and this event will cause the machine to learn in the wrong way. The prediction will be not as accurate as it could be.

From 2015 to 2016, there is a huge gap in Figure 5. During this period, one of the most critical technical reforms happened in the oil industry - rising efficiency gains in US shale oil. The increase in efficiency leads to a lower break-even price thus increasing the production of the oil [10]. However, OPEC, an oil and gas industry cartel, tried to increase the production of oil to pressure American producers out of the market. It accelerated the decrease in oil prices.

From 2019 to 2022, the world suffered from Covid-19. During the pandemic period, there is about a 10% reduction in oil consumption because of reduced air and road travel. The quarantine policy reduces the oil demand for the industries [11]. In the following year, the conflict between Russia and Ukraine had a profound effect on the oil price. Russa has a potential tendency to pull large volumes of oil and oil products from the European market by the end of 2022, and the price of crude oil could increase from $150 to $200 [12]. The huge up and down in the oil price between 2019 to 2022 are affected by those events.

Those sudden events are not predictable and will be hard to consider in the model due to the uncertainty of the future. That could be a reason why this model is not perfumed well.

The data from Brent is one of the most prestige data for the oil market. However, by looking at the first two decades of the Brent Oil Price, the fluctuation is not as huge as in the last two decades. Using the first few decades as train data will not be able to build a great model for prediction. Because the relationship is not strong enough to predict the future. While using the data does not have a certain pattern that can be traced, the model will not be able to predict accurately.

5. Summary

Oil is one of the most important resources to people and government, its price can have a profound effect. Predicting the price of oil can be one of the most important tasks. However, the price of oil can be affected by many factors and most of those factors are hard to be predicted and eliminated in the model. This will make the model hard to predict accurate data by using historical data like Brent Oil Price. Even though this study follows the steps of using the SARIMA model, using the difference to make the data stationarity, the prediction of this model is not performed well, its prediction has a very big difference from the real price. Using the historical data from 1987 to 2011, to predict the price from 2011 to 2022 only has an RMSE of 69.9 which is a large number. Discussing the model finds there are some reasons, why this model does not perform well. First, there are many outside events that can have a profound effect on the oil price such as the finical crisis, the Covid -19 pandemic, and the Russia-Ukraine conflict. Those events are not predictable and are hard to be considered and eliminated in the model. However, if these events are not considered, the model will not have a great performance. Secondly, the oil price data shows there is not a certain pattern that can be considered to predict the oil price. For the first few decades, oil prices do not have huge fluctuations, but in the recent two decades, the oil price changes a lot. The prediction will be hard to be accurate if using the non-fluctuation data to predict the fluctuation data.

In another word, there is a lot of research showing the prediction of oil prices, but according to this study, the prediction of the oil price will be unattemptable by using the Brent Oil Price. There are too many affections can affect the oil price and they are hard to be considered.

References

[1] Datopian. (n.d.). Brent and WTI spot prices. DataHub. Retrieved September 17, 2022, from https:// datahub. io/core/oil-prices.

[2] Zhang X, Liu Y, Yang M, et al. Comparative study of four time series methods in forecasting typhoid fever incidence in China. PloS one, 2013, 8(5): e63116.

[3] S&P Global. (2021, July). Dated Brent Price Assessment explained. Explained | S&P Global Commodity Insights. Retrieved September 17, 2022, from https://www.spglobal.com/commodityinsights/en/our-methodology/price-assessments/oil/dated-brent-price-assessment-explained.

[4] McCabe B P M, Tremayne A R. Testing a time series for difference stationarity. The Annals of Statistics, 1995: 1015-1028.

[5] Liu L M. Identification of seasonal ARIMA models using a filtering method. Communications in Statistics-Theory and Methods, 1989, 18(6): 2279-2288.

[6] Schaffer A L, Dobbins T A, Pearson S A. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. BMC medical research methodology, 2021, 21(1): 1-12.

[7] Glen S. Akaike's Information Criterion: Definition, Formulas, Statistics How To. 2019.

[8] Wang Z, Bovik A C. Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE signal processing magazine, 2009, 26(1): 98-117.

[9] Farmer, Roger EA. The stock market crash of 2008 caused the Great Recession: Theory and evidence. Journal of Economic Dynamics and Control, 2012, 36(5): 693-707.

[10] Weijermars R, Johnson A, Denman J, et al. Creditworthiness of North American oil companies and Minsky financing categories: assessment of shifts due to the 2014-2016 oil price shock. Journal of Finance and Accounting, 2019, 6(6): 162-180.

[11] Bildirici M, Guler Bayazit N, Ucan Y. Analyzing crude oil prices under the impact of covid-19 by using lstargarchlstm. Energies, 2020, 13(11): 2980.

[12] Khudaykulova M, Yuanqiong H, Khudaykulov A. Economic Consequences and Implications of the Ukraine-Russia War. International Journal of Management Science and Business Administration, 2022, 8(4): 44-52.