Forecasting Crude Palm Oil Price with Statistical Methods

Shahreedza Anissa Bin Mohd Saidi  
*Masters of Data Science (2021488378)  
College of Computing, Informatics and Media*  
*Universiti Teknologi MARA (UiTM)*Shah Alam, Malaysia  
shahreedzasaidi@gmail.com

# Introduction

Crude Palm Oil also known as CPO is among the widely trade commodities in recent years. Nothing to be surprised as, the competitive price of CPO resulted in higher demand each and every single year. Malaysia alone is the second largest exporter of CPO for 2021 (Kondalamahanty, 2021) behind Indonesia. CPO being traded on the daily basis on Bursa Malaysia with prices available to be download on Malaysia Palm Oil Council (MPOC) website.

CPO under the category of vegetable oil, other examples include soybean oil, corn oil and sunflower oil to name a few.

Over the past decade, the competition of the vegetables oil remains fierce as who can offer the most competitive price. Vegetable oils are widely used for cooking, cosmetics, and fuel source.

Goldstein Research forecast that the global vegetable oil market size is set to reach 252 million metric ton by 2025. Further expansion of demand would result in bigger number.

However, CPO has been criticized for its environmental impact. Oil palm plantations are often associated with deforestation, habitat loss, and biodiversity decline. There are also concerns about the use of forced labor and child labor in certain palm oil plantations. To be fair, other vegetable oils also give the same impact on the environment. Soya bean for instance also use the land and corn oil perform the same process. With the current expanding world, we cannot expect other commodities to be environmentally sound than the other.

As CPO price being traded as futures on the market. The ability to forecast the price enable traders or even manufacturers to anticipate demand changes. Manufacturers can plan their production and implement economic of scale to improve profitability. Traders on the other hand can take advantage of future prices with the advantage of knowing when to buy and sell. This shows the importance of CPO price forecasting.

# Problem Statements

Crude Palm Oil like any other traded commodities has price that move up and down the market. The strategies any person that take advantage of any movement would need to know what going to happen in the future. Higher demand results in higher prices while lower demand results in unfavorable situations for person that is holding the commodities. Manufactures or even trader would be happy to know what is going to happen in the future. They can adjust their position to ensure every cent of profit could be cash in. The rapid change in demand makes any forecasting study a problem. This study aim is to develop a reliable forecasting methodology that can be used to forecast the price or movement of CPO price on a timely basis.

# Research Questions and Research Objectives

## Research Questions

### What is the model that can forecast the CPO price reliably and accuractely?

### What is the best performing model with the lowest error to forecast the CPO price?

### What is the best parameters for each model to ensure the error is minimise in forecasting the CPO price?

## Research Objectives

### To compare the performance of each model in forecasting the CPO price.

### To analyse the performance indicator to find the best model with the lowest error in forecasting the CPO price.

### To analyse the data and conduct statistical test to find the best parameters for each model in forecasting the CPO price.

# Related Works

A study conducted by (Al-Khowarizmi, Nasution, Lubis, & Lubis, 2020) use data from August 2019 to September 2019. They use Simple evolving connection system (SECos) to forecast the CPO price. SECos is a version of deep learning with three neuron layers. The techniques improve the forecasting model by adjusting the weight of the coefficients.

After the model have been develop, they then validate the model by using Mean Absolute Percentage Error (MAPE) to check the performance and accuracy of the model. The MAPE is very low equal to 0.035%.

ARIMA, ARIMAX and ARDL was used to forecast the monthly CPO price from 2008 to 2017 conducted by (Khalid, Nur Ahmad Hamidi, & Thinagar, 2018). They found that ARIMAX model is the most accurate and the most efficient as compared to ARDL and ARIMA in forecasting the CPO price. They also point out that the CPO price is highly influenced by the stock of palm oil, crude petroleum oil price and soybean oil price.

(Kanchymalay, Salim, & Krishnan, 2019) take a different approach of using the weather variables, namely, temperature, rain amount, pressure, humidity, and radiation as well as past CPO price. CPO price of 10 years were collected from MPOC, and the environmental parameters were from meteorology department of Malaysia from a period of 2005 to 2016. LSTM (Long Short-Term Memory) a form of deep learning was used to forecast the CPO price. Their study shows an average accuracy of 90% in forecasting the CPO price.

On the other hand, Holt-Winters including multiplicative Holt-Winters and additive Holt-Winters were used to forecast the crude palm oil production and prices in Thailand. (Suppalakpanya, Nikhom, & Booranawong, 2019) collected data from January 2005 to February 2018 from the local government of Thailand. The results show that the EAHW method provides goods accuracy as indicated by the mean absolute percentage error (MAPE). Furthermore, applying the different initial trend values to the forecasting methods, the level of forecasting accuracy is significantly different.

(Khamis, Hameed, & Nor, 2018) try to forecast the price of palm oil in Malaysia for the next years based on the period of 31 years. Methods wise, ARIMA, ARCH and GARCH were selected to forecast the price. AIC and H-Q statistics were used to obtain the best model. They found that ARIMA(2,1,5) performed better compared to ARCH and GARCH models.

# Methodology

A diagram of a data flow

Description automatically generated

Figure V.1: Methodology of the Study

## Data Collection

Data of the series came from MPOC website. An excel file is downloaded then transformed into a Comma Separated Value (CSV) file. Transformation includes aggregating all of data in multiple column into two columns consisting of Data and Price.

## EDA (Explanotory Data Analysis)

The purpose of this EDA is to understand the data much better. Among the things that being watched at, are the trend, seasonality, cyclical events or even if there are any outliers inside the series.

## Statistical test and Correlogram

Statistical test conducted in this study include Augmented-Dickey-Fuller test. The Augmented Dickey-Fuller test is a statistical test to determine whether a time series dataset is stationary or exhibits a unit root, which implies non-stationarity. If the p-value of the ADF test is below the significance level, we may reject the null hypothesis and conclude the series is stationary.

Adding to that, an autocorrelation plot is plotted to determine the autocorrelation coefficient of the series. An ACF plot shows the correlation strength of the series with its past lags. The higher the nodes in the plot, the stronger the series with its past lags. Furthermore, if the nodes are greater than the blue region in the plot, the lags are considered to be statistically significant, indicating a higher likelihood that the observed autocorrelation is not due to random chance.

Lastly, would be plotting a Partial Autocorrelation plot (PACF). Different from and ACF plot, the PACF used to measure the direct relationship between observations at different lags of the series while managing the effects of intermediate lags. Same goes to PACF, if the nodes exceed the blue region, it indicates that the lags are statistically significant and implying that the observed partial autocorrelation is unlikely to random chance.

## Split the data.

The splitting of the data is because for testing and validating the model. For this study a ratio of 80% for training and 20% for testing is used. The 80% of the data will be used to train and develop the model with the right parameters. Then, the remaining 20% will be use as a testing parameter. The testing data set then will be feed through a performance indicator to determine if there is any way, the model could be improved further.

## Train the model

Since in this study is using statistical tools such as AR, MA, ARMA and ARIMA. The parameters such as p, d and q values are being set by same statistical test mentioned earlier. The p value by the PACF, the q by ACF while the q is set by differencing the series with lags.

## Test and Validation

Testing and validation are by using Root Mean Squared Error (RMSE). RMSE is calculated by taking the square too of the mean of the squared differences between the predicted values and the corresponding actual values. It is expressed in the same units as the variables being predicted, which makes it easily interpretable and comparable with the original data. The RMSE figure than will be a measuring stick to gauge the performance of the model in forecasting the training data.

## Model Comparison

The best model for each model will then be compared to each other with the RMSE. The model that has the lowest RMSE will be selected as the best model. Not just that the performance of the test and the predicted value but the performance of the training and the model itself. The purpose of this is not show if the best model is showing any overfitting issues. If the training performance is lower than the testing performance, it can be concluded that the model is overfitting. Overfitting is the issue when the model tries to fit the training data and not generalizing the overall formulation of the model.

# Data Understanding

## Description of the Data

The dataset used in this study was retrieved from the website of Malaysian Palm Oil Council (MPOC) (Malaysian Palm Oil Council, 2023) . MPOC used the data from Bursa Malaysia as Bursa is the place that the CPO being traded daily. Data from January 2022 to December 2022 was extracted into an excel file. From the excel file, it then transformed into a Comma Separated Value (CSV) file. There are two columns representing the date and the close CPO price for the day. A total of 234 rows of data were being used in this study. The data is less than a calendar year is because of public holidays and weekend that results in close counter for the CPO to be traded.

## Exploratory Data Analysis (EDA)

### Visual Plot

A graph of a graph

Description automatically generated

Figure VI.1: Daily Closing Price FYE 2022

### ADF test statistic

Figure V.1 is the plot of daily closing CPO price for the year ending 2022. From the year alone, the CPO price already fluctuated massively in the first half. The price drops from around RM 6500 to RM 4000 in a few months alone. This shows how volatile the CPO is. After the drop in the price, the CPO then comes to a steady halt. This shows how the market are adjusting to any changes and stabilize after the peak. Anticipating for the volatility is a difficult task adding to the fact the trend looks like this.



Figure VI.2

Local and global test shows the series are not stationary. Local means are not the same as global mean. From Figure V.1 alone, visually we can identify that. This assumption is supported by the ADF-statistics in Figure V.2. The p-value is larger than the significance level of 5%. is not rejected. Therefore, the original series is not stationary.

### Autocorrelation and Partial Autocorrelation

A graph of a graph

Description automatically generated

Figure VI.3

Figure V.3 is the autocorrelation of the original series. The series is correlated with lag 1 until lag 20. From lag 1 to 20, the height of the nodes shows the strength of correlation between the series itself and the past lags. If the notes are above the blue region, it indicates the lags are statistically significant and higher likelihood that the observed autocorrelation is not due to random chance.

A graph of a graph showing the same number of points

Description automatically generated

Figure VI.4: Partial Autocorrelation of the Series.

Figure V.4 is the PACF of the series. Different from autocorrelation, PACF in the other hand measures the relationship between an observation in a time series with observations at prior time stamp with the relationship of intervening observations removed. From this plot alone, the series is partially correlated with lag 1 and 2. For lag 1, the series is highly correlated while lag 2 is still correlated but at a lower degree as compared to lag 1. Nevertheless, both are statistically significant as the nodes exceed the blue line of the plot.

# Results

### Auto-regressive (AR)

A screenshot of a computer screen

Description automatically generated

Figure VII.1: AR (2)

The above Figure VII.1 is an AR (2) model. To make it simple, we just look at the p-value of the coefficient. After being tested with multiple *p* order, the AR (2) coefficient all have p-value lower than the significance level. It shows the coefficient are statistically significant to be accepted into the model.

A graph with numbers and lines

Description automatically generated

Figure VII.2

The AR (2) model plotted above shows the model try to fit to the training data. The model able to capture the first upward trend but failed to capture the downward trend. The model is great at modelling the training data with constantly following the blue line. This among the main concern of overfitting the data.

### Moving Average (MA)

A screenshot of a computer screen

Description automatically generated

Figure VII.3

Figure VII.3 is the summary of the MA (2) model. After being tested with multiple *q* order, the MA (2) coefficient all have p-value lower than the significance level. It shows the coefficient are statistically significant to be accepted into the model.

A graph of different colored lines

Description automatically generated

Figure VII.4

The MA (2) model plotted above shows the model try to fit to the training data. The model is not able to capture the movement of the testing data by a huge margin. It can be concluded that MA (2) is not for the data. Furthermore, with the training data, the model failed to capture the movement of the data itself. Therefore, it failed to forecast for the future value of the data.

### Auto-regressive Moving Average (ARMA)

A screenshot of a computer

Description automatically generated

Figure VII.5: ARMA (1,1)

The above Figure VII.5 is an ARMA (1,1) model. The p-values of the coefficient are lower than the significance level.

Therefore, the coefficient is statistically significant.

A graph with numbers and lines

Description automatically generated

Figure VII.6

The ARMA (1,1) model plotted above shows the model try to fit to the training data. The model able to capture the first upward trend but failed to capture the downward trend. The model is great at modelling the training data with constantly following the blue line. This among the main concern of overfitting the data. As the model able to capture the training data but failed for the testing data due to high error.

### ARIMA

A screenshot of a computer

Description automatically generated

Figure VII.7

ARIMA(1,2,1) is finalized as the coefficient p-value are lower than the significance level. Therefore, the coefficients are statistically significant.

A graph with numbers and lines

Description automatically generated

Figure VII.8: ARIMA(1,2,1)

The ARIMA (1,2,1) model plotted above shows the model try to fit to the training data. The difference being the order of d in the model. The ‘*d*’ in ARIMA remove the trend from the model to make it stationary. The model able to capture the first upward trend but failed to capture the downward trend. The model is great at modelling the training data with constantly following the blue line. This among the main concern of overfitting the data. As the model able to capture the training data but failed for the testing data due to high error.

### pdmarima

A screenshot of a table

Description automatically generated

Figure .

Figure VII.9 shows the model suggested by pmdarima which is ARIMA(3,2,1). arL1 coefficient are not statistically significant which may pose a threat in formally fitting the model.

A graph with numbers and lines

Description automatically generated

Figure .

Figure VII.10 shows a fitted model for model suggested by pmdarima. It is almost similar to the model fitted by ARIMA(1,2,1). Although seems very similar but the results may differ after computing the performance error.

# Results

Table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **AR(2)** | | **MA(2)** | |
|  | Train | Test | Train | Test |
| 80 | 172.44 | 231.88 | 470.81 | 1417.52 |
| 85 | 168.3 | 437.6 | 488.06 | 1338.61 |
| 90 | 164.62 | 203.19 | 484.36 | 1281.45 |
| 95 | 160.75 | 139.4 | 598.56 | 1237.35 |
| Weighted Error | 166.2507 | 249.3614 | 513.1586 | 1314.463 |

Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ARMA(1,1)** | | **ARIMA(1,2,1)** | |
|  | Train | Test | Train | Test |
| 80 | 168.75 | 315.26 | 171.26 | 435.65 |
| 85 | 164.7 | 489.45 | 167.62 | 139.48 |
| 90 | 161.3 | 261.76 | 164.2 | 80.17 |
| 95 | 157.66 | 167.85 | 160.41 | 76.46 |
| Weighted Error | **162.8406** | 303.7949 | 165.6156 | **174.8194** |

Table

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| 0.8 | 168.62 | 493.89 |
| 0.85 | 165.10 | 432.35 |
| 0.9 | 161.96 | 82.11 |
| 0.95 | 158.23 | 81.57 |
| Weighted Error | 163.23 | 261.15 |

Table 1, 2 and 3 are the results of CPO price modelling. The model performance is calculated based on different levels of train and test splitting size. The training size start from 80% until 95%. The error is then calculated based on this level of error. From the table above, we can see that **ARIMA(1,2,1)** is the best model as it has the lowest error rate of **174.8194**. Although the training data does not come first as ARMA has the best performance, but the error rate are almost similar with each other. This shows that the model is not overfitting the data.

# Conclusion

The main purpose of the model is to find the best model to forecast the Crude Palm oil price. The reliability of the model would become a main concern as data price would move out of trend, which there affect the model viability. There would be ineffective resource utilization across the organization. Other business operations would be affected by poor planning of resources. For example, futures trader would expect the model to move in a particular way, but the actual movement act different from the expected. This deviation would cost the return or even a loss.

In conclusion, ARIMA(1,2,1) demonstrates a high level of accuracy in forecasting the CPO price, exhibiting lower errors compared to the other models. This indicates that ARIMA provides a reliable estimation of the CPO prices, resulting in a more precise forecast. Therefore, based on its performance and lower error rates, ARIMA can be considered as the preferred choice for forecasting the CPO price forecasting.

Therefore, it is vital to prevent overfitting of the model. Although the model performance is the best as compared to the other, overfitting can cause the issue mentioned beforehand. Further testing and development need to be performed as future data made available.

For future research, this study can explore more on the application of Deep Learning such as LSTM (Long Short-Term Memory). Furthermore, exploring advanced techniques like machine learning-based regression models, such as Random Forest or Gradient Boosting, can help uncover nonlinear relationships and improve forecasting accuracy.

##### Acknowledgment

##### We would like to express our sincere gratitude to all those who have contributed to the successful completion of this case study project, especially to my lecturer, Dr. Siti Meriam binti Zahari for her invaluable guidance and support throughout the duration of this project.

# References

(2023). Retrieved from Malaysian Palm Oil Council: https://bepi.mpob.gov.my/admin2/price\_local\_daily\_view\_cpo\_msia.php?more=Y&jenis=1Y&tahun=2021

Al-Khowarizmi, Nasution, I., Lubis, M., & Lubis, A. (2020). The effect of a SECoS in crude palm oil forecasting to improve business intelligence. *Bulletin of Electrical Engineering and Informatics*, 1604-1611.

Kanchymalay, K., Salim, N., & Krishnan, R. (2019). Time Series Based Crude Palm Oil Price Forecasting Model with Weather Elements using LSTM Network. *International Journal of Engineering and Advanced Technology*, 3188-3192.

Khalid, N., Nur Ahmad Hamidi, H., & Thinagar, S. (2018). Crude palm oil price forecasting in Malaysia: An econometric approach (Peramalan Harga Minyak Sawit Mentah di Malaysia: Satu Pendekatan Ekonometrik). *Jurnal Ekonomi Malaysia*, 247-259.

Khamis, A., Hameed, R., & Nor, M. (2018). Comparative Study on Forecasting Crude Palm Oil Price using Time Series Models. *Scientific Research Journal*.

Kondalamahanty, A. (2021, 4 5). *Malaysia's 2021 crude palm oil exports to surge 24.4% on year: MPOC*. Retrieved from spglobal: https://www.spglobal.com/commodityinsights/en/market-insights/latest-news/agriculture/040521-malaysias-2021-crude-palm-oil-exports-to-surge-244-on-year-mpoc#:~:text=Speaking%20at%20a%20web-based%20seminar%20called%20POINTERS%2C%20Ismail,17%20million%20mt%

Suppalakpanya, K., Nikhom, R., & Booranawong, A. (2019). An evaluation of holt-winters methods with different initial trend values for forecasting crude palm oil production and prices in Thailand. *Suranaree Journal of Science and Technology* , 13-22.