Forecasting on the crude palm oil production in Malaysia using SARIMA Model

Forecasting on the crude palm oil production in Malaysia using SARIMA Model

Siti Amnah Mohd Tayib[1], **Siti Rohani Mohd Nor**[1,2], **and Siti Mariam Norrulashikin**[1]

[1]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

[2]**Email:** sitirohani@utm.my

Abstract. Today, accurate prediction on the seasonal trend of the crude palm oil production is critical for the government and agriculturist management to aid in decision-making. The study aims to forecast the Malaysia crude palm oil production by using the Seasonal Autoregressive Integrated Moving Average model. The monthly data of Malaysia crude palm oil production were obtained from Malaysian Palm Oil Board, from January 2014 until September 2019. The Seasonal Autoregressive Integrated Moving Average model was applied to the data by using the Box-Jenkins approach. Based on the adequacy checking and accuracy testing, SARIMA$(1,0,0)(0,1,1)_{12}$ is the best fitted model for the Malaysia crude palm oil production. As a result of the findings, the SARIMA$(1,0,0)(0,1,1)_{12}$ model appears to be the best choice for decision makers to make reliable and accurate long-term forecasts on Malaysia crude palm oil production.

1. Introduction

Palm oil production is crucial for the Malaysian economy since Malaysia is the second-largest producer in the world after Indonesia [1][2]. However, there are three problems related to the crude palm oil (CPO) production. First, the demand on the crude palm oil (CPO) production in Malaysia has decreased over time due to other countries' global boycott [3]. Second, according to Abdullah [4], Malaysia has limited land availability for further palm oil expansion. Third, country's oil palm industry is experiencing severe labour shortages due to the lack of interest among Malaysians to work in the palm oil industry. Hence, it is important to study the trend of the CPO production in the future so that the government and private organization can make strategy to solve these problems. More than that, accurate prediction of the CPO will help the agriculturist management to reduce unnecessary spending, schedule production and staffing, and avoid missed potential opportunities.

To make accurate prediction on the CPO production, it is important to analyze the historical trends of the CPO data in order to choose the suitable time series model that could be applied to the data. Previous studies have found that the trend of the historical data of the CPO production is seasonal. Abdullah [4] stated that the CPO in Malaysia is not random and is composed into four main components which are trend, cyclical, seasonal and irregular components. Ahmad [5] found that the crude palm oil production in Malaysia showed seasonal pattern from June 2001 until May 2011. Mah and Nanyan [6] applied Autoregressive Integrated Moving Average (ARIMA) model to the CPO production, however the seasonality trend was removed by differencing in order to apply the ARIMA

model to the data. This demonstrates that the CPO data in Malaysia follows a seasonal trend. Hence, the objective of this study is to model the crude palm oil production using Seasonal Autoregressive Integrated Moving Average (SARIMA) model for 2014 until 2019. The analysis results are expected to support management in planning the CPO Production in the future.

The study is organized as follows: Section 2 describes the CPO production and SARIMA model's structure. Section 3 discusses the results of the analysis. Section 4 gives conclusion remark on the study.

2. Data and Methodology

In this study, the monthly dataset of Malaysian crude palm oil production (in tonnes) from January 2014 until September 2019 were obtained from the Malaysian Palm Oil Board (MPOB) website. The data are composed of 5 years of monthly data with a total of 69 monthly data. This study aims to apply SARIMA model to the obtained Malaysian crude palm oil production data, by using the Box-Jenkins approach.

2.1. Box-Jenkins Methodology

The process of defining, fitting, and testing the seasonal integrated autoregressive moving average time series model to the data is known as Box-Jenkins analysis. The approach is suitable for data of at least 50 observations in a time series. The Box-Jenkins approach for the SARIMA model follows the same Box-Jenkins approach of the ARIMA model [5]. The steps of the Box-Jenkins approach of the SARIMA model are as follows:

2.1.1. The Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

The SARIMA model is a time series model that is commonly used. It's a variation on the well-known ARIMA model. However, ARIMA is not adequate for data with seasonality. Thus, in order to facilitate non-stationary seasonal data, the ARIMA model is extended to the SARIMA model. The multiplicative SARIMA$(p,d,q)(P,D,Q)_s$ model has the following equation:

$$\phi(B)\,\Phi_p(B^s)(1-B)^d(1-B^s)^D\,y_t = \delta + \theta(B)\Theta_Q(B^s)\,\varepsilon_t \quad (1)$$

where ε_t is Gaussian white noise, $\phi(B)$ is ordinary autoregressive and $\theta(B)$ stand for moving average components. For $\Theta_Q(B^s)$ and $\Phi_p(B^s)$ are seasonal autoregressive and moving average components, respectively, and $(1-B)^d(1-B^s)^D$ are the ordinary and seasonal difference components of order d and D.

Step 1: Model Identification

The Box-Jenkins methodology refers to Wei's [7] model identification, in which necessary transformations, including differential transformations, are described. According to Cooray [8], the first step of the Box-Jenkins process is to plot the data. The data plot is used to identify the patterns of the data, such that if the series includes trends, seasonality, unchangeable variation and non-standard phenomena. Next, the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the original series will also be plotted to examine for the non-stationary pattern and necessary differential series. The general formula to transform the time series data from non-stationary to stationary is given by $w_t = (1-B)^d(1-B)^D\,y_t$. If the seasonal time series data is non-stationary, then a first differencing need to be made. If the first differenced data appears to be non-stationary, then the second differencing need to be applied. The seasonal difference formula is:

$$w_t = y_t - y_{t-s} = (1-B^s)y_t.$$

More than that, ACF and PACF plots are also used to determine the order of p, and q where p is the

highest order in the self-regression polynomial, and q is the highest order in the moving average polynomial [9]. In addition, ACF and PACF plots could also help the researchers to classify the most appropriate model for the given data set [10].

Step 2: Estimating the Model Parameters
The significance of parameters is tested using standard *t*-test,

$$t_{stat} = \frac{\text{point estimate of parameter}}{\text{standard error of estimate}} \tag{2}$$

where the parameters model are significant if $|t_{test}| > 2$ for significance level $\alpha = 0.05$. We normally select significance level, α equal to 0.05, which corresponds to 95% of the confidence interval. If the value *p* is smaller than α and *t* is greater than 2, then the model parameter H_0 will be rejected and therefore the model parameter is significant to be used.

Step 3: Model Checking
A Ljung-Box test is used to investigate whether the residual of the model independently distributed. The diagnostic checking on the model is based on the tests and the residual plots. This experiment tests the residual autocorrelation sizes as a band. The equation of Ljung-Box test is:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{rk^2(e)}{n-k}\ . \tag{3}$$

The chi-square random variable with *m-r* of freedom is distributed roughly, where *r* is the estimated total number of parameters of the ARIMA model. Based on this formula, *rk(e)* is the residual autocorrelation at lag *k, n* is the number of residuals, *k* is the time lag, and *m* is the number of time lags to be tested. When the *p*-value of the Q estimate is small which mean less than 0.05, then the model is considered inadequate. A new or updated model should be considered by the researcher until a suitable model is found based on the diagnostic checking. The residuals values should be small and in general within $\frac{\pm 2}{\sqrt{n}}$ of zero.

Step 4: Forecasting with the Model
The last step of the Box-Jenkins method is to make forecast measurements, based on the best selected model chosen in Step 3. The forecasts values of the SARIMA model should be within the 95% of the confidence interval. In order to validate the accuracy of the SARIMA model, out-sample forecast measurements are calculated based on the accuracy measurement errors such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and mean absolute percentage error (MAPE).

2.2. Model Evaluation
The measurement errors that are used to measure the accuracy of the SARIMA's historical fitting and forecasts in this research are mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and mean absolute percentage error (MAPE).

$$\text{MAE} = \frac{1}{T}\sum_{t=0}^{T}|y_t - \hat{y}_t| \tag{4}$$

$$\text{MSE} = \frac{\sum_{t=0}^{T}(y_t - \hat{y}_t)^2}{T} \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{t=0}^{T}(y_t - \hat{y}_t)^2}{T}} \tag{6}$$

$$\text{MAPE} = \frac{100}{T}\sum_{t=0}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \quad (7)$$

where t is time period, T is total number of observations, y_t is the actual value, and $\hat{y}_t$ is the forecasted value at time t.

The model with the lowest measurement error in terms of the in-sample and out-sample data will be chosen as the best selected model for the crude palm oil production. 86% of the in-sample data for the model's evaluation is from January 2014 until November 2018, whereas the 14% of the out-sample data for the evaluation on the forecasting of SARIMA model is from December 2018 until September 2019.

3. Result and Discussion

The first step in Box-Jenkins methodology is model identification. In this step, the data are plotted to identify the pattern of the data. The observed data were covering the period from January 2014 to November 2018 with a total of 59 observations, as shown in figure 1. According to figure 1, the time series plot shows no stability over time since the CPO production increases from February to August and dramatically decreases by the end of the year from December to January. In addition, the data show an obvious seasonal trend and the series are fluctuated around the mean. Thus, this shows that the model that will be applied to this data should be able to capture the seasonality behaviour. Thus, SARIMA model is chosen for this study. In modelling the SARIMA model, the first step is to determine whether the data is stationary or non-stationary. A stationary process has the property that the mean, variance and auto-covariance structure do not change over time. Augmented Dickey Fuller (ADF) test is employed to determine whether the time series is stationary not. The p-value of the ADF test for the data is 0.2893, which eventually indicated that the null hypothesis of non-stationary is not rejected at 5% level of significant.

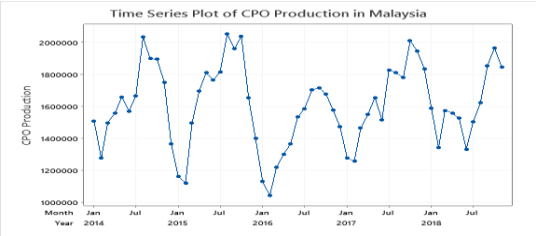

Figure 1. Time series plot of CPO production in Malaysia

The ACF plot in figure 2 agrees with the ADF test result of the CPO data. Based on figure 2, the ACF shows a pattern of a non-stationary seasonal series since the ACF at the seasonal lags 12 are large and failed to die out quickly. Other than that, PACF plot on figure 3 shows a large spike at lag 1 and the value is close to 1. These indicate that the time series is non-stationary. Therefore, to obtain the stationary series, the data should be differenced.

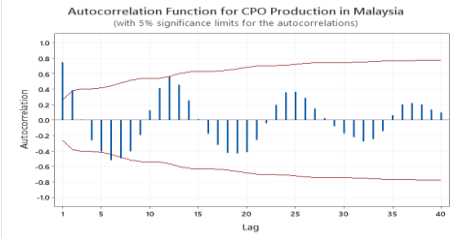

Figure 2. ACF plot of CPO Production in Malaysia

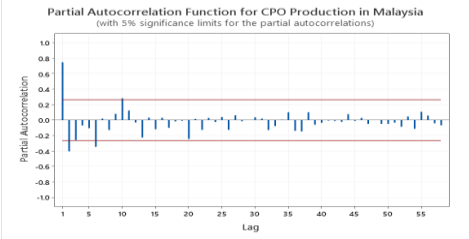

Figure 3. PACF plot of CPO Production in Malaysia

Figure 2 also shows that there is an annual seasonality of data as there is a large autocorrelation at lag 12. The ACF trailed off to zero rather quickly which means no trend exists and significant in the large value at the seasonal lag 12. Therefore, the next step is to difference the observation denoted by the order d=0 since there is no trend in the data and D=1 since data exhibit seasonal pattern.

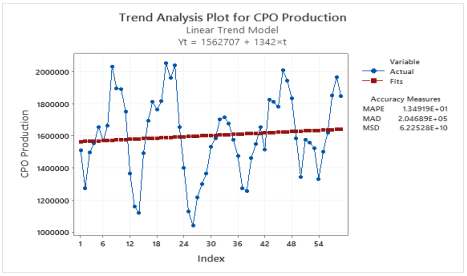

Figure 4. Trend Analysis plot of CPO Production in Malaysia

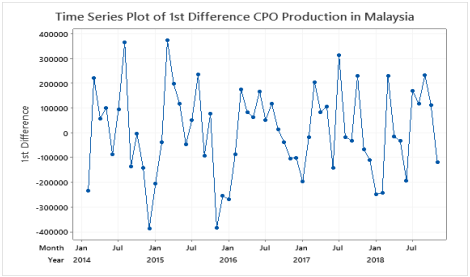

Figure 5. Time Series Plot of first difference CPO Production in Malaysia

After differencing, the differenced data were plotted again and was denoted as first difference of CPO Production in Malaysia. The graph of the differenced data is illustrated in figure 5. Figure 5 shows the data have been transformed from non-stationary to stationary, since the plotted data fluctuated below and above the horizontal lines. Next, according to Box-Jenkins method, the stationary differenced data were applied to the SARIMA model. To find the parameters of the SARIMA model, the ACF and PACF of the differenced data were plotted and are illustrated in figure 6 and figure 7.

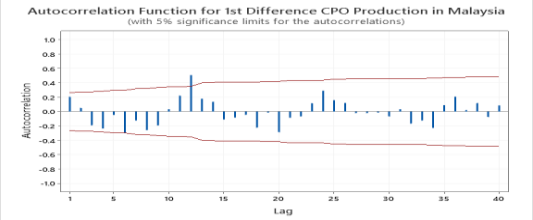

Figure 6. ACF plot of first difference CPO Production in Malaysia

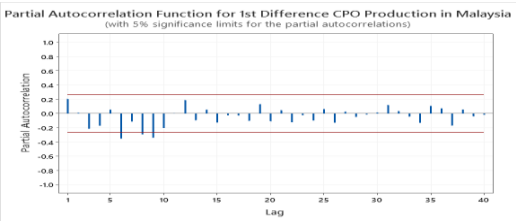

Figure 7. PACF plot of first difference CPO Production in Malaysia

Based on ACF, there are cuts off behaviour at lag 2 which present the order of Moving Average (MA) and PACF shows cut off at lag 1 which indicates the order of Autoregressive (AR). The Seasonal Moving Average (SMA) is equal to 1 since the lag spike at lag 12 while the Seasonal Autoregressive (SAR) is equal to 0 since there is no lag spike at lag multiple of 12. All the SARIMA models that can be consider from the plots of ACF and PACF in figure 6 and figure 7 are shown in table 1. Once the parameters have been estimated, the adequacy of the model will be checked to determine the adequate SARIMA model for the series. From the Ljung-Box statistic it is found that only five SARIMA models in table 1 are adequate since the *p*-value is larger than the significance level, alpha, 0.05. The performance of the five SARIMA models are compared by using Akaike Information Criteria (AIC) and numerical measurement errors, and are tabulated in table 1 and table 2. The model with the lowest Akaike information criteria (AIC) and measurement errors will be chosen to forecast.

Based on table 1, it appears that the estimated SARIMA$(1,0,0)(0,1,1)_{12}$ model outperformed the others since the model has the lowest AIC value which is 1242.81. More than that, table 2 shows that the model SARIMA$(1,0,0)(0,1,1)_{12}$ has the smallest values for the four measurement error tests which are RMSE, MSE, MAE and MAPE. Figure 8 also illustrates that the time series plot of original data and predicted data of SARIMA$(1,0,0)(0,1,1)_{12}$ are almost identical. **The** parameter **estimates** for the best fitted SARIMA$(1,0,0)(0,1,1)_{12}$ model is given in table 2

Table 1. White Noise Tests and AIC Values

SARIMA(p,d,q) x (P,D,Q)s	Adequate	Log Likelihood	AIC
$(1,0,0)(0,1,1)_{12}$	Yes, p-value > α=0.05	-618.41	1242.81
$(1,0,0)(1,1,0)_{12}$	Yes, p-value > α=0.05	-618.75	1243.50
$(1,0,0)(2,1,0)_{12}$	Yes, p-value > α=0.05	-618.68	1245.37
$(1,0,0)(0,1,2)_{12}$	Yes, p-value > α=0.05	-618.33	1244.67
$(1,0,2)(1,1,0)_{12}$	Yes, p-value > α=0.05	-618.11	1246.22

Table 2. Test for in sample CPO Production in Malaysia

SARIMA(p,d,q) x (P,D,Q)s	RMSE	MSE	MAE	MAPE
$(1,0,0)(0,1,1)_{12}$	104071	1.083×10^{10}	87379.4	0.0552 %
$(1,0,0)(1,1,0)_{12}$	116281	1.352×10^{10}	94213.4	1.2435 %
$(1,0,0)(2,1,0)_{12}$	111054	1.233×10^{10}	77947.3	1.8367 %
$(1,0,0)(0,1,2)_{12}$	109150	1.191×10^{10}	86658.7	1.8323 %
$(1,0,2)(1,1,0)_{12}$	114373	1.308×10^{10}	89233.0	2.8943 %

Table 3. Final estimates of parameter for SARIMA(1,0,0)(0,1,1)$_{12}$ model

Model	**SARIMA(1,0,0)(0,1,1)$_{12}$**		
Parameter	**Coefficient**	**Standard Error of the Coefficient**	**P-Value**
AR1	0.7659	0.0927	0.000
SMA12	0.772	0.163	0.000

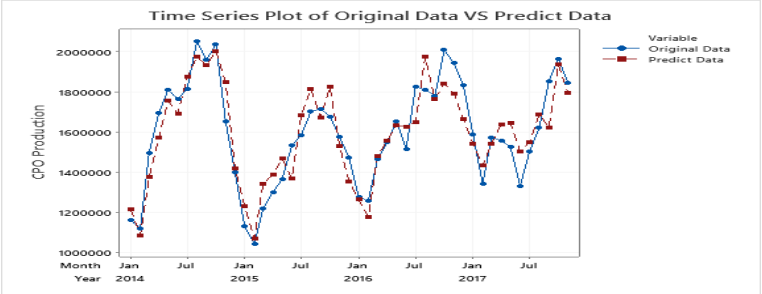

Figure 8. Time series plot of original and predict CPO Production in Malaysia

3.1. Diagnostic check

The residual ACF and PACF plots of CPO production for SARIMA(1,0,0)(0,1,1)$_{12}$ model fit are shown in figure 9 and figure 10. From figure 9 and figure 10, all ACF plots of residuals are within the 95 percent of confidence interval which is within the standard error limits that indicate residuals exhibit white noise. This gives strong evidence that the SARIMA(1,0,0)(0,1,1)$_{12}$ is an adequate model. To confirm with the conclusion drawn from the plot, Ljung-Box test result for residuals from fitted SARIMA model is given in fable 4. The *p*-values of the Ljung-Box statistic test are more than 0.05, which indicates that the null hypothesis of data points are independent distributed is not rejected at 5% level of significant.

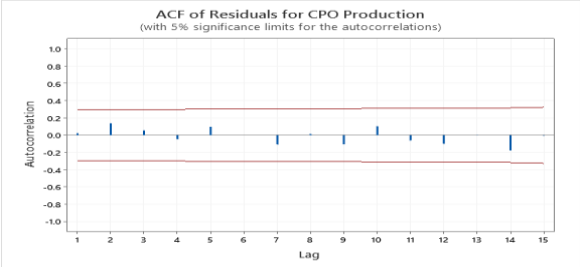

Figure 9. ACF of residuals from fitted SARIMA model.

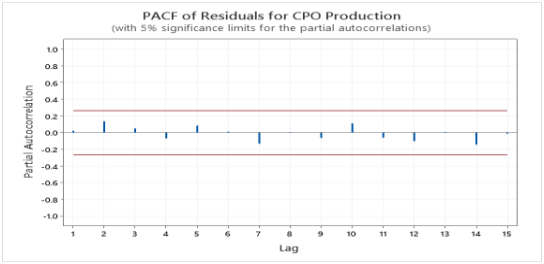

Figure 10. PACF of residuals from fitted SARIMA model.

Table 4. Ljung-Box statistic result for SARIMA(1,0,0)(0,1,1)$_{12}$ model

Lags	**Chi-Square**	**P-value**
12	12.26	0.268
24	20.73	0.538
36	35.30	0.407

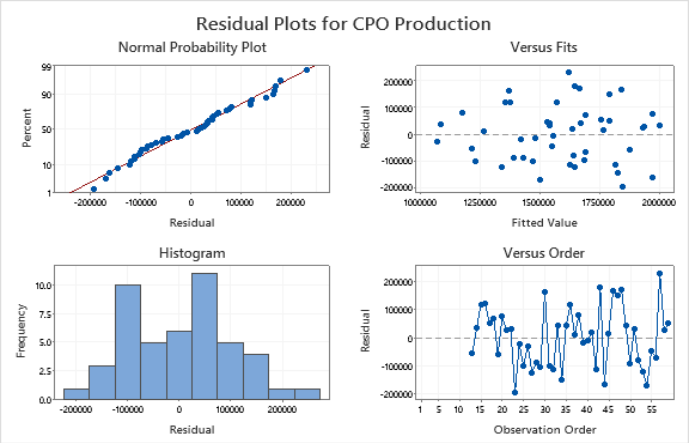

Figure 11. Residual Plots for CPO Production in Malaysia from fitted SARIMA model

Based on the normal probability plot and histogram chart in figure 11, it shown that the fitted SARIMA model which is SARIMA$(1,0,0)(0,1,1)_{12}$ is an adequate model as the errors are follow normal distribution. All the residual plots of SARIMA$(1,0,0)(0,1,1)_{12}$ model below satisfied the mean of errors is equal to zero and the errors follow normal distribution.

3.2. Forecast Measurements

Table 5 indicates the value of accuracy measurements for out-sample data. According to table 5, SARIMA$(1,0,0)(0,1,1)_{12}$ has the lowest value of measurement errors in terms of RMSE, MSE, MAE and MAPE. This concludes that SARIMA$(1,0,0)(0,1,1)_{12}$ is chosen as the best selected model to forecast the crude palm oil production in Malaysia since the residual of the model is normal, and the model has the smallest measurement errors in terms of in-sample and out-sample data.

Table 5. Accuracy Measurement for Out-sample

SARIMA(p,d,q) x (P,D,Q)s	RMSE	MSE	MAE	MAPE
$(1,0,0)(0,1,1)_{12}$	166320	2.766 x 1010	122506	0.073 %
$(1,0,0)(1,1,0)_{12}$	247253	6.113×10^{10}	207776	0.127 %
$(1,0,0)(2,1,0)_{12}$	398594	15.888×10^{10}	355547	8.631 %
$(1,0,0)(0,1,2)_{12}$	282892	8.003×10^{10}	237495	1.226 %
$(1,0,2)(1,1,0)_{12}$	266647	7.110×10^{10}	224000	0.572 %

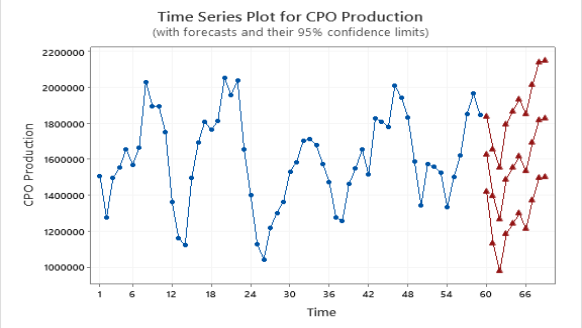

Figure 12. Forecasts for CPO Production in Malaysia

Once the fitted model has been chosen, it can be used to generate forecasts for future period. Forecasts for the next 10 months for CPO production by using SARIMA$(1,0,0)(0,1,1)_{12}$ are plotted in figure 12. The pattern of the forecast plot is look like historical pattern. In addition, the forecasts are within the 95% confidence interval. Thus, the forecast seems very reasonable. All the forecast values are shown in table 6.

Table 6. Forecasts from December 2018 to September 2019

		95% Limits		
Year	**Forecast**	**Lower**	**Upper**	**Actual**
December 2018	1627739	1419233	1836244	1808038
January 2019	1394916	1132281	1657550	1737461
February 2019	1265201	975483	1554920	1544518
March 2019	1488637	1184150	1793125	1672058
April 2019	1554832	1242005	1867659	1649368
May 2019	1615873	1298256	1933489	1671467
Jun 2019	1533173	1212780	1853567	1510957
July 2019	1692394	1370383	2014405	1740759
August 2019	1818447	1495491	2141403	1821548
September 2019	1826764	1503255	2150274	1842433

4. Conclusion

There are two objectives of this study which is to model the crude palm oil using SARIMA models and to forecast the upcoming production of crude palm oil by using SARIMA model. These two objectives can be achieved after performing the analysis based on the results in Section 3. Based on the in-sample error measurements, SARIMA model is adequate to be applied to the crude palm oil data since it could allow for seasonal and multiplicative seasonality data. Then the fitted SARIMA model which is SARIMA$(1,0,0)(0,1,1)_{12}$ will be used to forecast the crude palm oil production data. The forecast result of the SARIMA$(1,0,0)(0,1,1)_{12}$ shows almost similar pattern with the past year data and within the 95% upper and lower bound of the limits. In conclusion, the crude palm oil production can be forecasted accurately by using SARIMA$(1,0,0)(0,1,1)_{12}$ model.

Acknowledgement
The authors would like to acknowledge Persatuan Sains Matematik Malaysia (PERSAMA) for sponsoring the conference fees for Simposium Kebangsaan Sains Matematik ke-28 (SKSM28) and financial resources from Fundamental Research Grant Scheme (FRGS) grant with vote number 5F370 and UTM Encouragement Research (UTMER) grant vot 17J78.

References
[1] Pakiam R 2013 *Malaysia Keeps Palm Oil Export Tax Unchanged to Spur Shipments* Bloomberg News
[2] McClanahan P 2013 *Can Indonesia increase palm oil output without destroying its forest?* The Guardian
[3] Norhidayu A, Nur-Syazwani M, Radzil R, Amin I, & Balu, N 2017 The production of crude palm oil in Malaysia *Int. J. Econ. Manag* 11(3) 591-606.
[4] Abdullah R 2012 An analysis of crude palm oil production in Malaysia *Oil Palm Industry Economic Journal* 12(2) 36-43.
[5] Ahmad S 2011 Forecasting on the crude palm oil and kernel palm production: Seasonal ARIMA approach *IEEE Colloquium on Humanities, Science and Engineering* 939-944
[6] Mah P J W & Nanyan N N 2020 A Comparative Study Between Univariate and Bivariate Time Series Models For Crude Palm Oil Industry In Peninsular Malaysia. *Malaysian Journal of Computing* 5(1) 374-389.
[7] Wei W W 2006 *Time Series Analysis Univariate and Multivariate Methods* vol 2 (Boston: Pearson Addison Wesley)
[8] Cooray T M J A 2008 *Applied time series: analysis and forecasting* (England: Alpha Sci. Int. UK)
[9] Madsen H 2007 *Time Series Analysis* (New York: Chapman & Hall/CRC)
[10] Suppalakpanya K, Booranawong A, Booranawong T and Nikhom R 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* 26 13-22.