Fourier Regression and ARIMAX Model for Forecasting Monthong Durian Price Index

Kanittha Yimnak[1]* and Rungsarit Intaramo[2]

[1]Department of Applied Statistics, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi, Thailand
[2] Department of Mathematics, Faculty of Science, Thaksin University, Thailand

Received: 29 May 2020, Revised: 17 February 2021, Accepted: 6 April 2021

Abstract

The purposes of this study were to create models for forecasting the Monthong durian price index at the farm and to compare the forecasting efficiency of ARIMAX model (Autoregressive Integrated Moving Average with Exogenous Variable model) and Fourier Regression model. The data set was collected monthly between January 2014 and December 2020 (a total of 84 months). Production index and China consumer confidence were used as explanatory variables. The efficiency of both methods was compared by Mean Absolute Percentage Error (MAPE). From the result, we found that the ARIMAX (1,1,1) and Fourier regression models were both suitable for forecasting the Monthong durian price index. However, the MAPE value obtained from the ARIMAX model was 3.365 times higher than that obtained from Fourier regression model, suggesting that the Fourier regression model is more efficient.

Keywords: Monthong durian price index; China consumer confidence; Fourier regression; ARIMAX model
DOI 10.14456/cast.2021.57

1. Introduction

Durian is an important economic fruit of Thailand, especially Monthong durian. By around a decade ago, durian exports had been increased quickly. China has the greatest demand for durian [1] because it is a famous fruit in China. This increasing demand gave farmers the motivation to expand durian cultivation. The Office of Agricultural Economics in Thailand reported that durian production increased to 1,017,097 tons in year 2019 [2]. Monthong durian, a famous species that is grown in the eastern region of Thailand, has been increasingly cultivated because it has a good taste and attracts high prices in the market. Modelling for forecasting the Monthong durian price provides information that can help agricultural authorities and related parties to assist farmers. There are many factors (or exogenous variables) expected to affect the Monthong durian price index at the farm and these types of factors were used to construct the models such as the production index (PI) and China consumer confidence (CCC). PI and CCC are collected from the Office of Agricultural Economics, Thailand [3] and the National Bureau of Statistics of China [4], respectively. According

*Correspondence author: Tel.: +66 86-4182542
E-mail: kanittha_y@rmutt.ac.th

to the market mechanism, excessive production affects the product prices, and the production index was chosen for constructing the model. Moreover, China consumer confidence is used as an explanatory variable because durian is a popular fruit in China. There are many methods for forecasting time series. The most popular method is Box Jenkins method using ARIMA model (Autoregressive Integrated Moving Average model). This modelling method is taken into the consideration of the correlation or the behavior of the data in the past for forecasting future data [5]. In addition, sometimes time series data may be affected by other factors (exogenous variable) that can influence the accuracy of the forecast value. Therefore, models for forecasting time series data have been developed based on other factors called ARIMAX model (Autoregressive Integrated Moving Average with Exogenous Variable model) [5, 6].

Regression model, one of the choices, is purposed to compare the efficacy of modelling. It is used for studying the correlation between independent variables and a dependent variable. There are three approaches for regression modelling such as parametric, semiparametric and nonparametric regression. The parametric regression model has rigorous assumption. A data set that does not correspond to the assumptions causes lower model performance [7]. On the other hand, there are no assumptions in the nonparametric regression model. For an unknown form of regression function (or a function curve), there are many functions proposed such as spline kernel, wavelet, local polynomial and Fourier. In this study, we used Fourier regression approach, proposed by Bilodea [8] for nonparametric and semiparametric cases. It is used for modelling because it can recognize datasets that have trigonometric pervading in the sine and cosine cases [7]. In addition, the dataset of the Monthong durian price index corresponds to Fourier approach. Fourier birespon series was developed to estimate the nonparametric regression model by Semiati [9], and then Fourier series was applied for semiparametric regression [10].

2. Methodology

2.1 ARIMAX model

The ARIMAX model was presented by Tiao and Box [11]. This model was developed from the ARIMA model. The ARIMAX (p,d,q,r) model [6] can be written as:

$$(1-\mathrm{B})^d\left(1-\phi_1\mathrm{B}-\phi_2\mathrm{B}^2-\ldots-\phi_p\mathrm{B}^p\right)\mathrm{Y}_t=\theta_0+\left(1-\theta_1\mathrm{B}-\theta_2\mathrm{B}^2-\ldots-\theta_q\mathrm{B}^q\right)\varepsilon_t+\sum_{i=1}^{r}\alpha_i X_{it}, \quad (1)$$

Where B is the backward shift operator.

$\phi_1,\ldots,\phi_p$ are the parameters of the autoregressive part of model,

$\theta_0,\theta_1,\ldots,\theta_q$ are the parameters of the moving average part,

d is a number of times of order differencing,

$\nabla^d=(1-B)^d$,

$\left(1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q\right)$ is a moving average polynomial with order q,

$\left(1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p\right)$ is an autoregressive polynomial with order p,

ε_t are error terms,

X_t is the value of the independent variable X at time t.

The steps for constructing the ARIMAX model [12] are presented as follows:

1) Test the stationary of the dependent series by correlogram of r_k and r_{kk} that are shown in equations (2) and (3) [6], respectively. The nonstationary respond series can be transformed by $1^{st}, 2^{nd}, \ldots, n^{th}$ differencing steps.

$$r_k = \frac{\sum_{t=1}^{n-k}(Y_t - \bar{Y})\,(Y_{t+k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \quad , \tag{2}$$

$$r_{kk} = \begin{cases} r_1 \\ \dfrac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k} r_{k-1} r_j} \end{cases} \quad , \tag{3}$$

2) Determine the order of the parameters p and q by considering the sample autocorrelation function (SACF) and the sample partial autocorrelation function (SPACF).

3) Estimate the coefficients using least square method (or maximum likelihood method) and test the significance along with the residual series.

4) Operate the same process to the input series as was done for the respond series.

5) Assign the structure of the ARIMAX model by approximating the cross-correlation coefficient between the response series and the input series to set the structure of the ARIMAX model.

6) Verify that the model coincides with the attributes of the time series data by creating diagnostic analysis. The suitability tests of the model are considered as follows:

6.1) The mean absolute percentage error (MAPE) and the square root of the mean square error (RMSE) with the smallest values as in the equations (4a) and (4b)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{y_t} \times 100\,, \tag{4a}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad , \tag{4b}$$

where y_t is the actual value at time t, $\hat{y}_t$ is the forecast value at the time t and n is the number of observations.

6.2) R^2 is as follows:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\frac{1}{n}(y_t - \bar{y})^2} \tag{5}$$

6.3) Box and Ljung or Q-statistic for testing the suitability of the model is as follows:

$$Q_m = n(n+2)\sum_{j=1}^{m}\left\{\frac{r_k^2(e_t)}{(n-k)}\right\}, \tag{6}$$

where e_t is the forecasting error at time t,
n is the number of observations,
m is the number of lags being tested.

The calculated Q value is the Chi-square distribution and has the degree of freedom equal to $m - n$. Under the null hypothesis, it can be said that the model does not have autocorrelation (or the term of estimated residual has a white noise appearance).

2.2 Fourier regression

Multiple regression is an analysis of the correlation between independent variables ($x_{ji}; i = 1,2, \dots, n, j = 1,2, \dots, p$) and dependent variable ($y_i; i = 1,2, \dots, n$) as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_j x_{ji} + \varepsilon_i \tag{7}$$

Where ε is the residual of the model. The residual $\varepsilon \sim N(0, \sigma^2)$ is not reciprocally correlated. Parameter estimation can be solved by ordinary least square (OLS) method or maximum likelihood estimation (MLE) method [13]. The equation (7) can be written in matrix form as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8}$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} ; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)} ; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} ; \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

For the nonparametric regression method, the curve shape which is presumed to be put within a certain functional shape, is used for constructing the correlation model for dependent and independent variables [14]. The Fourier regression model is as follows:

$$y_i = \eta(x_i) + \varepsilon_i \tag{9}$$

where $\eta(x_i)$ that is the unknown shape is the regression curve. Moreover, it is presumed to be sleek on function space and x_i is explanatory variable. A trigonometric polynomial function, having a degree of flexibility to attune to the dataset, is a Fourier series as follows:

$$f(x) = \frac{1}{2}\alpha_0 + \gamma x + \sum_{k=1}^{k} \alpha_k \cos kx. \tag{10}$$

The equation (7) can be transformed as:

$$\mathbf{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon} \; ,$$

where

$$f(x) = \left[f(x_{11}) \; f(x_{12}) \cdots f(x_{1n}) \; f(x_{21}) \; f(x_{22}) \cdots f(x_{2n}) \cdots f(x_{p1}) \; f(x_{p2}) \cdots f(x_{pn}) \; \right]^T ,$$

and $f(x_{ji})$, curvilinear function can be solved by equation (11).

$$f(x) = \frac{1}{2}\alpha_0 + \gamma_j x_{ji} + \left(\alpha_{j1} \cos x_{ji} + \alpha_{j2} \cos 2x_{ji} + \dots + \alpha_{jk} \cos kx_{ji} \right) \tag{11}$$

Let $\mathbf{f(x)} = \mathbf{A}\boldsymbol{\theta}$,

$\boldsymbol{A} =$

$$\begin{bmatrix} 1 & x_{11} & cosx_{11} & cos2x_{11} & \cdots & cosKx_{11} & x_{21} & cosx_{21} & cos2x_{21} & \cdots & cosKx_{21} & \cdots & x_{p1} & cosx_{p1} & cos2x_{p1} & \cdots & cosKx_{p1} \\ 1 & x_{12} & cosx_{12} & cos2x_{12} & \cdots & cosKx_{12} & x_{22} & cosx_{22} & cos2x_{22} & \cdots & cosKx_{22} & \cdots & x_{p2} & cosx_{p2} & cos2x_{p2} & \cdots & cosKx_{p2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & cosx_{1n} & cos2x_{1n} & \cdots & cosKx_{1n} & x_{2n} & cosx_{2n} & cos2x_{2n} & \cdots & cosKx_{2n} & \cdots & x_{pn} & cosx_{pn} & cos2x_{pn} & \cdots & cosKx_{pn} \end{bmatrix}_{n\times(p(K+1)+1)},$$

$$\boldsymbol{\theta} = \begin{bmatrix} \emptyset & \gamma_1 & \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1K} & \gamma_2 & \alpha_{21} & \alpha_{22} & \cdots \alpha_{2K} \cdots & \alpha_{1K} & \gamma_p & \alpha_{p1} & \alpha_{p2} & \cdots \alpha_{pK} \end{bmatrix}^T_{1\times(p(K+1)+1)}"$$

where $\emptyset = \frac{n}{2}\alpha_0$ and α_0 are constant. The OLS method is used to estimate nonparametric regression. $\psi(\theta) = \varepsilon^T\varepsilon$ and $\hat{\theta} = (A^TA)^{-1}A^Ty$ are the estimations of nonparametric regression by the Fourier series [7]. Appropriate K values tend to be large. The most suitable K value indicates a better fit model.

3. Results and Discussion

The models of the ARIMAX and Fourier regression for forecasting Monthong durian price index are discussed as follows.

3.1 ARIMAX model for forecasting the Monthong durian price index

Figure 1 represents the correlogram of the Monthong durian price index as being white noise or the residuals is the normal distribution. The probable orders of autoregressive and moving average (p and q), which considered the values of ACF and PACF that are out of the 95% confidence range, are shown in Table 1. The suitable model for forecasting the Monthong durian price index was ARIMAX (1,1,1) due to having low errors and the highest stationary R^2. Considering the Q-statistics, the probability is greater than the significance level of 0.05, indicating that the error of the model has a normal distribution which means that it has a mean value of zero and constant variance.

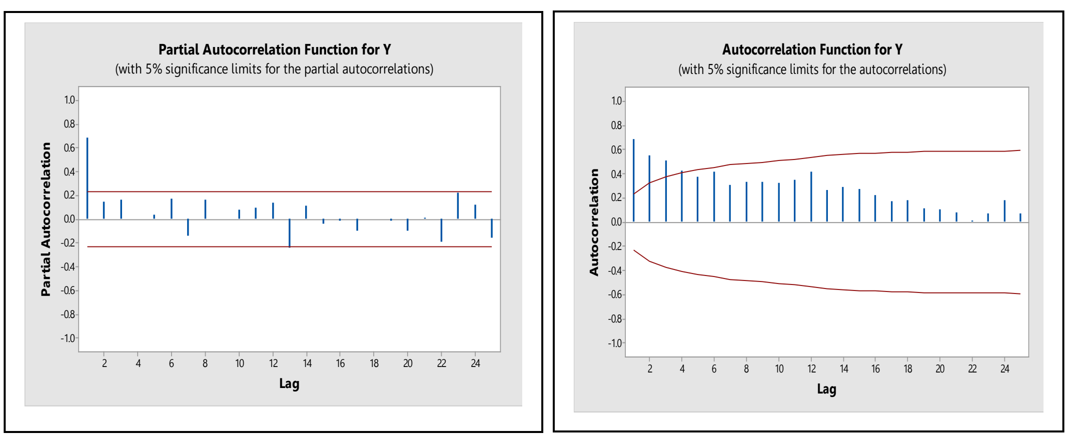

Figure 1. The correlogram of the Monthong durian price index

Table 1. ARIMAX parameter estimation

Model	Variable	Coefficient	t-statistic	p-value	R^2	RMSE	MAPE	Q-statistic	p-value
ARIMAX (1,0,1)	AR(1)	.823	9.712	.000	0.705	77.884	14.879	23.055	0.112
	MA(1)	.125	.839	.404					
	x_1	-.298	-2.951	.004					
	x_2	3.330	9.313	.000					
ARIMAX (2,0,1)	AR(1)	1.552	7.372	.000	0.667	76.377	12.758	26.520	0.033
	AR(2)	-.557	-2.860	.005					
	MA(1)	.896	5.480	.000					
	x_1	-.275	-2.503	.014					
	x_2	3.103	2.338	.022					
ARIMAX (1,0,2)	AR(1)	.782	6.804	.000	0.650	78.340	14.929	22.406	0.098
	MA(1)	.074	.445	.658					
	MA(2)	-.079	-.527	.600					
	x_1	-.325	-2.996	.004					
	x_2	3.356	9.851	.000					
ARIMAX (1,1,0)	AR(1)	-0.295	-2.759	0.007	0.657	76.650	15.881	26.713	0.062
	x_1	-0.270	-4.276	0.000					
	x_2	0.318	3.665	0.000					
ARIIMAX (0,1,1)	MA(1)	.620	6.413	.000	0.687	73.217	15.881	20.360	.256
	x_1	-.200	-4.256	.000					
	x_2	.247	4.361	.000					
ARIMAX (1,1,1)	**AR(1)**	**.342**	**2.050**	**.044**	**0.705**	**71.491**	**15.148**	**19.710**	**0.234**
	MA(1)	**.908**	**6.685**	**.000**					
	$\boldsymbol{x_1}$	**-.121**	**-1.979**	**.051**					
	$\boldsymbol{x_2}$	**.159**	**2.279**	**.025**					

3.2 Fourier regression for forecasting the Monthong durian price index

Table 2 shows the MAPE and R^2 values when *K*=5, 10, 15, 20, 25, 30, 35, and 40, respectively. The optimum K value is *K*=40 as it has the lowest MAPE and the highest R^2. Figure 2 shows the line of actual data and the line of the forecast of the Monthong durian price index by the Fourier regression method when *K*=10, 30, and 40, respectively. The forecast line with *K*=40 and the actual data line coincide closely.

Table 2. The values of *K*, MAPE and R^2 when $K = 5,10,15,20,25,30,35$ and 40

K	MAPE	R^2	*K*	MAPE	R^2
5	21.3369	0.5484	25	16.9538	0.7718
10	20.6831	0.6054	30	14.0312	0.8472
15	18.9405	0.6649	35	10.0862	0.9045
20	17.6595	0.7108	**40**	**4.5010**	**0.9776**

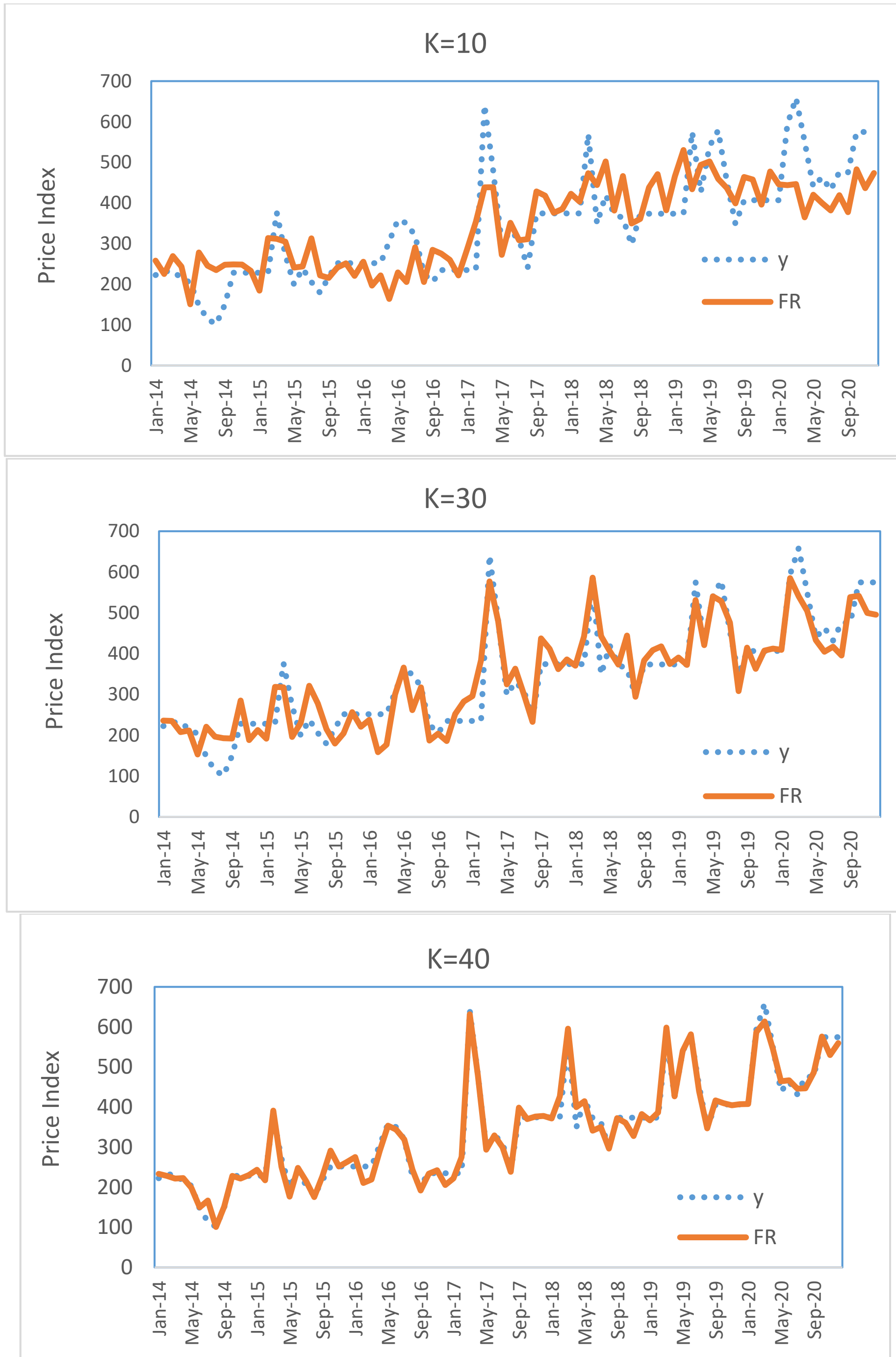

Figure 2. Graphs showing the comparison between the actual values (y) and the forecasted values using the Fourier regression (FR) when *K*=10, 30, and 40

Table 3 reveals the comparison of the MAPE between ARIMAX model and Fourier regression model and the line graphs of both models were coincided with the actual values (as can be seen in Figure 3). Although the ARIMAX model and Fourier regression model are the effective models for forecasting Monthong durian price index, the MAPE using ARIMAX model was 3.365 times more than Fourier regression model. Therefore, Fourier regression model is more accurate than ARIMAX model. However, Fourier regression model, nonparametric regression, used many parameters to fit the model.

Table 3. The comparison of MAPE between ARIMAX and Fourier regression models

Model	MAPE	Comparison ratio
ARIMAX	15.148	3.365 :1
Fourier Regression	4.5010	

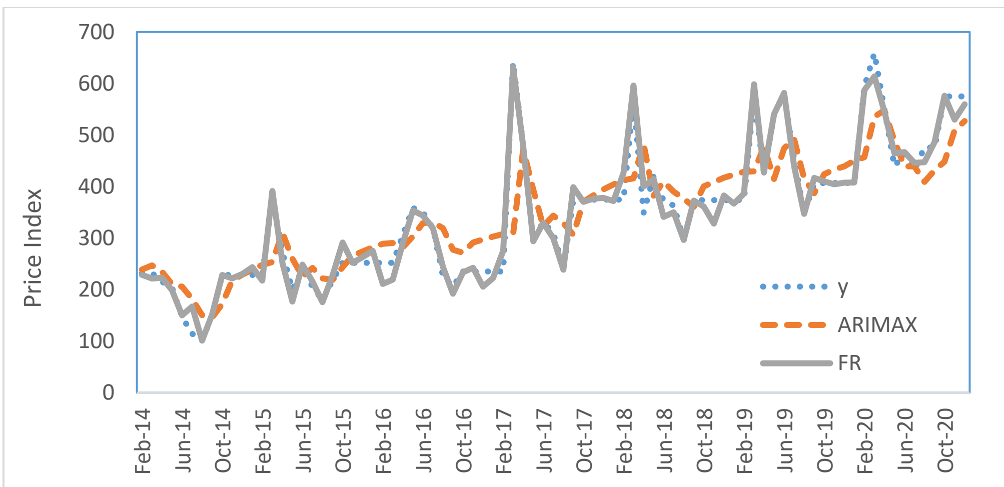

Figure 3. Graphs showing the actual values (y), the forecasted values of ARIMAX model (ARIMAX) and Fourier regression model (FR)

4. Conclusions

This paper presents the comparison of the forecasting model efficiency between ARIMAX model and Fourier regression model for forecasting Monthong durian price index using the data set between January 2014 and December 2020. Production index and China consumer confidence were explanatory variables. The results show that both models have the performance for forecasting Monthong durian price index. Nevertheless, Fourier regression model provides the lower errors.

References

[1] Isvilanonda, S., 2019. *The Situation of the World Production and Consumption of Durian and Thai Exports of Durian, Document Assembled in the Forum "Look at the future of Thai Export Durian Market"*. [online] Available at: https://www.slideshare.net/somporn isvilanonda/durian-production-and-consumption-and-thailand-export-70619

[2] Office of Agricultural Economics, 2020. *Agricultural Production Index*. [online] Available at: http://www.oae.go.th/assets/portals/1/fileups/aeocdata/files/Table1_th_agriPriceIndex_02_64.XLS

[3] Office of Agricultural Economics, 2020. *Agricultural Price Index*. [online] Available at: http://www.oae.go.th/assets/portals/1/fileups/aeocdata/files/Table2 _th_agriPriceIndex_02_64.XLS

[4] National Bureau of Statistics of China, 2020. *China Consumer Confidence*. [online] Available at: https://tradingeconomics.com/china/consumer-confidence

[5] Dhakonlayodhin, B. and Areepong, Y., 2018. A comparison of forecasting models of stock price using ARIMA and ARIMAX models. *Huachiew Chalermprakiet Science and Technology Journal*, 4(1), 44-55.

[6] Anggraeni, W., Vinarti, R. and Kurniawati, Y., 2015. Performance comparisons between Arima and Arimax method in moslem kids clothes demand forecasting: Case study. *Procedia Computer Science*, 72, 630-637.

[7] Prahutama, A., Suparti and Utami, T.W., 2018. Modelling fourier regression for time series data- a case study: modelling inflation in foods sector in Indonesia. *Journal of Physics: Conference Series,* 974, 012067, https://doi.org/10.1088/1742-6596/974/1/012067

[8] Bilodeau, M., 1992. Fourier Smoother and Additive Models. *The Canadian Journal of Statistics*, 3, 257-259.

[9] Semiati, R., 2010. *Fourier Birespon Series Nonparametric Regression.* Thesis. ITS, Surabaya.

[10] Asrini, L.J., 2012. Fourier series parametric regression. *Proceedings of the National Seminar FMIPA Universitas Negeri Surabaya*, 24 November 2012, Surabaya, Indonesia, pp. 77-80.

[11] Tiao, G.C. and Box, G.E.P., 1981. Modelling multiple time series with applications. *Journal of the American Statistical Association*, 76(376), 802-816.

[12] Yang, M., Xie, J., Mao, P., Wang, C. and Ye, Z., 2017. Application of the ARIMAX model on forecasting freeway traffic flow. *17th COTA International Conference of Transportation Professionals*, 7-9 July 2017, Shanghai, China.

[13] Drapper, N.R and Smith, H., 1992. *Applied Regression Analysis*. 2nd ed. New York: Marcel Dekker.

[14] Eubank, R.L.,1988. *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.