Economic and Technology Management Review, Vol. 12(2017): 11 – 17

Forecast of annual paddy production in MADA region using ARIMA (0,2,2) model

[Anggaran pengeluaran padi tahunan di kawasan MADA menggunakan model ARIMA (0,2,2)]

Aimi Athirah Ahmad*, Mahendran Shitan** and Fadhilah Yusof***

Keywords: paddy production, modelling, time series, ARIMA model

Abstract

Rice is the nation's staple food thus sustainable paddy production is crucial for food security. Forecasting paddy production is important to ensure Malaysians livelihood in the future. Based on MADA, rice production in the MADA region showed an increasing trend over a period of 35 years from 1979 to 2014. The objective of this study is to forecast annual paddy production in the MADA region from 2016 to 2020 by using five different ARIMA models and previous MADA data. The model also predicted that paddy production in the region will be at 1,257,400 tonnes in 2020. Forecasting results can be used as the foundation for national policies regarding food security.

Introduction

The current total population of Malaysia is 30.9 million and is expected to increase by 4.9% in year 2020. Rice is the nation's staple food thus sustainable paddy production is crucial for food security of the growing population. Forecasting paddy production is important to ensure Malaysians livelihood in the future. Through statistical modeling, forecasting results will provide policy makers the foundation to formulate or re-formulate and implement better policies to sustain paddy production in the country. Rice is mainly cultivated in Peninsular Malaysia which contributes to 85% of the nation's total paddy production. There are seven major rice granary areas which are in Kedah, Perlis, Kelantan, Terengganu, Pulau Pinang, Perak and Selangor. These areas are reserved by the government for wetland paddy cultivation (Firdaus et al. 2012). Amongst these, the largest is the Muda Agricultural Development Authority (MADA). In addition, the Muda granary area accounts for 31.7% and 24.1% of total rice production and rice cultivation area in Malaysia, respectively. Based on MADA data, rice production showed an increasing trend over a period of 35 years from 1979 to 2014.

Due to the availability of advanced computer technology, various sophisticated statistical models have been developed to forecast crop production. In general, there are two types of time series models; univariate and multivariate. Unlike multivariate models, univariate time series model excludes other variables that affect

*Economic and Social Science Research Centre, MARDI Headquarters, Persiaran MARDI-UPM, 43400 Serdang, Selangor

**Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor

***Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor

E-mail: aimiathirah@mardi.gov.my

production such as environment, water, and soil and focuses only on crop yield. An example of univariate model is the Autoregressive Integrated Moving Average (ARIMA) model.

Several studies have used ARIMA model to successfully forecast rice production. Biswas and Bhattacharyya (2013) predicted that paddy production in West Bengal, India would reach 16,500.5 thousand tonnes in 2016 by choosing ARIMA (2,1,1) as the best fitted model. By using ARIMA (2,1,0) model, Sivapathasundaram and Bogahawatte (2012) predicted an increase of rice production in Sri Lanka. The ARIMA model has also been used in forecasting production of other crops such as coconut (Cristina 2015) and coffee (Craparo et al. 2015). In the latter study, results showed a decreasing production trend due to climate change.

Besides forecasting crop production, the ARIMA model can also be used to predict wholesale price. Kim et al. (2015) compared various statistical models including ARIMA to estimate the future price for onions and cabbages. The study showed that Holt-Winters smoothing, ARIMA and VAR models have similar MAPE of estimate values of 11.3%, 11.5% and 11.4%, respectively. These three models predicted that the wholesale price for cabbages ranged between 450 and 490 won in March 2015.

Therefore, the objective of this study is to forecast annual paddy production in the MADA region from 2016 to 2020 by using ARIMA model and previous data on rice production of MADA.

Materials and methods

Data on annual paddy production in the MADA region were obtained from MADA's Planning and Information Technology Section. For each planting season, MADA conducted a 'Paddy Production Survey' in four districts of the MADA region namely Kangar, Jitra, Pendang and Kota Sarang Semut. The annual paddy production was based on the average rice yield collected during the surveys.

The annual production data were analysed using Integrated Time Series Modelling software (version ITSM2000). The time series plot of paddy production in MADA from year 1979 to 2014 consists of 36 observations is shown in *Figure 1*. From the figure, it is clearly showed that the time series plot is not stationary and the productions of paddy in MADA were increased over the years. However, there are two years that the productions are significantly decreased due to some events that may have caused to a drop of paddy production in those years. The model used in this study is the Auto Regressive Integrated Moving Average (ARIMA) time series model. Based on Brockwell and Davis (2002), ARIMA is the integration of Auto Regressive Moving Average (ARMA) process where ARIMA process can be reduced to ARMA process if and only if d=0. ARIMA model uses the lag and shift of historical information to predict future patterns. This model is represented as a regression model with a moving average to provide great detail and precision.

A zero mean stationary ARMA (p, q) process is defined as a sequence of random variable $\{X_t\}$ given by:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q} \quad (1)$$

where $\{Z_t\}$ is a sequence of uncorrelated random variables with zero mean and constant variance, denoted as $\{Z_t\} \sim WN(0,\sigma^2)$ where $\{X_t\}$ is ARMA (p, q) with mean μ if $\{X_t - \mu\}$ is an ARMA (p, q) process. If d is a non-negative integer, then $\{X_t\}$ is an ARIMA (p, d, q) process if $(1 - B)^d X_t$ is ARMA (p, q) process, where B is a usual backward shift operator.

Similar to other types of regression modeling such as linear and multivariate regression, ARIMA follows a 3-stage modeling process which is identification, estimation and diagnostic testing, and forecasting in order to model the paddy production.

Identification process

Firstly, a simple linear regression modeling was used to characterise the trend component of the data. Subsequently, the data was differenced twice at lag=1 so that *d*=2. After that, sample ACF and sample PACF for differenced data were plotted. In *Figure 1*, the sample ACF and PACF suggested that the model may contains the parameters p(0,1) and q(0,1,2), respectively.

Estimation and diagnostic testing

The suggested models were then estimated using Preliminary Estimation option in ITSM2000 software. By using the maximum likelihood estimation, the preliminary models that give the minimum value of Akaike's Information Corrected Criterion (AICC) were chosen to fit the model well.

The formula of AICC is given by:

$$AICC = -2 \ln likelihood\ (\hat{\phi}, \hat{\theta}, \hat{\sigma}^2) + \frac{[2n(p+q+1)]}{[n-(p+q)-2]} \qquad (2)$$

where

ϕ = a class of autoregressive parameters,
θ = a class of moving average parameters,
σ^2 = variance of white noise,
n = number of observations,
p = order of autoregressive component,
q = order of moving average component,

and *likelihood* $(\phi, \hat{\theta}, \hat{\sigma}^2)$ is the likelihood of the data under Gaussian ARMA model with parameters $(\phi, \hat{\theta}, \hat{\sigma}^2)$.

Forecasting process

Using the best selected model, paddy production in MADA region from year 2015 to 2020 was forecasted. The selected models were then compared using mean absolute error (MAE), the root mean square error (RMSE) and mean absolute percentage error (MAPE) to measure the accuracy of the forecast data.

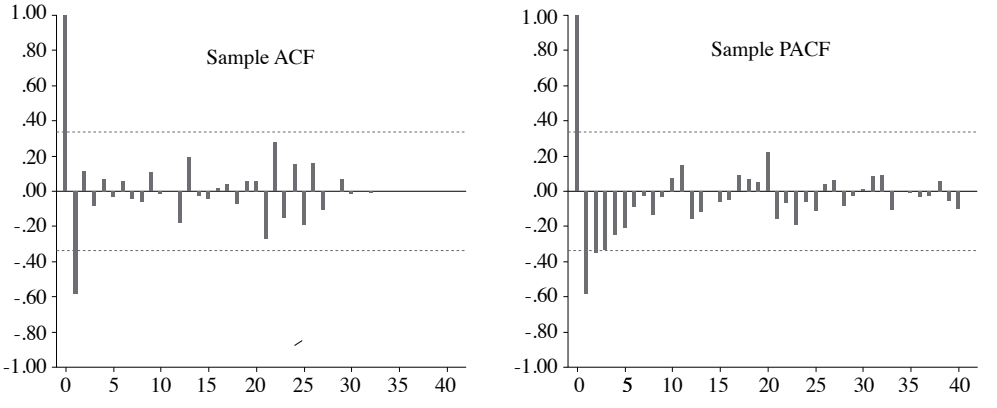

Figure 1. Stationarity of differenced data (d=2) for sample ACF and PACF

The criteria are explained by the following equations:

$$MAE = \frac{\Sigma_{i=1}^{n}|y_i - \hat{y}_t|}{n}, \qquad (3)$$

$$RMSE = \sqrt{\frac{\Sigma_{i}^{n}(y_i - \hat{y}_t)^2}{n}}, \qquad (4)$$

$$MAPE = \frac{\left|\frac{y_i - \hat{y}_t}{y_i}\right|}{n} x\ 100\%. \qquad (5)$$

Where y_i and $\hat{y}_t$ are the actual observed values and the predicted values respectively while n is the number of predicted value.

Results and discussion

Using historical data on paddy production in the MADA region from year 1979 to 2014, a time series plot that consists of 36 observations were produced (*Figure 2*). The plot shows that paddy production in the MADA region was not stationary and increased over the years. However, there was a significant decrease in production for a few years, most likely due to some events that may have caused to a drop of paddy production in those years.

A clearer method to determine the stationarity of the data was used by producing a sample ACF and sample PACF (*Figure 3*). Sample ACF showed that the auto-correlation function exhibited a slow decay as the numbers of lag increased indicating that the data was not stationary. Additionally, data transformation was done by differencing the data twice at lag 1 (d=2) which resulted in stationary data.

Since the series for paddy production $\{Y_t : t = 1,2, ...,36\}$ is non-stationary, transformation was needed in order to eliminate the trend component and to achieve stationarity. The time series $\{Y_t\}$ was differenced twice at lag-1 to obtain stationarity and thus gave:

$$V_t = (1 - B)^2 Y_t. \qquad (6)$$

The time series plot, V_t, is shown in Figure 4. The plot showed the absence of trend component when data transformation was applied.

To model the time series as zero mean stationary process, the mean of V_t to give

$$X_t = V_t - (2861)$$

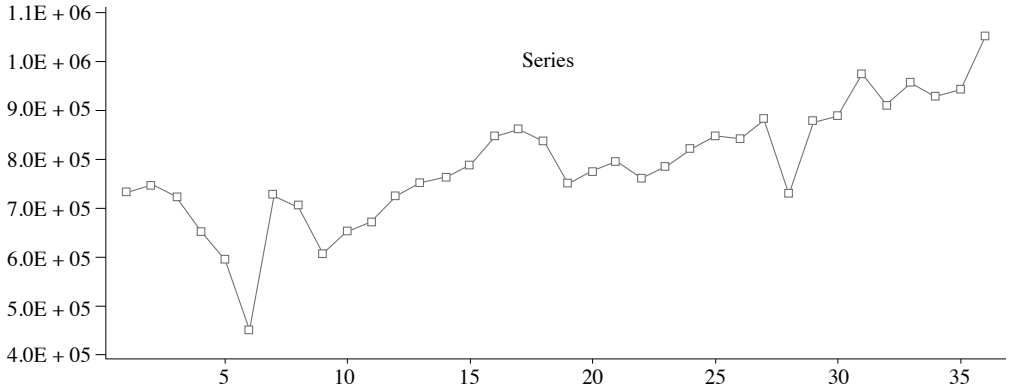

Figure 2. Annual paddy production in MADA region from the previous 35 years (1979 to 2014)

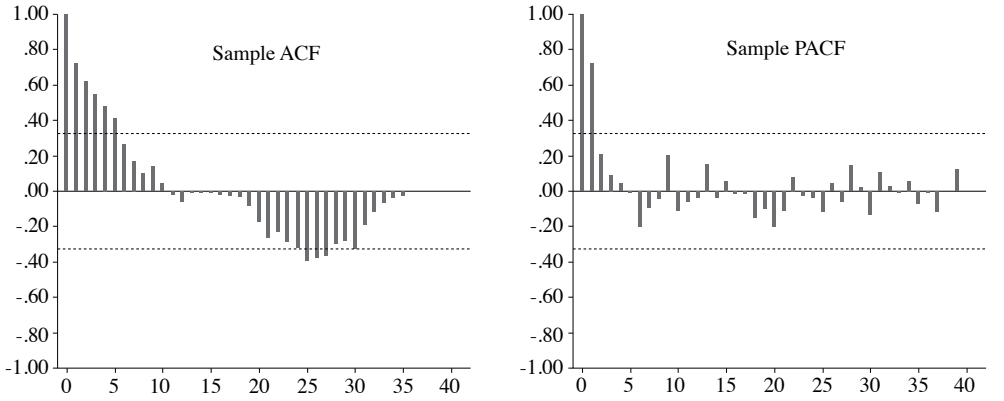

Figure 3. Stationarity of the original data for sample ACF and PACF

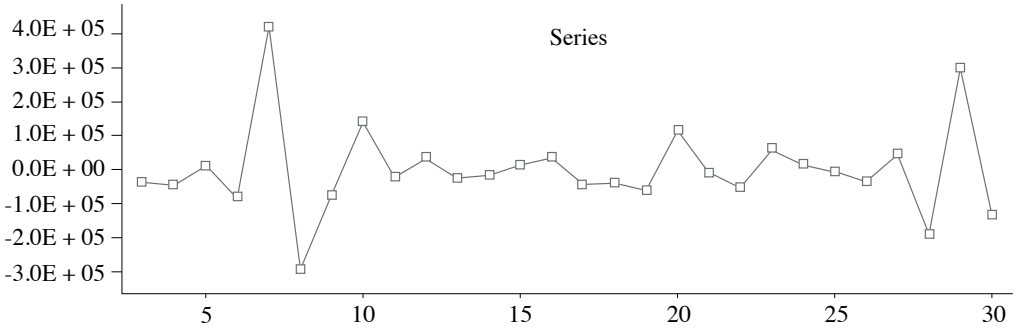

Figure 4. Plot of time series $\{V_t\}$

In this study, the forecasted values for paddy production from year 2015 to 2020 were tested using five different ARIMA models. These values were then compared to the true value obtained from MADA to evaluate the forecasting performance. The best model used to forecast paddy production in MADA region was chosen based on its forecasting performance. The criteria used to evaluate forecasting performance are mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE). Models with lower values of MAE, RMSE and MAPE indicate more accuracy in forecasting paddy production. The values of AICC, MAE, RMSE and MAPE for each possible model are summarised (*Table 1*).

Based on *Table 1*, ARIMA (0,2,1) model had the minimum value of AICC of 871.63. However, ARIMA (0,2,2) model was chosen as the best model to forecast paddy production in the MADA region based on its forecasting criteria. ARIMA (0,2,2) had the lowest MAE, RMSE, and MAPE values compared to the other models. where the value of MAPE is 14.472% when compare the true value in 2015 which is 936995 tonnes (MADA 2016).

Forecasting results are indications for the extent of contribution of the MADA granary region to the overall rice production in the country. *Table 2* shows the forecasted paddy production in the MADA region from 2016 to 2020 using ARIMA (0,2,2) model. In 2020, paddy production in the region was

Table 1. AICC, MAE, RMSE and MAPE values of five different ARIMA models when compared to true value of paddy production in 2015

Model	AICC	MAE	MAPE	MAPE
ARIMA (1,2,0)	884.75	174605	174605	18.635%
ARIMA (0,2,1)	871.63	168305	168305	17.962%
ARIMA (1,2,1)	872.28	141205	141205	15.070%
ARIMA (0,2,2)	872.15	135605	135605	14.472%
ARIMA (1,2,2)	874.83	144105	144105	15.380%

Table 2. Forecast of paddy production in MADA region from 2016 to 2020 using ARIMA (0,2,2) model

Year	Forecasted value (metric tonnes)	95% Confidence interval
2016	1,072,600	(916,540,1,228,700)
2017	1,114,500	(900,240,1,328,800)
2018	1,159,300	(883,360,1,435,200)
2019	1,206,900	(865,620,1,548,200)
2020	1,257,400	(847,020,1,667,700)

predicted at 1,257,400 tonnes with a 95% confidence interval of (847,020,1,667,700). Since the population is expected to increase by 1.6% yearly until 2020 (Anon. 2011), the forecasting results can be used as the foundation for national policies regarding food security.

This study, however, only considers the time series data of paddy production for forecasting purposes and excludes, however there could be other factors that might affect paddy production of paddy for example, such as climate variability, management practices, soil characteristics and water. Further studies will be conducted to include these production parameters in multivariate models thus providing a more accurate prediction of paddy production. As a result, the future study of this project is to apply other available model to forecast paddy production that incorporating more agriculture parameters and related information to predict future production.

Conclusion

Using a univariate time series forecasting approach, it was determined that ARIMA (0,2,2) is the best model to predict paddy production in the MADA region for year 2016 to 2020.

Acknowledgement

The authors would like to acknowledge the Department of Planning and Information Technology, MADA for providing data and the Agrobiodiversity and Environment Research Centre, MARDI for funding this project.

References

Anon. (2011). *National Agro-Food Policy 2011-2020*. Ministry of Agriculture and Agro-Based Industry, Putrajaya, Malaysia

Brockwell, P.J. and Davis, R.A. (2002). *Introduction to time series and forecasting, 2nd Edition*, Springer-Verlag, New York

Craparo, A.C.W., Van Asten, P.J.A., Laderach, P., Jassogne, L.T.P. and Grab, S.W. (2015). *Coffea arabica* yields decline in Tanzania

due to climate change: Global implications. *Agriculture and Forest Methodology* 207: 1 – 10

Cristina, T.L. (2015). Forecasting coconut production in the Philippines with ARIMA model. *AIP Conference Proceeding*,1643,86

Biswas, R. and Bhattacharyya, B. (2013). ARIMA modeling to forecast area and production of rice in West Bengal. *Journal of Crop and Weed* 9(2): 26 – 13

Sohyun, K., Kuk, H.N., Cheolho, S. and Youngchan, C. (2015). The development and evaluation of onion and cabbage wholesale price forecastingmodels *International Journal of Software Engineering and Its Applications* 9(8): 37 – 50

Sivapathasundaram, V. and Bogahawatte, C. (2012). Forecasting of paddy production in Sri Lanka: Time series analysis using ARIMA model *Tropical Agriculture Research* 24(1): 21 – 30

Abstrak

Beras merupakan makanan ruji rakyat negara ini, oleh itu kelestarian pengeluaran padi adalah penting untuk keselamatan makanan. Unjuran penghasilan padi adalah penting bagi memastikan kelangsungan hidup pada masa hadapan. Berdasarkan data daripada MADA, trend penghasilan beras menunjukkan peningkatan selama 35 tahun iaitu dari 1979 hingga 2014. Objektif kajian adalah untuk membuat ramalan hasil padi tahunan di MADA pada tahun 2016 hingga 2020 dengan menggunakan model ARIMA dan data MADA yang sedia ada. Model ini meramalkan peningkatan hasil padi sebanyak 1,257,400 tan pada tahun 2020. Hasil kajian ini dapat digunakan sebagai asas kepada polisi dan dasar negara dalam isu keselamatan makanan.