INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES
Vol 14, Issue 12, (2024) E-ISSN: 2222-6990

Predicting Paddy Production in Malaysia: A Comparative Analysis between Arima and Neural Network Autoregressive (NNAR) Models

Siti Nurhafizah Mohd Shafie[1], Nasuhar Ab. Aziz[2*], Mohd Noor Azam Nafi[3], Syazwani A.Malek @ Abdul Malek[4], Azzah Amran[5], Siti Nur Atiqah Mohd Shafie[6]
[1,2,3]College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Kelantan, Kota Bharu Campus, 15150 Kota Bharu, Kelantan, Malaysia, [4,5]Faculty of Business and Management, Universiti Teknologi MARA Kelantan, Kota Bharu Campus, 15150 Kota Bharu, Kelantan, Malaysia, [6]College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Kampus Mukah, 96400 Mukah, Sarawak, Malaysia
Email: sitinurhafizah@uitm.edu.my, norazam05@uitm.edu.my, waniey@uitm.edu.my, azzahamran@uitm.edu.my, sitinuratiqahms@uitm.edu.my
Corresponding Author Email: nasuha978@uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v14-i12/23351 DOI:10.6007/IJARBSS/v14-i12/23351

Published Date: 28 December 2024

Abstract

Forecasting behavior of economic and machine learning models has recently attracted much attention in the research sector. In this study an attempt has been made to compare the forecasting behavior of Autoregressive Integrated Moving Average (ARIMA) and Neural Network Autoregressive (NNAR) modes using univariate model time series data of annual paddy production (1980-2022) in Malaysia. The data was obtained from the open website of Department of Statistic Malaysia (DOSM). Through the evaluation of forecasting accuracy suing Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), the results showed that the eliminated error of NNAR is much smaller than the estimated error of ARIMA for paddy production. So, the best model to forecast paddy production is NNAR(1,1).

Keywords: Box-Jenkins Methodology, Forecasting, Neural Network Autoregressive, Paddy production, Time Series Model

Introduction

Rice and paddy are closely related terms within the context of agriculture. Rice refers to the starchy seeds of cereal grass, a staple food consumed globally. On the other hand, paddy specifically denotes the rice that still has its hull surrounding the inner edible kernel (Pallas, 2016). While rice is a consumable grain, paddy refers to the pre-harvest stage, representing the flooded cultivation area. The distinction between these two highlights the different stages of the rice cultivation process, emphasizing the harvested product and the field in which it grows.

In 2021, rice accounted for 90% of all cereal production and was ranked among the top five agricultural products produced worldwide (FAO, 2022). This emphasizes the critical role of paddy production in the agricultural system. On a global scale, China leads in paddy production, followed by India, Bangladesh, Indonesia, Vietnam, Thailand, Philippines, Burma, Pakistan, and Brazil. Notably, Malaysia also contributes significantly as the 26th country in global paddy cultivation (The United States Department of Agriculture, 2024). Majority of the paddy produced in Malaysia is manufactured in the Muda Agricultural Development Authority (MADA) region, the Kemubu Agricultural and Development Authority (KADA) region and the Integrated Agricultural Development Area (IADA) region (DOA, 2023).

Instead of exporting paddy, Malaysia imports rice from Thailand, Vietnam, Pakistan, India, and Myanmar (Padiberas Nasional Berhad, 2024). These countries are recognized as primary suppliers of rice to Malaysia. In terms of paddy exports, Malaysia has no substantial paddy export activity. Instead, the country imports rice, which is a milled and processed version of paddy. Malaysia faces challenges in achieving self-sufficiency in rice production due to various factors, including limited arable land and water resources. As a result, this country relies on rice imports to meet its domestic demand. According to statistics from the Department of Agriculture, Malaysia produced 1.677 million metric tonnes of rice in 2021 and imported 1.62 million metric tonnes of rice. In the same period, Malaysia exported 161,000 metric tonnes of rice (The Edge, 2023). It's important to note that while Malaysia is a major producer of other agricultural products, rice production doesn't currently suffice to fulfill its internal consumption needs.

The paddy production trend in Malaysia from 1980 to 2018 shows a general increase, peaking in 2014 at 2,848,560 metric tons. However, a noticeable decline is observed from 2015 to 2022, with production dropping from 2,741,4042 metric tons to 2,281,739 metric tons. The rice production trend in Malaysia from 1980 to 2018 show general increase, peaking in 2018 at 1699766 metric tons. The trend then continues to drop to 1574956 metric tons in 2022. This downward trend may indicate various factors affecting paddy cultivation during this period, requiring further analysis to understand and address the decline in production.

In light of the declining situation of paddy in Malaysia, few studies have been undertaken to forecast paddy production using variance methods such as Exponential Smoothing Techniques, Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) (Fauzi & Bakar, 2022; Setiawan & Fatekurohman, 2022; Mahat & Idris, 2018; Ahmad et al., 2017; Alimana et al., 2017; Saad & Ismail, 2009; Samsuddin et al., 2008). However, according to Khaishei and Bijari (2011) ARIMA approximation may not be sufficient for complex non-linear real-world issues. This is also supported by Zhang et al. (1998) that real world systems were frequently non-linear. The Neural Network AutoRegressive(NNAR) model was a type of ANN model over other non-linear statistical models. Various research has been done in comparing ARIMA and NNAR model for forecasting production of rice in India (Vijayalakshmi et al., 2023; Annamalai & Johnson, 2023).

Thus, to the best of our knowledge, there is no comparative study on forecasting paddy using Exponential Smoothing Techniques, ARIMA and NNAR. Therefore, the purpose of this study is to analyze and predict the annual future trend of paddy production by comparing Exponential Smoothing Techniques, ARIMA and NNAR model. Given the visible reduction in past statistics, we shall forecast paddy in Malaysia for the following ten years.

Data Sources and Method

The data was obtained from the open website of Department of Statistics Malaysia (DOSM). The dataset contains yearly Malaysia's paddy production data in tonnes from the period 1980 to 2022 and were used in the process of identification, estimation and forecasting. Research by Saad and Ismail (2009) divided the data acquired into in-sample evaluation and out-sample evaluation, allocating 80% of the data for in-sample evaluation and 20% for out-sample evaluation. Therefore, to construct potential models for paddy production forecasting, we employed the methodology from previous research and divided the data into training and validation sets using a ratio of 80:20. The R programming language was utilized to perform the analysis for both models, ARIMA and NNAR.

Box Jenkins Methodology

The Box-Jenkins method is a statistical technique used for time series analysis and forecasting. It was developed by George Box and Gwilym Jenkins and it is divided into three modeling stages which are identification, estimation and validation. The details of each stage were summarized in Table 1.

Table 1
Stages in ARIMA modelling

Stage	Process	Criteria
1.	Model Identification	Analyze historical statistics, including ACF and PACF. After analyzing the data, the most appropriate subclass of the general model is identified.
2.	Model Estimation	The parameters are estimated after order of p, d, and q have been determined
3.	Model Validation	Models are analyzed for adequacy and inadequacy. If the model cannot satisfy the test criteria, it may need to be respecified.
4.	Model Application	After meeting all test criteria and confirming the model's fitness, it is ready to predict the forecast value.

Autoregressive Integrated Moving Average (ARIMA)

The moving average (MA) and autoregressive (AR) models are combined to create ARIMA. The amount of differencing that must be done on the time series data in order for it to be stationary is indicated by the Integrated (I) value. ARIMA (p,d,q) is the notation for this model, which has three parameters (p,d,q). The moving average components are denoted by the parameter q, the number of autoregressive lags is denoted by d, and the number of data differencing operations to make the data series stationary is indicated by p. The AR (p) model is generally expressed as (1).

yt = $\Phi tyt{-}1$ + … + $\Phi pyt{-}p$ + et
(1)

where et is the error term at time t, yt is the variable value at time t and Φj is the parameter of the AR coefficient with j = 1, 2, …, p. The AR model is dependent on the value of the previous data, as shown by the above equation. In the meantime, both the previous error terms and the present error value have an impact on the MA (q) model. The MA model is typically expressed as (2).

yt = et − $\theta 1et{-}1$ − … − $\theta qet{-}q$
(2)

where et is the error term at time t, yt is the variable value at time t, and θi is the parameter of MA coefficient, i = 1, 2, …, q. As a result, the equation of the AR and MA models that was previously mentioned is combined to create the ARMA model. Equation (3) represents the general expression of the ARIMA(p,q) model where the description of this equation is the same as that of the AR and MA models.

yt = $\Phi 1yt{-}1$ + … + $\Phi pyt{-}p$ + et − $\theta 1et{-}1$ − … − $\theta qet{-}q$
(3)

Since the ARIMA model is for non-stationary series, it requires the process of differencing towards the data set in order to achieve a stationary series. The differencing is performed by using the backward shift operator, which indicates the number of backward steps a time-series value may take. The backward shift operator is denoted by β. When the operator is applied to an equation yt, then βyt = $yt{-}1$. This indicates that the current period of time has been shifted backwards by one period. Therefore, for the first order of differencing, Δyt can be written as (4).

$$\begin{aligned}\Delta\, y_t &= y_t - \ y_{t-1}\\ &= (1-\beta)y_t\end{aligned}$$
(4)

To sum up, the orders of differencing can generally be written as $(1-\beta)^d y_t$ at dth lag.

Mathematically, ARIMA (p,d,q) is written as (5) and (6)

$$w_t = \varphi_1 w_{t-1} + \varphi_2 w_{t-2} + \cdots + \varphi_p w_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} + e_t$$
(5)

$$\left(1 - \varphi_1\beta - \varphi_2\beta^2 - \cdots - \varphi_p\beta^p\right)w_t = \left(1 - \theta_1\beta - \theta_2\beta^2 - \cdots - \theta_q\beta^q\right)e_t$$
(6)

From the equation (6), substituting $w_t = (1-\beta)y_t$ as an example for first differencing and the result after simplifying, then the equation becomes equation (7).

$$\left(1-\varphi_1\beta-\varphi_2\beta^2-\cdots-\varphi_p\beta^p\right)(1-\beta)y_t=\left(1-\theta_1\beta-\theta_2\beta^2-\cdots-\theta_q\beta^q\right)e_t \quad (7)$$

Neural Network Autoregressive (NNAR)

In the modeling process, we focused on the NNAR model for machine learning. We used a computerized approach to determine the most suitable number of hidden layers. We produced the most accurate models by deliberately altering the number of hidden layers and neurons. It's worth noting that neural networks without hidden units are practically identical to linear statistical forecasting methods. Hidden units are important in neural networks because they allow the mapping of input and output variables along with introducing nonlinearity. They also help to identify trends in the dataset. Lagged values for time-series data can be used as input data for a neural network, just as they would be in a linear autoregressive model.

An NNAR (p,k) model denotes that there are k nodes and p delayed inputs in the hidden layer. Furthermore, a NNAR (p,0) model lacks the parameter constraints that guarantee stationarity, making it equivalent to an ARIMA (p,0) model. The production of the model is developed in two steps. The K activations happen first. The hidden layer is determined as a function of the input characteristics $X_j = X_{t-1},\ldots,X_{t-p}$ in the activation A(k) for k = 1,..., K,

$$(k)=h(k)=g\left(w_{k0}+\sum_{j=1}^{p}w_{kj}X_j\right) \quad (8)$$

where g is a predefined nonlinear activation function. You can think of each A(k) as a distinct $h_k(x)$ transformation of the unique features. The K activations from the hidden layer are transferred to the output layer.

$$f(x)=\beta_0+\sum_{k=1}^{K}\beta_k A(k) \quad (9)$$

NNAR analysis employs the sigmoid activation function, being equivalent to the logistic regression function. This activation function converts a linear function to a probability that ranges from 0 to 1. The sigmoid activation function can be expressed mathematically as (10).

$$g(z)=\frac{\exp(z)}{1+\exp(z)}=\frac{1}{1+\exp(-z)} \quad (10)$$

Model Evaluation

Mean Absolute Error (MAE)

Without accounting for the direction of the forecasts, the MAE computes the errors for average magnitude in a set of predictions. This refers to the difference that exists between the observation and the measured value. The following is the formula to determine the MAE value where N is the size of the test set, and $\hat{y}_t$ is the predicted value of y_t. The MAE is given as (11).

$$\text{MAE} = \frac{1}{N}\sum|y_t - \hat{y}_t| \quad (11)$$

Root Mean Squared Error (RMSE)
The error also can be assessed using the Root Mean Square Error. It is also claimed to be especially sensitive to outliers and to be the most widely used measure. The model's prediction is more accurate the lower the RMSE is. Nevertheless, Hyndman & Athanasopoulos (2018) noted that when time series data has a different scale, this scale cannot be used to compare predicting accuracy levels. The following determine the RMSE value where N is the size of the test set, and $\widehat{y_t}$ is the predicted value of y_t. The RMSE is given as (12).

$$\text{RMSE}= \sqrt{\frac{1}{N}\sum(y_t - \widehat{y_t})^2} \quad (12)$$

Mean Absolute Percentage Error (MAPE)
MAPE is also the test that will be applied to estimate the accuracy of the forecasting model. A low or small MAPE value indicates that the forecasting model has a better and good performance. Widmark (2022) stated that using MAPE values enables researchers in making comparisons of the accuracy value between different datasets. It can be measured by using the formula that is written below where N is the size of the test set, and $\widehat{y_t}$ is the predicted value of y_t. It usually expresses accuracy as a percentage and it is defined as (13).

$$\text{MAPE} = \frac{1}{N}\sum\left|\frac{y_t - \widehat{y_t}}{\widehat{y_t}}\right| \quad (13)$$

Results and Discussion
Figure 1 shows the trends in paddy and rice production from 1980 to 2022. Paddy production, represented by the straight line, shows significant fluctuation especially in the early years, with noticeable dips and recoveries until the mid-1990s. From that point onwards, there is a more consistent upward trend with periods of minor fluctuations, peaking around 2017 before experiencing a slight decline. This is due to Several factors that have contributed to the decline in paddy production in Malaysia, including the reduction in available land and the limited adoption of mechanization technology.

Rice production denoted by dash line, on the other hand, exhibits a more consistent trend with less volatility compared to paddy production. From 1980 to around 1995, rice production showed a slight downward trend, but from the late 1990s onwards, there is a steady increase. There was a sudden positive shock in 2017, leading to a peak in 2018 at 1,699,766 metric tons. The peak in Malaysia's rice production in 2018 was partly due to increased imports from India. During the same period, total rice production in India increased by 3.21 million tonnes. The increase in rice production in India resulted in a higher number of rice imports to Malaysia. The rise continues until around 2018, after which there is a slight decline, but the overall trend remains positive.

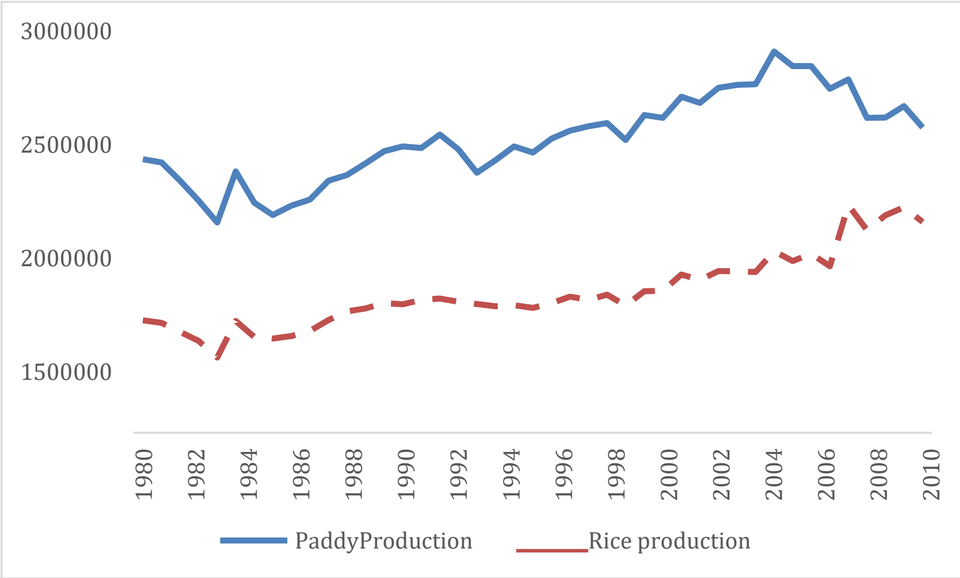
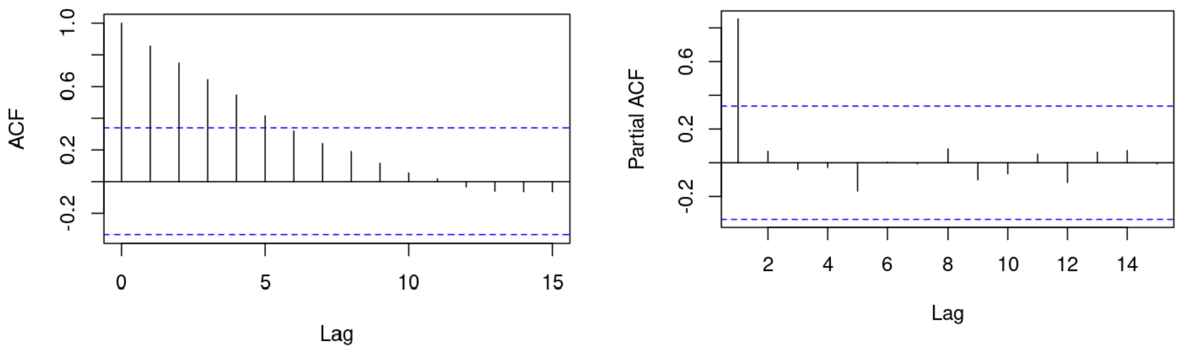


Figure 1: Trend Pattern of Paddy Production and Rice Production

Figure 2 depicts the ACF and PACF plot for paddy production time series data. The plot



illustrates an obvious decay pattern; hence the series is non-stationary. The PACF plot emphasizes only one spike at lag one indicating that only the first order of differencing is needed to make the series stationary.
Figure 2: ACF and PACF Plot of Original Paddy Production Time Series

The appropriate form of ARIMA(p,d,q) model can only be determined in cases where the data series is stationary. Since the data is proven to be not-stationary, differencing was performed to fulfill the scenario. Figure 3 and Figure 4 show the trend analysis for paddy production after first differencing. The time series plots show the stationary series since the fluctuation of the series around mean value zero. The plots show that the difference series are stationary by looking at the spike of ACF and PACF that cut off quickly. Figure 4.6 shows that none ACF and PACF values are significant because there are no spikes beyond the upper or lower confidence limit for paddy production series. Therefore, it can be concluded that the random walk ARIMA (0,1,0) model yielded a reasonably well fit for paddy production difference series for ARIMA model. Moreover, following Pooja et al. (2023), the study also assumed the ARIMA (0,1,0) model for paddy production.

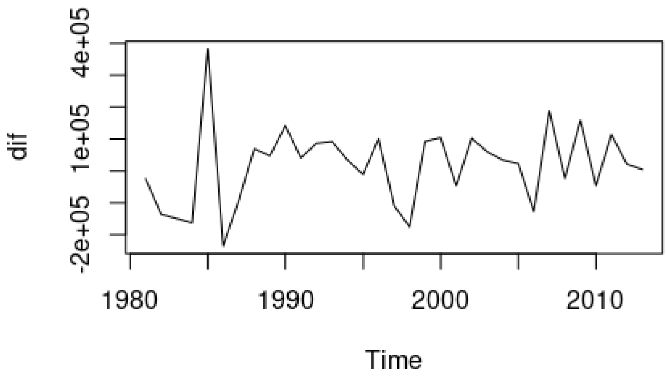


Figure 3: Trend Analysis for Paddy Production After First Differencing

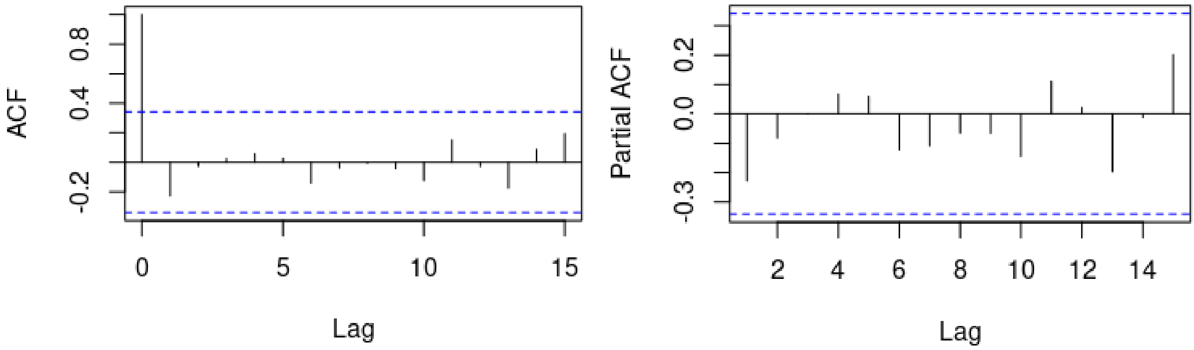


Figure 4: ACF and PACF Plot of Paddy Production Time Series After First Differencing

ARIMA and ANN Performance

Table 2 summarized the error measure of in-sample and out-sample evaluation used for the best model of ARIMA(0,1,0) and NNAR(1,1). Aligning with the objective of this study, all the previously analyzed models were compared to choose the most appropriate model to forecast the paddy production in Malaysia. Among the models, NNAR(1,1) has the smallest value of error measure for paddy production. In general, the model that performs better on the out-sample evaluation part is considered more reliable for future predictions. Thus, the NNAR (1,1) is the most suitable model for estimating the future prediction for paddy production.

Table 2

Error Measure of Paddy Production for in-sample and out-sample evaluation using ARIMA and NNAR

ARIMA(0,1,0)	MAE	RMSE	MAPE
In-sample evaluation	95118.25	121909.5	4.761581
Out-sample evaluation	98277.95	133173.8	4.007766
NNAR(1,1)	MAE	RMSE	MAPE
In-sample evaluation	94723.57	113528.5	4.742611
Out-sample evaluation	89922.63	107422.1	3.604858

Conclusion

This study suggested an appropriate time series model by extending the benchmark for forecasting paddy in Malaysia. It is important to highlight that the period of the study represents the only period for which data are available. The data set consists of annual paddy and rice production from 1980 to 2022. The study fitted the model using data from 1980 - 2013 and evaluation part for 2014 - 2022. The forecasted paddy was compared to the actual values for years 1980 - 2013 in out sample evaluations. First, the performance of the ARIMA model in fitting and forecasting the paddy in Malaysia, was examined. The result shows that the random walk ARIMA (0,1,0) model is the only model suited in predicting the trend of paddy production. Moving to the proposed NNAR models in which the paddy was fitted and forecasted in order to evaluate the ability of the proposed NNAR model. The result found that NNAR (1,1) is the only appropriate model to fit paddy production in Malaysia. The overall forecasting performance of all models in this study was evaluated using one-step-ahead forecasts. Evaluations were carried out using out-sample forecasts. Overall, it is concluded that the NNAR (1,1) model was found to be the most reliable for forecasting paddy production.

References

Ahmad, A. A., Shitan, M., & Yusof, F. (2017). Forecast of annual paddy production in MADA region using ARIMA (0, 2, 2) model. Economic and Technology Management Review, 12, 11-17.

Alimana, K., Abasb, N., & Zakariac, W. N. W. (2017). Modelling of Paddy Production in Malaysia. In Master Project Symposium on Systems Engineering and Professional Science (Vol. 13, p. 12-18).

Annamalai, N., & Johnson, A. (2023). Analysis and forecasting of area under cultivation of rice in India: univariate time series approach. SN Computer Science, 4(2), 193.

DOA. (2023). Booklet Statistik Tanaman: Sub-sektor Tanaman Makanan. https://www.doa.gov.my/doa/resources/aktiviti_sumber/sumber_awam/maklumat_pertanian/perangkaan_tanaman/booklet_statistik_tanaman_2023.pdf

DOSM. (2023). Indikator Pertanian Terpilih. https://newss.statistics.gov.my/newss-portalx/ep/epFreeDownloadContentSearch.seam?cid=75975

Fauzi, F. D., & Bakar, A. S. A. (2022). Rice production forecasting in Malaysia: A Box-Jenkins and ARIMA model approach. In 8th Annual ECOFI Symposium, 212-220.

FAO. (2022). Agricultural production statistics. 2000–2021. FAOSTAT Analytical Brief Series No. 60. Rome. https://doi.org/10.4060/cc3751en

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Applied soft computing, 11(2), 2664-2675.

Mahat, N., Alias, R., & Muhamad Idris, S. (2018). Comparative study of fuzzy time series and artificial neural network on forecasting rice production. In Proceedings of the Second International Conference on the Future of ASEAN (ICoFA) 2017–Volume 2: Science and Technology (pp. 165-173). Springer Singapore.

Padiberas Nasional Berhad. (2024). Rice Importation. BERNAS. https://www.bernas.com.my/commitment/upstream/rice-importation

Pallas, L. (2016). Rice Processing: Beyond the Farm Gate. In Encyclopedia of Food Grains. Second Edition. Academic Press: Oxford. 446-452.

Pooja, B. N., Ajitha, T. K., Prema, A., Ayyoob, K. C., & Syama, S. M. (2023). Forecast models for area, production and productivity of paddy in Kerala.

Saad, P., & Ismail, N. (2009). Artificial neural network modelling of rice yield prediction in precision farming. Artificial Intelligence and Software Engineering Research Lab, School of Computer & Communication Engineering, Northern University College of Engineering (KUKUM): Jejawi, Perlis.

Samsudin, R., Saad, P., & Shabri, A. (2008). A Comparison of Neural Network, Arima Model and Multiple Regression Analysis in Modeling Rice Yields. Editorial Advisory Board, 113.

Setiawan, F., Dewi, Y. S., & Fatekurohman, M. (2022). Comparison of Arima Method and Artificial Neural Network Method to Predict Productivity Rice In Panti District. Edumaspul: Jurnal Pendidikan, 6(2), 2481–2494. https://doi.org/10.33487/edumaspul.v6i2.4681

The Edge. (2023). Cut rice exports until domestic supply stabilises, govt told. The Edge Malaysia. https://theedgemalaysia.com/node/682568

The United States Department of Agriculture. (2024). Rice Explorer. Ipad.fas.usda.gov. https://ipad.fas.usda.gov/cropexplorer/cropview/commodityView.aspx?startrow=21&cropid=0422110&sel_year=2023&rankby=Production

Vijayalakshmi, G., Pushpanjali, K., & Mohan Babu, A. (2023). A comparison of ARIMA & NNAR models for production of rice in the state of Andhra Pradesh. International Journal of Statistics and Applied Mathematics, 8(3), 251-257.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35-62.