Jurnal Teknologi, 52 Mei 2010: 135–147

HYBRID NEURAL MODELS FOR RICE YIELDS TIMES FORECASTING

RUHAIDAH SAMSUDIN[1], PUTEH SAAD[2] & ANI SHABRI[3]

Abstract. In this paper, time series prediction is considered as a problem of missing value. A model for the determination of the missing time series value is presented. The hybrid model integrating autoregressive integrated moving average (ARIMA) and artificial neural network (ANN) model is developed to solve this problem. The developed models attempts to incorporate the linear characteristics of an ARIMA model and nonlinear patterns of ANN to create a hybrid model. In this study, time series modeling of rice yield data in Muda Irrigation area, Malaysia from 1995 to 2003 are considered. Experimental results with rice yields data sets indicate that the hybrid model improve the forecasting performance by either of the models used separately.

Keywords: ARIMA; Box and Jenkins; neural networks; rice yields; hybrid ANN model

1.0 INTRODUCTION

Rice is the world's most important food crop and a primary source of food for more than half of the world's population. Almost 90% of rice is produced and consumed in Asia, and 96% in developing countries. In Malaysia, the Third Agriculture Policy (1998–2010) was established to meet at least 70% of Malaysia's demand a 5% increase over the targeted 65%. The remaining 30% comes from imported rice mainly from Thailand, Vietnam and China [8]. Raising level of national rice self-sufficiency has become a strategic issue in the agricultural ministry of Malaysia. In this study, time series modeling of rice yield data in Muda Irrigation area from 1995 to 2003 are presented. However, there are missing value exist in these data with 27 observations in 2003 are not recorded. Hence, to model time series from 1995 to 2003, the unrecorded data need to be predicted. Models are developed to solve the problem and fill the missing values. The ability to predict the future enables the farm managers to take the most appropriate decision in anticipation of that future.

Thus, various kinds of forecasting models have been developed and researchers have relied on statistical techniques to predict rice yields. The accuracy of time series

[1&2] Department of Software Engineering, Faculty of Computer Science and Information System, University Technology of Malaysia

[3] Department of Mathematic, Science Faculty, University Technology of Malaysia

forecasting is fundamental to many decisions processes and hence the research for improving the effectiveness of forecasting models has never been stopped. Of the various forecasting models, the ARIMA model is one of the most popular models in traditional time-series forecasting and is often used as a benchmark model to compare with other models [12]. The popularity of the ARIMA model is due to its statistical properties as well as the well-known Box-Jenkins methodology in the model building process. Preciado, *et al.* used ARIMA model to forecast the missing value for fisheries historical series. However, the ARIMA model is only a class of linear model and thus it can only capture linear feature of data time series. But many data time series are often full of nonlinearity and irregularity.

Recent research activities in artificial neural network (ANN) have shown powerful pattern classification and pattern recognition capabilities. One major application area of ANN is forecasting [9]. ANN provides an attractive alternative tool for both forecasting researchers and has shown their nonlinear modeling capability in data time series forecasting. ARIMA models and ANN are often compared with mixed conclusions in terms of superiority in forecasting performance. Although the ANN model achieves success in many time series forecasting, they have some disadvantages. Since the real world is highly complex, there exits some linear and nonlinear patterns in the time series simultaneously. It is not sufficient to use only a nonlinear model for time series because the nonlinear model might miss some linear features of time series data [13].

In the real world, rice yield time series are rarely linear or nonlinear and are often seasonal. Hence, it is hypothesized that by combining time series models, such as ARIMA with ANN models in the rice yields can be characterized more accurately. Thus, the objective of this paper is to develop a hybrid model for rice yield forecasting. This model combines a time series linear model (ARIMA) and nonlinear ANN model. This is because the ANN model and ARIMA model are complementary. The hybrid model is then compared with individual models considering the forecast accuracy for 27 observations in 2003 are not to be recorded.

2.0 LINEAR MODEL

The Box-Jenkins, linear model was developed by Box and Jenkins [1] almost three decades ago, and has been one of the most popular approaches to forecasting. The general ARIMA models are compound of a seasonal and non-seasonal part are represented by the following way:

$$\text{ARIMA}\ (p, d, q)\ (P, D, Q)_s \tag{1}$$

where (p, d, q) nonseasonal part of the model and (P, D, Q) seasonal part of the model. This model can be written as

$$\phi_p(B)\Phi_p(B^s)\ \nabla^d\ \nabla_s^D z_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{2}$$

where $\phi(B)$ and $\theta(B)$ are polynomials of order p and q, respectively; $\Phi(B^s)$ and $\Theta(B^s)$ are polynomials in B^s of degrees P and Q, respectively; p order of nonseasonal autoregression; d number of regular differencing; q order of the nonseasonal moving average; P order of seasonal autoregression; D number of seasonal differencing; Q order of seasonal moving average; and s length of season. Their approach for a building a time series model is a method of finding, for a model (ARIMA) that adequately represents a data process for a given a set data. The approach divides the forecasting problem into three stages: identification, estimation and diagnostic checking. The identification stage involved transforming the data if necessary to improve the normality and the stationary time series. The next step is choosing the suitable model by analyzing both the autocorrelation (ACF) and partial autocorrelation function (PACF) of the stationary series. Once a model is identified, the parameters of the model are estimated. It is necessary to check if the assumptions are satisfied. Diagnostic checking using the ACF and PACF of residuals was carried out, which can be referred to Brockwell & Davis [2]. The forecasting model was then used to compute the fitted values and forecasts values.

3.0 THE NEURAL NETWORK FORECASTING MODEL

The greatest advantage of a neural network is its ability to model complex nonlinear in the data series. The ANN model performs a nonlinear functional mapping from the input observations $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ to the output value (y_t), i.e.,

$$y_t = a_0 + \sum_{j=1}^{q} a_j f\left(w_{0j} + \sum_{i=1}^{p} w_{ij} y_{t-1}\right) + \varepsilon_t \tag{3}$$

where a_j $(j = 0, 1, 2, \ldots, q)$ is a bias on the jth unit, and w_{ij} $(i = 0, 1, 2, \ldots, p; j = 0, 1, 2, \ldots, q)$ is the connection weights between layers of the model, $f(\bullet)$ is the transfer function of the hidden layer, p is the number of input nodes and q is the number of hidden nodes [5].

Training a network is an essential factor for the success of the neural networks. Among the several learning algorithms available, back-propagation has been the most popular and most widely implemented learning algorithm for all neural network paradigms [13]. In this paper algorithm of back-propagation is used in the following experiment.

Actually, the ANN model in (3) performs a nonlinear functional mapping from the past observation $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ to the future value (y_t) , i.e.,

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, w) + \varepsilon_i \tag{4}$$

where w is a vector of all parameters and f is a function determined by the network structure and connection weights. Thus, in some senses, the ANN model is equivalent to a nonlinear autoregressive (NAR) model. A major advantage of neural networks is their ability to provide flexible nonlinear mapping between inputs and outputs. They can capture the nonlinear characteristics of time series well.

4.0 THE HYBRID FORECASTING METHODOLOGY

In real life, data time series forecasting is far from simple due to high volatility, complexity, irregularity and noisy. Furthermore, real-world time series are rarely pure linear or nonlinear. They often contain both linear and nonlinear patterns. If this is the case, there is no universal model that is suitable for all kinds of time series data. Both ARIMA models and ANN models have achieved success in their own linear or nonlinear domains, but neither ARIMA nor ANN can adequately model and predict time series since the linear models cannot deal with nonlinear relationships while the ANN model alone is not able to handle both linear and nonlinear patterns equally well [12]. On the other hand, as previously mentioned, for time series forecasting the relationship between ARIMA and ANN is complementary. ARIMA model is a class of linear models that can capture time series' linear characteristics, while ANN models are a class of general function approximator capable of modeling non-linearity and which can capture nonlinear patterns in time series. Hybridizing the two models may yield a robust method, and more satisfactory forecasting results may be obtained by incorporating a ARIMA model and a ANN model.

Thus a hybrid model consists of an ARIMA model for the linear part of the time series data and an ANN model for the nonlinear part can be defined as

$$Y_t = L_t + N_t \tag{5}$$

where Y_t is original time series, Y_t = linear part based on the ARIMA model, and N_t = nonlinear part based on the neural network model.

In this study a hybrid methodology was applied to forecast rice yield series. First, ARIMA models were applied to rice yield series using the method of Box and Jenkins [1]. Second, a residual series obtained after fitting an ARIMA model to the actual data was obtained. Third, a neural network model was fitted based on the back-propagation algorithm to the residuals. It is considered that an ARIMA model cannot capture the nonlinear structure of the data, and its residuals will contain information about non-linearity. After fitting an ARIMA model, the residuals are obtained, and since the residuals are nonlinear in nature, an ANN is found to be useful for fitting the residuals. The results from an ANN are used for prediction of the error term for the ARIMA model.

Let e_t denote the residual at time t from the linear model, then

$$e_t = y_t - \hat{L}_t \tag{6}$$

where $\hat{L}_t$ is the forecast value for time t from the estimated relationship. The ANN model for the residuals will be

$$e_t = f\left(e_{t-1}, e_{t-2}, \ldots, e_{t-n}\right) + \varepsilon_t \tag{7}$$

where f is a nonlinear function determined by the neural network and ε_t is the random error. Therefore, the correct model identification is critical. Denote the forecast from (7) as $\hat{N}_t$, the combined forecast will be

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \tag{8}$$

In summary, the proposed methodology of the hybrid system consists of two steps. In the first step, an ARIMA model is used to analyze the linear part of the problem. In the second step, a neural network model is developed to model the residuals from the ARIMA model. Since the ARIMA model cannot capture the nonlinear structure of the data, the residuals of linear model will contain information about the nonlinearity.

5.0 EMPIRICAL RESULTS

5.1 Data Sets

The data is collected from Muda Agricultural Development Authority (MUDA) Kedah, Malaysia ranging from 1995 to 2003. There are 4 areas with 27 locations.

There are two types of season symptom that influenced the rice yield in Malaysia. The rice yield series data is used in this study to demonstrate the effectiveness of the hybrid method. These time series come from different location and have different statistical characteristics. The rice yields data contains the yields data from 1995 to 2003, giving a total of 432 observations, 27 observations are missing. The time series plot is given in Figure 1, suggests that there is a non-stationary pattern with one outlier was observed.

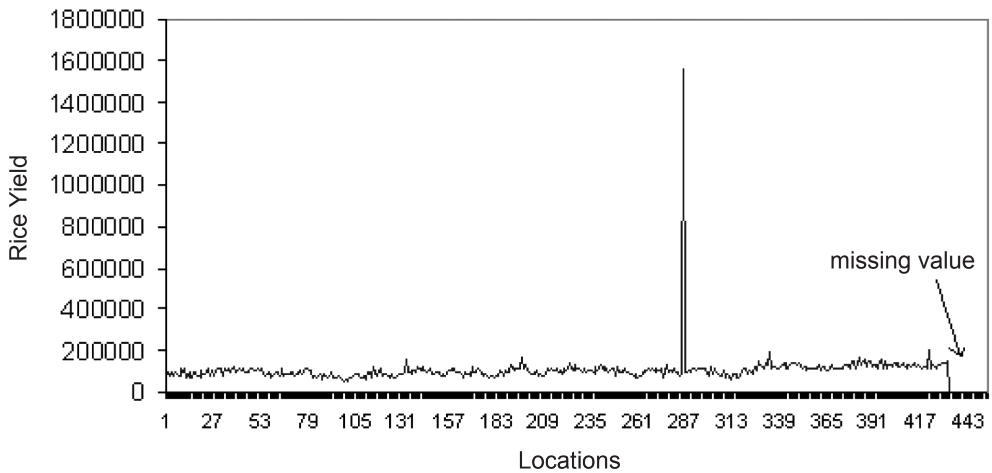

Figure 1 Rice yield data series (1995- 2003)

5.2 Fitting ARIMA Models to the Data

To assess the forecasting performance of different models, each data set is divided into two samples. The first series was used for training the network (modeling the time series) and the remaining is used for testing the performance of the trained network (forecasting). The data is collected from 1995 to first season in 2003 producing 372 observations for training purpose.The remainder as the output sample data set with 27 observations for testing purpose. The plots in Figure 1 shows the plots of the rice yield time series indicate that the time series are non-stationary in mean and variance. The natural logarithm is taken to reduce the variance, and then the first difference was applied in order to remove the trend. Figure 2 shows the natural logarithmic (to the base 10) transformed data that are used in the modeling and forecasting analysis.

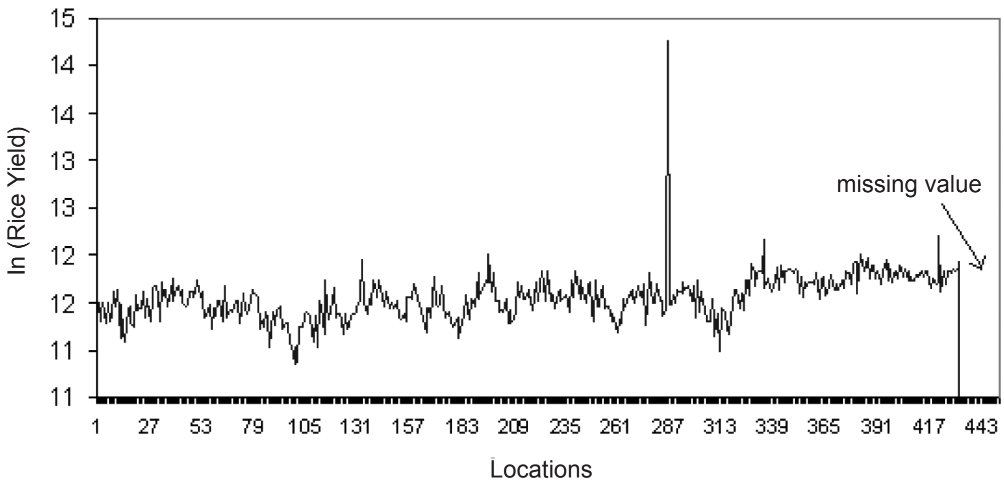

Figure 2 The natural logarithms of rice yield data series

The sample AutoCorrelation Function (ACF) and sample Partial AutoCorrelation Function (PACF) for the transformed series are plotted in Figure 3 and 4, respectively. The plot shows that there is seasonality in the series. We find the first-difference series becomes stationary. For ACF, there are major spikes at lags 1 and 27, while for PACF, we observe major spikes at lags 1, 2, 3, 4 and 27.

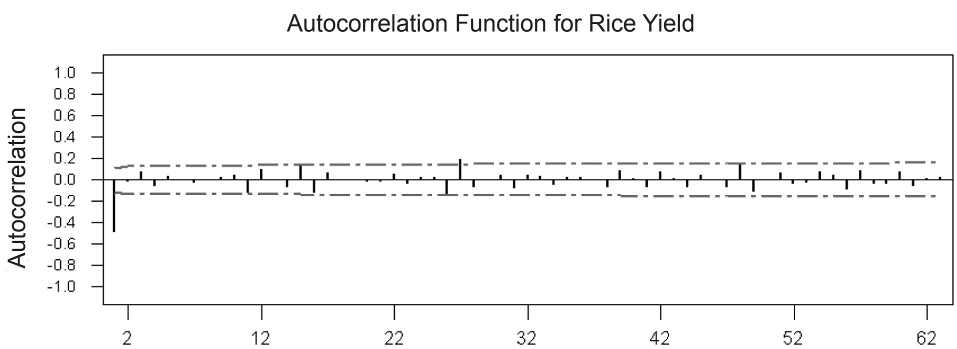

Figure 3 Sample PACF for the differenced series of natural logarithm

The performance of all models developed in this study are evaluated using wide variety of standard statistical performance evaluation measures: mean square error (MSE) and mean absolute relative error (MARE). MSE are used as the primary performance measure and MARE as the supplementary measure in evaluating the modeling ability as well as the forecasting ability.

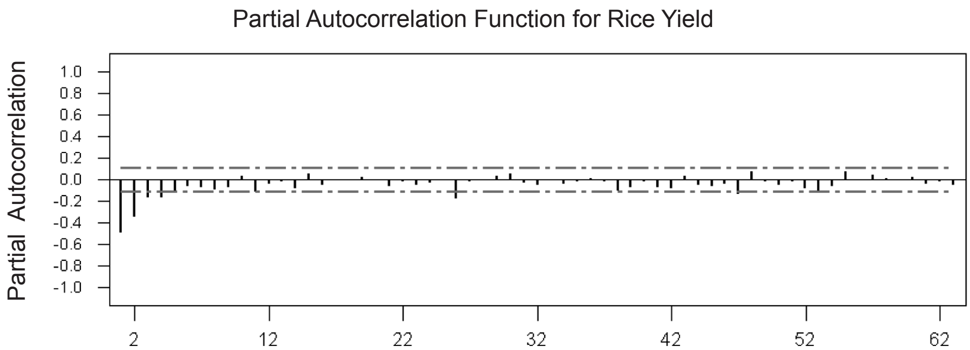

Figure 4 Sample PACF for the differenced series of natural logarithm

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2,$$

$$\text{MARE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|.$$

Several ARIMA models are identified, and the statistical results are compared in the following Table 1. The criterions to judge for the best model are relatively small of MSE and relatively small of MARE. Therefore, ARIMA (0, 1, 1) × $(1, 0, 0)_{27}$ model is reached as the optimum model from this class.

Table 1 Comparisons of ARIMA models statistics results

ARIMA MODEL	MSE	MARE
(4,1,0) × (1,0,0)	0.0475	0.0106
(4,1,1) × (1,0,1)	0.0459	0.0097
(0,1,1) × (0,0,1)	0.0464	0.1071
(0,1,1) × (1,0,1)	**0.03409**	**0.0093**

The ACF and PACF plots of residuals for the rice yields series are shown in Figure 5. From the residual plot of the best ARIMA model, it was observed that the selected ARIMA model passed the diagnostic checks and they were all white noise.

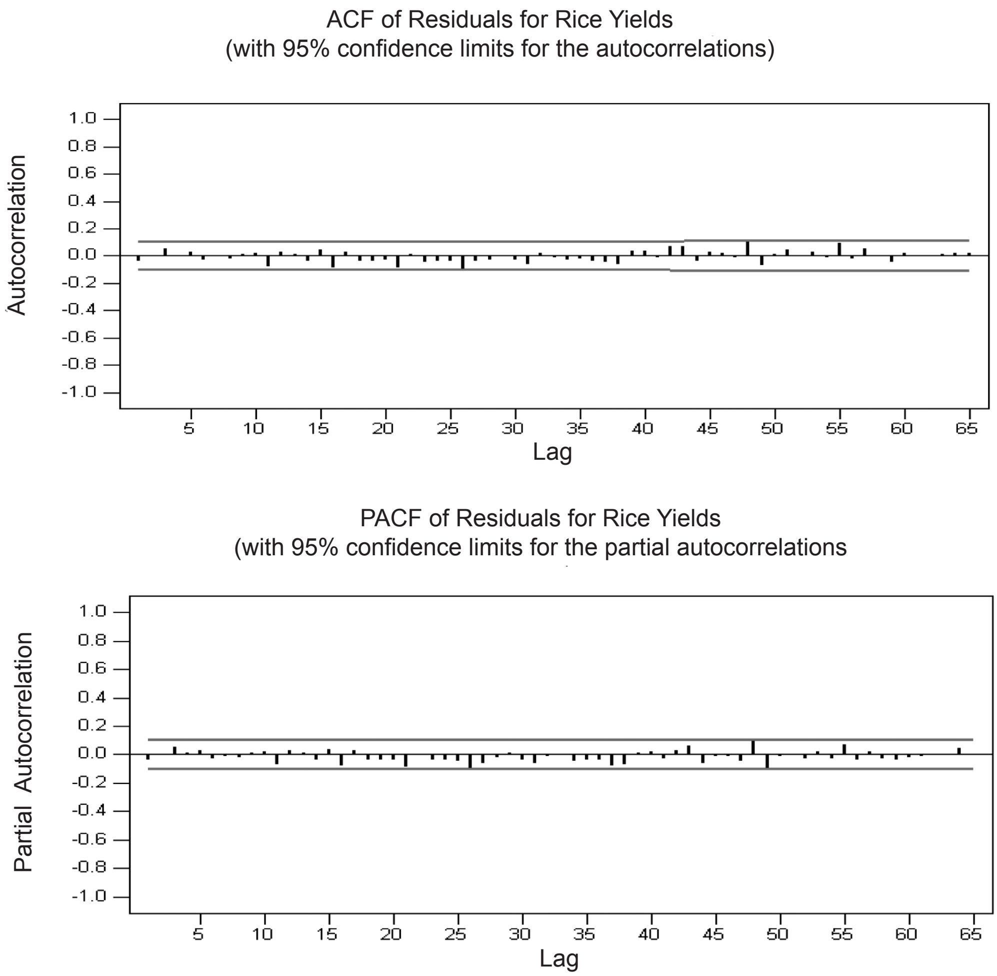

Figure 5 ACF and PACF of residuals for rice yields

5.3 Fitting Neural Network Models to the Data

In this investigation, we only consider the situation of one-step-ahead forecasting with 27 observations. Before the training process begins, data normalization is often performed. The linear transformation formula to [0, 1] is used

$$x_n = \frac{x_0 - x_{\min}}{x_{\max} - x_{\min}}$$

where x_n and x_0 represent the normalized and original data; and $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values among the original data. In order to conform the neural network used in the forecast, ACF and PACF were used to determine the maximum number of input neurons used during the training [3]. The input nodes are 3, 9, 18 and 27.

The hidden layer plays very important roles for many successful applications of neural network. The hidden layer allow neural network to detect the feature, to capture the pattern in the data, and to perform complicated nonlinear mapping between input and output variables. The most common way in determining the number of hidden nodes is via experiments or by trial-and-error. In this study, for each input layer, the number of hidden nodes were determined using formula "$I/2$" [4], "I"[10], "$2I$" [11] and "$2I+1$"[6], where I corresponding input neurons. The sigmoid activation function was used as the transfer function at both hidden and output layers.

The network was trained for 5000 epochs using the back-propagation algorithm with a learning rate of 0.001 and a momentum coefficient of 0.9. Table 2 list overall summary statistics for each ANN model results for the rice yield.

Table 2 Comparisons of various ANN models together with the performance statistics for the rice yields

Input	Hidden	MSE	MARE	Input	Hidden	MSE	MARE
3	1	0.6280	79.5213	18	9	0.3950	77.7039
	3	0.5311	78.8769		18	0.3423	77.2194
	6	0.3767	77.6057		36	0.4136	77.8110
	7	0.6125	79.4660		37	0.3105	76.7369
9	5	0.2078	75.6454	27	14	0.3318	76.9951
	9	0.4720	78.2811		27	0.2507	75.9907
	18	0.4159	77.8301		54	0.2521	76.1632
	19	0.2939	76.5933		**55**	**0.1450**	**74.2551**

Note: The data in boldface means the prediction performance of NN is better than those of others, respectively

The results in terms of MSE and MARE statistics for all the ANN models are presented in Table 3. Analyzing the results during training, it can be observed that the structure of ANN (27, 55, 1) gives slightly better forecasts the others of ANN.

Next, comparison was made between ARIMA model, neural network and hybrid model for 27 leads times. The performance of different model also was evaluated using MSE and MARE. In this study, all ARIMA modeling is implemented via MINITAB software while the ANN model is built with Neural Network Toolbox of MATLAB 14. Only the one-step-ahead forecasting with 27 observations were considered. A subset ARIMA (0, 1, 1) × $(1, 0, 1)_{27}$ has been found to be the most parsimonious among all ARIMA models that are also found adequate judged by the residual analysis. A neural network architecture of (27, 55, 1) is used to model the nonlinear patterns. Finally, by combining the ARIMA (0, 1, 1) × $(1, 0, 1)_{27}$ and neural network architecture of (27, 55, 1) hybrid models were obtained. The ARIMA, neural network and hybrid models for rice yields are discussed here. Table 3 gives the training and forecasting results for the rice yield data. The comparison between actual and predicted values is given in Figure 6.

Table 3 Comparison between different techniques over the whole 27 forecasted periods

	Performance Measure	ANN	ARIMA	Hybrid
Training	MSE	0.0576	0.0341	0.0322
	MARE	0.0138	0.0093	0.0092
Forecasting	MSE	0.0237	0.0125	0.0102
	MARE	0.0110	0.0072	0.0059

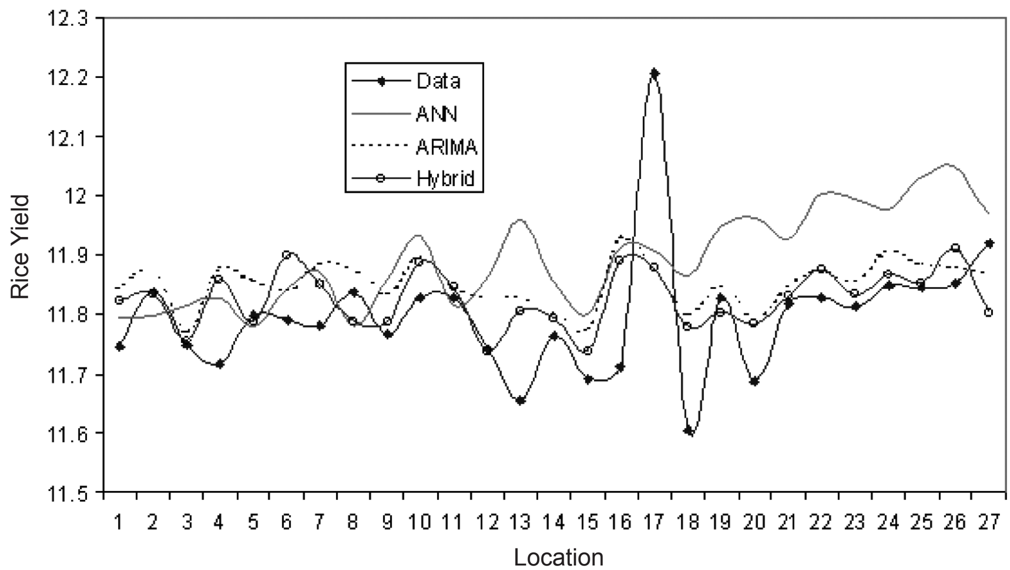

Figure 6 Comparisons of the prediction made with ARIMA, ANN, Hybrid and data of rice yields

It is observed that the best forecasting result in term of MSE and MARE, the performance of hybrid model was better than ARIMA and ANN models. This shows that the performance improves for hybrid models.

6.0 CONCLUSIONS

This study compares the performance of ANN model, the statistical model (ARIMA) and the hybrid model in forecasting the missing values of the rice yields of Malaysia. Results show the ANN model forecasts are considerably less accurate than the traditional ARIMA model which used as a benchmark. On the other hand, the hybrid model using the ARIMA model and the error of NN model is an effective way to improve the forecasting performance in error measures in term of MSE and MAPE. The hybrid model takes advantage of the unique strength of ARIMA and ANN in linear and nonlinear modeling. For complex problems that have both linear and nonlinear correlation structures, the combination method can be an effective way to improve forecasting performance. The empirical results with the rice yields data set clearly suggest that the hybrid model perform better than other two models explored in forecasting the missing values of the rice yields.

REFERENCES

[1] Box, G. E. P. & Jenkins, G. 1970. *Time Series Analysis. Forecasting and Control*. Holden-Day. San Francisco: CA.

[2] Brockwell, P. J. & Davis, R. A. 2002. *Introduction to Time Series and Forecasting*. Berlin: Springer.

[3] Cadenas, E. & Rivera, W. 2007. Wind Speed Forecasting in the South Coast of Oaxaca, Mexico. *Renewable Energy*. 32: 2116–2128.

[4] Kang, S. 1991. An investigation of the Use of Feedforward Neural Network for Forecasting. Ph.D. Thesis, Kent State University.

[5] Lai, K. K., Yu, L., Wang, S. and Huang, W. 2006. Hybridizing Exponential Smoothing and Neural Network for Financial Time Series Predication. ICCS 2006, Part IV, LNCS. 3994: 493–500.

[6] Lippmann, R. P. 1987. An Introduction to Computing with Neural Nets. IEEE ASSP Magazine, April, 4–22.

[7] Preciado, I., Punzon, A, Gallego, J. L. and Vila, Y. 2006. Using Time Series Methods for Completing Fisheries Historical Series. Bol. Inst. Esp. Oceanografia. 22(1–4): 83–90.

[8] Saad, P., Bakri, A., Kamarudin, S. S. and Jaafar, M. N. 2006. Intelligent Decision Support System for Rice Yield Prediction in Precision Farming. IRPA Report.

[9] Sharda, R. 1994. Neural Networks for the MS/OR analyst: An Application Bibliography. Interfaces 24(2): 116–30.

[10] Tang, Z. and Fishwick, P. A. 1993. Feedforward Neural Nets as Models for Time Series Forecasting. ORSA Journal on Computing. 5(4): 374–385.

[11] Wong, F. S. 1991. Time Series Forecasting Using Backpropagation Neural Network. Neurocomputing, 2: 147–159.

[12] Zhang, G. P. 2003. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. Neurocomputing 50: 159–175.

[13] Zou, H. F., Xia, G. P., Yang, F. T. and Wang, H. Y. 2007. An Investigation and Comparison of Artificial Neural Network and Time Series Models for Chinese Food Grain Price Forecasting. *Neurocomputing.* 70: 2913–2923.