Int.J.Curr.Microbiol.App.Sci (2021) *10*(01): 3421-3428

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 10 Number 01 (2021)**
Journal homepage: http://www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2021.1001.403

A Study on the Performance of the ARIMAX-ANN Hybrid Forecasting Model over the other Time Series Forecasting Models ARIMAX and ANN in Forecasting the Rice Yield

K. Supriya*

Department of Statistics & Mathematics, College of Agriculture, Rajendranagar, Hyderabad, India

**Corresponding author*

Keywords

ARIMA, ARIMAX, Forecasting and Production

Article Info

Accepted:
20 December 2020
Available Online:
10 January 2021

A B S T R A C T

Rice being the staple food for more than 50% of the world population, it is very essential to forecast the production of rice so as to meet the need of the rapidly growing population. Forecasting is also essential for better planning and decision making. Many forecasting techniques have evolved and it is the matter of prediction accuracy. In this study, the performance of ARIMAX-ANN hybrid forecasting model is compared with two other time series forecasting models, Autoregressive Integrated moving average with exogenous variables (ARIMAX) and Artificial Neural Networks (ANN) in forecasting the rice yield during both kharif and rabi seasons of Telangana state. The exogenous variables used in the study are percentage of dead hearts and percentage of white ears which are the damage symptoms of rice yield due to yellow stem borer (*Scirpophaga incertulas*). To compare the effectiveness of these three models 27 years rice yield data of both kharif and rabi seasons pertaining telangana state was used i.e., from 1990-2016 (both years inclusive). The results showed that ARIMAX-ANN hybrid model performed reasonably well compared to the other models i.e., Autoregressive integrated moving average model (ARIMA) and Autoregressive integrated moving average model with exogenous variables (ARIMAX).

Introduction

Rice (*Oryza sativa* L.) is the most important cereal crop of the world both in respect to area and production. It is the important staple food for more than 50% of the world population and provides 60-70 per cent body caloric intake to the consumers. Asia is the largest producer and consumer of rice in the entire world. The total Rice production in the world is 501.96 million metric tonnes as estimated by the United states Department of Agriculture in november 2020 (USDA). India ranks second in rice production in the world with the production of 118.0 million metric tonnes where as China ranks first with 149.0 million metric tonnes (Statistica, the statistical portal, 2020). India is a developing country with limited input requirements, soil-enriching properties and suitability for growing in areas, rice occupies a unique place in our agriculture system. Rice finds a

prominent place in Indian meals and remains a primary source of nutrition for the majority of population of our country.

Telangana State is the newly formed state in India bifurcated from Andhra Pradesh during June 2nd 2014. It borders Maharashtra on North West, Karnataka on West and Rayalaseema region of Andhra Pradesh state on south. The region has an area of 114.84 lakh ha and a population of 352.87 lakhs as per 2011 census. It is 12th largest state in the country. It has 31 districts. The Krishna and Godavari rivers flow through the state from West to East. Agriculture in Telangana is dependent on rainfall and agricultural production depends upon the distribution of rainfall. The influence of South-West monsoon is predominant. South-West Monsoon (79%) is spread over the period from June to September, North-East Monsoon (14%) from October to December and the rest 7% rainfall is received during the winter and summer months. Telangana (31 districts) receive a normal rainfall of 906.6 mm in a year.

National rice self sufficiency has become a strategic issue and the ability to forecast the future enables the farm managers to take the most appropriate decision in anticipation of the future. Many forecasting techniques have evolved but accuracy of the time series forecasting is fundamental to many decision processes. One of the most popular and commonly used models for the forecasting research and practice is the autoregressive integrated moving average (ARIMA) model. In the ARIMA models, the desired forecasting is generally expressed as a linear combination. But real world time series are often full of nonlinearity and are influenced by many exogenuous variables.. Hence, it is necessary to consider the effect of these exogenous variables which is taken care by the ARIMAX model. When the data is noisy exhibiting a non-linear pattern then machine learning techniques such as Artificial Neural Network is used. In this paper, a comparative study of the forecasting models ARIMAX-ANN hybrid model has been done with ARIMAX and ANN and it was found that the Hybrid model outperformed all other forecasting models.

Materials and Methods

The main purpose of this study is to investigate the forecasting ability of the three forecasting models i.e., ARIMAX-ANN hybrid model, Autoregressive integrated moving average with exogenous variables (ARIMAX) and Artificial Neural network (ANN) to determine which is the model performs better. For this study, the data pertaining to the rice yield for both kharif and rabi seasons pertaining to the Telangana state, has been taken for the past 27 years i.e., from 1990-2016 and the exogenous variables are damage caused by yellow stem borer (*Scirpophaga incertulas*) which is expressed in terms of percentage of dead hearts and percentage of white ears which are the damage symptoms caused by ysb. The above said secondary data has been taken from the annual progress reports of AICRP reports, ICAR- Indian Institute of Rice Research, Rajendranagar, Hyderabad.

Auto Regressive Integrated Moving Average (ARIMA)

ARIMA model has been one of the most popular approaches to forecasting. The ARIMA model is basically a data-oriented approach that is adapted from the structure of the data themselves. An auto-regressive integrated moving average (ARIMA) process combines three different processes namely an autoregressive (AR) function regressed on past values of the process, moving average (MA) function regressed on a purely random

errors and an integrated (I) part to make the data series stationary by differencing. In an ARIMA model, the future value of a variable is supposed to be a linear combination of past values and past errors. Generally, a non seasonal ARIMA model, denoted as ARIMA (p,d,q), is expressed as

$$Y_t = F_0 + F_1 Y_{t-1} + F_2 Y_{t-2} + F_2 Y_{t-3} + \ldots + F_p Y_{t-p} + e_t - G_1 e_{t-1} - G_2 e_{t-2} \ldots\ldots - G_q e_{t-q}$$

Where Y_{t-i} and e_t are the actual values and random error at time t respectively. F_i (i = 1,2,…p) and Gj (j = 1,2,…,q) are the model parameters. Here 'p' is the number of autoregressive terms, 'd' is the number of non seasonal differences and 'q' is the number of lagged forecast errors. Random errors e_t are assumed to be independently and identically distributed with mean zero and the common variance σ_e^2.

Basically, this method has three phases:

Model Identification
Parameter estimation and
Diagnostic Checking.

The auto-regressive integrated moving average (ARIMA) model deals with the non-stationary linear component. However, any significant nonlinear data set limit the ARIMA.

Autoregressive Integrated moving Average with Exogenous variables (ARIMAX) model

Autoregressive integrated moving average with exogenous variable (ARIMAX) is the generalization of ARIMA (Autoregressive Integrated moving average) models. Simply an ARIMAX model is like a multiple regression model with one or more autoregressive terms and one or more moving average terms. This model is capable of incorporating an external input variable. Identifying a suitable ARIMA model for endogenous variable is the first step for building an ARIMAX model. Testing of stationarity of exogenous variables is the next step. Then transformed exogenous variable is added to the ARIMA model in the next step. (Bierens 1987).

An ARIMA model is usually stated as ARIMA (p,d,q), where 'p' stands for the order of autoregressive process (Box and Jenkins, 1970). The general form of the ARIMA (p,d,q) can be written as

$$\Delta^d Y_t = \delta + \theta_1 \Delta^d Y_{t-1} + \theta_2 \Delta^d Y_{t-2} + \cdots\cdots\cdots\cdots + \theta_p Y_{t-p} + e_{t-1}\alpha_1 - \alpha_2 e_{t-2}\alpha_q e_{t-2}$$

Where as Δ^d gives the differencing of order d i.e., $\Delta = y_t - y_{t-1}$ and $\Delta^2 = \Delta y_t - \Delta y_{t-1}$

In Arimax model we just add exogenous variable on the right hand side

$$\Delta^d Y_t = \delta + \beta X_t + \theta_1 \Delta^d Y_{t-1} + \theta_2 \Delta^d Y_{t-2} + \cdots\cdots\cdots\cdots + \theta_p Y_{t-p} + e_{t-1}\alpha_1 - \alpha_2 e_{t-2}\alpha_q e_{t-2}$$

Where X_t is the exogenous variable and β is the coefficient.

Forecasting model

In the present study, the models have been developed on the basis of the secondary data of past 27 years for Rajendranagar region. The model is applicable for the areas having agro climatic conditions similar to Rajendranagar region. The data on the best check varieties have been used to nullify varietal differences. This is the standard practice while using the Time series data analysis.

Data pertaining to the rice yield, damage data caused by yellow stem borer which is in the form of percentage of dead hearts and

percentage of white ears for the past 27 years is divided into two groups, they are training data and testing data. The training data is a set of data that will be used to perform analysis and determine the model. The testing data is a set of data that will be used to test the accuracy of the forecast results. Hence out of 27 years, 24 years data is taken as training data and 3 years data is taken for testing data. The data was analyzed using the software SPSS 20.

The ARIMA Model

Arima modeling process begins with Estimation ARIMA model. This estimation includes degree of autoregressive (p), differencing (d), and moving average (q) and seasonality factors. This data is tested for stationarity using sequence plots, because this autocorrelations approach zero exponentially after the second or third time lag. So degree of d for Arima models can be decided accordingly. After we find degree of d then we have to find degree of p and q to make ARIMA model. Prediction degree of p and q can be seen from the correlogram plot of autocorrelation function (ACF) and partial autocorrelation function (PACF).

The ARIMAX Model

Arimax model requires independent variables that acts as an additional variable. Independent variables that are used in this study are the percentage of dead hearts and percentage of white ears which are the damage symptoms caused by yellow stem borer (*Scirpophaga incertulas)* during different stages of rice crop.

Artificial neural networks

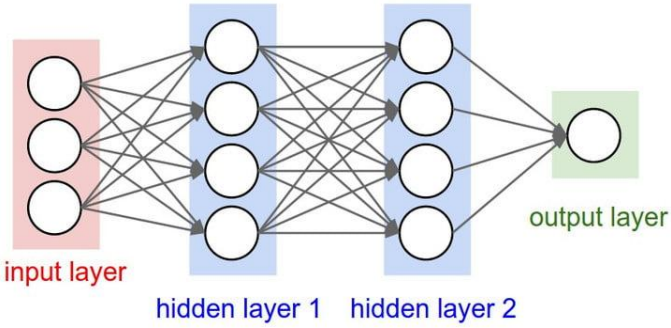
An Artificial neural network is a computer system that simulates the learning process of human brain. The greatest advantage of Neural networks is its ability to model nonlinear complex data series. The basic architecture consists of three types of neuron layers: input, output and hidden layers. The ANN model performs a nonlinear functional mapping from the input observations (y_{t-1}, y_{t-2}, y_{t-3}, ……..y_{t-p}) to the output value y_t.

$$Zyt = a_0+\Sigma a_j\, f(W_{oj}+\Sigma\, W_{ij}\, y_{t-1}) +e_i \qquad \textbf{(4.1)}$$

Where a_j (j=0,1,2,3,….. q) is the bias on the j^{th} unit and W_{ij} (i=0,1,2,……p, j=0,1,2,…….q) is the connection weights between layers of the model, f(.) is the transfer function of the hidden layer, p is the number of input nodes and q is the number of hidden nodes (Lai et. al., 2006). The activity function utilized for the neurons of the hidden layer was the logistic sigmoid function that is described by

$$f(x) = 1/1+e^{-x} \qquad (4.2)$$

This function belongs to the class of sigmoid functions which has advantages characteristics such as being continuous, differentiable at all points and monotonically increasing.



ARIMAX-ANN Hybrid Model

When the time series data contains both linear and non-linear components a hybrid approach (proposed by Zhang (2003) decomposes the time-series data into its linear and non-linear component. The hybrid model considers the

time series yt as a combination of both linear and nonlinear components.

That is yt = Lt+Nt+et (1)

Where Lt is the linear component present in the given data and Nt is the nonlinear component.

These two components are to be estimated from the data. The hybrid method of ARIMAX and ANN has the following steps.

First, a linear time-series model, ARIMAX is fitted to the data. At the next step residuals are obtained from the fitted linear model. The residuals will contain only the nonlinear components. Let et denotes the residual at the time t from the linear model, then

et = yt – Lt (2)

where Lt is the forecast value for the time t from the estimated linear model.

Diagnosis of residuals is done to check if there is still linear correlation structures left in the residuals then further we will go for nonlinearity check. The residuals are tested for nonlinearity by using BDS test. Once the presence of the nonlinearity is conformed in the residuals then the residuals modelled using a nonlinear model ANN. Finally the forecasted linear (ARIMAX) and nonlinear (ANN) components are combined to obtain the aggregated forecast values as

Yt = Lt+Nt (3)

Bayesian Information criteria (BIC)

It is a criterion for model selection among a finite set of models and is based on likelihood function. In case of model fitting it is possible to increase the likelihood by adding parameter, which may results in over fitting. BIC resolve this problem by introducing penalty term for the number of parameters in the model.

$$BIC = -2*\log(L) + m*\log(n)$$

Where, *L*: Likelihood of the data with a certain model

n: Number of observations
m: Number of parameters in the model

Root Mean squared error (RMSE)

It is or the estimated white noise standard deviation in ARIMA analysis. It is expressed as: square root of mean squared error and is also known as standard error of estimate in regression analysis

$$RMSE = (1/T)\sqrt{(\Sigma(P_t - A_t)^2)}$$

Where,

P_t: Predicted value for time t
A_t: Actual value at time t and
T: Number of predictions.

In conclusion, it is seen from the comparison of RMSE and R^2 values that the ARIMAX-ANN Hybrid forecasting model has the least values of RMSE values and highest values of R^2 than ARIMAX and ANN. Therefore ARIMAX-ANN hybrid model has outperformed the other two forecasting models.

Table.1 Zone wise performances of Forecasting models and forecasted values for rice Yield (kg/ha)

Zone	Model Accuracy and forecasted value	ARIMAX	ANN	ARIMAX-ANN
Southern Telangana Zone	**Kharif season**			
	2016*	5495.00	5495.00	**5495.00**
	2017	4723.21	4635.21	**4851.33**
	2018	4312.35	4631.02	**4652.67**
	2019	4021.32	4140.34	**4569.28**
	2020	4021.26	4139.23	**4568.12**
	2021	4021.13	4138.12	**4081.22**
	2022	4021.02	4137.09	**4081.21**
	RMSE	723.36	867.59	396.58
	R^2	0.26	0.32	0.69
	Rabi season			
	2016*	5485.00	5485.00	**5485.00**
	2017	4825.12	4926.33	**5065.85**
	2018	4523.25	4913.22	**4826.32**
	2019	4496.01	4866.62	**4824.64**
	2020	4495.32	4866.35	**4826.53**
	2021	4494.12	4866.33	**4326.35**
	2022	4494.01	4865.23	**4326.12**
	RMSE	907.31	933.16	700.01
	R^2	0.12	0.63	0.59
Central Telangana Zone	**Kharif season**			
	2016*	4995.00	4995.00	**4995.00**
	2017	4513.02	4478.89	**4607.18**
	2018	4478.36	4489.15	**4672.13**
	2019	4443.22	4498.67	**4596.36**
	2020	4399.54	4507.47	**4544.25**
	2021	4375.88	4515.60	**4526.33**
	2022	4374027	4523.11	**4525.12**
	RMSE	1156.837	724.96	615.682
	R^2	0.26	0.79	0.85
	Rabi season			
	2016*	5096.00	5096.00	**5096.00**
	2017	5028.36	5158.14	**5057.19**
	2018	5035.90	5163.07	**5055.17**
	2019	5043.11	5167.62	**5053.22**
	2020	5045.19	5171.84	**5042.31**
	2021	5046.27	5175.78	**5033.21**
	2022	5047.33	5179.49	**5028.24**

	RMSE	1440.65	857.00	706.14
	R^2	0.15	0.72	0.89
Northern Telangana Zone	**Kharif season**			
	2016*	5982.00	5982.00	**5982.00**
	2017	5799.21	6809.32	**5888.20**
	2018	5786.06	6812.77	**5891.21**
	2019	5721.42	6818.69	**5826.58**
	2020	5718.63	6822.91	**5823.79**
	2021	5711.90	6825.91	**5817.06**
	2022	5707.83	6828.04	**5813.00**
	RMSE	975.63	724.58	660.56
	R^2	0.63	0.83	0.95
	Rabi season			
	2016*	6111.00	6111.00	**6111.00**
	2017	6156.31	6877.51	**6365.27**
	2018	6018.37	6878.03	**6311.32**
	2019	6148.84	6882.34	**6441.80**
	2020	6250.21	6885.40	**6543.18**
	2021	6357.28	6887.58	**6650.25**
	2022	6464.24	6889.13	**6757.22**
	RMSE	911.40	418.69	237.71
	R^2	0.47	0.83	0.92

References

Box G.E.P. and Jenkins G. (1970). Time series analysis, Forecasting and control, Holden-Day, San Francisco, CA.

Wiwik Anggraeni, Retno Aulia Vinarti, Yuni Dwi Kurnia wati; (2015). Performance comparison between Arima and Arimax method in Moslem kids clothes demand forecasting: case study; *procedia computer science*, 72(2015):630-637.

Valipour, Mohammad; (2012). Parameters Estimate of Autoregressive Moving Average and Autoregressive Integrated Moving Average Models and compare their ability for inflow forecasting, *Journal of Mathematics and Statistics*, 8 (3): 330-338.

Herui cui and zu peng; (2015). Short term city electric load forecasting with considering temperature effects, An improved Arimax model; *Hindawi publishing corporation*, Article ID 589374, 10 pages.

Suleman Nasiru, Albert luguterah, Lea Anzagra;(2013). The efficacy of Arimax and Sarima models in predicting monthly currency in circulation in Ghana, Mathematical theory and modeling. ISSN 2224-5804. Vol(3), No.5.

Wiwik Anggraeni, Kuntoro Boga Andri, Sumaryanto, Faizal mahananto; (2017). The performance of Arimax model with VAR model in forecasting strategic commodity price in Indonesia. Procedia Computer Science, 124:189-196.

Kumari Prity, Mishra G.C., Anil kumar pant, Garima shukla and Kujur S. N (2014), Autoregressive Integrated Moving Average (ARIMA) approach for prediction of Rice (oryza sativa L.) yield in India, The Bioscan 9(3): 1063-1066.

Kumari Prity, Mishra G.C., Anil kumar pant, Garima Shukla and S.N.Kujur (2014). Comparison of forecasting ability of different statistical models for productivity of Rice (oryza sativa L.) in India, The ecoscan 8 (3 & 4): 193-198.

How to cite this article:

Supriya, K. 2021. A Study on the Performance of the ARIMAX-ANN Hybrid Forecasting Model over the other Time Series Forecasting Models ARIMAX and ANN in Forecasting the Rice Yield. *Int.J.Curr.Microbiol.App.Sci.* 10(01): 3421-3428.
doi: https://doi.org/10.20546/ijcmas.2021.1001.403