

Original Research Article

<https://doi.org/10.20546/ijcmas.2020.909.326>

## A Study on the Performance of ARIMAX-ANN Hybrid Model over the Other Forecasting Models while Forecasting the Damage Caused by Brown Plant Hopper (*Nilaparvata lugens*) in Telangana State

K. Supriya\*

Department of Statistics & Mathematics, PJTSAU, College of Agriculture,  
Rajendranagar, Hyderabad – 500 030, India

\*Corresponding author

### ABSTRACT

#### Keywords

ANN, ARIMAX,  
ARIMAX-ANN  
Hybrid model,  
Forecasting and  
undulating  
topography A

#### Article Info

Accepted:  
18 August 2020  
Available Online:  
10 September 2020

Agriculture is the most important sector of Indian Economy. Indian agriculture sector accounts for 18 per cent of India's gross domestic product (GDP) and provides employment to 50% of the country's workforce. Among cereals, Rice (*Oryza sativa L.*) is the most important cereal crop of the world both in respect to area and production. The total Rice production in the world is 496.22 million metric tons as estimated by the United States Department of Agriculture in April 2020 (USDA). India ranks second in rice production in the world with the production of 116.42 million metric tons. Certain biotic, abiotic reasons are resulting in low productivity. Among the biotic stresses insect pests constitute the key factor. In Telangana state, among the key insect pests of rice, Brown planthopper (*Nilaparvata lugens*) is one of the pests which cause major damage to the crop yields. In this study, three time series forecasting models, Artificial Neural Network (ANN), ARIMAX and ARIMAX-ANN Hybrid models were compared to forecast the damage caused by Brown Planthopper (*Nilaparvata lugens*) during both kharif and rabi seasons of Telangana state. To compare the effectiveness of these three models 30 years data during both kharif and rabi seasons pertaining to Telangana state was used i.e., from 1990-2019. The results showed that the ARIMAX-ANN Hybrid model outperformed the ARIMAX and ANN Forecasting models.

### Introduction

Agriculture plays a vital role in Indian economy. It continues to play a pivotal role to the sustainable growth and development of the Indian economy. Not only does Agriculture play a pivotal role to the sustainable growth and development of the

Indian economy, it also contributes significantly to production, employment and demand generation through various backward and forward linkages. Among the cereals, Rice has shaped the culture, diet and economy of thousands of millions of people. For more than half of the humanity "rice is life". Considering the importance and position of Rice in the entire world, the United Nation

has designated the year 2004 as the "International Year of rice". In order to achieve the target of increasing the productivity levels to meet the future demand, there is a need to raise productivity particularly in the rainfed production system. In our country, the average yield levels in the rainfed production systems are low ranging from 0.6 to 0.8 t/ha mainly due to vagaries of monsoon, poor soil fertility, undulating topography, biotic stresses and lack of adoption of improved technologies. Among the biotic stresses insect pests constitute the key factor since the warm and humid climate in the rainfed rice areas is conducive for the pest incidence and multiplication as well. The average yield losses in rice have been estimated to vary between 21-51 per cent. There are about more than 100 varieties of insect pests which cause damage to the rice crop. Among them planthoppers are the key insect pests causing great loss to the farmer. There are two plant hoppers viz., brown plant hopper (BPH), *Nilaparvata lugens* Stal and whitebacked plant hopper (WBPH), *Sogatella furcifera* Horvath which have gained major pest status on rice crop in recent years. The brown plant hopper (BPH) is common in rainfed and irrigated wetland environments especially during the reproductive stage of the rice. Both adults and nymphs suck the sap from the base of the tillers, resulting in yellowing and drying of the rice plants. The symptoms spread as patches of infestation in concentric circles within the field. This condition is known as 'hopper burn'.

Since pest forewarning provides lead time for managing impending pest attacks, and optimizes selection of pest control options for minimizing crop loss and reducing cost of plant protection, the sustainable growth, overall development and poverty alleviation, the present study has been taken up to forecast the damage caused by Brown planthopper using time series forecasting

models. Many techniques have been developed to predict the future behavior of a particular series of events but no simple answer as to which is the 'best'. The primary advantage of forecasting is that it provides various valuable information that can be used to make some strategies or alternate plans to design and shape the future. In the present study, time series data pertaining to the damage caused due to Brown planthopper (*Nilaparvata lugens*) which was given in terms of no. of bph /10 hills. The data was collected for 30 years i.e., 1990-2019 during both kharif and rabi seasons and forecasting of damage was done for the years 2020, 2021 and 2022 with the help of the Artificial Neural Network (ANN) model, Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) model and ARIMAX-ANN Hybrid model and comparison of the prediction accuracies of these three models using  $R^2$  (statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model) and Root Mean Square Error (RMSE) values.

## Materials and Methods

The main purpose of this study is to compare the forecasting abilities of the three time series forecasting models i.e., Artificial Neural Network (ANN) model, ARIMAX model and ARIMAX-ANN Hybrid model and to determine which model performs better. For this study, the data pertaining to the damage measured in terms of no. of bph/10 hills i.e., pertaining to Telangana state has been taken for the past 30 years i.e., from 1990-2019. The above said secondary data has been taken from the annual progress reports of AICRP, ICAR- Indian Institute of Rice Research, Rajendranagar, Hyderabad, RARS Jagtial and RARS Warangal.

## **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

$$Y_t = F_0 + F_1 Y_{t-1} + F_2 Y_{t-2} + F_3 Y_{t-3} + \dots + F_p Y_{t-p} + e_t - G_1 e_{t-1} - G_2 e_{t-2} - \dots - G_q e_{t-q}$$

Where  $Y_{t-i}$  and  $e_t$  are the actual values and random error at time t respectively.  $F_i$  ( $i = 1, 2, \dots, p$ ) and  $G_j$  ( $j = 1, 2, \dots, q$ ) 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  $\sigma_e^2$ .

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

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

Where as  $\Delta^d$  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

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

Where  $X_t$  is the exogenous variable and  $\beta$  is the coefficient.

## **Artificial neural network**

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

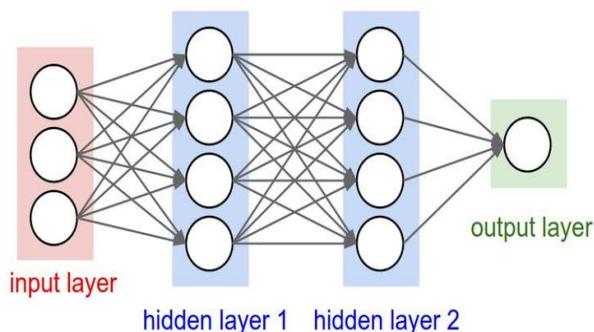
nonlinear complex data series. The basic architecture consists of three types of neuron layers: input, output and hidden layers. The ANN model performs a nonlinear functional mapping from the input observations ( $y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-p}$ ) to the output value  $y_t$ .

$$Zy_t = a_0 + \sum a_j f(W_{oj} + \sum W_{ij} y_{t-1}) + e_t \quad (4.1)$$

Where  $a_j$  ( $j=0,1,2,3,\dots, q$ ) is the bias on the  $j^{\text{th}}$  unit and  $W_{ij}$  ( $i=0,1,2,\dots, p, j=0,1,2,\dots, q$ ) is the connection weights between layers of the model,  $f(\cdot)$  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

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

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  $y_t$  as a combination of both linear and nonlinear components.

$$\text{That is } y_t = L_t + N_t + e_t \quad (3.3.5.1)$$

Where  $L_t$  is the linear component present in the given data and  $N_t$  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  $e_t$  denotes the residual at the time  $t$  from the linear model, then

$$e_t = y_t - L_t \quad (3.3.5.2)$$

where  $L_t$  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 confirmed 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

$$Y_t = L_t + N_t \quad (3.3.5.3)$$

The graphical representation of hybrid methodology is given in figure 1.

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

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

Where,

$P_t$ : Predicted value for time t

$A_t$ : Actual value at time t and

T: Number of predictions.

### Coefficient of determination ( $R^2$ )

R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable. In investing, R-squared is generally considered the percentage of a fund or security's movements that can be explained by movements in a benchmark index. It can be given by the formula:

A data set has  $n$  values marked  $y_1, \dots, y_n$  (collectively known as  $y_i$  or as a vector

$y = [y_1, \dots, y_n]^T$ , each associated with a predicted (or modeled) value  $f_1, \dots, f_n$  (known

as  $f_i$ , or sometimes  $\hat{y}_i$ , as a vector  $f$ ). Define the residuals as  $e_i = y_i - f_i$  (forming a vector  $e$ ).

$$\bar{y} = 1/n \sum_{i=1}^n y_i \quad (1)$$

If  $\bar{y}$  is the mean of the observed data then the variability of the data set can be measured using three sum of squares formulas

The total sum of squares (proportional to the variance of the data):

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad (2)$$

The regression sum of squares, also called the explained sum of squares:

$$SS_{reg} = \sum_i (f_i - \bar{y})^2 \quad (3)$$

The sum of squares of residuals, also called the residual sum of squares:

$$SS_{res} = \sum_i (y_i - f_i)^2$$

$$SS_{res} = \sum_i (y_i - f_i)^2 \quad (4)$$

The most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX), Artificial Neural Network (ANN) and ARIMAX-ANN Hybrid model were the main statistical methods to be used for analysis. Data analysis was performed using different statistical packages

### Artificial Neural Network (ANN) Forecasting

MATLAB 13.0 software was used to generate the best neural networks to predict

and get the best results.

### **Autoregressive Integrated Moving Average Models with Exogenous variables (ARIMAX) Forecasting**

**SPSS 20** software was used software to find the best ARIMA model to fit the time series and get the best possible forecasts.

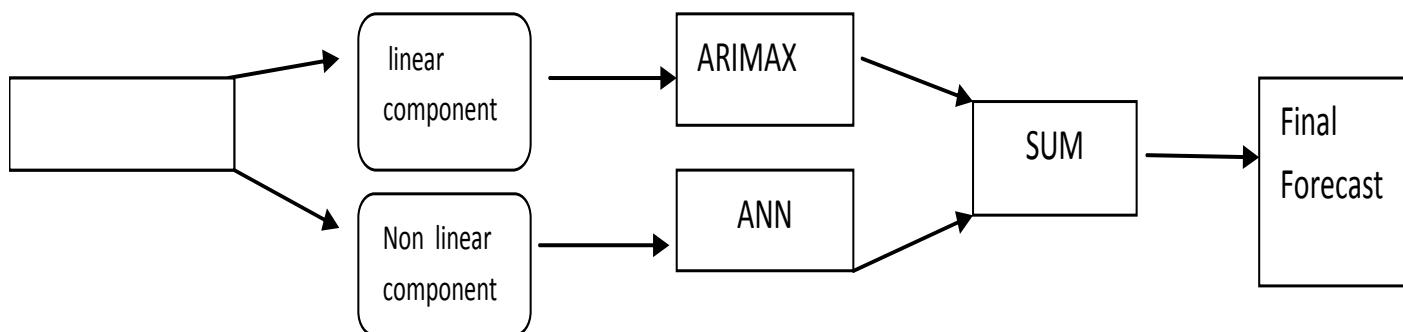
### **Results and Discussion**

The study was carried out to compare the effectiveness of the forecasting models for forecasting the damage caused by percentage due to key insect pest of rice i.e., Yellow stem borer in Telangana state in India which was measured in terms of percentage of dead hearts and percentage of white ears.

The forecasting techniques used in developing

the models were Artificial Neural Network, Auto regressive Integrated Moving Average with Exogenous variables and ARIMAX-ANN Hybrid model. The models have been developed on the basis of the secondary data for the past 30 years i.e., from 1990-2019 (both years inclusive) for the three different zones of the Telangana state. The three different zones of the state are a) Southern Telangana Zone b) Northern Telangana zone and c) Central Telangana zone. The data on the best check varieties has been used in the present study to nullify the varietal differences. This is the standard practice while using the time series data. The Root mean square error and  $R^2$  were used to compare prediction accuracies. A comparative study of the three zones is given below. Also, forecasted values for the years 2020, 2021 and 2022 using different forecasting techniques is also given in Table 1.

**Fig.1** The graphical representation of hybrid methodology



**Table.1** Zone wise performances of forecasting models and forecasted values for damage due to Brown planthopper (*Nilaparvata lugens*)

Zone	Forecasting Model and forecasted value	ANN		MLR		ARIMAX		ARIMA		ARIMAX-ANN	
		kharif	Rabi	kharif	Rabi	kharif	rabi	kharif	Rabi	kharif	Rabi
Southern Telangana Zone	2016*	270.80	257.30	270.80	257.30	270.80	257.30	270.80	257.30	<b>270.80</b>	<b>257.30</b>
	2017	267.82	247.41	283.27	277.34	217.59	251.39	258.62	266.38	<b>260.13</b>	<b>234.57</b>
	2018	261.43	252.27	255.66	269.06	232.49	261.25	256.92	266.01	<b>245.72</b>	<b>231.94</b>
	2019	274.67	263.72	246.95	264.48	226.52	263.72	256.78	265.52	<b>243.26</b>	<b>233.97</b>
	2020	253.70	262.65	242.75	262.57	235.77	259.63	254.68	265.48	<b>238.64</b>	<b>236.45</b>
	2021	282.36	266.38	240.96	261.70	281.32	266.35	252.32	265.50	<b>238.27</b>	<b>235.62</b>
	2022	274.06	263.10	240.30	261.38	272.56	261.32	252.28	265.32	<b>236.95</b>	<b>234.29</b>
	RMSE	118.72	102.25	700.74	449.10	125.40	103.38	174.62	104.88	115.62	85.30
	R <sup>2</sup>	0.70	0.97	0.43	0.30	0.65	0.27	0.17	0.16	0.79	0.76
Central Telangana Zone	2016*	212.70	283.50	212.70	283.50	212.70	283.50	212.70	283.50	<b>212.70</b>	<b>283.52</b>
	2017	238.24	247.48	200.54	260.20	206.23	252.61	188.52	282.55	<b>198.68</b>	<b>226.35</b>
	2018	244.01	246.47	209.97	265.96	200.65	245.11	189.64	269.55	<b>196.72</b>	<b>198.66</b>
	2019	248.89	238.49	219.28	266.39	199.62	202.38	189.26	267.88	<b>178.55</b>	<b>142.52</b>
	2020	253.00	236.54	198.55	267.89	196.31	192.14	189.55	267.66	<b>182.32</b>	<b>142.44</b>
	2021	256.45	228.62	196.12	265.99	195.45	189.52	189.60	267.63	<b>180.32</b>	<b>141.96</b>
	2022	259.31	212.74	197.73	265.96	193.66	189.25	189.61	267.62	<b>178.62</b>	<b>140.99</b>
	RMSE	14.94	28.29	377.56	447.76	74.27	92.19	101.45	98.72	13.42	15.78
	R <sup>2</sup>	0.98	0.95	0.46	0.19	0.17	0.28	0.22	0.28	0.98	0.98
Northern Telangana zone	2016*	196.00	164.00	196.00	164.00	196.00	164.00	196.00	164.00	<b>196.00</b>	<b>164.00</b>
	2017	213.85	181.99	194.2	164.41	198.26	172.22	211.08	190.23	<b>201.92</b>	<b>168.62</b>
	2018	213.78	181.39	192.20	164.30	199.12	178.31	210.83	189.92	<b>201.23</b>	<b>167.36</b>
	2019	213.95	181.66	189.34	164.28	208.11	192.23	208.67	192.02	<b>201.86</b>	<b>168.42</b>
	2020	214.13	181.93	188.35	163.98	209.12	192.64	208.45	192.19	<b>202.63</b>	<b>169.64</b>
	2021	214.30	182.20	187.41	163.77	209.66	195.21	208.36	192.20	<b>202.99</b>	<b>170.36</b>
	2022	214.47	182.46	186.99	162.98	210.01	196.01	208.11	192.20	<b>203.10</b>	<b>172.35</b>
	RMSE	49.70	33.29	229.84	265.73	51.52	46.94	66.22	61.58	34.01	17.92
	R <sup>2</sup>	0.70	0.82	0.18	0.20	0.17	0.41	0.09	0.21	0.73	0.89

It is observed that in all the three zones percentage of white ears is more than the percentage of dead hearts which shows that more care has to be taken in the reproductive stage than vegetative stage to avoid damage due to white ears. Compared to other zones in Northern Telangana zone the damage percentages are more which shows that the climate of this particular zone is more congenial for the pest outbreak than other zones. In all the three zones the Hybrid model has the lowest value of RMSE and highest value of R<sup>2</sup> which showed that ARIMAX-

ANN Hybrid model outperformed ARIMAX and ANN models in all the three zones.

## References

- Anderson, J.A. and Rosenfeld, E. (1988). Neurocomputing, Foundations of Research, Cambridge, MA, MIT Press.
- Bruce, Curry. (2007). Redundancy in parameters in neural networks: an application of Chebyshev polynomial

**How to cite this article:**

Supriya, K. 2020. A Study on the Performance of ARIMAX-ANN Hybrid Model over the Other Forecasting Models while Forecasting the Damage Caused by Brown Plant Hopper (*Nilaparvata lugens*) in Telangana State. *Int.J.Curr.Microbiol.App.Sci.* 9(09): 2609-2616. doi: <https://doi.org/10.20546/ijcmas.2020.909.326>