Int.J.Curr.Microbiol.App.Sci (2020) *9*(7): 1490-1497

International Journal of Current Microbiology and Applied Sciences
***ISSN: 2319-7706* Volume 9 Number 7 (2020)**
Journal homepage: http://www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2020.907.171

Comparative Study of ARIMAX-ANN Hybrid Model with ANN and ARIMAX Models to Forecast the Damage Caused by Yellow Stem Borer (*Scirpophaga incertulas*) in Telangana State

K. Supriya*

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

**Corresponding author*

Keywords

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

Article Info

Accepted:
14 June 2020
Available Online:
10 July 2020

A B S T R A C T

Agriculture plays a vital role in Indian economy. Among the cereals, Rice has shaped the culture, diet and economy of thousands of millions of people. The total Rice production in the world is 496.22 million metric tonnes as estimated by the United states Department of Agriculture in 2019 (USDA). India ranks second in rice production in the world with the production of 115 million metric tones. In India, Rice productivity is low 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. In Telangana state, among the key insect pests of rice, Yellow stem borer (*Scirpophaga incertulas*) is one of the pests which causes 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 Yellow Stem borer (*Scirpophaga incertulas*) during both kharif and rabi seasons of Telangana state. To compare the effectiveness of these three models 30 years data 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

Rice (*Oryza sativa I.*) 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 496.22 million metric tonnes as estimated by the United states Department of Agriculture in July 2019 (USDA). India ranks second in rice production in the world with the production of 115 million metric tonnes where as China ranks first with 146.73 million metric tonnes (Statistica, the statistical portal, 2019). 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. The region has an area of 114.84 lakh ha and a population of 352.87 lakhs as per 2011 census. 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. Telangana (31 districts) receive a normal rainfall of 906.6 mm in a year. Based on the Agro-climatic conditions, the state has been divided into three agro-climatic zones. They are northern Telangana zone, Southern Telangana zone and Central Telangana zone.

Further, rice crop is prone to the attack of weeds, several insect pests and diseases causing crop losses to the extent of 30 – 40% which further adds to the complexity to achieve high yield potential. Among the biotic stresses insect pests cause major damage to the crop yields. 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 Yellow stem borer is one of the key insect pests in rice causing approximately 25-60% of the yield loss to the farmer. The larvae of the borers enter the tiller to feed, grow and cause the characteristic symptoms of 'dead hearts' or 'white ears' depending on the stage of the crop. During the vegetative stage, the feeding frequently results in severing the apical parts of the plant from the base. When such type of damage occurs during stem elongation, the central leaf whorl does not unfold, turns brownish and dries out although the lower leaves remain green and healthy. This condition is known as 'dead heart'. Affected tillers dry out without bearing panicles. Similarly, during reproductive stage, severing of growing part from the base results in the drying out of panicles. The empty panicles are very conspicuous in field as they remain stiff, straight, whitish and are called 'white ears'. Infestation results in partial/ total chaffiness of the glumes and ill-filled grains.

Dead hearts

White ears

Materials and Methods

The main purpose of this study is to compare the forecasting abilities of the three 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 percentage i.e., percentage of dead hearts and percentage of white ears during both kharif and rabi seasons pertaining to the 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

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

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 G_j (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:

1) Model Identification
2) Parameter estimation and
3) 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.

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

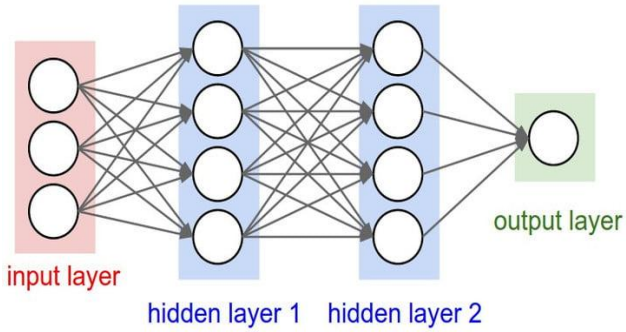
architecture consists of three types of neuron layers: input, output and hidden layers. The ANN model performs a nonlinear functional mapping from the input observations (y_{t-1}, y_{t-2}, y_{t-3}, ……..y_{t-p}) to the output value y_t.

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

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

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

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.

That is

$$y_t = L_t+N_t+\mathbf{e_t} \qquad \textbf{(3.3.5.1)}$$

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

$$e_t = y_t - L_t \qquad \textbf{(3.3.5.2)}$$

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

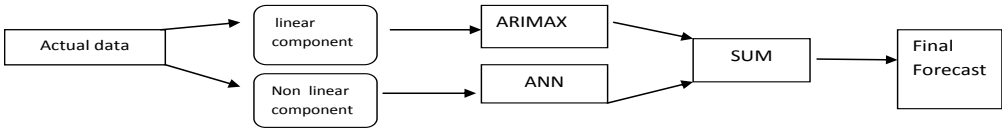
Diagnosis of residuals is done to check if there is still linear correlation structures left in the residuals then further we will go for nonlinearity check. The residuals are tested for nonlinearity by using BDS test.

Once the presence of the nonlinearity is 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

$$Y_t = L_t+N_t \qquad \textbf{(3.3.5.3)}$$

The graphical representation of hybrid methodology is given in the following figure.



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) $\sqrt{(\Sigma(P_t - A_t)^2)}$

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

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 $y_1,...,y_n$ (collectively known as y_i or as a vector $y = [y_1,...,y_n]^T$), each associated with a predicted (or modeled) value $f_1,...,f_n$ (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*).

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

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):

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

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

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

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

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

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

The most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad \textbf{(5)}$$

Results and Discussion

The study was carried out to compare the effectiveness of the forecasting models for forecasting the damage 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 below.

Table.1 Zone wise performances of Forecasting models and forecasted values for damage due to Yellow stem borer.

Zone	Forecasting Models and forecasted values	ANN		ARIMAX		ARIMAX-ANN	
		DH%	WE%	DH%	WE%	DH%	WE %
Southern Telangana Zone	**Kharif season**						
	2020	9.41	9.82	8.22	8.76	**9.89**	**9.89**
	2021	9.43	9.90	8.18	8.75	**9.84**	**9.88**
	2022	9.45	9.97	8.20	8.99	**9.86**	**10.11**
	RMSE	3.08	3.11	3.78	3.53	**1.44**	**2.47**
	R^2	0.90	0.30	0.11	0.37	**0.97**	**0.58**
	Rabi Season						
	2020	10.22	11.75	8.22	8.76	**9.89**	**9.89**
	2021	10.19	11.77	8.18	8.75	**9.84**	**9.88**
	2022	10.16	11.79	8.20	8.99	**9.86**	**10.11**
	RMSE	3.26	3.50	3.95	4.98	**2.87**	**1.69**
	R^2	0.41	0.29	0.17	0.07	**0.72**	**0.91**
Central Telangana Zone	**Kharif season**						
	2020	9.45	8.78	9.53	9.42	**9.60**	**9.72**
	2021	9.71	8.70	9.54	9.54	**9.80**	**9.66**
	2022	9.74	8.62	9.53	9.32	**9.42**	**9.01**
	RMSE	2.45	1.13	2.50	2.60	**1.45**	**1.09**
	R^2	0.26	0.83	0.60	0.24	**0.88**	**0.94**
	Rabi Season						
	2020	9.86	9.69	9.61	10.12	**10.20**	**10.10**
	2021	9.86	9.71	9.58	10.23	**10.10**	**9.98**
	2022	9.86	9.73	9.53	10.36	**10.60**	**9.81**
	RMSE	2.84	3.19	2.20	3.20	**1.51**	**2.57**

	R^2	0.46	0.26	0.33	0.10	**0.98**	**0.57**
Northern Telangana Zone	**Kharif season**						
	2020	9.98	20.05	10.05	20.72	**10.56**	**20.42**
	2021	9.93	20.04	10.06	20.85	**10.26**	**20.05**
	2022	9.92	20.04	10.05	20.92	**9.86**	**20.03**
	RMSE	1.48	3.30	1.34	1.44	**1.31**	**1.32**
	R^2	0.56	0.54	0.32	0.36	**0.76**	**0.70**
	Rabi Season						
	2020	11.26	20.08	10.46	20.65	**10.18**	**21.92**
	2021	11.24	19.92	10.48	20.54	**10.05**	**21.65**
	2022	11.25	19.89	10.52	20.32	**10.01**	**21.28**
	RMSE	1.20	2.40	1.25	3.30	**1.12**	**1.82**
	R^2	0.73	0.61	0.27	0.35	**0.79**	**0.91**

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^2 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 polynomials. *Computational Management Science*, 4(3), 227-242.

Christian, Schittenkop; Gustavo, Deco and Wilfried, Brauer. (1997). Two Strategies to Avoid overfitting in Feed forward Networks, *Neural networks*, 10(3), 505-516.

Gao Jiti and Lking Maxwel. (2015). ARIMAX-GARCH-Wavelet model for forecasting volatile data. *Model Assisted statistics and applications*, 10(3), 243-252.

Gorr, W.; Nagin, D. and Szczypula, J. (1994). Comparative study of artificial neural network and statistical models for predicting student grade point averages. *International Journal of Forecasting*, 10, 17–34.

Halbert, White. (2008). Learning in Artificial Neural Networks: A Statistical Perspective, *Neural Computation*, 1(4), 425-464.

Kalita H, Avasthe RK and Ramesh K. (2015). Effect of weather parameters on population build up of different insect pests of rice and their natural enemies. *Indian Journal of Hill farming*, 28(1), 69-72.

Rathod, S., Singh, K. N., Paul, R.K., Meher, R.K., Mishra, G.C., Gurung, B., Ray, M. and Sinha, K. (2017). An improved ARIMA Model using Maximum

Overlap Discrete Wavelet Transform (MODWT) and ANN for Forecasting Agricultural Commodity Price. *Journal of the Indian Society of agricultural Statistics*, 71(2), 103–111.

Sang, Hoon Oh (2010). Design of Multilayer Perceptrons for Pattern Classifications. *The Journal of the Korea Contents Association*, 10(5), 99-106.

Zhang, G.P and Min, Qi. (2003). Neural network forecasting for seasonal and trend time series. *European Journal of Operation Research*, 160(2), 501-514.

Zhang, G.P. (2003). Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing*, 50(17), 159-175.

Zhang, G.P. (2007). Avoiding Pitfalls in Neural Network Research. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 37(1), 3-16.

How to cite this article:

Supriya, K. 2020. Comparative Study of ARIMAX-ANN Hybrid Model with ANN and ARIMAX Models to Forecast the Damage Caused by Yellow Stem Borer (*Scirpophaga incertulas*) in Telangana State. *Int.J.Curr.Microbiol.App.Sci*. 9(07): 1490-1497.
doi: https://doi.org/10.20546/ijcmas.2020.907.171