Indian Journal of Agricultural Sciences **88** (8): 1275–79, August 2018/Article

Hybrid linear time series approach for long term forecasting of crop yield

WASI ALAM[1], KANCHAN SINHA[2], RAJEEV RANJAN KUMAR[3], MRINMOY RAY[4], SANTOSHA RATHOD[5], K N SINGH[6] and PRAWIN ARYA[7]

ICAR-Indian Agricultural Statistics Research Institute, Pusa, New Delhi 110 012

Received: 25 August 2017; Accepted: 01 May 2018

ABSTRACT

Long term forecasting of crop production is required to establish long term vision, say by 2025, to meet growing demand of population at that point of time. Existing univariate linear time series ARIMA approach is valid for short term forecast only. In this paper, a technique for long term yield forecast has been proposed. Initially, we have tried to improve short term forecast of yield by using hybrid ARIMA through ANN approach. The forecast values of yield through hybrid approach was considered as baseline data for long term forecast of yield. Time series data on rice yield was considered for Aligarh district of Uttar Pradesh for the study. Through ARIMA (2,1,0), we got short term forecast of yield by 2020 and the residuals obtained by 2013 were used to model and forecast through ANN approach. For the residuals, 05:04s:1l (05 time delay and 04 hidden nodes) model was identified as suitable one as it has minimum values of mean absolute percentage error (MAPE) for training and testing sets. Using 05:04s:1l model, residuals were forecasted by 2020, forecast values of yield obtained through ARIMA (2,1,0) were corrected by forecasted residuals and eventually get forecast of yield through hybrid approach. The estimated MAPE for ARIMA (2,1,0) and hybrid approach were 17.677% and 4.65%, respectively. Significant reduction in MAPE through hybrid approach indicates it's much better performance as compared to ARIMA alone. Using hybrid approach, we got forecast of yield by 2020 and considering this forecasted yield as baseline data, we got forecast by 2025 through the proposed approach.

Key words: ARIMA, Artificial neural network, Hybrid ARIMA, Long term yield forecast

In a developing country like India, food security means making available minimum quantity of food grains to the entire population. Despite the fact that India has made a satisfactory achievement in food grains production, its population growth has nullified the benefits of production. The FAO forecasts that global food production will need to increase by over 40% by 2030 and 70% by 2050 (FAO 2009). Among food grains, rice is the most important crop of the developing world and the staple food for more than 60% of the Indian population. In India, the annual compounded growth rate of rice production has declined from 3.55% during 1981-90 to 1.74% during 1991-2000. Projection of rice demand by 2030 mentioned in vision 2030 of Central Rice Research Institute (Adhya *et al*. 2011) has been computed on the basis of fixed historical growth rate. This approach is quite adhoc and having no sound statistical foundation. Forecasting the future demand/supply of crop production to meet the need of corresponding future growing population is a major concern for policy planners. Production is simply multiplication of cropped area with yield. Here, we have tried to forecast yield. In order to get more reliable future crop production forecast, we need more precise time series forecast of yield. Traditionally, classical autoregressive integrated moving average (ARIMA) model (Box *et al.* 2007, Cogger 1988, Clements 2003) has been widely used for short term time series forecasting. For other works of authors, one may refer to Paul *et al.* (2011, 2014), Naveen *et al.* (2012), Joshi *et al.* (2015), Alam *et al.* (2016), Chaturvedi and Alam (2010), Sinha *et al.* (2018) and Singh *et al.* (2011). In ARIMA approach, the future value of a variable is assumed to be a linear function of several past observations and random errors. Classical ARIMA models are typically well-suited for short-term forecasts, but not for long term forecasts due to the convergence of the autoregressive part of the model to the mean of the time series. Moreover, this approach does not explain the nonlinear component of residuals obtained through ARIMA model. Here, we have tried to improve the performance of ARIMA through the approach of Zhang (2003) in first instance and the improved forecast values have been used for long term forecast through the proposed technique.

MATERIALS AND METHODS

ARIMA model does not explain the nonlinearity component, i.e. errors. Here, we have tried to improve the

[1]Senior Scientist (e mail: wasi@iasri.res.in)

performance of ARIMA model through artificial neural network (ANN) approach (Vapnik 2000) by explaining residuals.This method consists of two phases. In the first phase, the time series was analyzed by using ARIMA models. An ARIMA model is given by:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t$$

where,

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ (Autoregressive parameter)

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_P B^q$ (Moving average parameter)

ε_t, white noise or error term; d, differencing term; B, Backshift operator, i.e. $B^a Y_t = Y_{t-a}$

In the next phase, the residuals obtained in the previous phase were examined by ANN and then forecast values obtained from the ARIMA model were summed. In this paper, time delay neural network approach was used to develop a new hybrid model to overcome the limitations of ARIMA in an attempt to yield more accurate results. A typical time delay neural network structure with one hidden layer is denoted by I: Hs: O*l*, where I is the number of nodes in input layer, s denotes the logistic sigmoid transfer function, O denotes number of nodes in the output layer and *I* indicates linear transfer function. Here, ANN approach was used to develop a new hybrid model to overcome the limitations of ARIMA in an attempt to yield more accurate results. In our proposed approach, an ARIMA model was first used to model the linear patterns of time series data (yield). The residuals of the linear model will then contain only the nonlinear relationship. Therefore, in the second phase, the ANN was used to model the nonlinear patterns of ARIMA residuals. This hybrid approach was used to get better forecasts as compared to classical time series models. ANNs are flexible computing frameworks for modeling a broad range of nonlinear problems. One significant advantage of the ANN models over other classes of nonlinear models is that ANNs are universal approximators that can approximate a large class of functions with a high degree of accuracy. Their power comes from the parallel processing of the information from the data. No prior assumption of the model form is required in the model building process. The network model is largely determined by the characteristics of the data. Single hidden layer feed forward network is the most widely used model form for time series modeling and forecasting. The model is characterized by a network of three layers of simple processing units connected by a cyclic links. The relationship between the output and the inputs has the following mathematical representation:

$$y_t = w_0 + \sum_{j=1}^{Q} w_j g(w_{0j} + \sum_{i=1}^{P} w_{ij} y_{t-i}) + e_t$$

The logistic function is often used as the hidden layer activation function. Data normalization is often performed before the training process begins. When nonlinear transfer functions are used at the output nodes, the desired output values must be transformed to the range of the actual outputs of the network. Even if a linear output transfer function is used, it may still advantageous to standardize the outputs as well as the inputs to avoid computational problems, to meet algorithm requirement and to facilitate network learning. In general data normalization is beneficial in terms of classification rate and mean squared errors, but the benefit diminishes as network and sample size increase. In addition data normalization usually slows down the training process. Normalization of the output values (targets) is usually independent of the normalization of the inputs. For time series modeling problems, however, the normalization of targets is typically performed together with the inputs. The choice of range to which inputs and targets are normalized depends largely on the activation function of output nodes, with typically [0, 1] for logistic function. It should be noted that, as a result of normalizing the target values, the observed output of the network should be corresponding to the normalized range. Thus, to interpret the results obtained from the network, the outputs must be rescaled to the original range. From the user's point of view, the accuracy obtained by ANNs should be based on the rescaled data sets. Performance measures is also be calculated on the rescaled outputs.

Let y_t denote the value of the time series at time point t, then we assume that

$$y_{t+1} = f(y_t,\ldots,y_{t-n+1}) + \varepsilon_{t+1}$$

for some autoregressive order n, where ε_t represents some noise at time t and f is an arbitrary and unknown function. The goal is to learn this function f from the data and obtain forecasts for t+h, where $h \in \{1,\ldots,H\}$. Hence, we are interested in predicting the next H data points, given the time series data. Here, we want to predict multiple time periods ahead (H>1), as we know univariate linear time series approaches like ARIMA provides short term direct H-step ahead forecast. In direct H-step ahead forecasting, we learn H different models of the form

$$y_{t+h} = f_h(y_t,\ldots,y_{t-n+1}) + \varepsilon_{t+h}$$

To forecast h>H, we have proposed the following iterative steps for long term forecast through hybrid time series models through machine learning approaches: 1. Select the suitable ARIMA model and obtain the fitted values of yield along with residuals. 2. Apply BDS (Brock Dechert Scheinkman) test for testing the nonlinearity of the residuals. If residuals are non-linear, apply nonlinear machine learning technique like ANN on the residuals. 3. Select the best ANN model for the residuals on the basis of minimum values of forecast accuracy measure and correct the fitted values of ARIMA model through the fitted residuals obtained through the selected ANN model. 4. Compute the MAPE (Mean Absolute Percentage Error) for the fitted values of ARIMA and corrected yield using fitted residuals through selected ANN model. 5. If MAPE for hybrid approach is less than ARIMA model, use the hybrid approach for long term forecast in the following way:

(i) Obtain short term out of sample forecast of yield

through the selected ARIMA model using corrected / baseline data. (ii) Forecast the residuals up to the desired forecast horizon by the suitable ANN model. (iii) Obtain baseline data by correcting the short term forecast values of yield (obtained by ARIMA model) through the forecasted residuals using the selected ANN model. (iv) Select suitable ARIMA model on the basis of baseline/ corrected data and obtain short term forecast of the yield up to the desired forecast horizon. (v) Consider the baseline data obtained as above for further long term forecast. (vi) Repeat steps i-v until we get the forecast of the desired forecast horizon.

In this article, Zhang's hybrid approach (Zhang 2003) has been employed. This approach considers time series (y_t) as a function of linear and nonlinear components. Hence

$$y_t = f(L_t, N_t)$$

where L_t and N_t represents the linear and nonlinear component, respectively. As needs be the relationship between linear and nonlinear components, it can be written as following

$$y_t = L_t + N_t$$

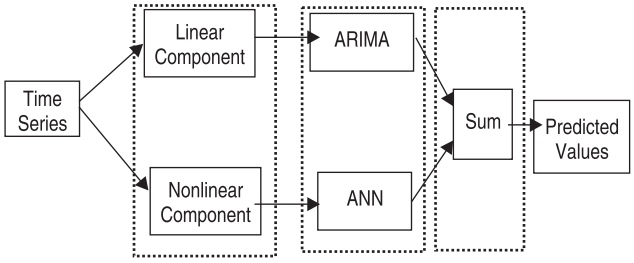
The main strategy of this approach is to model the linear and nonlinear components separately by different model. The methodology comprises of three steps. Initially, an ARIMA model is employed to fit the linear component. Let the prediction series provided by ARIMA model denoted as $\hat{L}_t$. In the second step, rather than predicting the linear component, the residuals denoted as $\hat{L}_t$ which are nonlinear in nature are predicted. The residuals can be gotten by subtracting the predicted value $\hat{L}_t$ from actual value of the considered time series y_t.

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

Now the residuals are predicted employing an ANN model. Let the prediction series provided by ANN model denoted as $\hat{N}_t$. Eventually, the predicted linear and nonlinear components are combined to generate aggregate prediction.

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

The ARIMA-ANN hybrid approach is graphically shown below



RESULTS AND DISCUSSION

Time series data on rice yield of Aligarh district of Uttar Pradesh has been taken from Directorate of Economics and Statistics from the year 1975 to 2013. Sequence charts for yield is shown in Fig 1.

ARIMA approach was applied on the yield data. On the basis of minimum values of goodness of fit (AIC=18.54, BIC=23.21), insignificant p-value of Ljung-Box Q test for residuals (p=0.372) and significant p-values (Table 1) of the parameters, we selected ARIMA (2,1,0) as suitable model. ARIMA (2,1,0) is also selected as suitable model by *auto. arima* option of R-software. Using ARIMA (2,1,0) model, fitted values of yield were obtained along with the residuals by 2013. By using ARIMA (2,1,0) model, we forecasted

Table 1 Parameter estimate

Model	Parameter estimate	SE	t	p-value
ARIMA (2,1,0)				
AR Lag1	-.714	.163	-4.385	<.01
AR Lag 2	-.283	.163	-1.740	.091

Table 2 Out of sample forecast

Model	2014	2015	2016	2017	2018	2019	2020
Forecast	2.39	2.42	2.43	2.48	2.51	2.54	2.58
UCL	2.99	3.03	3.11	3.25	3.32	3.41	3.50
LCL	1.80	1.80	1.74	1.72	1.70	1.68	1.67

Table 3 Actual and forecast values by ARIMA and hybrid approach

Year	Actual yield	ARIMA (2,1,0)	Hybrid approach
2003	2.06	1.97	2.07
2004	2.25	1.97	1.91
2005	2.09	1.97	2.19
2006	2.06	2.1	2.17
2007	2.01	2.22	2.10
2008	2.15	2.2	2.16
2009	1.73	2.12	1.82
2010	2.04	2.13	2.40
2011	2.2	2.06	2.23
2012	2.46	2.01	2.33
2013	2.24	2.07	1.88
2014		2.3	2.06
2015		2.39	2.83
2016		2.42	2.24
2017		2.43	2.66
2018		2.48	2.68
2019		2.51	2.46
2020		2.54	2.32
2021		2.53	2.52
2022		2.54	2.64
2023		2.58	2.54
2024		2.62	2.47
2025		2.64	2.63

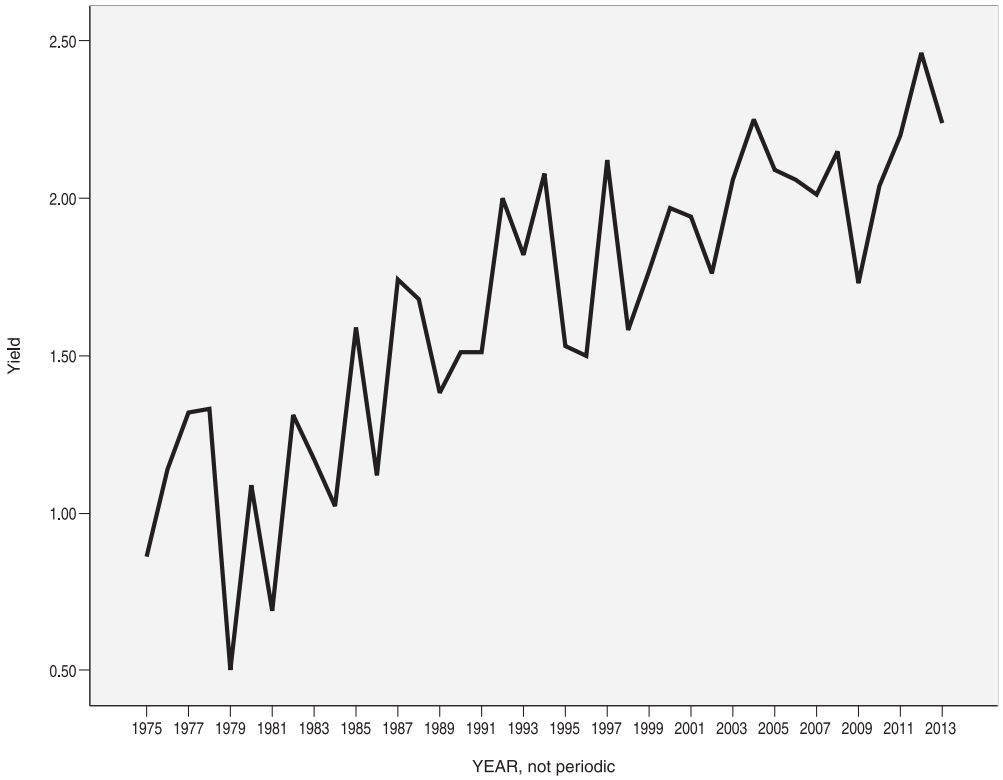


Fig 1 Sequence chart for yield.

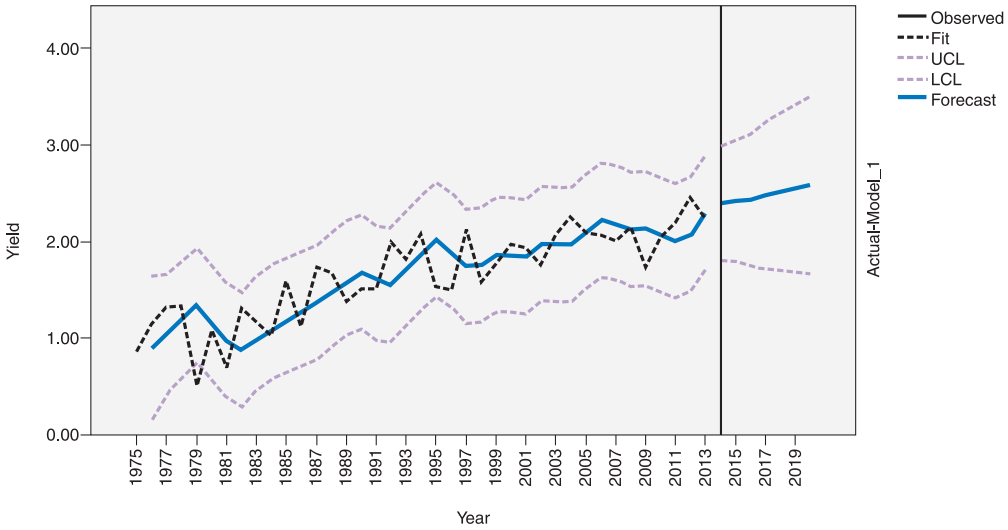


Fig 2 Forecast of crop yield by 2020 using ARIMA (2,1,0).

crop yield by 2020 which are mentioned in Table 2 and are pictorially presented in Fig 2.

BDS test of residuals obtained through ARIMA (2,1,0), supports that residuals are not identically independently distributed as P<0.01, hence, we applied artificial neural network approach for modeling and forecasting of residuals. ANN approach was applied on the residuals obtained by ARIMA (2, 1, 0). We tried different models with the combinations of different time delays and hidden nodes using logistic function as activation function on residuals from the year 1975 to 2002 as training data set. Keeping in view of the validity of the model and their performance, the neural network model with 05:04s:1l (05 time delay and 04 hidden nodes) was found to be the best one with the minimum MAPE for training=7.21 and testing=59.76. Through this trained model, we got the fitted values of residuals and corrected the fitted values of yield obtained by ARIMA (2,1,0) and eventually obtained forecast of crop yield through hybrid approach. MAPE for hybrid approach was found to be 4.65% as compared to 17.677% of ARIMA (2,1,0) alone. For the purpose of validation on the basis of trained ANN model, we got forecast of yield through hybrid approach and are presented in Table 3 along with the forecast values by ARIMA (2,1,0). Performance of hybrid approach was found to be better than that of ARIMA (2,1,0) alone, both under training as well as validation part of the data sets. Forecasted yield through hybrid ARIMA approach by 2020 are shown in Fig 3.

Considering these forecast values of yield by 2020 as a baseline data obtained through hybrid linear time series approach as mentioned earlier, we again applied ARIMA approach and selected ARIMA (2,1,0) model as suitable one and forecasted the yield by 2025 as mentioned in Fig. 4. Forecast of yield by 2025 obtained by ARIMA (2,1,0) and hybrid approach are reported in Table 3. This work can further be improved by introducing prediction interval in future. Through the proposed approach, we can forecast yield up to any desired year for example 2030, 2035 etc. which is desired in formulating long term vision viz. vision 2030 of CRRI, Cuttack instead of traditional way of forecasting of rice demand/supply using historical growth rate.

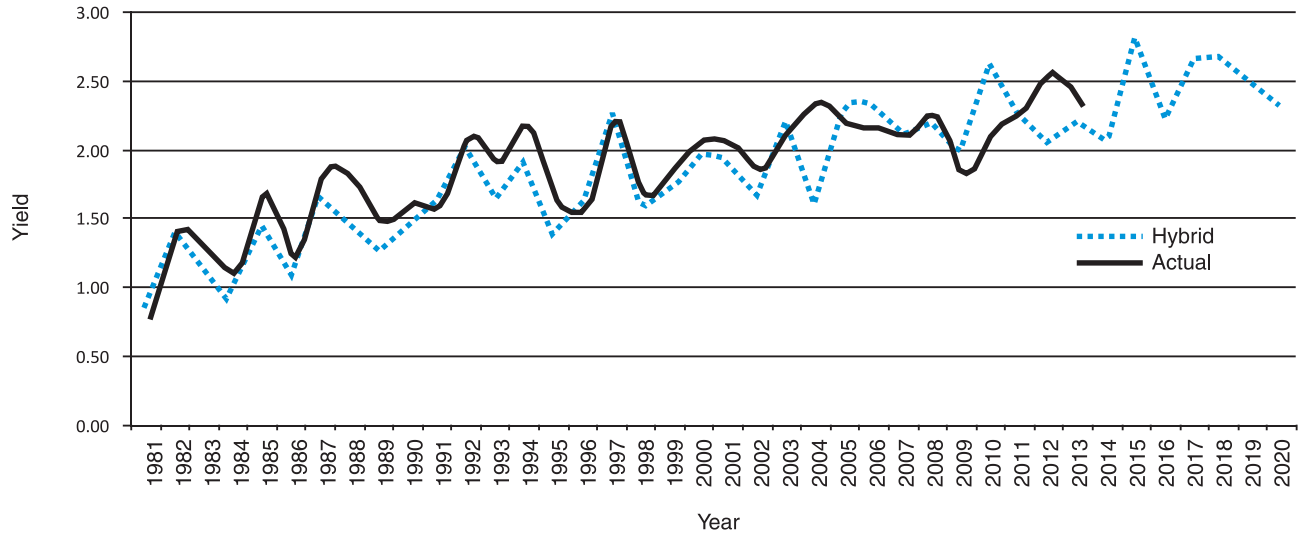


Fig 3 Forecast of crop yield by 2020 using hybrid ARIMA approach.

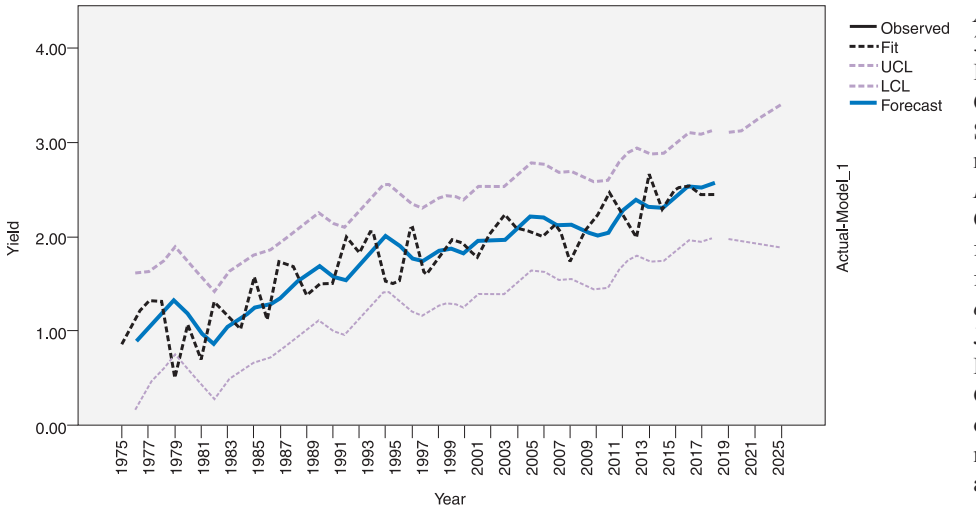


Fig 4 Crop yield forecast by 2025 through linear time series approach.

ACKNOWLEDGEMENT

Authors would like to thank ICAR-Indian Agricultural Statistics Research Institute for funding the project entitled Development of hybrid time series models using machine learning techniques for forecasting crop yield with covariates, from which this paper is drawn.

REFERENCES

Adhya T K, Kar P and Sinha S K. 2011.Vision 2030. ICAR-Central Rice Research Institute. [http://www.crri.nic.in/crri_vision2030_2011.pdf]

Alam W, Chaturvedi A, Kumar A, Sinha K and Singh K N. 2016. Sequential testing for decision making in the management of mustard aphid using size-biased negative binomial distribution. *International Journal of Agricultural and Statistical Sciences*, **12**(2): 531-5.

Chaturvedi A and Alam W. 2010. UMVUE and MLE in a family of lifetime distributions. *Journal of Indian Statistical Association* **48**(2): 189-213.

Box G E P, Jenkins G M and Reinsel G C. 2007. *Time-Series Analysis: Forecasting and Control*, 3rd edition. Pearson Education, India.

Clements M P. 2003. Editorial: Some possible directions for future research. *International Journal of Forecasting* **19**:1–3.

Cogger K C. 1988. Proposals for research in time series forecasting. *International Journal of Forecasting* **4:** 403–10.

Joshi D, Aggarwal M A, Pandey D, Bind A and Alam W. 2015. Generation of distinct profiles of rice varieties based on agro-morphological characters and assessment of genetic divergence. *Research on Crops* **16**(2): 311-9.

Naveen N C, Kumar D, Alam W, Chaubey R, Subramanian S and Raman R. 2012. A model study integrating time dependent mortality in evaluating insecticides against Bemisia tabaci (Hemiptera:Aleyrodidae). *Indian Journal of Entomology* **74**(4): 384-8

Paul R K, Alam W and Paul A K. 2014. Prospects of livestock and dairy production in India under time series framework. *Indian Journal of Animal Sciences* **84**(4): 462-6.

Paul A K, Alam W and Singh P. 2011. Average linkage method for clustering rice producing states of India. *Indian Journal of Agricultural Sciences* **81**(8): 75-8.

Singh N O, Paul, A K, Singh G N, Singh P and Alam W. 2011. Modeling seasonal growth of fish using modified Gompertz model with sine wave function. *Indian Journal of Animal Science* **81**(6): 648-50.

Sinha K, Panwar S, Alam W, Singh K N, Gurung B, Paul R K and Mukherjee A. 2018. Price volatility spillover of Indian onion markets: A comparative study. *Indian Journal of Agricultural Sciences* **88**(1): 114–20.

Vapnik V. 2000. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York

Zhang G P. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neuro-computing* **50:** 159 –75.