Statistical models for forecasting pigeonpea yield in Varanasi region

PRITY KUMARI[1]*, G.C.MISHRA[1]and C.P. SRIVASTAVA[2]

[1]Section of Agricultural Statistics, Departmentof Farm Engineering
[2]Departmentof Entomology and Agricultural Zoology
Institute of Agricultural Sciences, Banaras Hindu University ,Varanasi-221005, India
Email: psingh2506@aau.in

ABSTRACT

Present study deals with different linear and non-linear statistical models like multiple linear regression (MLR) model, autoregressive integrated moving average (ARIMA) model and artificial neural network (ANN) for forecastingpigeon pea yield grown in Varanasi region of Uttar Pradesh using 27 years of data (1985-86 to 2011-12). The performance of the model was assessed by root mean squared error (RMSE). On the basis of empirical studies, ANN was found to be best suitable model having lowest RMSE with forecasted yield during the year 2012-13 for Varanasi region.

Keywords: Artificial neural network (ANN), autoregressive integrated moving average (ARIMA) model, regression model andpigeonpea yield.

Indian agriculture is known throughout the world for its multi-functional success in generating employment, livelihood, food, nutritional and ecological security. Pigeonpea finds a prominent place in Indian meals and remain a primary source of protein for the majority of vegetarian population of the country. It occupies an important place in human nutrition due to its high protein content than cereal grains. Therefore its availability to common man is a major challenge. Due to high variability in yield from year to year, there is a need to provide reliable yield forecast which will be helpful in decision making as well as future planning. Various studies are existing in the literature, for forecasting crop yield with linear and non-linear techniques but the prominent one among linear are Regression and Autoregressive Integrated Moving Average (ARIMA) Model and for non-linear, Artificial Neural Network (ANN) architecture (Agrawal *et al.*, 1986;Zhang, 1998;Sharma *et. al.*, 2012; Kumari, *et. al.*,2013 and Kumari, *et. al.*,2014).

Multiple linear regressions (MLR) are widely suitable for short or intermediate term forecasting. In the present study, MLR was used for developing forecasting models using predictors as appropriate un-weighted and weighted weather indices (Kumar *et al.*,1999; Varmola *et. al.*,2004:Agrawal and Mehta, 2007; Chauhan *et al.*,2009**)**.

The ARIMA model, also known as the Box-Jenkins model(Box and Jenkins, 1970)is commonly used as the most efficient forecasting technique. ARIMA essentially relies on past values of the data series as well as previous error terms for forecasting. However, ARIMA models are relatively more robust and efficient than more complex structural models in relation to short-run forecasting (Gorantiwar *et. al.*,2011; Kumar *et. al.*,2013).

On the other side, ANNhas many distinguishing features that make it attractive to researcher.This is in contrast to many traditional techniques for time series predictions, such as Regression and ARIMA, which assume that the existing relation in the problem under study is generated from linear processes and so might be inappropriate for most real-world problems that are nonlinear. Therefore, there is need to solve nonlinear, time-variant problems also as many applications such as in agriculture and other field, which are basically uncertain in their behaviour and changes with time. ANNs are known to provide competitive results to various traditional time series models such as ARIMA model (George *et. al.*,2001; Ho *et. al.,*2002; Mishra and Singh, 2013; Meena *et. al.*,2016).

MATERIALS AND METHODS

In the present study, time series secondary data on pigeonpea yield for Varanasi were collected for the period 1985-86 to 2011-12 from All India Coordinated Research Project on Pigeonpea (Indian Council of Agricultural Research) and weekly weather data for region of Varanasi

* Present address : College of Horticulture, Anand Agriclutural University, Anand-388110, Gujarat

were collected from All India Coordinated Research Project on Dry Land Agriculture, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi. Five main weather variables maximum temperature (X_1), minimum temperature (X_2), rainfall (X_3), maximum relative humidity (X_4) and minimum relative humidity (X_5) were considered for building regression model.

Multiple linear regressions (MLR) model

In the present study, MLR technique was used for developing crop weather based forecasting models using predictors as appropriate un-weighted and weighted weather indices. Here weekly weather data from July 1[25th standard meteorological week(SMW)] to March 15[11thSMW] in each year from 1985-86 to 2011-12 of Varanasi were utilized for development of multiple regression models. Out of 27 years data, 24 years data were utilized for development of regression model and 3 years data were used to validate the forecasting ability of developed model.

Agarwal and Mehta (2007) model was followed as given below:

$$Y = A_0 + \sum_{i=1}^{p} \sum_{j=0}^{1} a_{i,j} Z_{i,j} + \sum_{i \neq i'=1}^{p} \sum_{j=0}^{1} a_{i.i'.j} Z_{i,i',j} + cT + e$$

Where, $Z_{i,j}$, $Z_{iI',j}$: Weather indices; *i,i'*: 1, 2, …p; *Y*: Dependent Variable; *T*:Year number; A_0:Intercept ; *p*:Number of weather variables under study. 'e' *e*rror term, is normally distributed with mean zero and constant variance.Stepwise regression technique was used to select the important weather indices.

Autoregressive integrated moving average (ARIMA) model

ARIMA is one of the most traditional methods of non-stationary time series analysis. In contrast to the regression models, the ARIMA model allows time series data to be explained by its past or lagged values and stochastic error terms. ARIMA model is usually stated as ARIMA (p, d, q) and is expressed in the form:

$$Y_t = \theta_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots\ldots\ldots\ldots + \Phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots\ldots\ldots - \theta e_{t-q},$$

Where Y_t and e_t are the actual values and random error with mean zero and the constant variance σ_e^2 at time t, respectively, Φ_i(i= 1,2,…….,p) and θ_j(j = 1,2,……,q) are model parameters, p and q are referred to as orders of autoregressive and moving average polynomials respectively (Box and Jenkins, 1970).

In order to construct the best ARIMA model order of autoregressive (*p*), differencing (d) and moving average (*q*) parameters have to be effectively determined. The model having relatively small Root Mean Squared Error (RMSE), relatively high R^2 and adjusted R^2 was considered to be the best among all.

Artificial neural network (ANN)

Artificial neural network (ANN) by name itself says that it is a network of artificial neurons which follows the concept of function as in human brain neurons. The structure of artificial neural network consists of several layers of processing units /neurons/ nodes (Haykin, 2001).

Modeling with ANN involves two important tasks, namely, *topology* and *learning* algorithm of network. The topology of a networks involves (i) fixing the number of layers, (ii) the number of neurons for each layer, (iii) the node function for each neuron, (iv) whether feedback or feed-forward, and (v) the connectivity pattern between the layers and the neurons. All these adjustments are to be taken care of for improved performance of the system. The learning phase involves adjustments of weights as well as threshold values. (Haganand Menhaj,1994).

Usually, the data is divided into three non-overlapping sets: the so-called *training, validation* and *testingset*. The training set, consisting lager portion of data, is used to teach the network in order to get the desired target function. Then the validation set is used to decide when to stop training process, to avoid *overfitting*, a situation where the network memorizes the training data rather than learning the law that governs them. The testing data set, which exposed to theunseen data, is used tomeasure performance of trained network bymean square error (MSE) or root mean square error (RMSE).

In present case, Neural Network architectures were developed by using Levenberg Marquardt (LM) Algorithm (Ranganathan, 2004; Hao and Bogdan 2011) as a training algorithm of weight matrix.

RESULTS AND DISCUSSIONS

Multiple linear regression (MLR) model

The stepwise multiple linear regression analysis results presented in Table 1 showed that all the generated variables entered in three different modelsaffected significantly the yield but Model 3 was considered better than the remaining three models because of greater value R^2/Adjusted R^2value. Model 3 was explained by the variables *viz.* constant, Z251, Z20 and Z11. The constituent of each

Table 1: MLR model estimate of the predictors yield

Model		Coefficients			
		B	Std. Error	t	Sig.
1	(Constant)	402.36	274.02	1.46	.156
	Z251	.60	.15	4.03	.001
2	(Constant)	3413.98	1421.74	2.40	.026
	Z251	.70	.14	4.80	.000
	Z20	-7.04	3.27	-2.15	.043
3	(Constant)	4736.30	1452.75	3.26	.004
	Z251	.583	.14	3.98	.001
	Z20	-8.89	3.14	-2.82	.010
	Z11	76.16	35.63	2.14	.045

Table 2: ANN model parameters

Weights	H_1	H_2	Biases	Values
I_1	WI_1H_1=0.76	WI_1H_2= -1.36	BH_1	-1.64
I_2	WI_2H_1= 1.50	WI_2H_2= -2.05	BH_2	0.56
O	WOH_1= -0.46	WOH_2= -0.99	B_O	-0.34

Table 3: Performance of models

Model Accuracy and forecasted value	ANN	ARIMA	MLR
Forecast	980.84	Non Significant	1205.76
RMSE	299.93		884.02
MSE	89961.49		781497.00
R square	0.63		0.59

of these generated variables is as follows:

$$Z_{2,5,1} = \sum_{w=1}^{37} r_{25w} X_{2w} X_{5w}$$

$$Z_{2,0} = \sum_{w=1}^{37} X_{1w} \qquad Z_{1,1} = \sum_{w=1}^{37} r_{1w} X_{1w}$$

Where,

r_{25w}= Correlation coefficient between yield (Y) and product of 2nd and 5th weather parameter (viz. minimum temperature(X_2) and minimum relative humidity(X_5)and)

r_{1w}= Correlation coefficient between yield (Y) and 1st weather parameter (viz.maximum temperature (X_1))

The estimates of the constant and independent variables entered in the Model 3, were 4736.30, 0.58, -8.89 and 76.16 with standard error of 1452.75, 0.14, 3.14 and 35.63 respectively. Also they are statistically significant (Table 1). Since the models were developed only on the basis of 24 years data while three years data were taken as holdout in order to check the forecasting ability of the models. MSE of the Model 3 was calculated on the basis of three years data which were used to explain the error in the forecasting model and so MSE for that model was 781497.00. R-square and Adjusted Rsquare were 0.59 and 0.51 respectively (Table 1).The forecasted value of yield of late maturing pigeonpea during the year 2012-13 was obtained as 1205.76 kg ha^{-1} by this regression model.

ARIMA Model

In the present study, attempts were made to forecast the pigeonpea yield in Varanasi region with the help of ARIMA model. Since, with different combinations of parameters such as autoregressive terms (p), differencing terms (d) and Moving Average terms (q), none of the ARIMA models were found to be significant hence, ARIMA was not considered to be appropriate for this situation.

Artificial neural network architecture

Neural Network architecture was developed by using time series yield data of pigeonpea where lag values are taken as independent variable and MATLAB Neural Network Toolbox 2010 was used to develop thesearchitectures. The network used was a two-layer feed-forward network.

Out of various architectures of Neural Network, the best architecture (having relatively small MSE/RMSEand high R^2 value) was chosen with following topology: a) two-layer feed-forward network (one Input & one Hidden Layer) b) Input layer having two lag value of time series yield data as inputs c) Hidden layer having two node with sigmoid activation function and d) Output layer having one node

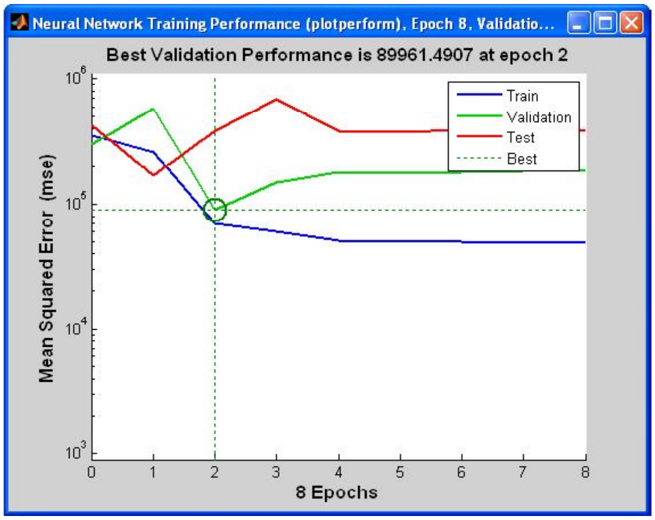
Neural Network Training Performance (plotperform), Epoch 8, Validatio...
Best Validation Performance is 89961.4907 at epoch 2
Mean Squared Error (mse)
10^6
10^5
10^4
10^3
0
1
2
3
4
5
6
7
8
8 Epochs
Train
Validation
Test
Best

Fig. 1: Performance of LM algorithm

with Linear activation function.

Therefore, four weights for input to hidden neurons and two weights for hidden to output neurons and three bias values were chosen. For training 70%, for each of validation and testing 15 % data were used by using Random Data Division Process.

Let the two input lag value in input layer were denoted by notation I_i (i=1,2), two hidden node of hidden layer were denoted as H_j(j=1,2) and output node is denoted as O then the weights among input & hidden neurons are denoted by WI_1H_1, WI_2H_1, WI_1H_2, WI_2H_2 and among hidden & output neurons WOH_1, WOH_2. Similarly, bias values of three nodes (two hidden nodes and one output node) were denoted as BH_1, BH_2 and B_O. The performance of the proposed network when trained with Levenberg-Marquardt (LM) algorithm was accessed by their Mean Squared Error (MSE) value along with multiple correlation coefficient (R) between observed and predicted outputs. Here parameters of ANN model *i.e.* weights among different nodes and biases value of each node were mentioned in the Tables 2.

From Fig.1, it was observed that the best validation performance MSE 89961.49 or RMSE 299.93was obtained at epoch 2. The regression analysis plot shown in Fig.2, displayed a linear regression between network outputs and the corresponding targets with the R value as 0.79 (R^2= 0.63) showing the fit was reasonably good for the data sets.The forecasted value of yield of pigeonpea during the year 2012-13 was obtained as 980.84kg ha^{-1} by this architecture.

Comparison of ANN, ARIMA and MLR Model:

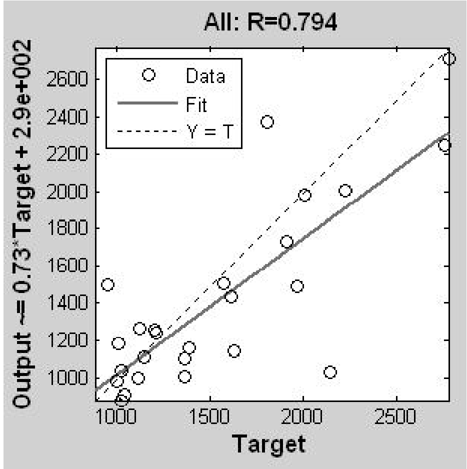

Fig. 2: Regression analysis of LM algorithm

Table 3 reflects that the forecasted value of yield was best explained by ANN model during 2012-13 for pigeonpea yield in Varanasi region with having relatively small value of root mean squared error (RMSE) 299.93 and relatively high value of R^2 0.63.

CONCLUSION

This paper aimed to evaluate the performance of Artificial Neural Network (ANN) by comparing it with Multiple Linear Regression (MLR) and Autoregressive Integrated Moving Average (ARIMA) Model. As compared to both linear model, ANN architecture, was found to be more appropriate for forecasting yield of pigeonpea for Varanasi region.

ACKNOWLEDGEMENT

The authors are thankful to the All India Coordinated Research Project on Pigeonpea (Indian Council of Agricultural Research) and All India Coordinated Research Project on Dry Land Agriculture, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, for providing data to carry out the present study.

REFERENCES

Agarawal, Ranjana; Jain R. C. and Jha, M.P (1986). Model for studying rice crop-weather relationship. *Mausam*, **37(1):**67-70.

Agarawal, R. and Mehta, S.C. (2007). Weather based forecasting of crop yields, pest and diseases – IASRI Models. *J. Ind. Soc. Agril. Stat.*, **61 (2):** 255-263.

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

Chauhan V.S., Shekh A.M., Dixit S.K., Mishra A.P. and Kumar

Sanjay (2009), Yield prediction model of rice in Bulsar district of Gujarat, *J. Agrometeorol*, **11 (2):** 162-168.

George R. K., Rammohan S., Kulshretha M. S., Shekh A.M. and Jaita H. (2001). Prediction of soil temperature using artificial neural network, *J. Agrometeorol*; **3 (1&2):** 169-173.

Gorantiwar S.D., Meshram D.T. and Mittal H. K. (2011). Seasonal ARIMA model for generation and forecasting evapotranspirtion of Solapur district of Maharashtra, *J. Agrometeorol;* **13 (2):** 119-122.

Hagan M. T. and Menhaj M. (1994). Training feedforward networks with the Marquardt algorithm, *IEEE Trans on Neural Networks*, **5(6)**: 989–993.

Hao Yu and Bogdan M. Wilamowski (2011). Levenberg–Marquardt Training, [web page] http://www.eng.auburn.edu/~wilambm/pap/2011/K10149_C012.pdf

Haykin, S. (2001). Neural Networks – A Comprehensive Foundation. IEEE Press, New York.

Ho, S.L., Xie, M. and Goh, T.N. (2002). A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction, *Comp. Indus. Eng.*, **42**: 371–375.

Kumar J. Ashok, Muralidhar M., Jayanthi M. and Kumaran M. (2013). Trend analysis of weather data in shrimp farming areas of Nagapattinam district of Tamil Nadu, *J. Agrometeorol,* 15 (2): 129-134.

Kumar R., Gupta B.R.D., Athiyaman B., Singh K.K. and Shukla R.K. (1999). Stepwise regression technique to predict pigeonpea yield in Varanasi district, *J. Agrometeorol*, **1 (2):** 183-186.

Kumari Prity, Mishra G.C. and Srivastava C.P. (2013). Forecasting of Productivity and Pod Damage by *Helicoverpaarmigera*using Artificial Neural Network Model in Pigeonpea (*Cajanuscajan), Int. J. Agri., Env.& Biotech.,* **6(2):** 187-193.

Kumari Prity, Mishra G.C, Pant Anil Kumar, ShuklaGarima and Kujur S. N. (2014), Autoregressive Integrated Moving Average (Arima) Approach for Prediction of Rice (Oryza Sativa L.) Yield in India, *The Biosc.,* 9(3): 1063-1066.

Meena P. K., Khare Deepak and Nema M. K. (2016). Constructing the downscale precipitation using ANN model over the Kshipra river basin, Madhya Pradesh, *J. Agrometeorol*, 18(1): 113-119.

Mishra, G. C. and Singh, A. (2013). A study on forecasting price of groundnut oil in Delhi by ARIMA Methodology and Artificial Neural Network, AGRIS On-line Papers in Economics and Informatics, **5(3):** 25-34.

RanganathanAnanth, (2004). The Levenberg-Marquardt Algorithm [web page] http://www.ananth.in/Notes_files/lmtut.pdf.

SharmaVidushi,RaiSachin and DevAnurag (2012). A Comprehensive Study of Artificial Neural Networks, *Int.J.Adv. Res. in Comp. Sci.& Soft. Eng.,* 2: 10.

Varmola S. L., Dixit S. K., Patel J. S. and Bhatt H. M. (2004). Forecasting of wheat yields on the basis of weather variables, *J. Agrometeorol*; 6 (2): 223-228.

Zhang G., Patuwo B. and HuM. Y. (1998). Forecasting with artificial neural networks: the state of the art," *Int.J.Forecas.*, 14(1):35–62.

Received : June 2016 ; Accepted : October 2016