Predicting the area and production of sugarcane in Tamil Nadu, India using neural networks

P. Dinesh Kumar*

Department of Agriculture, Karunya University, Coimbatore 741 114, India

Sugarcane is a major cash crop in India, grown in almost 5 million hectares with a production of 339 million tonnes. Tamil Nadu contributes significantly to the production of sugarcane. Data from the past year show a huge fluctuation in the area and production of sugarcane in the state. Predicting the area and production employing traditional modelling techniques fails because the assumptions are never attained in the field. To overcome this, soft computing techniques like artificial neural networks (ANNs) are used. In this study, a multilayer perceptron neural network (MLP-NN) with back-propagation was used to predict the area and production of sugarcane in Tamil Nadu. The MLP-NN (2,2) model predicts the area with minimum mean absolute error (MAE; 18.139) and root mean squared error (RMSE; 23.058) values and with high accuracy (99%). For production, the MLP-NN (2,1) model estimates minimum MAE (24.875) and RMSE (31.199) values with high accuracy (99%). So, MLP-NN (2,2) and MLP-NN (2,1) are the best ANN models to predict the area and production of sugarcane in Tamil Nadu respectively. Additionally, ANN models perform better in predicting nonlinear stochastic data.

Keywords: Back propagation, multilayer perceptron, neural network, nonlinear stochastic data, sugarcane area and production.

SUGARCANE, a major cash crop in India, is cultivated on almost 5 million hectares, which is about 3% of the total crop cultivated area in the country. The by-products of sugarcane, viz. sugar, molasses and bagasse, play a vital role in its production at a global level. Sugar supplies almost 10% of the daily calorie intake in India. Due to its versatile usage in many industries, sugarcane is also known as 'wonder cane'[1]. The sugarcane production in India is almost 339 million tonnes, coming second at the global level after Brazil. The average sugarcane productivity in India is 70.24 tonnes, which is higher than the average world productivity. Sugarcane is generally cultivated more in the tropical states of India than the sub-tropic ones. The tropical states such as Gujarat, Karnataka, Tamil Nadu and Andhra Pradesh grow almost 60% of the total sugarcane crop[2]. Uttar Pradesh is the leading sugarcane-producing state, followed by Maharashtra and Tamil Nadu. In 2017, the total sugarcane production in Tamil Nadu was about 189.876 million tonnes in 2.183 lakh hectares.

Previously many traditional linear and nonlinear methods were used to model the area and production of sugarcane. They demand certain assumptions which should be met to model the trend better. However, the natural process generally does not follow such assumptions. To overcome such complex situations, soft computing techniques like neural networks have been introduced. The neural network models are gaining popularity in recent decades due to their accuracy. In the present study, artificial feedforward neural networks have been used to estimate the area and production of sugarcane in Tamil Nadu.

Many studies have reported on the use of artificial neural networks (ANNs) in the field of agriculture[3,4]. Most of them are used for modelling the yield[5,6]. ANNs predict the time-series data better than the traditional autoregressive integrated moving average (ARIMA) models[7]. Thus ANNs were preferred in this study to predict the change in the area and production of sugarcane in Tamil Nadu.

Secondary data on the area and production of sugarcane in Tamil Nadu for the period from 1986 to 2016 were collected for this study from the Department of Economics and Statistics, Government of Tamil Nadu. A multilayer perceptron neural network (MLP-NN) with back-propagation (BP) was preferred for predicting the time-series data.

ANN is a synthetic miniature of the human brain which learns from the data using neurons. A typical feedforward neural network comprises at least one input, a hidden layer and an output layer. Each hidden layer contains at least one neuron. Figure 1 shows a typical feedforward neural network topology. The neuron is the basic unit of a neural network. It receives information from the input layer and passes it to the output layer. Each neuron has an activation function with a threshold value. So, the input should have the minimum threshold value to activate the neuron.

For the given inputs $x_1, x_2, \ldots, x_n$, a weight w_{ij} is associated with them. Then,

$$f(u) = \sum_{i}^{n} w_{ij}\, x_j + b_j,$$

where b_j is the bias.

For estimating the area and production of sugarcane, MLP-NN with BP was used. The activation function of

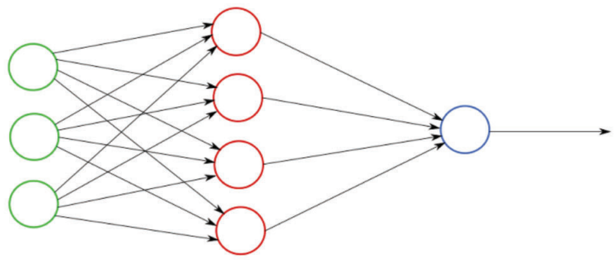

Figure 1. A basic multilayer perceptron neural network (MLP-NN) structure.

*e-mail: pdkstatistics@gmail.com

the node was a sigmoidal function with threshold values 0 to 1. The function is given by

$$f(u)=\frac{1}{1+\mathrm{e}^{-cu}}.$$

BP is the learning technique that computes errors and propagates the values back into the hidden layer to improve the model. Repeated BP reduces the error with precise output.

ANN models will not represent the system unless they are adequately trained. Initially, random values are taken as initial weights; then, the models are trained by adjusting weights iteratively to get the desired output. Training the ANN model is an iterative process where BP plays a major role in providing desired output. The ANN modelling procedure is explained stepwise as follows.

Step 1: In time-series modelling, the input variables are selected based on the lags that are related to them. Based on partial autocorrelation, the lag variables are decided. The data are divided into training set and test set.

Step 2: The selected lag variables are centred between 0 and 1 by removing the means of the variables. The lag variables are scaled to lie between minimum and maximum values using the following expression

$$z_i=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}}.$$

Step 3: There is no specific procedure to fix the number of layers and the number of nodes per layer in a model. The appropriate numbers are selected based on the iteration process. Various options are attempted and the one with the minimum error value is considered appropriate for the model. While fitting the model, the activation functions are selected. The initial values should be fixed to regenerate the model outputs.

Step 4: The model will be fitted for the training set and used to predict the test set. Based on the prediction accuracy, the best model will be selected.

Several statistical parameters are used to evaluate the model. The following parameters are estimated for the test set to evaluate the performance of the model:

(1) Mean absolute error,

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|.$$

(2) Mean absolute percentage error,

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{|y_i|}.$$

(3) Symmetric mean absolute percentage error,

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{\frac{|y_i|+|\hat{y}_i|}{2}}.$$

(4) Mean squared error,

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

(5) Root mean squared error,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}.$$

(6) Accuracy,

$$A = 1-\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|.$$

Here n is the number of observations, y_i the observed value and $\hat{y}_i$ is the value estimated by the model. The model with minimum MAE and RMSE values, and higher accuracy is considered as the best.

Table 1 shows the descriptive statistics of the area and production of sugarcane in Tamil Nadu. Descriptive statistics helps visualize the nature of the data. The range statistics is almost equal to the mean value of area and production, showing a large fluctuation in the data with time. The area and production data are slightly right-skewed and platykurtic in nature. This skewness and kurtosis statistics prove the non-normal nature of the data. So, using traditional linear and nonlinear models cannot provide a good model estimate.

While using the ANN model, selecting the input variables and the number of hidden layers plays a vital role. There were no hard rules for selecting the number of hidden layers in a neural network. It is more art than science. In time-series forecasting, selecting the input variables was based on the partial autocorrelation function (PACF). The PACF value

Table 1. Descriptive statistics of area and production of sugarcane

Parameters	Area (‘000 hectar)	Production (million tonnes)
Mean	275.90	293.38
Standard deviation	56.10	65.82
Median	272.96	294.90
Trimmed mean	275.28	291.92
Minimum	191.07	176.56
Maximum	391.20	451.68
Range	200.13	275.13
Skewness	0.05	0.19
Kurtosis	–1.26	–0.65
Standard error	9.92	11.63

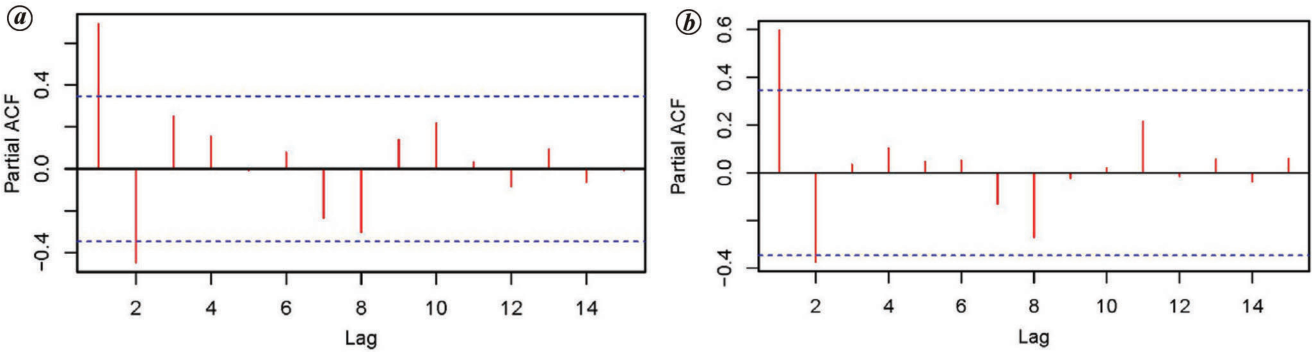

Figure 2. ***a***, Partial autocorrelation function (ACF) for the area of sugarcane. ***b***, Partial ACF for the production of sugarcane.

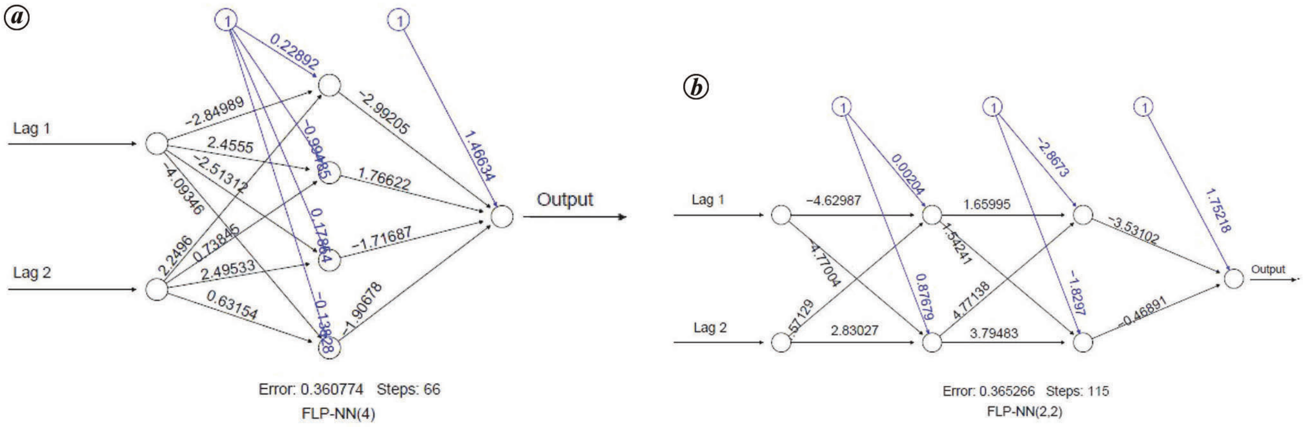

Figure 3. The (***a***) FLP-NN (4) model and (***b***) FLP-NN(2,2) model for the area of sugarcane.

Table 2. Performance statistics of single-layered artificial neural network (ANN) for the area of sugarcane

No. of neurons	MAE	MAPE	SMAPE	MSE	RMSE	Accuracy
1	22.051	0.090	0.087	780.047	27.929	0.988
2	23.411	0.093	0.090	778.081	27.894	0.984
3	21.184	0.086	0.083	715.046	26.740	0.990
4	18.739	0.076	0.074	552.238	23.500	0.996

MAE, Mean absolute error; MAPE, Mean absolute percentage error; SMAPE, Symmetric mean absolute percentage error; MSE, Mean squared error; RMSE, Root mean squared error.

of area and production of sugarcane revealed that the first two lags had a significant correlation (Figure 2). Thus, the first two lag variables will be considered as the input variables for the ANN models to predict the area and production of sugarcane. According to the universal approximation theorem, any neural network with one hidden layer with a sufficient number of neurons can predict the output with an acceptable level of accuracy[8]. In order to avoid overfitting the models, the number of hidden layers with sufficient neurons was restricted to two[9].

The network was trained using the BP algorithm with a threshold value of 0.01. The logistic function was used as the activation function of the model for both area and production. The network that predicted the test set with the minimum error was considered the best neural network model. Tables 2 and 3 show the performance statistics of the single- and two-layered neural network models for the area of sugarcane respectively. To find the best MLP-NN model, a trial and error method was used to find the optimum number of neurons in the hidden layer. Among the 12 trained models, the MLP-NN model with two hidden layers and two neurons each had the minimum MAE (18.139) and RMSE (23.058) values, followed by MLP-NN with one hidden layer and four neurons. So, MLP-NN(2,2) and MLP-NN(4)

Table 3. Performance statistics of two-layered ANN for the area of sugarcane

No. of neurons in layer I	No. of neurons in layer II	MAE	MAPE	SMAPE	MSE	RMSE	Accuracy
1	1	23.300	0.093	0.091	834.761	28.892	0.990
2	1	50.448	0.189	0.182	3222.161	56.764	0.986
3	1	24.041	0.096	0.094	872.623	29.540	0.991
4	1	27.568	0.110	0.107	1110.411	33.323	0.980
1	2	50.308	0.188	0.181	3209.143	56.649	0.986
2	2	18.139	0.074	0.071	531.684	23.058	0.991
3	2	23.622	0.091	0.089	660.519	25.701	0.991
4	2	22.552	0.092	0.088	810.375	28.467	0.988
1	3	50.393	0.189	0.182	3213.207	56.685	0.986
2	3	25.962	0.098	0.096	765.800	27.673	0.987
3	3	23.302	0.095	0.091	835.126	28.899	0.986
4	3	24.339	0.096	0.093	818.297	28.606	0.983

Table 4. Performance statistics of single-layered ANN for the production of sugarcane

No. of neurons	MAE	MAPE	SMAPE	MSE	RMSE	Accuracy
1	29.684	0.114	0.108	1158.123	34.031	0.980
2	28.039	0.108	0.103	1070.158	32.713	0.984
3	24.875	0.099	0.093	973.401	31.199	0.980
4	31.539	0.119	0.113	1256.526	35.448	0.974

Table 5. Performance statistics of two-layered ANN for the production of sugarcane

No. of neurons in layer I	No. of neurons in layer II	MAE	MAPE	SMAPE	MSE	RMSE	Accuracy
1	1	26.868	0.104	0.100	1044.420	32.317	0.994
2	1	24.169	0.094	0.091	881.533	29.691	0.996
3	1	47.683	0.176	0.164	3615.573	60.130	0.982
4	1	47.229	0.175	0.163	3589.613	59.913	0.981
1	2	46.808	0.174	0.162	3522.841	59.354	0.980
2	2	46.968	0.174	0.162	3547.216	59.559	0.981
3	2	46.968	0.174	0.162	3536.296	59.467	0.981
4	2	27.021	0.104	0.099	986.675	31.411	0.983
1	3	30.860	0.117	0.111	1202.603	34.679	0.980
2	3	26.439	0.103	0.099	1005.873	31.716	0.991
3	3	45.851	0.171	0.158	3377.152	58.113	0.978
4	3	29.705	0.112	0.108	1108.549	33.295	0.985

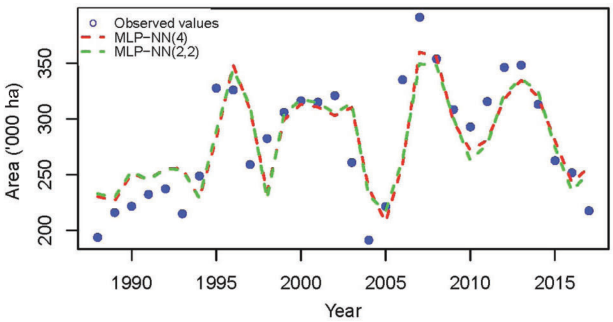

Figure 4. Observed values and predicted models for the area of sugarcane.

are the two best models to predict the area of sugarcane. Figure 3 represents the models MLP-NN(4) and MLP-NN (2,2) respectively, with connection weights. Figure 4 shows the observed values and the two predicted models.

Similarly, Tables 4 and 5 list the statistics used to test the model performance of single- and two-layered ANN respectively for the production data. Among the trained models, MLP-NN(2,1) had the minimum MAE (24.875) and RMSE (31.199) with high accuracy (99.6%), followed by MLP-NN(3). So, MLP-NN(2,1) and MLP-NN(3) are the best models which predicted the observed values with minimum error. Figure 5 shows the connection weights of MLP-NN(3) and MLP-NN(2,1) models. Figure 6 shows the selected prediction models with the observed values.

A series of ANN models was developed to predict the area and production of sugarcane in Tamil Nadu. To test the model performance, statistics such as MAE, MAPE, RMSE and accuracy were used. Models MLP-NN (2,2) and MLP-NN(4)

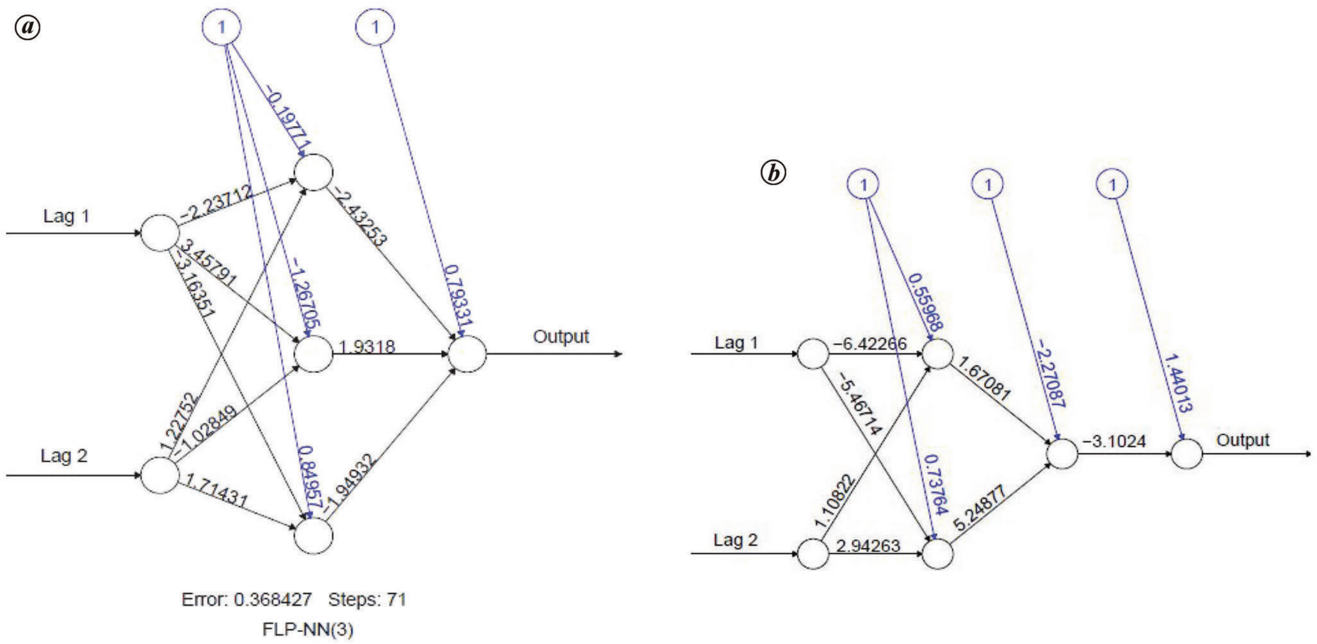

Figure 5. The (***a***) FLP-NN(3) and (***b***) FLP-NN(2,1) model for the production of sugarcane.

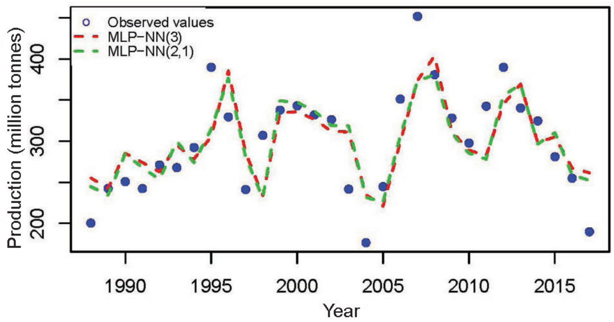

Figure 6. Observed values and predicted models for the production of sugarcane.

predicted the area of sugarcane with minimum error. Among the models used to predict the production of sugarcane, MLP-NN(2,1) and MLP-NN(3) performed better. Collectively, MLP-NN with BP will be a better choice to predict the nonlinear time-series data, as it does not follow any traditional assumptions.

Conflicts of interest: The authors declare that they have not conflict of interest.

1. Suresh, K. K. and Krishna Priya, S. R., Forecasting sugarcane yield of Tamil Nadu using ARIMA models. *Sugar Tech.*, 2011, **13**, 23–26.
2. Solomon, S., Sugarcane agriculture and sugar industry in India: at a glance. *Sugar Tech.*, 2014, **16**, 113–124.
3. Bala, B. K., Ashraf, M. A., Uddin, M. A. and Janjai, S., Experimental and neural network prediction of the performance of a solar tunnel drier for drying jackfruit bulbs and leather. *J. Food Process Eng.*, 2005, **28**, 552–566.
4. Ehret, D. L., Hill, B. D., Raworth, D. A. and Estergaard, B., Artificial neural network modelling to predict cuticle cracking in greenhouse peppers and tomatoes. *Comput. Electron. Agric.*, 2008, **61**, 108–116.
5. Savin, I. Y., Stathakis, D., Negre, T. and Isaev, V. A., Prediction of crop yields with the use of neural networks. *Russ. Agric. Sci.*, 2007, **33**, 361–363.
6. Rahman, M. M. and Bala, B. K., Modelling of jute production using artificial neural networks. *Biosyst. Eng.*, 2010, **105**, 350–356.
7. Sanghani, A., Bhatt, N. and Chauhan, N. C., A review of soft computing techniques for time series forecasting. *Indian J. Sci. Technol.*, 2016, **9**, 93–102.
8. Hornik, K., Stinchcombe, M. and White, H., Multilayer feedforward networks are universal approximators. *Neural Networks*, 1989, **2**, 359–366.
9. Maier, H. R. and Dandy, G. C., Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Model. Softw.*, 2000, **15**, 101–124.

Received 16 June 2022; accepted 9 December 2022

doi: 10.18520/cs/v124/i4/500-504