Comparison Study Using Arima and Ann Models for Forecasting Sugarcane Yield

Ali J. Ramadhan[1*], *S. R.* Krishna Priya[2], *N* Naranammal[2], *S.*Pavishya[2], *K.* Naveena[3], *Soumik* Ray[4], *P.*Mishra[5], *Mostafa* Abotaleb[6]*, Hussein* Alkattan[6] and *Zainalabideen* Albadran[1]

[1]University of Alkafeel, Najaf, Iraq
[2]PSG College of Arts and Science, Tamil Nadu, India
[3]Centre for Water Resources Development and Management, Kozhikode, India
[4]Centurion University of Technology and Management, Odisha, India
[5]JNKVV College of Agriculture, Rewa, India
[6]South Ural State University, Chelyabinsk, Russia

Abstract. Sugarcane is the largest crop in the world in terms of production. We use sugarcane and its byproducts more and more frequently in our daily lives, which elevates it to the status of a unique crop. As a result, the assessment of sugarcane production is critical since it has a direct impact on a wide range of lives. The yield of sugarcane is predicted using ARIMA and ANN models in this study. The models are based on sugarcane yield data collected over a period of 56 years (1951-2017). Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) have been used to analyze and compare the performance of different models to obtain the best-fit model. The results show that the RMSE and MAPE values of the ANN model are lower than those of the ARIMA model and that the ANN model matches best to this data set.

1 Introduction

An estimated 70 percent of India's population is engaged in some form of agriculture, earning a living through farming. During the past few years, the population has grown significantly, which poses the question of how much food is available for everyone. Therefore, the production slope must be on an ascending scale. For the first few years, improvements in agricultural technology led to higher yields. Fertilizer and pesticide use have made the issue of crop production more difficult after a few years. As a result, increasing crop yields is a top priority. Sugarcane is a major source of income in India, both directly and indirectly. Sugarcane is a difficult crop to cultivate because of a variety of factors, includingthetype of soil, climate, irrigation, fertilizers, insects, disease control, and varieties, which all affect its productivity [1].

As a result, determining how much sugarcane can be produced in a given environment is critical because it affects so many different kinds of life. In this study, the accuracy of ARIMA and ANN models in predicting Sugarcane yield was compared.

[*] Corresponding author: ali.j.r@alkafeel.edu.iq

2 RELATED WORKS

A comparative study was made between ARIMA and ANN models with published stock data [2]. ANN and ARIMA models were used to compare and forecast the electricity price [3]. A similar comparison has been made between ANN and ARIMA models in terms of both modeling and forecasting [4-6]. The ANN and ARIMA techniques compared the two developed hybrid models [7].

Conjugate Gradient is one of the popular optimization practices used in Artificial Neural Networks to improve learning algorithms [8]. In [9] discussed the various advantages of Artificial Neural networks over Conventional approaches.The monthly wholesale prices of soybeans and rapeseeds were used to demonstrate the effectiveness of artificial neural networks compared to linear model methods [10]. A model based on the ANN has been used to predict the yield of rapeseeds in winter at the beginning of the season [11]. ANN architectures such as BPN and RBFN are well adapted to forecast monsoon rain and other weather factors [12]. A non-linear Artificial Intelligence technique like a neural network can be an excellent strategy to improve predicting ability for time series with a non-linear structure [13].

DATA

All India sugarcane data for 1951-2017 (tonne/hectare) were collected from “www.indiastat.com” The 1951–2012 data were used for modelling, while the remaining five years from 2013–2017 data were used for model validation.

2.1 FORECASTING MODELS

Auto-Regressive Integrated Moving Average (ARIMA) Model

Box and Jenkins introduced the ARIMA model in 1970. For the ARIMA model, identifying the stochastic process and properly predicting the future values are the primary goals. The development of discrete-time series and dynamic system models also benefits from the application of these methodologies. The ARIMA model’s symbols are (p,d,q), and the sequence of auto-regression (AR), integration (difference), and motion average (MA) are (p,d,q).

ARIMA modeling is only applicable to stationary series, hence stationarity is the initial step. Mean and variance is stationary when they remain constant in a non-stationary series. Using differencing, logarithmic transformation stabilizes variance. After conversion and difference, several ARMA models are selected that correspond closely to the data. ACF and PACF determine p and q. In ARIMA, ACF and PACF determine q and p. ACF degrades quickly at zero latency. For stationary time series, ACF will disappear. Selecting the orders of p and q from the non-negative values, creating several ARIMA models, and determiningφ (B) and θ (B) are then determined.

Once the model parameters are determined, a diagnostic test is performed.If the residuals from the adapted model are normally distributed and uncorrelated, then the structure of the model and all the coefficients can be used to estimate the prediction.The ARIMA method allows for evaluating only a few metrics viz. RMSE and MAPE.

2.2 Artificial Neural Network (ANN) Model

Neural networks are virtual networks of basic processor neurons that are interconnected and model the CNS of the brain. Depending on the retention time, a dynamic neural network's topology integrates long-term or short-term memory. Using time delay at the neural network's input layer can enhance short-term memory.

The use of transit detection, pattern recognition, approximation, and time sequence prediction are some of the uses of ANN.ANNs are basic processing units connected to networks that are alternatives to traditional computer approaches. ANNs learn from sample data instead of modeling computing processes to display related things. For time series prediction, determine the number of layers and nodes in each layer. Such parameters can only be found by hands-on testing. A one-layer neural network may approximate any non-linear function with enough hidden nodes and training data. This study used a hidden-layer neural network. In time series analysis, the number of lagged input nodes represents the data's autocorrelation structure. Easy to compute output nodes. The study used one output node. Most neural networks use multilayer perceptrons (MLP). It has any number of inputs, hidden units, outputs, and activation function links between input layers. Artificial neural networks rely on learning and training algorithms to update connections and establish weights. Learning algorithms change a network's weight and bias. Adjustment algorithms help it do what you desire. Both learning and training algorithms drive future changes in an ANN (ANN). Synaptic weights are determined using a scaled conjugate gradient. Models are analyzed using RMSE and MAPE.

3 RESULTS AND DISCUSSION

The yearly data of yield of sugarcane are from 1931 to 2012 for model building. The later five years till 2017 data were used for the validation of the result.

3.1 ARTIFICIAL NEURAL NETWORK

The MLP was used to create the neural network model (Multi-layered Perceptron). The input layer consists of 10 nodes, and the hidden layer consists of 1 to 5 nodes and the structure of the neural network. The model of neural networks consisting of 10 input nodes and 5 hidden nodes is the most efficient. (10:5:1i). No examples were omitted from the 78-case study since 62 (79.5%) of the cases were used for training and 16 (20.5%) for testing.

Table 1 shows 10 node input layers, four hidden layers, and an output layer with only one node. The hidden layer's activation function is a hyperbolic tangent. The network's output layer uses the identity activation function and the total square error function.

Table 1. Network Information

Input Layer	Covariates	1	yone
		2	ytwo
		3	ythree
		4	yfour
		5	yfive
		6	ysix
		7	yseven
		8	yeight
		9	ynine
		10	yten
	Number of Units[a]		10
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		4
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	y
	Number of Units		1
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares
a. Excluding the bias unit			

Synaptic weights larger than zero are shown in bright color, while synaptic weights less than zero are shown in dark color in figure 1 depicting the network architecture. Table 2 details the results of the training and how they were put to use on the actual testing sample. The goal of the network is to minimize the sum of squares,the error function in the training phase. SS Error for the Sugarcane Yield Variable / SS Error for the "Null" Model, where the Mean of the Sugarcane Yield Variable is utilized as the anticipated values. 0.046 represents the training sample's relative error and 0.011 represents the testing sample's relative error.

Table 2. Model Summary

	Training	Sum of Squares Error	1.398	
		Relative Error	.046	
		Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]	
		Training Time	0:00:00.00	
	Testing	Sum of Squares Error	.099	
		Relative Error	.011	
Dependent Variable: Y				
a. Error computations are based on the testing sample.				

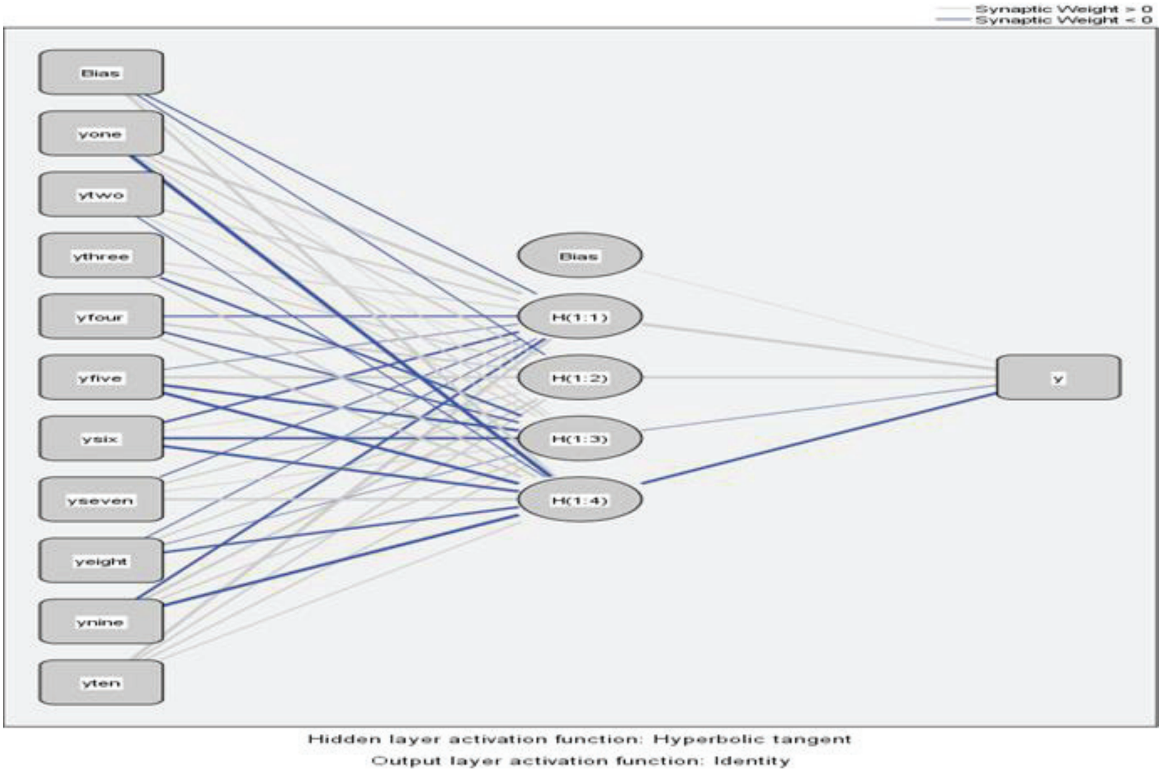

Fig. 1. Architecture of the network

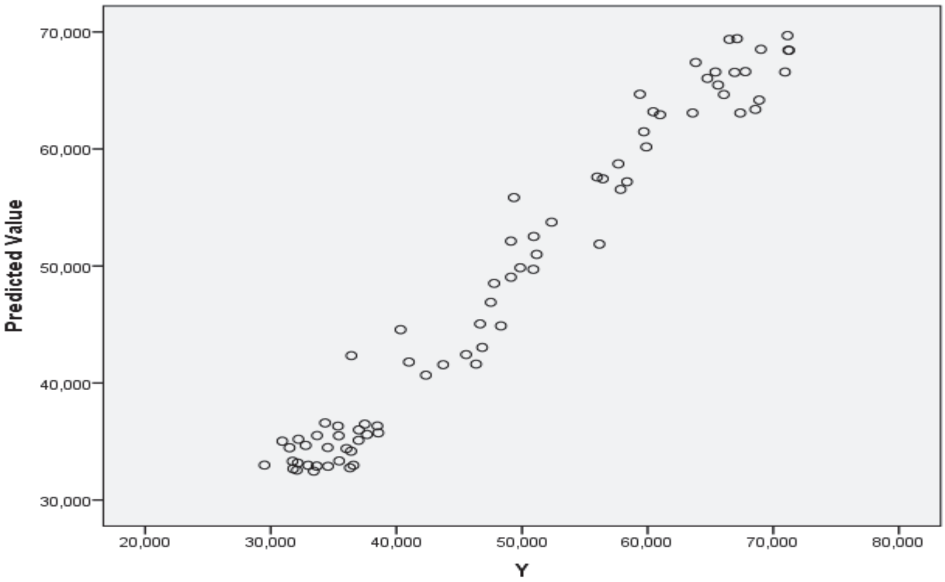

Fig. 2. Predicted value Vs Observed value

Figure 2 depicts the scatterplot's one-to-one relationship between predicted y-axis values and actual x-axis values. Although there are some outliers, this graph indicates a straight line, and results from both trainings as well as testing samples are combined. We can infer from the data shown above that the ANN's performance meets our standards. As shown in Figure 3, there is no clear correlation between the residual and the expected values.

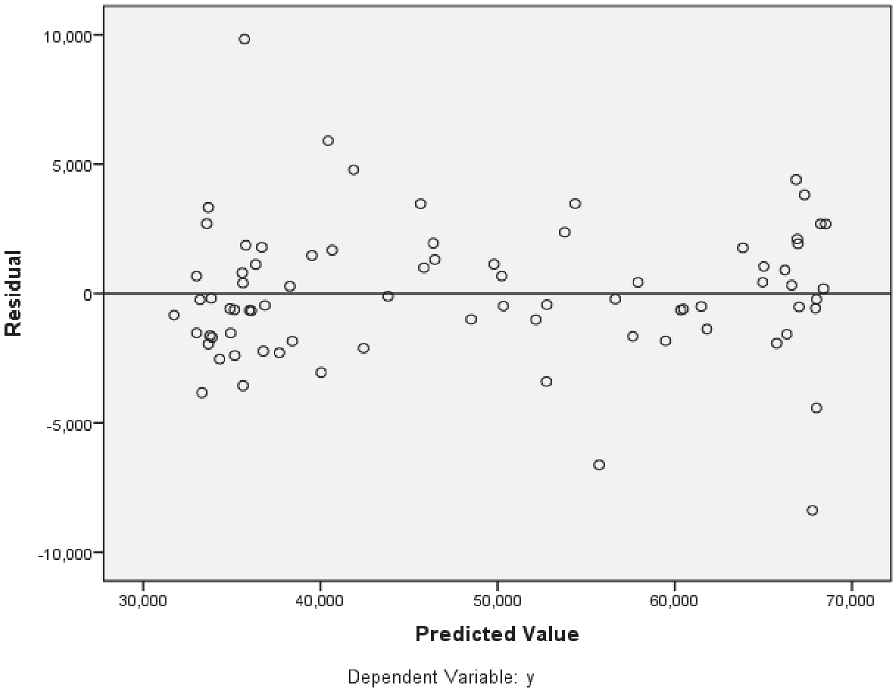

Fig. 3. Predicted value Vs Residual

3.2 ARIMA

A five-year out-of-sample forecast was made using the ARIMA model. Using a sequential numbering system, Figure 4 depicts the productivity of sugarcane from 1931 to 2008. The positive trend over time in the figure shows that the series is non-stationary. Both ACF and PACF (partial autocorrelation functions) confirm this. Figure 5 demonstrates a linear decrease of the autocorrelation coefficients in the ACF time series, while the PACF time series indicates a large PACF at lag 1. It is a sign of time series nonstationarity. The first-order difference has been used to keep the series stationary.

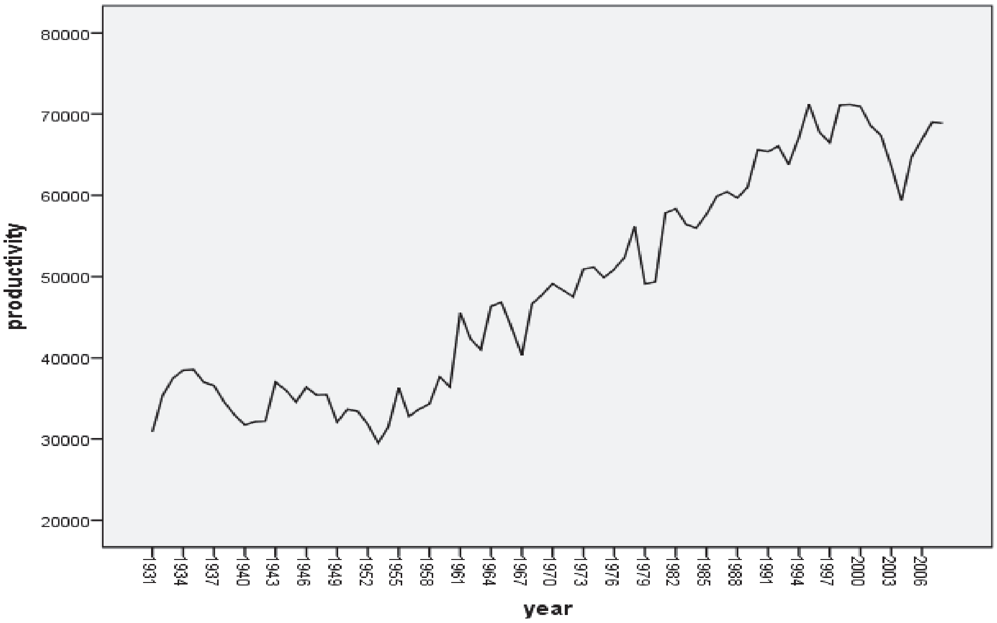

Fig. 4. The time plot of yield of Sugarcane

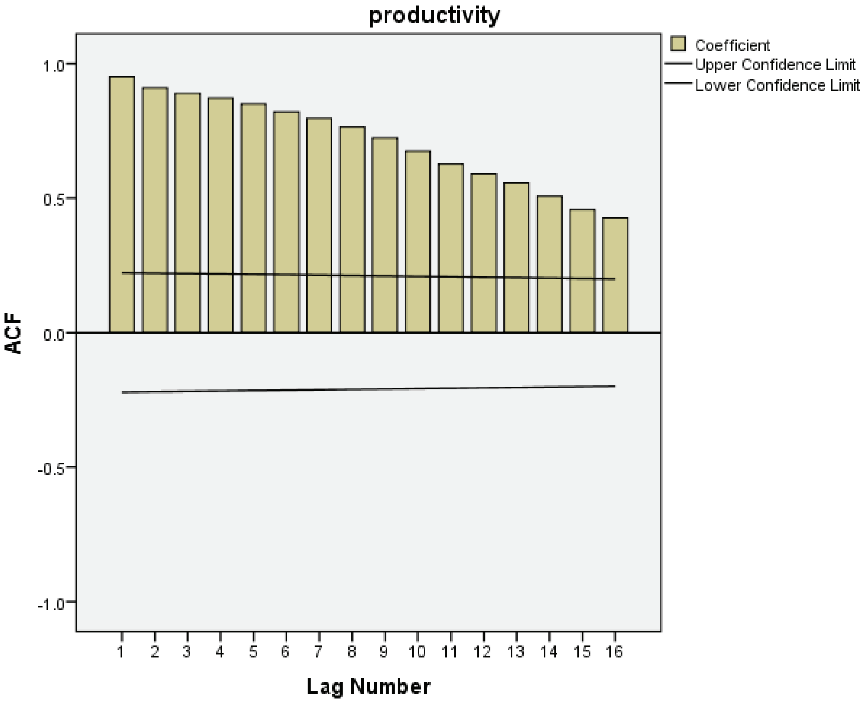

Fig. 5. Autocorrelations at different lags of productivity of sugarcane

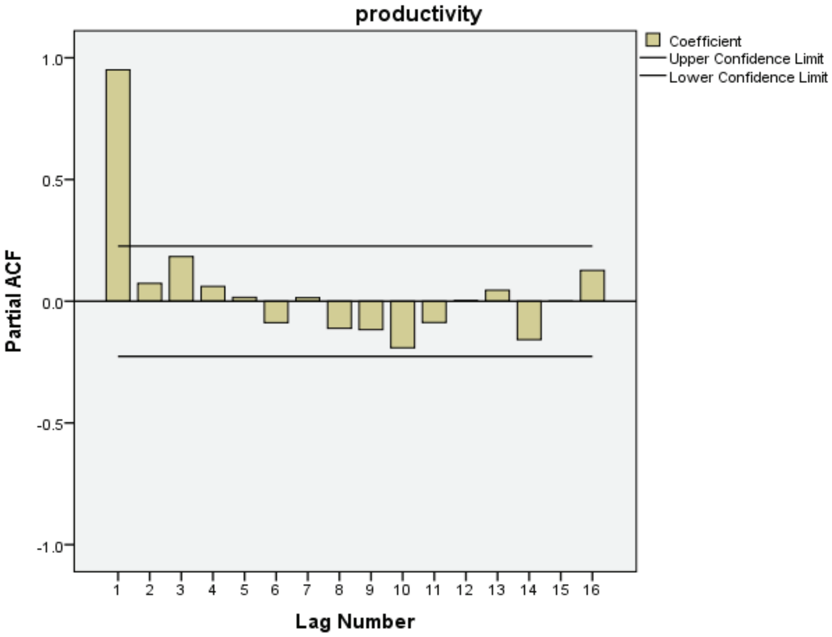

Fig. 6. Partial Autocorrelations at different lags of sugarcane in India

Figure 7 shows a time plot of the series, clearly showing that the series has no automation.

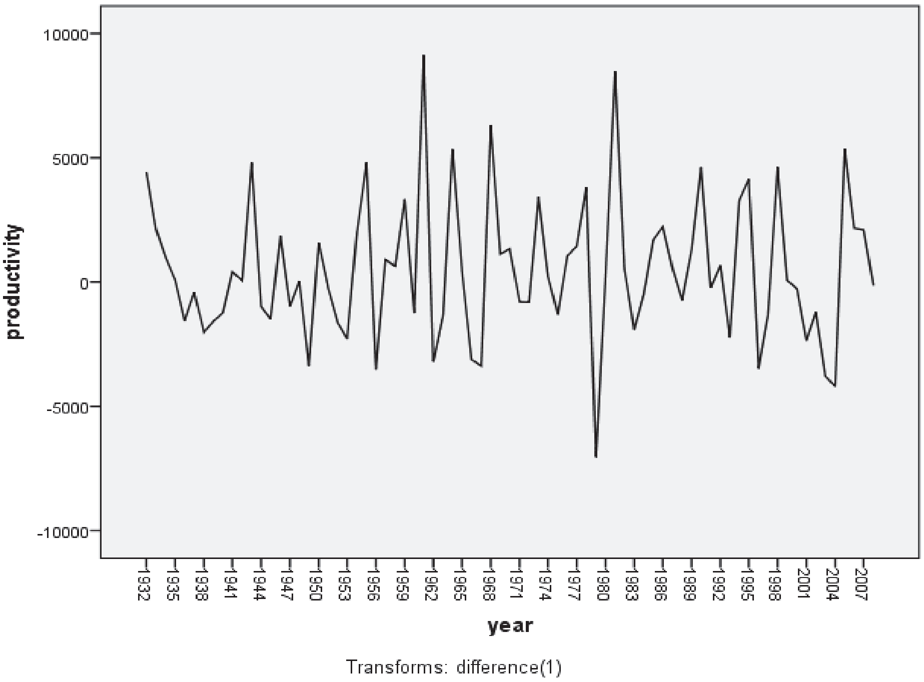

Fig. 7. The time plot of the differenced series of productivity of sugarcane

Figures 8 and 9 show the autocorrelation functions and the partial autocorrelation functions of differenced series. Here in both the ACF and PACF plot most of the lags are non-significant, so it is concluded that mean of the series is stationary.

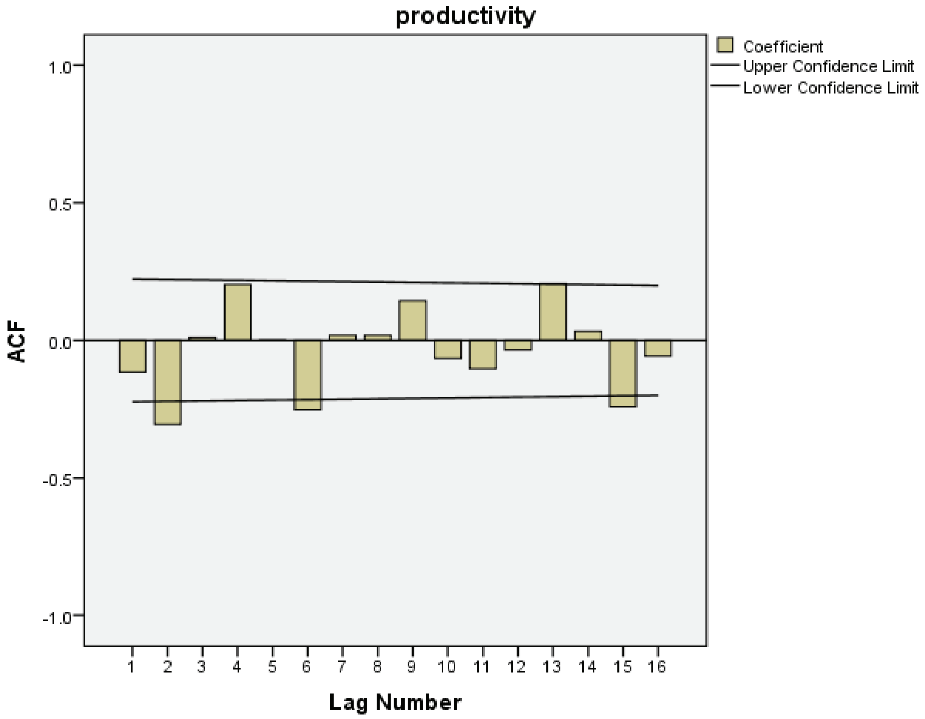

Fig. 8. Autocorrelations at different lags of 1st differenced time series of productivity of sugarcane

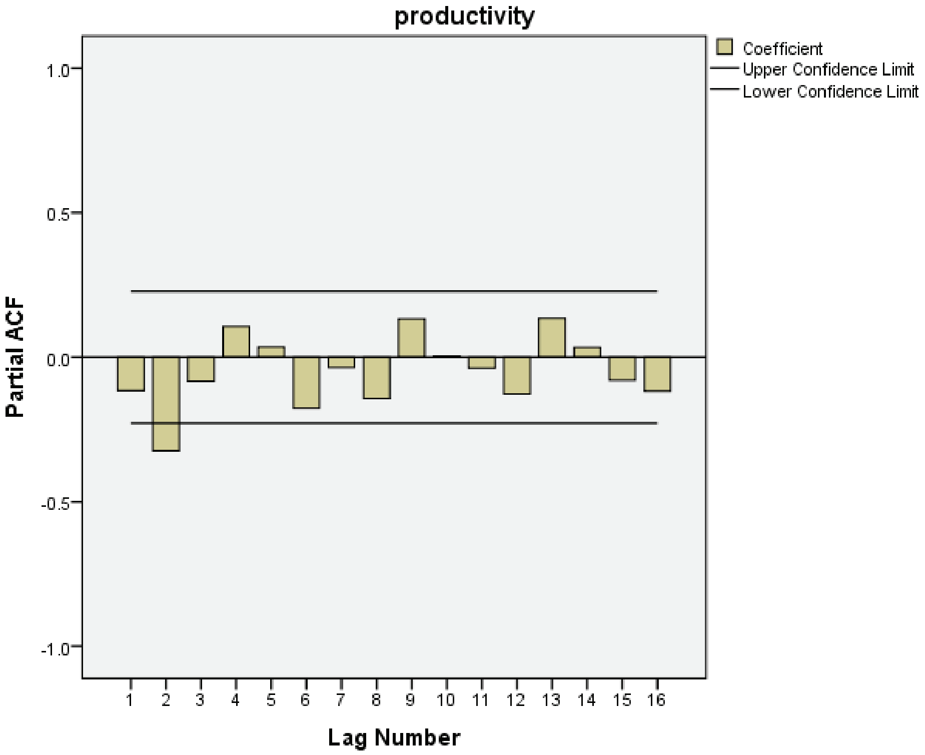

Fig. 9. Partial Autocorrelations at different lags of 1st differenced time series of productivity of sugarcane

Once the time series becomes stationary, the ARIMA model is estimated. In particular, the ACF has one significant spike, while the PACF has no significant spike. Therefore, the difference series shows the ARIMA model (0, 1, 1) as the best of the ARIMA model family. Model parameters are given in Table 4. In model-fit statistics (table 3), ARIMA

(0,1) was rated the best model. This model satisfies the invertibility requirement and the moving average (MA) was found to be statistically significant at the given level of significance.

Table 3. Model fit statistics and Ljung-Box Q statistics for the yield of sugarcane

Model Fit Statistics				Ljung-Box Q(18)			No Outliers
Stationary R-squared	RMSE	MAPE	Normalized BIC	Statistics	DF	Sig.	
.035	2886.995	4.854	16.049	32.215	17	.011	0

Table 4. Model Parameters for the yield of sugarcane

	Estimates	SE	t	Sig.
Constant Difference MA Lag 1	476.135	234.680	2.029	0.46
	290	0.111	2.610	0.011

Residual analysis have been conducted to verify the model adequacy.The residuals of the ACF and PACF were obtained from an experimentally identified model, they were found most of all lags are nonsignificantgiven in the figure 10. So we can state that model is adequate.

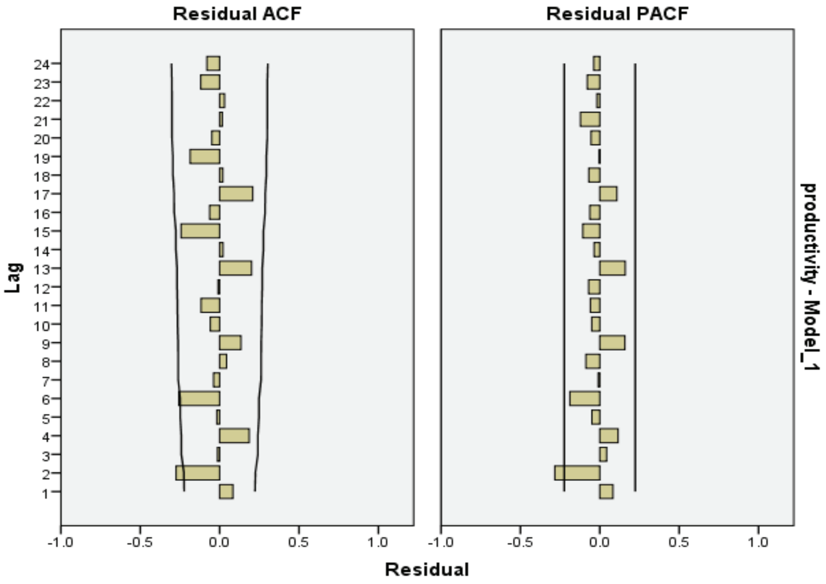

Fig. 10. Residual auto and partial autocorrelations for Indian Export of Indian coffee

The model's adequacy was also determined based on Box-Pierce Q statistics values in Table 5 and was found to be non-significant. Thus, ARIMA (0,1,1) models can be considered satisfactory among different ARIMA models and can be used to forecast sugarcane yields in India.

COMPARISON OF ARIMA AND ANN:

Table 5. Comparison of ARIMA and ANN

YEAR	OBSERVED VALUES	PREDICTED VALUES OF ANN MODEL	PREDICTED VALUES OF ARIMA MODEL
2013	71668	68016	70764
2014	68254	67561	71240
2015	70522	68280	71716
2016	71511	68171	72192
2017	70720	68929	72668
MAPE		2.48	3.155
RMSE		2017.99	2520.506

The RMSE and MAPE values of ANN are 2017.99 and 2.48respectively.The RMSE and MAPE values of ARIMA are 2520.506 and 3.155 respectively. The values above are determined to find the best model. The data was forecasted for the next five years(2013-2017 validation data).In the ANN model, the deviation between the observed and predicted value is minimum.Also,ANN has the least MAPE and RMSE values when compared to the ARIMA model.

4 CONCLUSION

For the prediction of sugarcane yields in India, the ARIMA and ANN models have been described and evaluated in this study. The ANN model outperforms the ARIMA model in terms of RMSE and MAPE forecast accuracy in this study.

REFERENCES

1. Adebiyi, A.A., Adewumi, A.O., and Ayo, C.K., (2014). Comparison of ARIMA and Artificial Neural Networks Models. Journal of Applied Mathematics.
2. Farizawani, A.G., Puteh, M., Marina, Y., Rivaie, A., (2020). A review of artificial neural network learning rule based on multiple variants of conjugate gradient approaches. Journal of Physics: Conference Series.
3. Al-Mahdawi, H. K., Albadran, Z., Alkattan, H., Abotaleb, M., Alakkari, K., & Ramadhan, A. J. (2023, December). Using the inverse Cauchy problem of the Laplace equation for wave propagation to implement a numerical regularization homotopy method. AIP Conference Proceedings (Vol. **2977**, No. 1). AIP Publishing.
4. Hansen, J.V., Mcdonald, J.B., Nelson, R.D.,(1999). Time series prediction with genetic-algorithm designed neural networks. Computational Intelligence. **15**(3): 171-184.
5. Ehsan khodadadi, S. K. Towfek, Hussein Alkattan. (2023). Brain Tumor Classification Using Convolutional Neural Network and Feature Extraction. Fusion:Practice and Applications, **13**(2), 34-41.
6. Kumar, P., Sharma, P., (2014). Artificial Neural Networks-A Study. International Journal of Emerging Engineering Research and Technology. **2**(2): 143-148.
7. Mahalingaraya, Rathod, S., Sinha, K., Shekhawat, R.S., Chavan, S.,(2018). Statistical Modelling and Forecasting of Total Fish Production of India: A Time Series Perspective. International Journal of Current Microbiology and Applied Sciences. **7**(03): 1698-1707.
8. Merh, N., Saxena, P.V., Pardasani, K.R., (2010). A comparison between hybrid approaches ofANN and ARIMA for Indian stock trend forecasting. Business Intelligence Journal. **3**(2): 22-43.
9. Akbari, E., Mollajafari, M., Al-Khafaji, H. M. R., Alkattan, H., Abotaleb, M., Eslami, M., & Palani, S. (2022). Improved salp swarm optimization algorithm for damping controller design for multimachine power system. IEEE Access, **10**, 82910-82922.
10. Niedbala, G., (2019). Simple model based on Artificial Neural Network for early prediction and simulation of winter rapeseed yield. Journal of Integrative Agriculture. **18**(1): 54-61.
11. Niedbala, G., Kozlowski, R.J., (2019). Application of Artificial Neural Networks for Multi-Criteria Yield Prediction of Winter Wheat. Journal of Agricultural Science and Technology. **21**: 51-61.

12. Al-Nuaimi, B. T., Al-Mahdawi, H. K., Albadran, Z., Alkattan, H., Abotaleb, M., & El-kenawy, E. S. M. (2023). Solving of the inverse boundary value problem for the heat conduction equation in two intervals of time. Algorithms, **16**(1), 33.
13. Yao, J.T., Tan, C.L., Poh, H.L.,(1999). Neural networks for technical analysis: a study on KLCI. International Journal of Theoretical and Applied Finance. **2**(2): 221-241.