

International Journal of Agriculture Extension and Social Development

Volume 9; Issue 4; April 2026; Page No. 186-192

Received: 18-01-2026
Accepted: 21-02-2026

Indexed Journal
Peer Reviewed Journal

A hybrid time series modeling framework for banana area and production in Assam

¹Akankhya Sandilya, ¹Borsha Neog, ²Horindra Gogoi, ³Pallabi Bora, ¹Smrita Barua, ⁴Tonmoy Neog

¹Department of Agricultural Statistics, Assam Agricultural University, Jorhat, Assam, India

²Department of Agricultural Economics, Assam Agricultural University, Jorhat, Assam, India

³Department of Extension Education, Assam Agricultural University, Jorhat, Assam, India

⁴Department of Geophysics, Banaras Hindu University, Varanasi, Uttar Pradesh, India

DOI: <https://www.doi.org/10.33545/26180723.2026.v9.i4c.3468>

Corresponding Author: Borsha Neog

Abstract

Prediction generation is very crucial in the field of agriculture too because it enables to get the idea of future events. Such an accurate forecast is very useful for policy making and farm planning. In this research, banana area, banana production data of Assam were taken as variables, their trend analysis was done and ultimately forecasting models were developed for all four series using one linear and two nonlinear models. Individual model building was succeeded by hybrid model development. The performances of all models were compared using root mean square error (RMSE) measure. The performance of hybrid ARIMA-ANN model was much superior than other models for both banana area, banana production series. Therefore, hybrid ARIMA-ANN model was considered as the best model for banana area & banana production. Finally, forecast results were generated using the best forecasting models for the subsequent five years. The forecast results unveiled that both production of banana and cultivable area of banana will increase in future.

Keywords: ARIMA, ANN, NLSVR, BDS, hybrid

1. Introduction

Application of forecasting models for crop production becomes a very trending topic now a days. A good forecasting model enables to take accurate decisions at appropriate time. Many studies had been conducted earlier in this area that provided very precise and accurate results of forecasting. A good forecasting model is not only useful for policy makers but also for farmers to take appropriate action during farming activities. In this research, area and production data of Assam of banana were attempted to fit with linear and nonlinear models. Banana is an important fruit crops of India as well as Assam. India is the leading banana producing country in the world whereas Assam is among the major banana producing state of India. The climatic condition of Assam is also suitable for banana production. However, the banana production of Assam is comparatively less than other banana producing states of the country. The main objectives of this research were to analyze the area and production series of banana through trend analysis, enumerate Compound Annual Growth Rate and at last building of forecasting model for both series. Trendline analysis helps to identify the trend pattern exhibited by the series. This was followed by forecasting model development for Assam because a region-specific model can give more accurate results as compared to a nationwide model. For forecasting model development, one linear model (ARIMA) and two non-linear model (ANN & NLSVR) were selected. Many studies have been conducted earlier that proved the usefulness of these models to get

accurate forecasts. Ramadhan *et. al* (2024) carried out research comparing the ARIMA and ANN models for forecasting sugarcane yield. This study proved that individual ANN model outperformed the ARIMA model in capturing the sugarcane yield data series. Although individual model's performances were observed to give good performance, however they have certain limitations as well. For instance, linear model can explain only those datasets well which follows approximately a linear relationship whereas neural networks may sometimes tend to overfit a series. Zhang *et. al* developed hybrid approach in 2003, which revolutionized this field. Hybrid approach combined the linear model and nonlinear model simultaneously which overcame the drawbacks of individual model. Neog *et al.* (2023) proved that hybrid ARIMA-SVM model outperformed individual models in explaining the autumn rice production series. Rathod *et al.* (2017) compared the performance of ARIMA, TDNN, NLSVR model and their hybrid for forecasting banana production in Karnataka. The results revealed that hybrid ARIMA-NLSVR model performed superiorly than remaining models. Thus, in this research hybrid approach was also adopted to build good forecasting models.

Data Source: The two-variables considered in this study are banana area, banana production. The required data of area and production of Assam was collected from "www.indiastatagri.com", "Statistical Handbook of Assam"

(Directorate of Economics and Statistics, Assam). Banana area and production data was collected from 1961-62 to 2023-24. The whole analysis was conducted using two software- Rstudio and SPSS. 'Forecast' package of Rstudio software was utilized to build the ANN model whereas 'TSSVM' package was used to develop NLSVR models. BDS test was conducted using 'tseries' package of Rstudio software.

2. Research Methodology

Auto-Regressive Integrated Moving Average Model (ARIMA)- ARIMA is a linear time series model, was developed by Gorge Box and Gwilyn Jenkins in 1970 (Box and Jenkins, 1970). It mainly consists of three components namely autoregressive (AR), integrated (I) and moving average (MA). AR component explains the effect of previous lags upon current observation while MA component refers to the influence of random shocks of previous lags on current observation. On the other hand, differencing component records the times of differencing conducted on non-stationarity series to transform it to a stationary series. These three components were combined to get the typical ARIMA (p, d, q) model is given by,

$$\phi_p(B)(1-B)^d X_t = \alpha + \theta q(B)\varepsilon_t$$

α = constant

X_{t-i} = lag values

ϕ = coefficients of AR model

ε_t = random shock values

θ = coefficients of MA mode

Box and Jenkins put forwarded four steps of ARIMA modelling; identification stage, estimation stage, diagnostic checking stage and forecasting stage. In the identification step, appropriate orders of AR, I, MA (i.e. p, d and q) were identified. The value of 'd' was determined through Augmented Dicky Fuller (ADF) test. Values of p and q were determined through PACF (Partial Auto-Correlation Function) and ACF (Auto-Correlation Function) plot of the differenced data. In the estimation phase, coefficients of the identified model were estimated using method of maximum likelihood or iterative least square method. Here performance of selected model was compared to some other

related models and best model was promoted to next step. Third stage includes the adequacy checking of model through residual analysis. If the residual series follows the characteristic of white noise, that ARIMA model is regarded as good fit and was promoted to fourth step. In the fourth step, forecasts were generated using the selected ARIMA model.

Artificial Neural Network (ANN): It is a generalized model of non-linear functions. The artificial neural network specially used for time series analysis is called time delay neural network (TDNN). It is a supervised model comprises of three layers; input layer, hidden layer, output layer. Each layer comprises of one or more nodes or neurons. Neurons or nodes are the fundamental block of neural networks. Node of ANN is composed of synapse, adder & activation function. Three important hyperparameters of ANN are; number of hidden layers, activation function & training epochs. The primary goal of this model is to minimise the error by adjusting the weights and hyperparameters. This model operates using Marquardt Levenberg backpropagation algorithm. The structure of time delay neural network is defined by I:Hs:Ol where I is the number of nodes in the input layer, s is the sigmoid activation function of input layer, O is the number of nodes of the output layer and its activation function is linear denoted by l. The mathematical form of this model is given by-

$$y_t = f \left(\sum_{j=0}^s w_j g \left(\sum_{i=0}^q w_{ij} y_{t-1} \right) \right)$$

Where, Y_t = Time series

- w_{ij} = weight from i th node to j th node
- w_j = weight added on input to i th node
- p = number of input nodes
- q = number of hidden nodes
- g and f = activation function

As there is no predefined procedure or hard & fast rule to determine best ANN model for a series, thus trial & error method was adopted. For this 80% data was used to develop the model and 20% data was used evaluate the model. Different models were developed and their performances were recorded through RMSE and MAPE values for each model. The model, exhibiting lowest error was selected as best ANN model for that series.

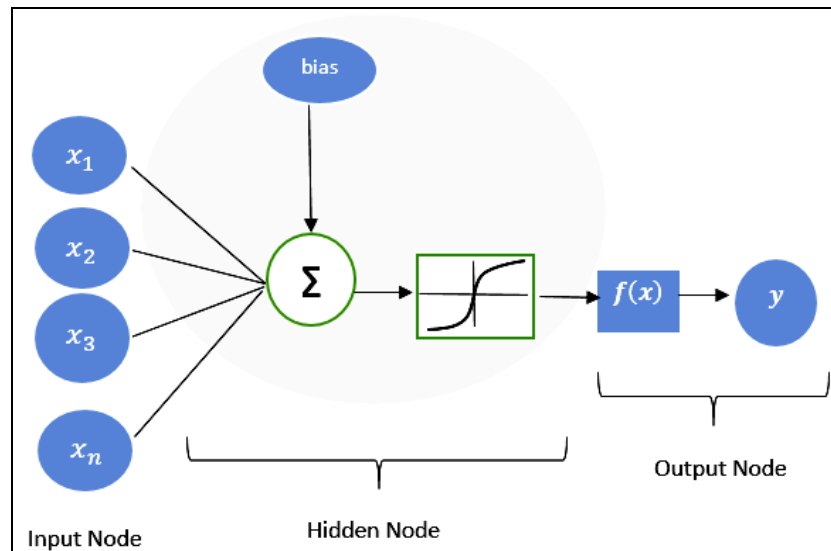


Fig 1: ANN Architecture

Non-linear Support Vector Regression (NLSVR): Non-Linear Support Vector Regression (NLSVR) is a Supervised Machine Learning model, which was developed by Vapnik in 1997 (Vapnik, 1997). In this model initially the series was transferred to a high dimensional feature space by kernel function and then assumes a non-linear regression plane in that feature space. This regression plane is called optimal hyperplane and the data points were called as support vectors. The tolerance area around the plane is called epsilon tube. The gap between epsilon tube boundary to the hyperplane is called margin. The distance of those support vectors that fall outside the epsilon tube from the tube boundary is defined as slack variables. The principle of this model is to determine the coefficients of optimal hyperplane in such a way that it fits most of the support vectors within the epsilon tube by balancing the length of margin (represented by regularised term) and prediction error outside the epsilon tube (called empirical error). Regularised term and empirical error term is combined to form the regularised risk function which is minimised with

respect to the unknown parameters of optimal hyperplane to get the optimised estimates of the coefficients. The general expression for optimal hyperplane function be,

$$f(x) = W^T \phi(x) + b$$

Where, $\phi(\cdot): R_n \rightarrow R^{nh}$ = non-linear kernel function which map the original input space into high dimensional feature space

- $W \in R^{nh}$ = weight vector
- W^T = transpose of weight vector
- b = bias

The most common kernel function used in NLSVR models is Radial Basis Function (RBF). The mathematical form of RBF kernel function is given as,

$$k(x_i, x_j) = \exp(-\gamma \|x - x_i\|^2); \text{ where } \gamma = \text{kernel bandwidth parameter.}$$

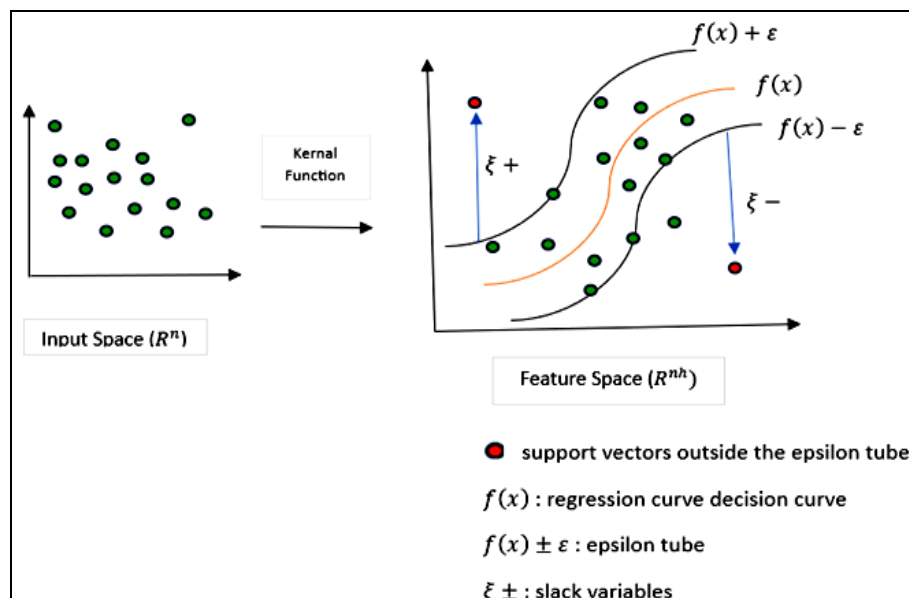


Fig 2: NLSVR Model

Hybrid model: hybrid approach was developed by Zhang in 2003. This model assumes that each series is a combination of linear and nonlinear component. In this approach linear model and nonlinear model were combined to capture the linear as well as nonlinear component simultaneously. Let us consider Y_t is a time series. The overall methodology of this model is divided in three parts. In the first part, linear model is applied over Y_t . This model fits the linear part and gives the prediction. Let the estimated series provided by the linear model is denoted by $\hat{L}_t$. In the

second part, the non-linear residuals ($e_t = Y_t - \hat{L}_t$) found from the linear model are modelled using nonlinear models. Predictions of the nonlinear model are also collected and are denoted by $\hat{N}_t$. Finally, the linear and non-linear series are combined to make the hybrid forecast. This relationship is given by,

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

$\hat{Y}_t$ is the required hybrid series of the initial series Y_t .

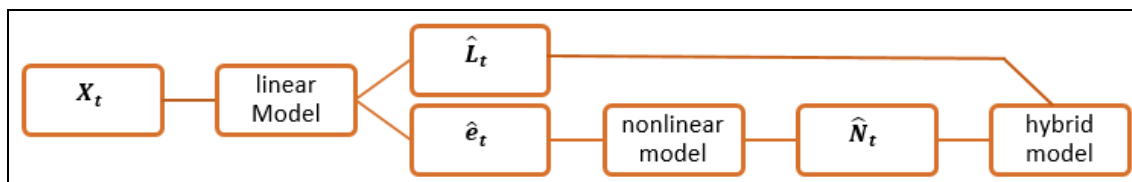


Fig 3: Hybrid Model

Compound Annual Growth Rate (CAGR): CAGR is defined as an average rate at which a quantity grows every year. In other words, CAGR is an annual growth rate at which a variable grows constantly over a number of years. The formula of CAGR is given by,

$$CAGR = \left(\frac{V_e}{V_b} \right)^{1/n} - 1$$

Where, V_e = ending value

V_b = beginning value

n = number of years

3. Results and Discussion

Initially trend analysis was conducted on the two series namely banana area, banana production. Method of curve fitting was adopted using four different curves namely

linear, quadratic, logarithmic and exponential. Each curve represents different types of relationship. As for example, those variables that increases approximately at a constant rate generally follows a linear trend, whereas variables increasing at a moderate or slightly higher growth rate (increasing) than linear model is well explained by a quadratic model. Similarly, the growth rate of a variable (increasing) following an exponential model is generally considered as faster than both linear and quadratic model. On the other hand, if trendline of a variable is logarithmic, then it indicates that the variable acquired an increasing trend with a faster decreasing rate. Each series were fitted with all the aforesaid functions, then performance of each model was recorded through coefficient of determination (R^2) and root mean square error (RMSE) value. Finally, the trendline model possessing lowest error and highest R-square value were selected as best model. The best trendline model of variables are given in Table-1.

Table 1: Description of best trendline model of Banana Area and Production series

Variables	Equation	R-square	RMSE
Banana Area	$y = -0.0022x^2 + 0.7926x + 12.685$	0.938	3.057
Banana Production	$y = 0.064x^2 + 7.1812x + 183.34$	0.7986	62.674

Trendline analysis was followed by calculation of Compound Annual Growth Rate of both the series. The CAGR percent of banana area was found 1.82 while 2.48% was found for banana production during the assessed period of 63 years.

This was succeeded by forecasting model development. All the four series were fitted with ARIMA initially. In the identification step, orders of ARIMA model were identified. Stationarity of each series was tested through ADF test and ACF-PACF plot. Results revealed that second differenced series of the two variables (banana area, banana production) are stationary in nature. Therefore, the order of differencing 'd' is taken as 2. To determine the order of AR and MA, the ACF-PACF plot of second differenced series was observed to get the number of significant spike present in each plot. Value of 'p' was determined through PACF plot while value

of 'q' was selected through ACF plot. One significant lag was found both in ACF and PACF plot. Therefore, from all the visual observation ARIMA (1,2,1) was identified as the potential model for banana area and production series. In the estimation step, coefficients of potential ARIMA models were estimated. After this, the performances of the selected models were compared to some other related ARIMA model to examine the validity of selection. The best models were subsequently promoted to next step of diagnostic checking. Here, residual analysis was conducted using Ljung box test and residual ACF-PACF plot. The results of Ljung box test of all the models were given in table-4. These results (p-value > 0.05) depicted that the residual series of each model is white noise (no autocorrelation), thus the respective models were regarded as good fit. Similar result was also found in residual ACF-PACF plot where no significant lags

were observed. Therefore, both the good fit models were promoted to the final step of forecasting. In this step,

forecasts were generated for the upcoming five years using the models.

Table 2: ADF test results of Banana Area and Production Series

Variables	Series	Dicky Fuller Statistic	lag	p-value
Banana Area	1 st differenced series	-2.459	3	0.39
	2 nd differenced series	-6.645	3	0.01
Banana Production	1 st differenced series	-1.8968	3	0.62
	2 nd differenced series	-6.6281	3	0.01

Table 3: Description of best ARIMA models of Banana Area and Production series

Variables	Best Models	AR coefficient	p-value	MA coefficient	p-value
Banana Area	ARIMA (1,2,1)	-0.325	0.05	1.000	0.9
Banana Production	ARIMA (1,2,1)	-0.311	0.1	0.866	0.0

Table 4: Ljung Box test result of ARIMA models of Banana Area and Production series

Variables	Best Models	Ljung Box	lag	p-value
Banana Area	ARIMA (1,2,1)	9.803	16	0.88
Banana Production	ARIMA (1,2,1)	12.374	16	0.79

After ARIMA modelling, the BDS (Brock Dechert Scheinkman) test was conducted on each residual series, to test the significance of presence of nonlinear component in each residual series. BDS test was conducted using the 'tseries' package of Rstudio software. This test was conducted considering different epsilon value (based on variability of the residual series) and embedding dimension.

The p-values of the test result was smaller than 0.05, thereby null hypothesis was rejected. This indicated that the series was nonlinearly dependent (in different embedding dimension). This test specifies the importance of nonlinear model development, because the residual series exhibits the nonlinear component which was unable to capture by the linear ARIMA model.

Table 5: BDS Test Result of Banana Area

Dimension	Epsilon	Statistic	p-value
2	0.099	-4.8785	0
2	0.297	-5.1165	0
3	0.099	-7.0672	0
3	0.297	-7.2650	0

Table 6: BDS Test Result of Banana Production

Dimension	Epsilon	Statistic	p-value
2	1.753	2.2978	0.02
2	5.259	1.7576	0.08
3	1.753	-8.0596	0
3	5.259	-7.4335	0

For ANN modelling, trial and error method was adopted. Each original series was divided into two parts in 8:2 ratio

to make the training as well as testing set. Thereby, thirty models were developed considering different numbers of input node (1 to 6) and hidden node (2 to 10 at an increment of 2), and their performances were recorded through RMSE and MAPE values. The model, which exhibited lowest error both in training as well as testing set is considered as the best ANN model for that series. Best ANN models of both series were given in table-6.

Table 7: Description of Best ANN model of Banana Area and Production series

Variables	Best ANN Models	Number of weights	No. of input node	No. of hidden node
Banana Area	5:2s:1l	15	5	2
Banana Production	2:10s:1l	41	2	10

ANN modelling was followed by NLSVR modelling. 'TSSVM' package of Rstudio software was used to build the NLSVR model for each series. It is an automated package that used the training data to develop the model.

Testing data set was used to get the idea of performance of the suggested NLSVR model by the package. At last forecasts were generated for each variable using their respective best NLSVR model.

Table 8: Best NLSVR description of Banana Area and Production series

Variables	Kernal Function	C	γ	ϵ
Banana Area	RBF	1	1	0.1
Banana Production	RBF	1	0.5	0.1

At the end of the analysis, hybrid models were developed by combining the linear and nonlinear model simultaneously. The residual series of ARIMA models were fitted with both ANN and NLSVR model. Ultimately, estimated residual series given by ANN and NLSVR model were added with

ARIMA estimated series of each variable. Thus, the new series obtained is the required hybrid series. The performances of hybrid models were also calculated through RMSE values of both training and testing set.

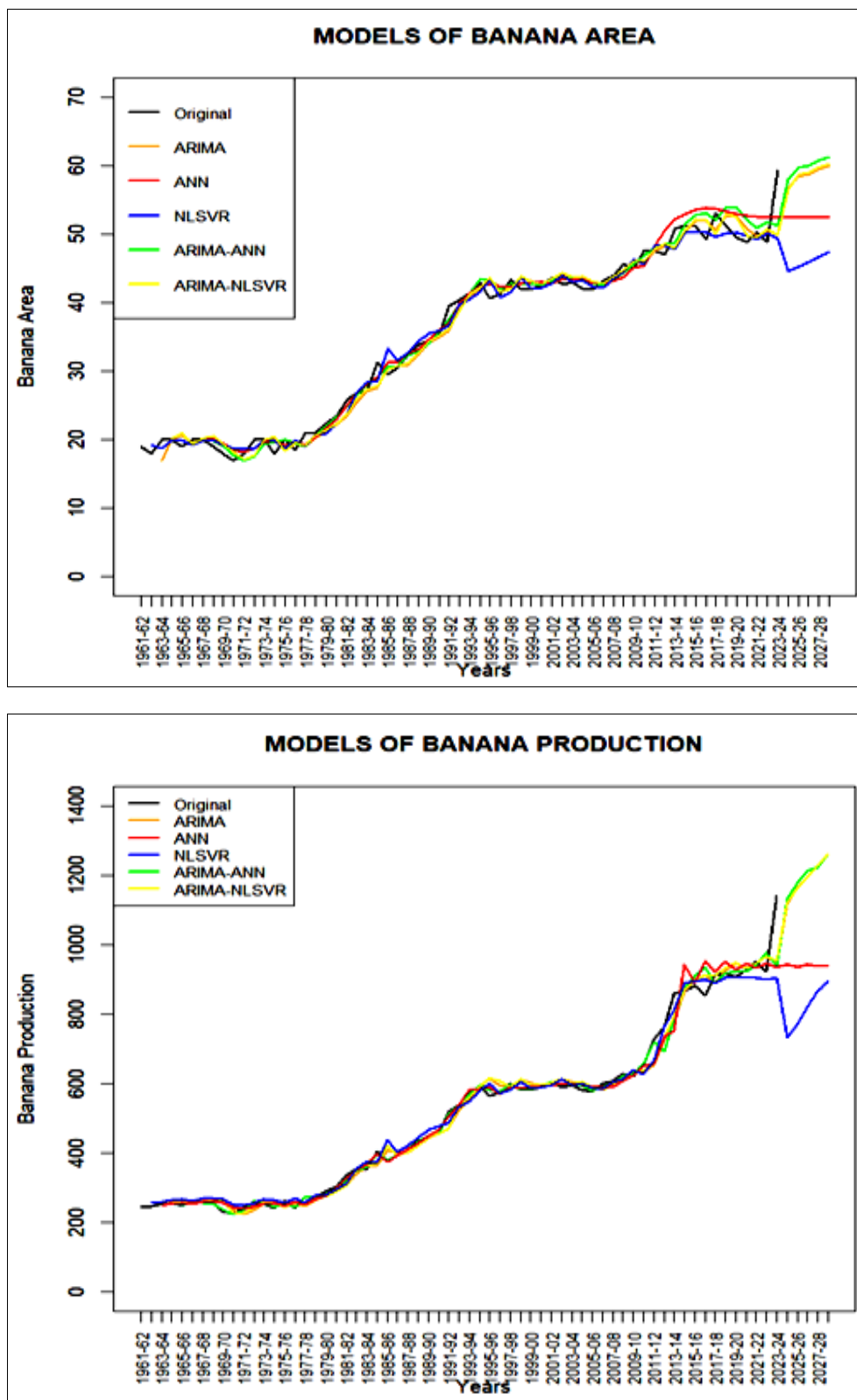


Fig 4: Graphical Representation of All Models of Banana Area and Banana Production Variables

At the end, the performances of each model were compared through RMSE and MAPE values of training and testing set.

Table 9: RMSE results of all developed model of four variables

Variables	Measures	ARIMA	ANN	NLSVR	ARIMA-ANN	ARIMA-NLSVR
Banana Area	Training	1.504	1.381	1.122	1.031	1.414
	Testing	3.176	3.327	3.186	3.250	3.254
Banana Production	Training	19.475	11.087	19.416	5.726	19.581
	Testing	66.201	77.307	72.479	69.031	63.192

Conclusion

The trendline analysis results disclosed that the banana area series is following a downward parabolic quadratic trendline (negative x^2 coefficient) whereas banana area series is following an upward parabolic quadratic trendline (positive x^2 coefficient). Thus, it could be concluded that the banana area is following an increasing trendline with a moderately decreasing growth rate (i.e. growth is slowing gradually). Alternatively, banana production is acquiring an increasing trendline with a moderate increasing growth rate. CAGR results disclosed that the cultivable banana area was growing by 1.82% on average per year since the values were found positive. Similarly, results signified that the banana production of Assam was growing annually at a rate of 2.48% on average. Therefore, it could be concluded that the area of banana cultivation was growing at a comparatively slower rate than the banana production. Results of forecasting models revealed that among the individual models, ANN model performed much better than other models for explaining banana area whereas in case of banana production variable, the ARIMA (1, 2, 1) model performed better than individual ANN and NLSVR model. Regarding the hybrid models, ARIMA-ANN model outperformed the ARIMA-NLSVR model in case both area and production variables. Comparing the all developed model, it was observed that hybrid ARIMA-ANN model showed superior results for both area and production variable and therefore this model was declared as best forecasting model. Thus, all best models were recommended for forecasting generation. At the end, predictions were generated for subsequent five years. It was expected from the forecast results that production of banana will increase in the future if the series continues to follow the similar trendline. Although the rate of growth was found slower for banana area as suggested by the results of trend analysis, however the forecast results reflected that the area under banana cultivation would also increase in future even though the rate of growth is moderate. However, much scope is still available for enhancing the area of banana cultivation.

Reference

1. Anonymous. Horticultural Statistics at a Glance. New Delhi: Ministry of Agriculture and Farmers Welfare, Government of India; 2020-2021.
2. Anonymous. Statistical Handbook of Assam. Assam: Directorate of Economics and Statistics, Government of Assam; 2023-2024.
3. Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. New York: John Wiley & Sons; 2015.
4. Neog B, Gogoi B, Barua S, Mahanta S, Sarmah K. Development of hybrid time series models for forecasting autumn rice production in Assam. Pharma

Innovation Journal. 2023;12(2):3376-3384.

5. Ramadhan AJ, Priya SRK, Naranammal N, Pavishya S, Naveena K, Ray S, *et al.* Comparison study using ARIMA and ANN models for forecasting sugarcane yield. BIO Web Conf. 2024;97:00078.
6. Rathod S, Mishra GC, Singh KN. Hybrid time series models for forecasting banana production in Karnataka State, India. J Indian Soc Agric Stat. 2017;71(3):193-200.
7. Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag; 2013.
8. Vapnik VG. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer M, Jordan M, Petsche T, editors. Advances in Neural Information Processing Systems. Vol 9. 1997. p. 281-287.
9. Varghese E, Gills R, Anuja AR, Jayasankar J, editors. Manual on advanced analytical tools for social science research. Vol II. Cochin: ICAR-Central Marine Fisheries Research Institute; 2023. p. 1-324.
10. Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003;50:159-175.