Asian Journal of Agricultural Extension, Economics & Sociology

39(11): 168-174, 2021; Article no.AJAEES.74117
ISSN: 2320-7027

Hybrid Statistical Models for Forecasting Yield of Mango and Banana in Tamil Nadu, India

P. Sujatha[1*]

[1]AEC & RI, TNAU, Trichy, India.

Author's contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/AJAEES/2021/v39i1130738
Editor(s):
(1) Dr. Kwong Fai Andrew Lo, Chinese Culture University, Taiwan.
Reviewers:
(1) Jesser Roberto Paladines Amaiquema , Universidad Técnica de Machala, Ecuador.
(2) Md. Hayder Khan, Sher-e-Bangla Agricultural University, Bangladesh.
Complete Peer review History: https://www.sdiarticle4.com/review-history/74117

Original Research Article

Received 09 August 2021
Accepted 18 October 2021
Published 25 October 2021

ABSTRACT

Horticulture sector plays a prominent role in economic growth of India. India is the second largest producer of fruits and vegetables in the world next to China. Among the horticultural crops, fruit crops are cultivated in majority of the area in India. Fruit crops play a significant role in the economic development, nutritional security, employment generation, and total growth of country. India is major producer of mango and banana, among fruit crops. The objective of this research paper is to predicate the yield of mango and banana in Tamil Nadu using different models such as linear and nonlinear, parametric, and non-parametric statistical models. In this research, a hybrid model had been proposed, which consists of linear and nonlinear models. In this hybrid model, combination of the Autoregressive Integrated Moving Average (ARIMA) and Regression model were used. The present study was conducted in Tamil Nadu. Since, area and production of Mango and Banana are higher in Tamil Nadu. Based on results obtained production and yield of Mango and Banana were predicted for next four years.

Keywords: ARIMA; regression model; hybrid model; time series.

**Corresponding author: E-mail: sujathap@tnau.ac.in;*

1. INTRODUCTION

Agriculture is the backbone of the Indian economy for 19.9 percent of the nation's Gross Domestic Product (GDP) in 2020. Nearly about 69 percent of the country's population still depends on the agriculture and horticulture sectors. Through the contribution of the agricultural sector to GDP is increasing from 2019. But a significant change in the composition of agriculture, showing shifting from cropping towards horticulture, livestock, and fisheries, is an important role. The horticulture sector contributed 2.5 percent higher production than 2019-20. India witnessed the shift in the area from food grains towards horticultural crops over the last ten years from 2010-11 to 2019-20. The area under the horticultural crops increased by more than 18 percent but the augmentation of an area of food grains is only 8 percent of the last ten years. The area under horticultural crop's annual production is increasing by 7 percent. The horticultural crops are cultivated in an area of 25.74(Million ha) with the production of 311.05 Million tons) [1]. Especially Fruit crops play a significant role in the economic development, nutritional security, employment generation, and overall production and income generation of our country. Fruit crops have climatic specificity and excellent fruits having delicacy, nutritive value, and good market acceptability are grown widely in temperate, tropical, and sub-tropical parts of the country. The large size of population in India is engaged in fruit production, distribution, and marketing [2]. Fruits being the major source of vitamins and minerals are aptly called protective foods and are an indispensable part of humans and animals. Although India contributes 11.80 percent of the total fruits of the world, the availability of fruits in the country has been estimated to be only 182 grams per day per person, amounting to the deficit of 48 grams per day per person (Anonymous, 2015a). There exists an increase in the yield of fruit crops. Hence, the objective of this research paper is to study the trend in the area, production, and productivity of mango and banana in Tamil Nadu using different models such as linear and nonlinear, parametric, and non-parametric statistical models [3].

2. MATERIALS AND METHODS

2.1 Data Source and Study Area

The present study was conducted in Tamil Nadu. This study was based on secondary data collected. This study was purposively conducted in Tamil Nadu because fruits production was 5.77 million metric tonnes in Tamil Nadu during 2019-20 (statista.com). The major fruits produced in Tamil Nadu were Mango, Banana and Jack fruits (called as Mukkani in Tamil). Secondary data on area and production of mango and banana were collected for the period of 38 years, from 1982-83 to 2017-18 from Tamil Nadu from National Horticulture Board (NHB) data base and indiastat.com. Statistical forecasting techniques were employed to predict the production and yield of mango and banana in Tamil Nadu. Statistical forecasting method is the approximation of event taking place in value of future production [4]. Statistical forecasting is used for decision-making and planning the future more effectively and efficiently and develop country economically. Methodology of this research paper involves prediction of present series based on behavior of past series over a period of 1982-83 to 2017-18. Annual data on yield (MT ha-1), area (ha) and production (MT) of mango and banana were collected for regression analysis and model building. Based on statistical forecasting models, the production and yield of mango and banana in Tamil Nadu were predicted for a period of 2018-2022. Study area was shown in Fig. 1.

2.2 Brief Description of the Statistical Models Employed

2.2.1 Auto Regressive Integrated Moving Average (ARIMA) model

The most important and widely used classical time series statistical model is the Auto Regressive Integrated Moving Average (ARIMA) model. The popularity of the ARIMA model is due to its linear statistical properties as well as the popular Box- Jenkins methodology (Box and Jenkins, 1970) for model constructing procedure. To obtain the stationary time series, the differencing term 'd' used to make the non-stationary series to stationary series. The general form of ARIMA model is ARIMA (p, d, q). p is order of autoregressive term, d is the order of differencing term and q is the order of moving average term in ARIMA(p, d, q) model. The process Y_t is said to follow integrate ARIMA model.

The ARIMA model is expressed as follows:

$$\left.\begin{aligned} \varphi(B)(1-B)^d Y_t &= \theta(B)\varepsilon_t \\ \Delta Y_t &= (1-B)^d \varepsilon_t \end{aligned}\right\} \quad \text{---------} \quad (1)$$

The Box-Jenkins ARIMA model building consists of three steps viz., identification, estimation, and diagnostic checking. First step in model building is to identify the model i.e. to determine the model order. Second step is to estimate the parameters of model based on identified model order. Finally, the third step is diagnostic checking of residuals. Brief description of the statistical models employed given in Fig. 2.

2.2.2 Time Delay Neural Network (TDNN)

The ANN for time series analysis is termed as Time Delay Neural Network (TDNN).

For the time series prediction problem, TDNN are possibly the simplest choice for representing a wide range of mappings between past and present values. In the case of a TDNN the training data are samples of a function that maps a set of past values of a time series on a future value. A TDNN trained on such data becomes an approximator to the (unknown) mapping expressed by the data. The time series phenomenon can be mathematically modeled using neural network with implicit functional representation of time, whereas static neural network like multilayer perceptron is presented with dynamic properties [5]. One simple way of building artificial neural network for time series is the use of time delay also called as time lags. These time lags can be considered in the input layer of the ANN. The TDNN is the class of such architecture. Following is the general expression for the final output Y_t of a multi- layer feed forward time delay neural network

$$Y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} Y_{t-p}) + \varepsilon_t \quad \text{----------(2)}$$

$$\beta_{ij}(i = 0,1,2 \dots \dots, p, j = 0,1,2,3 \dots \dots . q)$$

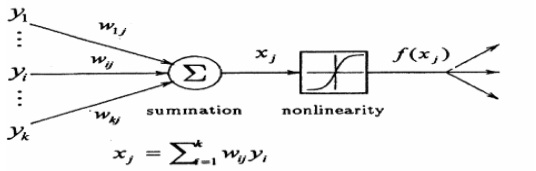
Where, α_j (j=0,1,2,3……,q) are model parameter

2.2.3 Non linear Support Vector Regression Model (NLSVR)

Support Vector Machine (SVM) is a advanced machine learning technique which was developed for linear classification problems. Latterly, the support vector machine for regression problems were developed by Vapnik by developing ε-insensitive loss function [6] and liner model has been extended to the nonlinear regression model [7] Modeling of in this kind of problems is called as Non Linear Support Vector Regression (NLSVR) model. The basic principle involved in NLSVR is to transform the original input time series in to a high dimensional feature space and then build the regression model in a new feature space. Consider a vector of data set $Z=\{x_i, y_i\}_{i=1}^{N}$ where$x_i \in R^n$ is the input vector, y_i is the scalar output and N is the size of data set.

Fig. 1. Map showing the study area

The general equation of Nonlinear Support Vector Regression estimation function is given as follows: f(x)=WT$\varphi(x)$+b, where, φ(x) : $R^n \rightarrow R^nh$ is a nonlinear mapping function which maps the original input space into a higher dimensional feature space vector. W∈ R^{nh} is weight vector, b is bias term and superscript T denotes the transpose.



2.2.2.4 Proposed hybrid methodology

The hybrid method considers the time series as a combination of both linear and non linear components. This approach follows the Zhang"s (2003) hybrid approach. Accordingly, the relationship between linear and nonlinear components can be expressed as follows

$$y_t = L_t + N_t \quad (3)$$

Where, L_t and N_t represent the linear and nonlinear component, respectively. In this research paper, the linear part is modeled using ARIMA model and non-liner part by TDNN and NLSVR. The methodology consists of three steps. Firstly, an ARIMA model is employed to fit the linear component. Let the prediction series provided by ARIMA model be denoted as $\hat{L}_t$. In the second step, the residuals ($e_t = y_t - \hat{L}_t$) obtained from ARIMA model are tested for non-linearity by using BDS test [8] Once the residuals confirm the non-linearity, then, they are modeled and predicted using TDNN and NLSVR. Finally, the forecasted linear and nonlinear components are combined to generate aggregate forecast represent the predicted linear and nonlinear component, respectively. The graphical representation of hybrid methodology is expressed in Fig. 3. The performances of the models under consideration were compared by using Mean Absolute Percentage Error (MAPE).

3. RESULTS AND DISCUSSION

3.1 Regression Analysis

Regression analysis has been carried out to know the factors influencing yield of mango and banana in Tamil Nadu. Regression model was fitted for yield of mango and banana in Tamil Nadu. Yield of mango and banana were considered as dependent variables in the study, whereas production and area were considered as independent variables. Secondary data on these variables were collected from 1982-1983 to 2017-2018. Data from 1982-83 to 2016-2017 were used for model building and data from 2017-2018 to 2019- 2020 were used for model validation. Time series data on yield of banana were highly heterogeneous as Coefficient of Variation (CV) was very high. Regression analysis was carried out for all the data sets separately to fit the model. In regression analysis, the value of R^2 was decreased and Variance Inflating Factor (VIF < 30) was low and most importantly number of significant variables were also increased. Results of regression analysis clearly indicated that yield of mango was dependent on area and production of mango as indicated by the coefficients of area and production of mango. Thus yield of mango is dependent on two important factors such as area and production. It could be inferred that the coefficient of multiple determination (R^2) for dependent variables yield of mango and banana were 0.92, 0.73 respectively, which indicated that 92 and 73 per cent of the variation in the dependent variables were explained by the independent variables included in the respective model. Similar results were also obtained by Sellam and Poovammal (2016) (Table 1).

3.2 Auto Regressive Integrated Moving Average (ARIMA) Model

The Auto Regressive Integrated Moving Average (ARIMA) model was built separately for time series data on yield of both mango and banana using SAS 9.4 software. The Box-Jenkins methodology was followed. Maximum likelihood method was used for estimation of parameter. Results of the study showed that residuals were non-correlated as it was evident from the probability of residuals obtained. Since, the model satisfied Box-Jenkins methodology of model building, forecasting on time series data on yield of mango and banana were done. For each model, forecasts were started after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors were available or at the end date of the requested forecast period, whichever was earlier. Results of the analysis were shown in Table 2 & 3. Generally MAE, RMSE and MAPE values were used to find the best model structure. Each

error was different. In this study, we received the lowest MAE value (1.11) when the model was trained. Hence, it was appropriate to use the ARIMA model for prediction of yield of mango and banana. Similar results were also obtained by Energetika [9].

3.3 Time Delay Neural Network (TDNN)

Time Delay Neural Network (TDNN) Model Specifications was used to find out the forward time delay neural network for time series data on yield of mango and banana and it was done using R software [10]. The Levenberg-Marquardt back propagation programme was used for neural network model based on secondary data. In TDNN models, learning rate 0.02 and 0.001, Sigmodial, linear functions were used as output layers. In this 90 per cent of data observations in data set were used for model development and remaining 10 percent of data were used for validation and yield was predicted for next four years (Table 4).

3.4 Nonlinear Support Vector Regression Model (NLSVR Model)

The nonlinear support vector regression model used for kernel function and set of hyper parameters. The kernel function in NLSVR requires optimization of the parameters. The regularization parameter α, the kernel bandwidth parameter μ. μ defines the variance of kernel function, these two parameters α and μ are tuning parameter (Table 5). Hybrid model is one of the promising and potential methods for time series prediction. Results of the study showed that hybrid model in general provides better forecast accuracy in terms of conventional root mean square error values (Table.6).

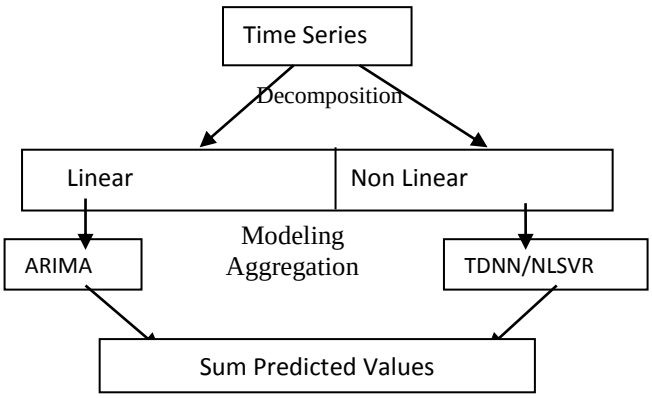


Fig. 2. Brief description of statistical models employed

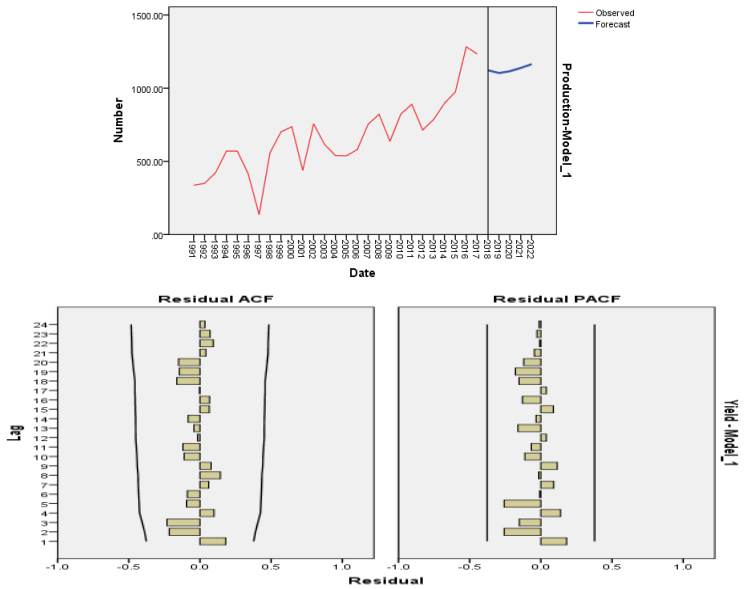


Fig. 3. Graphical representation of hybrid methodology

Table 1. Regression model

Regression Model/ Dependent Variables	R2 Value	Number of variables having <4
Mango Yield	0.92	30
Banana Yield	0.73	39

Table 2. ARIMA Model Statistics

Model	Number of Predictors	Model Fit statistics				Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	MAPE	MAE	Max APE	Statistic	DF	Sig.	
Yield-Model_1	3	-1.004E-013	29.39	1.11	346.07	11.14	18	.88	0

Table 3. Forecasted value

Model		2018	2019	2020	2021	2022
Yield-Model_1	Forecast	6.80	7.08	7.12	7.30	7.29
	UCL	11.83	10.90	11.50	11.57	11.91
	LCL	2.60	2.54	2.70	2.76	2.73

Table 4. Time delay neural network

Time Series	Function value		Time delay	No. of hidden nodes	Total no. of parameters
	Hidden	Output			
Mango yield	Sigmoidal	Linear	2	4	3
Banana yield	Sigmoidal	Linear	2	4	3

Table 5. NLSVR model

Time series	Kernel function	No.of years	α	μ	ε	K fold cross validation (K)	Error
Mango yield	RBF	25	1.09	1.00	0.01	3	0.037
Banana yield	RBF	23	1.078	1.62	0.01	3	0.044

Table 6. Comparison of forecasting performance of ARIMA, TDNN, NLSVR in testing data set of Mango yield

Year	Actual	Stepwise Regression	ARIMA	TDNN	NLSVR	ARIMA-TDNN	ARIMA-NLSVR
2018	8.31	8.03	6.80	7.08	7.12	7.30	7.29
2019	7.08	6.90	7.08	8.02	8.00	7.55	7.54
2020	7.12	7.01	7.12	8.00	7.15	7.56	7.58
2021	7.85	6.50	7.30	8.51	7.95	7.905	8.23
2022	8.22	6.50	7.29	8.50	7.96	7.895	8.23

4. CONCLUSIONS

Based on the results obtained in this research work, machine intelligence techniques such as time delay neural network and nonlinear support vector regression models performed better as compared to classical time series models when time series data was heteroscedastic and noisy. The regression analysis of the study showed that variables included in the models were strongly influenced the yield of mango and banana. Hybrid model is one of the promising and potential methods for time series prediction. Hybrid model in general provides better forecast accuracy in terms of conventional root mean square error values. Results of the study showed that hybrid model performed better when compared to single time series or machine

learning techniques for modeling and forecasting the time series data on yield of mango and banana in Tamil Nadu. Among the hybrid models, the ARIMA with Nonlinear Support Vector Regression model performed superior as compared to all forecasting models. Based on the results obtained, it was concluded that the farmers or policy makers can plan well in advance to increase the productivity of these crops by adopting suitable management practices.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. National Horticultural Board (NHB) Data Base. 2019-2020. Current Scenario of Horticulture in India.
2. Yadav AS, Pandey DC. Geographical perspectives of mango production in India. Imperial J. Interdisciplinary Res. 2016; 2(4):257-265.
   Available:http://nhb.gov.in/area-pro/NHB_Database_2020.pdf
3. Mayer DG, Stephenson RA. Statistical forecasting of the Australian macadamia crop. Acta Hortic. 2016;1109:265-270. DOI: 10.17660/ActaHortic.2016.1109.43
4. Kumar TLM, Prajneshu. Development of hybrid models for forecasting time-series data using nonlinear SVR enhanced by PSO. J. Stat. Theor. Pract. 2015;9(4):699-711.
5. Haykin S. Neural Networks: A Comprehensive Foundation. New York. Macmillan, ISBN 0-02-352781-7; 1999.
6. Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. In: “Advances in Neural Information Processing Systems”, (Eds.): Mozer, M., Jordan, M and Petsche, T. MIT Press, Cambridge, MA. 1997;9:281-287.
7. Rathod S, Mishra GC. Statistical models for forecasting mango and banana yeild of Karnataka, India. J. Agr. Sci. Tech. 2018; 20:803-816.
8. Brock WA, Dechert WD, Scheinkman JA, Lebaron B. A Test for independence based on the correlation dimension. Econ. Rev. 1996;15:197-235.
9. Grigonytė Ernesta, Butkevičiūtė Eglė. Short-term wind speed forecasting using ARIMA model, Energetika, T. 2016;62(1–2):45–55.
10. Jha GK, Sinha K. Time-Delay neural networks for time series prediction: An application to the monthly wholesale price of oilseeds in India. Neural Comput. Appl. 2014;24(3):563-571.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/74117