Acta Scientific AGRICULTURE (ISSN: 2581-365X)

Volume 6 Issue 9 September 2022

Research Article

Prediction of Mango Production Using Machine Intelligence Techniques: A Case Study from Karnataka, India

Santosha Rathod[1], Vijayakumar S[1*], Nirmala Bandumula[1] and Gayathri Chitikela[2]

[1]ICAR-Indian Institute Rice Research, Hyderabad, Telangana, India

[2]Professor Jayashankar Telangana State Agricultural University, Hyderabad, Telangana, India

***Corresponding Author:** Vijayakumar S, Scientist, ICAR-Indian Institute Rice Research, Hyderabad, Telangana, India.

DOI: 10.31080/ASAG.2022.06.1174

Received: July 15, 2022

Published: August 05, 2022

© All rights are reserved by **Vijayakumar S., *et al.***

Abstract

Mango is the largest producing fruit crop in India. On the other hand, Karnataka is called the horticultural state of India, where mango is the highest producing fruit crop. A developing economy relies heavily on forecasting for effective planning and long-term sustainable growth. The most common technique used for forecasting in several fields for many years is autoregressive integrated moving average (ARIMA). The assumptions of linearity and stationarity are major key flaws in this model. As many time series phenomenon in the real world are not purely linear, therefore there is an opportunity to enhance the prediction ability in time series analysis by employing nonlinear machine intelligence techniques like Autoregressive Neural Network (NAR: Neural Network Autoregressive) and non-linear support vector regression (NLSVR) model. In this study, an attempt is made to forecast the mango production of Karnataka using ARIMA, NAR and NLSVR. According to empirical evidence, the predicting accuracy of the time series machine intelligence technique is clearly superior than the traditional ARIMA model.

Keywords: Mango Production; Time Series; ARIMA; NAR; NLSVR

Abbreviations

NLSVR: Non-linear Support Vector Regression; ARIMA: Autoregressive Integrated Moving Average; NAR: Neural Network Autoregressive; NHB: National Horticultural Board; SVM: Support Vector Machine; SVR: Support Vector Regression; ADF: Augmented Dickey Fuller

Introduction

The horticulture sector is playing a prominent role in the economic development of the country by improving the income of the rural people. India is one of the top producers of horticulture crops. Next to Brazil and China, India is the world's largest producer of fruits and vegetables. Mango (*Mangifera indica* L.) is the most important fruit crop of India and belongs to the family Anacardiaceae. It is regarded as the king of fruits because of its mouthwatering flavour and alluring aroma. It is also a good source of vitamins A and C. Mango occupies 22 percent of the total area under fruits comprising 1.2 million hectares, with a total production of 11 million tons in the world. Raw mango fruits are used to make chutney, pickles, and juices at all stages of their development, including when they are immature and ripe. The ripe fruits are used to make desserts and a variety of goods, including jams, jellies, syrups, nectars, and squashes. Additionally, the mango kernel contains 8-10% high-quality fat that can be utilised to make soap and to replace cola in confections. Mango pulp and fresh mangoes are significant agricultural exports from India.

In India, the mango crop occupies 34.9 and 20.7 percent of the total area and total fruit production of the country (NHB data base 2014-15). UAE, Kuwait, and other Middle Eastern nations, as well as the European market, are India's primary mango export markets. Despite being the world's largest producer of mangoes—producing roughly 60% of all mangoes—India exclusively exports only the Alphonso and Dashehari kinds of fresh fruit. About 15% of the global mango market is accounted for by India, which also exports 40% of its total fruit production. Every year, the demand for mango fruit increases, but the rate of production cannot keep up with the demand [1]. Therefore, there is potential to expand the area and output of mango in the nation. This can be achieved by adopting techniques to increase productivity by making necessary policy implications. Forecasting is an important aspect of a developing economy so that adequate planning is undertaken for sustainable growth, overall development and poverty alleviation [2]. Statistical models are used to develop an appropriate forecast methodology by using past data to predict the future with the help of identifying the trends and patterns within the data [3]. In other words, statistical forecasting is the prediction of the probability of an event occurring in the future using the data that is currently available.

The autoregressive integrated moving average (ARIMA) model is one of the most fundamental and popular time series models. The well-known Box-Jenkins model-building methodology and the statistical features of the ARIMA model account for its popularity [4]. Previously, the ARIMA model was used in modeling and forecasting various agricultural commodities viz., India's pigeon pea production [5,6], groundnut oil [7], coconut production of India [8], rice yield of India [9-11], maize yield [12], mango area and production [13], oil seed production [14], total fish production of India [15], tomato supply and prices [16] and drought prediction [17,18].

The presumptive linear nature of the ARIMA model is its primary flaw. In other words, the ARIMA model cannot detect any nonlinear patterns since it assumes a linear correlation pattern among the time series. The major advantage of the Autoregressive Neural Network (NAR: Neural Network Autoregressive) is its flexible non-linear modeling capability and no need to specify model form as it is a data-driven approach [6,7,9,19]. By introducing the ε-insensitive error loss function, the support vector machine (SVM) which was initially proposed for classification problems, has been successfully extended to regression problems by Vapnik., *et al.* [20], and it is called Support Vector Regression (SVR). Many findings in the literature show that by employing SVR in time series modeling and forecasting, one can improve the prediction accuracy of the univariate models [21-23]. Machine intelligence techniques like NAR and SVR were previously employed for modeling and forecasting various agricultural commodities viz., rice yield [24], coffee price [25,26], Banana and Mango yield [27,28]. With these backgrounds, efforts have been made to forecast the mango production of Karnataka using ARIMA, NAR and NLSVR models.

Material and Methods

Data description

Yearly data on production (000' MT) of mango crop from 1980-81 to 2014-15 for Karnataka state was collected from the database of the National Horticulture Board (NHB) and www.indiastat.com. Data from 1980-1981 to 2011-2012 were used to develop the models, while data from 2012-2013 to 2014-2015 were used to assess how well the models performed in terms of predictions.

ARIMA model building

Generally, the ARIMA model [4], denoted as ARIMA (p,d,q), is expressed as follows

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t \quad \text{-------------------------------------- (1)}$$

Where,

$$\phi(B) = 1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p \text{ (Autoregressive parameter) -------------- (2)}$$

$$\theta(B) = 1-\theta_1 B-\theta_2 B^2-\cdots-\theta_p B^q \text{ (Moving average parameter) --------------- (3)}$$

ε_t = white noise or error term

d = differencing term

B = Backshift operator i.e. $B^a Y_t = Y_{t-a}$

Three steps make up the ARIMA model construction process: identification, estimation, and diagnostic testing. The identification stage of this model involves the experimental selection of its parameters. To change a non-stationary series into a stationary one, d must be identified. The test of the unit-root hypothesis, a statistical test, can be used to determine whether stationarity exists. The stationarity is tested using the Augmented Dickey Fuller (ADF) test. At the estimation stage, iterative least-squares or maximum likelihood approaches are used to estimate the parameters. The diagnostic checking stage uses the Ljung-Box test to determine the effectiveness of the chosen model. The three processes are repeated if the model is deemed to be inadequate until a suitable ARIMA model is chosen for the time-series under consideration.

NAR: Neural network autoregressive model

The ANN for time series analysis is termed an autoregressive neural network (NAR) model. The time series phenomena can be mathematically modelled using a neural network with implicit functional representation of time, while static neural networks like multi-layer perceptron are presented with dynamic properties. Using time delay, often known as time lags, is one straight forward method of creating an artificial neural network for time series. The input layer of the ANN can take these time gaps into account. The NAR is the class of such architecture. The general formulation for a multi-layer feed forward time delay NAR model's final output, Yt, is given below.

$$Y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j\, g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} Y_{t-p}\right) + \varepsilon_t \text{ ------------ } (4)$$

Where, $\alpha_j\,(j = 0,1,2, \ldots, q)$ and $\beta_{ij}(i = 0,1,2, \ldots, p, j = 0,1,2, \ldots, q)$ are the model parameters, also called as the connection weights, p is the number of input nodes, q is the number of hidden nodes and g is the activation function. The architecture of neural network is represented in figure 1.

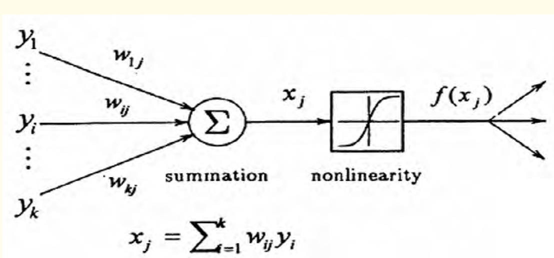

Figure 1: Neural network structure.

Nonlinear support vector regression (NLSVR) model

A supervised machine learning method called the support vector machine (SVM) was initially created to solve linear classification issues. By including the insensitive loss function [20] later in the year of 1997, Vapnik developed the support vector machine for regression problems. It has since been extended to include non-linear regression estimation problems and the modelling of such problems is known as Nonlinear Support Vector Regression (NLSVR) model. The fundamental idea behind NLSVR is to first create a high-dimensional feature space from the original input time series before creating the regression model in that space. Consider a data set with the formula Z=x (i) y (i=1) N, where x i is the input vector, yi is the scalar output, and N is the size of the data set. The general equation of the Nonlinear Support Vector Regression estimation function is given as follows;

$$f(x) = W^T \phi\,(x) + b \text{ ---------------- } (5)$$

Where $\phi(.)$: $Rn \rightarrow R^{nh}$ is a nonlinear mapping function which map the original input space into a higher dimensional feature space vector. $W \in R^{nh}$ is weight vector, *b* is bias term and superscript T denotes the transpose.

Result and Discussion

Summary statistics of mango production time series is given in table 1 and time series plots of the same is exhibited in figure 2. The investigated time series can be said to be stationary based on the plots obtained in figure 3, which is further supported by the results of the Augmented Dickey-Fuller unit root test presented in table 2. These results show that the series is stationary. Last but not least, the ARIMA (1, 0, 0) was found to be adequate for the time series under consideration based on the lowest Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Maximum Likelihood criteria's given in table 3 and parameter estimates of the same are given in table 4. The residuals from the ARIMA model of the Mango Production time series were not auto correlated, with a probability of chi-square of 0.5755. Additionally, the model's performance in the training and testing data set is given in tables 7 and 8.

Statistic	Mango Production	Statistic	Mango Production
Observation	35	Maximum	1868.3
Mean	1034.69	Standard Deviation	436.29
Median	1105.9	Skewness	0.38
Mode	-	Kurtosis	-0.92
Minimum	481.2	Coefficient of Variation (%)	42.17

Table 1: Summary statistics of Mango Production time series.

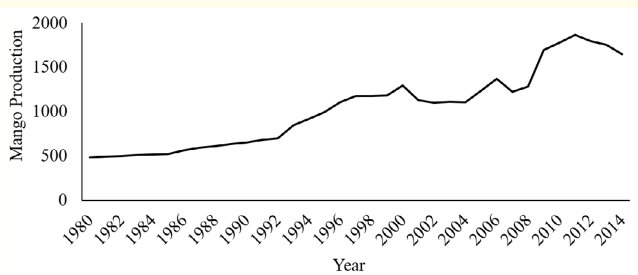

Figure 2: Time series plot of Mango Production of Karnataka.

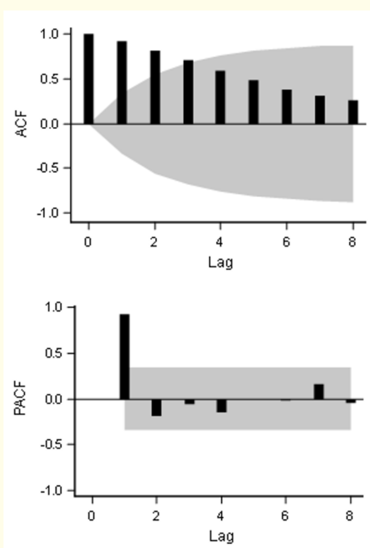

Figure 3: ACF and PACF plots for Mango Production time series.

ADF test statistic				PP test statistic			
Single mean	With trend	Probability		Single mean	With trend	Probability	
		Single mean	With trend			Single mean	With trend
-0.59	10.66	0.048	0.039	-0.6182	-10.039	0.0501	0.049

Table 2: Stationary test of Mango Production time series.

Models	Log-likelihood	AIC	BIC
ARIMA(0,0,1)	-241.16	488.33	492.99
ARIMA(0,0,2)	-232.13	472.26	478.48
ARIMA(1,0,0)	-211.65	429.30	433.48
ARIMA(1,0,1)	-211.89	431.78	437.99
ARIMA(1,0,2)	-211.85	441.48	433.70
ARIMA(2,0,0)	-211.66	429.31	435.53
ARIMA(2,0,1)	-211.12	432.24	440.02
ARIMA(2,0,2)	-210.96	433.93	443.25

Table 3: Log likelihood, AIC and BIC values of different ARIMA models.

Parameter	Estimate	Standard Error	t Value	Approx. Pr > |t|	Lag
MU	1055.5	625.15	1.69	0.0914	0
AR1,1	0.98	0.03	26.76	<0.0001	1

Table 4: Parameter estimation of ARIMA (1, 0, 0) by Maximum Likelihood Estimation method for Mango Production time series.

Based on the lowest training RMSE of NAR model the five models viz. 2:8s:1l, 2:10s:1l, 3:10S:1l, 4:8s:1l and 4:10s:1l, are selected (Table 5). These five models were further assessed based on their holdout forecasting performance. Out of total 29 neural network structures, a NAR model with two tapped delay and ten hidden nodes (2:10s:1l), was selected for forecasting Mango production of Karnataka. Based on repetitive experimentation, the learning rate and momentum term for all NAR model is set as 0.03 and 0.01 respectively (Table 5). Tables 7 and 8 show the model's performance on the training set and the testing data set.

The support vector regression model for Mango production time series was built with following parameter specifications (Table 6). For the time series under consideration, cross validation was done

Model	Parameters	RMSE	
		Training	Testing
2:2S:1L	9	90.10	127.58
2:4s:1l	17	98.68	166.12
2:6s:1l	25	78.03	112.95
2:8s:1l	33	70.54	108.99
2:10S:1l	41	84.63	112.12
3:2s:1l	11	95.35	126.32
3:4s:1l	21	105.10	160.00
3:6s:1l	31	82.20	109.72
3:8s:1l	41	88.96	149.88
3:10S:1l	51	90.55	125.70
4:2s:1l	13	97.94	179.06
4:4s:1l	25	155.64	162.26
4:6s:1l	37	85.16	233.77
4:8s:1l	49	107.47	158.09
4:10S:1l	61	123.73	171.70
5:2s:1l	15	111.23	191.85
5:4s:1l	29	215.09	293.80
5:6s:1l	43	169.39	146.63
5:8s:1l	57	137.57	160.24
5:10S:1l	71	177.77	171.04
5:12s:1l	85	143.80	187.28
5:14S:1l	99	138.04	121.27
6:2s:1l	17	105.17	162.72
6:4s:1l	33	104.87	197.23
6:6s:1l	49	142.10	126.29
6:8s:1l	65	212.74	226.85
6:10S:1l	81	175.08	252.14
6:12s:1l	97	156.04	156.29
6:14S:1l	113	194.11	289.05

Table 5: Model parameter selection in NAR model.

and the lowest cross validation error recorded was 0.077. Tables 7 and 8 also show the model's performance on the training set and the testing data set.

Based on the lowest MSE, RMSE and MAPE values of all models obtained for both training (Table 7) and testing data set (Table 8) considered, one can infer that both machine intelligence techniques viz., NAR and NLSVR outperformed over ARIMA model. Further, the value of MSE was reduced up to one-third in NLSVR in comparison to that of the ARIMA model which indicates that the performance of the NLSVR model was superior as compared to other models.

Kernel function	No. of SVs	C			K fold cross validation (K)	Cross Validation Error
RBF	9	7.24	3.76	0.20	10	0.077

Table 6: Model specification of SVR for Mango Production time series.

Criteria	ARIMA	NAR	SVR
MSE	23913.59	4975.68	1409.45
RMSE	154.64	70.54	37.54
MAPE	10.16	3.88	3.73

Table 7: Model performance of Mango Production time series for training data set.

Year	Actual	Forecast		
		ARIMA	NAR	SVR
2012	1795.10	1857.89	1879.09	1741.21
2013	1755.60	1847.64	1840.06	1701.69
2014	1646.50	1837.52	1792.95	1798.75
Criteria	MSE	16300.86	11878.47	9662.88
	RMSE	127.67	108.99	98.30
	MAPE	6.78	6.13	5.11

Table 8: Model performance of Mango Production time series for testing data set.

Similar results were reported by Kumar and Prajneshu [21] for banana yield data, Rathod., *et al.* [29] (spatiotemporal rice yield prediction), Rathod., *et al.* [30] (rice gall midge population prediction), and Rainfall prediction [31] where machine learning models performed better than classical ARIMA models.

Conclusion

For time series with both linear and non-linear components, ARIMA models are not always suitable. Non-linear machine learning methods, such as neural networks and support vector machines, can be a useful strategy to boost forecasting ability in this situation. Based on the findings of this study, it can be concluded that using artificial intelligence methods such as neural network autoregressive models and nonlinear support vector regression techniques in time series modelling and forecasting can improve forecasting accuracy. In particular, the nonlinear support vector regression model outperformed other models in terms of predicting the production of mangoes in Karnataka. Even though, the coefficient of variation is very high then also the machine intelligent techniques like NAR and NLSVR performed better. The reason could be that the nonlinear machine learning approaches outperform the autoregression model and can capture the heterogeneous trend in the data set. The use of additional machine learning methods for altering the autoregressive and moving average orders can advance this strategy.

Acknowledgments

All authors acknowledge the support provided by Indian Council of Agricultural Research, New Delhi.

Conflict of Interest

All authors acknowledge the support provided by Indian Council of Agricultural Research, New Delhi.

Bibliography

1. Pardhi R., *et al*. "Effect of Price of Other Seasonal Fruits on Mango Price in Uttar Pradesh". *Economic Affairs* 61.4 (2016): 1-5.
2. Vijayakumar S., *et al*. "Artificial Intelligence (AI) and its Application in Agriculture". *Chronicle of Bioresource Management* 6.1 (2022): 25-31.
3. Vijayakumar S., *et al*. "Rainfall and temperature projections and their impact assessment using CMIP5 models under different RCP scenarios for the eastern coastal region of India". *Current Science* 121.2 (2021): 222.
4. Box GEP and Jenkins G. "Time series analysis, Forecasting and control, 1970". Holden-Day, San Francisco, CA.
5. Iquebal MA and C Chattopadhyay. "Modelling and forecasting of pigeonpea (*Cajanus cajan*) production using autoregressive integrated moving average methodology". *Indian Journal of Agricultural Sciences* 81.6 (2011): 520-523.
6. Kumari Prity GC., *et al*. "Forecasting of productivity and pod damage by Helicoverpa armigera using artificial neural network model in pigeonpea (*Cajanus Cajan*)". *International Journal of Agriculture, Environment and Biotechnology* 6.2 (2013): 335-340.
7. Mishra GC and A Singh. "A study on forecasting prices of groundnut oil in Delhi by ARIMA methodology and artificial neural networks". *Agris on-line Papers in Economics and Informatics* 5.665-2016-44959 (2013): 25-34.
8. Naveena K., *et al*. "Forecasting of coconut production in India: A suitable time series model". *International Journal of Agricultural Engineering* 7.1 (2014): 190-193.
9. Kumari Prity., *et al*. "Comparison of forecasting ability of different statistical models for productivity of rice (*Oryza sativa* L.) in India". *The Ecoscan* 8.3 (2014): 193-198.
10. Kumari Prity., *et al*. "Autoregressive integrated moving average (arima) approach for prediction of rice (*Oryza sativa* l.) yield in India". *The Bioscan* 9.3 (2014): 1063-1066.
11. Alam Wasi., *et al*. "Improved ARIMAX modal based on ANN and SVM approaches for forecasting rice yield using weather variables". *Indian Journal of Agricultural Sciences* 88.12 (2018): 1909-1913.
12. Rathod Santosha., *et al*. "Forecasting maize yield using ARIMA-Genetic Algorithm approach". *Outlook on Agriculture* 46.4 (2017): 265-271.
13. Rathod Santosha and G C Mishra. "Weather based modeling for forecasting area and production of mango in Karnataka". *International Journal of Agriculture, Environment and Biotechnology* 10.1 (2017): 149.
14. Rathod SANTOSHA., *et al*. "Modeling and forecasting of oilseed production of India through artificial intelligence techniques". *Indian Journal of Agricultural Sciences* 88.1 (2018): 22-27.
15. Mahalingaraya, S., *et al*. "Statistical modeling and forecasting of Total fish production of India: a time series perspective". *International Journal of Current Microbiology and Applied Sciences* 7.3 (2018): 1698-1707.
16. Chitikela Gayathri., *et al*. "Artificial-Intelligence-Based Time-Series Intervention Models to Assess the Impact of the COVID-19 Pandemic on Tomato Supply and Prices in Hyderabad, India". *Agronomy* 11.9 (2021): 1878.
17. Patil R., *et al.* "Drought modelling and forecasting using arima and neural networks for ballari district, Karnataka". *Journal of the Indian Society of Agricultural Statistics* 74 (2020): 149-157.
18. Gorlapalli Amuktamalyada., *et al*. "Characterization and Prediction of Water Stress Using Time Series and Artificial Intelligence Models". *Sustainability* 14.11 (2022): 6690.

19. Rathod S and V Paramesha V. "Time Series Analysis using Machine Learning Techniques". In Vadivel, A., Paramesh, V., Uthappa, A. and Kumar, R.P. (Ed.). Ecosystem service analysis: concepts and applications in diversifed coconut and arecanut gardens. ICAR- Central Coastal Agricultural Research Institute, India (2022): 262-292.

20. Vapnik Vladimir., *et al*. "Support vector method for function approximation, regression estimation and signal processing". *Advances in Neural Information Processing Systems* 9 (1996).

21. Kumar TL. "Development of hybrid models for forecasting time-series data using nonlinear SVR enhanced by PSO". *Journal of Statistical Theory and Practice* 9.4 (2015): 699-711.

22. Alonso Jaime., *et al*. "Support Vector Regression to predict carcass weight in beef cattle in advance of the slaughter". *Computers and Electronics in Agriculture* 91 (2013): 116-120.

23. Chen Kuan-Yu and Cheng-Hua Wang. "Support vector regression with genetic algorithms in forecasting tourism demand". *Tourism Management* 28.1 (2007): 215-226.

24. Alam Wasi., *et al*. "Hybrid linear time series approach for long term forecasting of crop yield". *International Journal of Agriculture Sciences* 88 (2018): 1275-1279.

25. Naveena K., *et al*. "Hybrid ARIMA-ANN modelling for forecasting the price of robusta coffee in India". *International Journal of Current Microbiology and Applied Sciences* 6.7 (2017): 1721-1726.

26. Naveena K and Singh Subedar. "Hybrid time series modelling for forecasting the price of washed coffee (Arabica Plantation Coffee) in India". *International Journal of Agriculture Sciences* (2017): 0975-3710.

27. Rathod Santosha., *et al*. "Hybrid time series models for forecasting banana production in Karnataka State, India". *Journal of the Indian Society of Agricultural Statistics* 71.3 (2017): 193-200.

28. Rathod S and GC Mishra. "Statistical models for forecasting mango and banana yield of Karnataka, India". *Journal of Agricultural Science and Technology* 20.4 (2018): 803-816.

29. Rathod Santosha., *et al*. "Two-Stage Spatiotemporal Time Series Modelling Approach for Rice Yield Prediction and Advanced Agroecosystem Management". *Agronomy* 11.12 (2021): 2502.

30. Rathod Santosha., *et al*. "Climate-Based Modeling and Prediction of Rice Gall Midge Populations Using Count Time Series and Machine Learning Approaches". *Agronomy* 12.1 (2021): 22.

31. Saha Amit., *et al*. "A hybrid spatio-temporal modelling: An application to space-time rainfall forecasting". *Theoretical and Applied Climatology* 142.3 (2020): 1271-1282.