J. Agr. Sci. Tech. (2018) Vol.20: 803-816

Statistical Models for Forecasting Mango and Banana Yield of Karnataka, India

S. Rathod[1]*, and G. C. Mishra[2]

ABSTRACT

Horticulture sector plays a prominent role in economic growth for most of the developing countries. India is the largest producer of fruits and vegetables in the world next only to China. Among the horticultural crops, fruit crops are cultivated in majority of the area. Fruit crops play a significant role in the economic development, nutritional security, employment generation, and overall growth of a country. Among fruit crops, mango and banana are largest producing fruits of India. Generally, Karnataka is called as the horticultural state of India. In Karnataka, mango and banana are highest producing fruit crops. With these prospective, yield of mango and banana of Karnataka have been chosen as study variables. Forecasting is a primary aspect of developing economy so that proper planning can be undertaken for sustainable growth of the country. In this study, classes of linear and nonlinear, parametric and non-parametric statistical models have been employed to forecast yield of mango and banana of Karnataka. The major drawback of linear models is the presumed linear form of the model. In most of the cases, the time series are not purely linear or nonlinear as they contain both linear and nonlinear components. To overcome this problem a hybrid model has been proposed which consists of linear and nonlinear models. The hybrid model with the combination of Autoregressive Integrated Moving Average (ARIMA) and Support Vector Regression model performed better in both model building as well as in model validation as compared to other models.

Keywords: ARIMA, Hybrid models, NLSVR, Regression model, Time Series, TDNN.

INTRODUCTION

Agriculture is backbone of Indian economy accounting for 14 percent of the nation's Gross Domestic Product (GDP) and about 11 percent of the country's exports. Nearly about 65 percent of country's population still depends on agriculture for employment and livelihood. Though the contribution of agricultural sector to GDP is decreasing from last decade, but a significant change in the composition of agriculture, showing shifting from cropping towards horticulture, livestock, and fisheries, is noticeable. The horticulture sector contributed 30 percent of agricultural GDP, while the contribution of livestock sector is 4 percent. India witnessed the shift in area from food grains towards horticultural crops over the last five years (2010-2011 to 2014-2015) (NHB data base, 2014-2015). The area under horticultural crops has been increased about 18 percent but the augmentation of area of food grains is only 5 percent during this period. Since last decade, area under horticultural crops is increasing by 3 percent per year and the annual production is also increasing by 7 percent.

Among the horticultural crops, fruit crops are cultivated in an area of 7,216.00 ('000 ha) with production of 88,977.21 ('000 tons) (NHB data base, 2014-2015). Fruit crops play a significant role in the economic development, nutritional security, employment generation, and overall

[1] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India-11002.
*Corresponding author; e-mail: santosha.rathod@icar.gov.in
[2]Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, UP, India-221005.

growth of the 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. A large size of the population in India is engaged in fruit production, distribution and marketing (Yadav and Pandey, 2016). Fruits being the major source of vitamins and minerals are aptly called 'protective foods' and are indispensable part of human diet. Although India contributes 11.80 percent of 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). Hence, there exists a great gap to increase the yield of fruit crops.

India is the largest producer of fruits next to China. The annual production of fruits has been estimated to be 88.97 MT from an area of 06.38 Million ha. Over the decades, increase in area and production accounts to around 30.00 percent and 54.00 percent, respectively (Anonymous, 2015). With availability of diversified climatic and soil conditions, it is possible to grow an assorted range of tropical, subtropical, temperate, and arid zone fruit crops in the country (Radha and Mathew, 2007). India has emerged as leader in production of several horticultural crops *viz.*, mango, banana, cashew nut, areca nut, potato, papaya, okra, etc. Among the fruit crops, Mango is cultivated in an area of 2,500.10 ('000 ha) with a production of 18,002 ('000 tons) with average productivity of 7.3 MT ha^{-1}. Banana contributes an area of 776.50 ('000 ha) with a production of 26,509.10 tons ('000 tons) (NHB data base, 2014-2015) with average productivity of 37 MT ha^{-1}. Among fruit crops, mango and banana are largest producing fruits of India. On the other hand, Karnataka is called as the horticultural state of India. In Karnataka, banana is second in area (101,532 ha) and first in production (2,581,752 MT) with average productivity of 25.43 MT ha^{-1}. Mango is first in area (173,080 ha) and second in production (1,641,165 MT) with average productivity of 9.48 MT ha^{-1}) (Anonymous, 2015b). The Mango and banana are cultivated in almost all districts of Karnataka. With these prospective, yields of mango and banana of Karnataka have been chosen as study variables.

Statistical forecasting is used to provide assistance in decision making and planning the future more effectively and efficiently. Forecasting is a primary aspect of developing economy so that proper planning can be undertaken for sustainable growth of the country. Mainly there are two approaches of forecasting viz., (i) Prediction of present series based on behavior of past series over a period of time called as the extrapolation method, (ii) Estimation of future phenomenon by considering the factors which influence the future phenomenon, *i.e.*, the explanatory method (Diebold and Lopez, 1996). Statistical forecasting is the likelihood approximation of an event taking place in future. (Box and Jenkins, 1970)

Considering the above mentioned facts, a study was conducted to model and forecast the yield of mango and banana in Karnataka. Most commonly used classical linear time series models are ARIMA and linear regression models. Rathod *et al.* (2011), Narayanaswamy *et al.* (2012a), Narayanaswamy *et al.* (2012b), and Pardhi *et al.* (2016) applied regression analysis to study the effect of agricultural inputs and weather parameters on agricultural and horticultural crops. Naveena *et al.* (2014) used different time series models to forecast the coconut production of India. Khan *et al.* (2008) and Qureshi (2014) forecasted mango production of Pakistan using different statistical models. Omar *et al.* (2014) carried out price forecasting and spatial co-integration of banana in Bangladesh. Soares *et al.* (2014) compared different techniques for forecasting yield of banana plants. Olsen and Goodwin (2005) carried out a statistical survey on Oregon hazelnut production. Peiris *et al.* (2008) predicted coconut production in Sri Lanka using seasonal climate information. Mayer and Stephenson (2016) carried out statistical forecasting of Australian macadamia crop.

The major drawback of regression and AutoRegressive Integrated Moving Average (ARIMA) models is the presumed linear form of the model, i.e. a linear correlation pattern is assumed among the time series, and hence no nonlinear patterns can be captured by these models. The time series which contain both linear and nonlinear components, rarely are pure linear or nonlinear. Under such condition Neither ARIMA nor Artificial Neural Network (ANN) and Nonlinear Support Vector Regression (NLSVR) models are adequate in modeling the series which contains both linear and nonlinear patterns. To overcome this difficulty, hybrid time series method was evolved. Applications of hybrid methods in the literature (Zhang, 2003; Chen and Wang, 2007; Jha and Sinha, 2014; Kumar and Prajneshu, 2015; Ray *et al.*, 2016; Naveena *et al.*, 2017a; Naveena *et al.*, 2017b; Rathod *et al.*, 2017) shows that amalgamating different methods can be an efficient and effective way to improve time series forecasting. With these motivations, attempt has been made to develop hybrid forecasting models for forecasting yield of mango and banana in Karnataka by combining ARIMA with ANN and ARIMA with NLSVR models. The details methodology is explained in subsequent sections.

MATERIALS AND METHODS

Data Description

Yearly data on yield (MT ha^{-1}) of mango and banana were collected from data base of National Horticulture Board (NHB) and http://www.indiastat.com. Daily data on weather variables *viz.*, maximum temperature (^{0}C), minimum temperature (^{0}C), relative humidity (fraction), precipitation (mm) and wind speed (miles per second) and solar radiation (mega Joules per square meter) were obtained from http://www.indiawaterportal.org, a secondary website of India meteorological department and from http://globalweather.tamu.edu. Annual data on other exogenous variables (Table 1) were collected from "Agricultural Statistics at a Glance 2014-2015" published

Table 1. Variables considered for regression analysis.

Sl No	Variables	Units
1	Mango yield	MT ha^{-1}
2	Mango area	Hectares
3	Mango production	Million tons
4	Banana yield	MT ha^{-1}
5	Banana area	Hectares
6	Banana production	Million Tons
7	Maximum temperature (Index 1)	Degree Celsius (^{0}C)
8	Minimum temperature (Index 1)	Degree Celsius (^{0}C)
9	Relative humidity (Index 1)	Fraction
10	Precipitation (Index 1)	Millimeter (mm)
11	Wind speed (Index 1)	Miles per second (mps)
12	Solar radiation (Index 1)	Megajoules per square meter (MJ m^{-2})
13	Maximum temperature (Index 2)	Degree Celsius (^{0}C)
14	Minimum temperature (Index 2)	Degree Celsius (^{0}C)
15	Relative humidity (Index 2)	Fraction
16	Precipitation (Index 2)	Millimeter (mm)
17	Wind speed (Index 2)	Miles per second (mps)
18	Solar radiation (Index 2)	Mega Joules per square meter (MJ m^{-2})
19	Avg. size of operational holdings	Hectares
20	Area sown	Hectares
21	Net area irrigated	Hectares
22	Fertilizer distribution	Tons
23	Argil. credit cooperative societies	Numbers
24	Regulated markets	Numbers
25	Rural road length	Kilo meters (Kms)
26	No of IP sets	Numbers

by Department of Economics and Statistics, Karnataka. For regression analysis i.e. weather based forecasting, data on yield (MT ha^{-1}), area (ha) and production (MT) of mango and banana from 1985 to 2011 were used for model building and data from 2012 to 2014 were used to check the forecasting performance of the model. The information on weather variables (daily data) and other agricultural variables (annual data) were also used from 1985 to 2014.

In this study, information on exogenous variables are not available for longer period of time, however, time series models yield better results when we consider data on longer period. To overcome this constraint, only univariate data on yield of mango and banana has been considered. For mango (1980-1981 to 2013-2014) and banana (1954-1955 to 2014-2015), yearly data on yield (MT ha^{-1}), area (ha), and production (MT) were collected from data base of National Horticulture Board (NHB) 2014-15 and http://www.indiastat.com. To forecast yield of mango of Karnataka, data from 1980 to 2011 were used for model building and 2012 to 2014 were used to check the forecasting performance of the models. In the case of banana of Karnataka, data from 1954 to 2011 were used for model building and 2012 to 2014 were used for model validation. The list of variables considered for regression based forecasting is listed in Table 1.

Statistical Methodologies

In this work, number of statistical techniques viz., regression model, weather indices, ARIMA, ANN, NLSVR and proposed hybrid methodology are used to forecast the yield of mango and banana of Karnataka, India.

Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \cdots, + \beta_p X_p + \varepsilon \qquad (1)$$

Where, Y is the dependent (response) variable, X are independent (predictor or stimulus) variables, $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients and ε is the error term. An important issue in regression modeling is the selection of explanatory variables which are really influencing the dependent variable. There are many methods for selection; stepwise regression analysis is frequently used variable selection algorithm in regression analysis (Montgomery *et al.*, 2003). In this work, step wise regression analysis has been used because number of exogenous variables is more.

Weather Indices

For the daily data (d) on p variables, new weather variables and interaction components can be generated with respect to each of the weather variables using the below mentioned procedure (Agrawal *et al.*, 2001). In order to study the individual effect of each weather variable, two new variables from each variable can be generated as follows:

Let X_{id} be the value of the i^{th} weather variable at the d^{th} day, r_{id} is the simple correlation coefficient between weather variable X_i at the d^{th} day and yield Y_i over a period of n years, which is expressed as follows;

$$r_{id} = \frac{n\sum_{d=1}^{n} XY - \sum_{d=1}^{n} X \sum_{d=1}^{n} Y}{\sqrt{[n\sum_{d=1}^{n} X^{2} - (\sum_{d=1}^{n} X)^2] - [n\sum_{d=1}^{n} Y^{2} - (\sum_{d=1}^{n} Y)^2]}} \qquad (2)$$

The generated variables are given as follows;

$$Z_{ij} = \frac{\sum_{d=1}^{n} r_{id}^{\,j} x_{id}}{\sum_{d=1}^{n} r_{id}^{\,j}}, \; j = 0,1 \qquad (3)$$

For j= 0, we have unweighted generated variable as:

$$Z_{ij} = \frac{\sum_{d=1}^{n} x_{id}}{n} \qquad (4)$$

and weighted generated variables as:

$$Z_{i1} = \frac{\sum_{d=1}^{n} r_{id} x_{id}}{\sum_{d=1}^{n} r_{id}} \qquad (5)$$

Weather indices were constructed using daily weather variables. After calculating these indices [weighted (Index 2) as well as unweighted (Index 1) indices], they were used as independent variables in regression models [Equation (1)].

AutoRegressive Integrated Moving Average (ARIMA) Model

One of the most important and widely used classical time series models is the AutoRegressive 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 building procedure. Often, the time series are non-stationary in nature. To obtain the stationary time series, we need to introduce the differencing term *d.* to make the non-stationary series to stationary one. Then, the general form of ARMA model is represented as 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 integrated ARMA model if $\Delta Y_t = (1 - B)^d \varepsilon_t$.

The ARIMA model is expressed as follows:

$$\emptyset(B)(1 - B)^d Y_t = \theta(B)\varepsilon_t \qquad (6)$$

Where, $\varepsilon_t \sim WN\ (0, \sigma^2)$ and WN is the white noise. 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.

Time Delay Neural Network (TDNN)

The ANN for time series analysis is termed as Time Delay Neural Network (TDNN). The time series phenomenon can be mathematically modelled using neural network with implicit functional representation of time, whereas static neural network like multilayer perceptron is presented with dynamic properties (Haykin, 1999). 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.

$$Y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} Y_{t-p}\right) + \varepsilon_t \qquad (7)$$

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 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.

Nonlinear Support Vector Regression (NLSVR) Model

Support Vector Machine (SVM) is a supervised machine learning technique which was originally developed for linear classification problems. Later in the year 1997, the support vector machine for regression problems were developed by Vapnik by introducing ε-insensitive loss function (Vapnik *et al.*, 1997) and it has been extended to the nonlinear regression estimation problems. Modeling of such problems is called as NonLinear Support Vector Regression (NLSVR) model. The basic principle involved in NLSVR is to transform the original input time series into

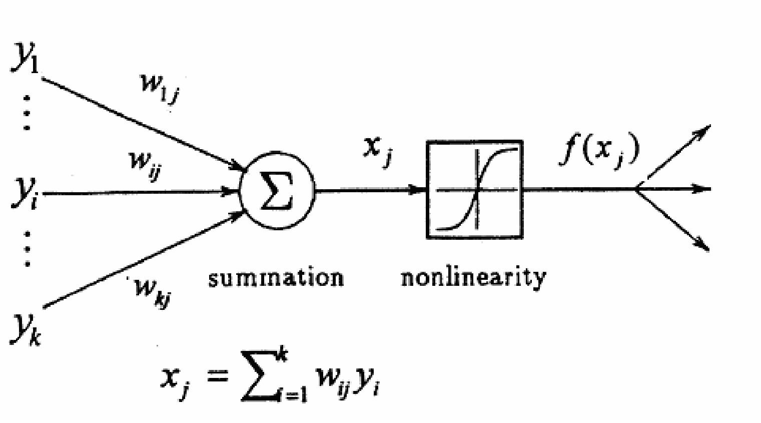

Figure 1. Neural network structure.

a high dimensional feature space and then build the regression model in a new feature space. Let us 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. The general equation of Nonlinear Support Vector Regression estimation function is given as follows:

$$f(x) = W^T \phi\,(x) + b \tag{8}$$

Where, $\phi(.)$: $R^n \rightarrow R^{nh}$ is a nonlinear mapping function which maps 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.

The Proposed Hybrid Methodology

The hybrid method considers the time series y_t 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

$$y_t = L_t + N_t \tag{9}$$

Where, L_t and N_t represent the linear and nonlinear component, respectively. In this work, 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 (Brock *et al*., 1996) Once the residuals confirm the non-linearity, then, they are modelled and predicted using TDNN and NLSVR. Finally, the forecasted linear and nonlinear components are combined to generate aggregate forecast.

$\hat{y}_t = \hat{L}_t + \hat{N}_t$ (10) Where, $\hat{L}$ and $\hat{N}$ represent the predicted linear and nonlinear component, respectively. The graphical representation of hybrid methodology is expressed in Figure 2.

Finally, the performance of the models under consideration is compared by using Mean Absolute Percentage Error (MAPE).

RESULTS AND DISCUSSION

Regression Analysis

Regression analysis has been carried out to know the factors influencing yield of mango and banana in Karnataka. Regression model was fitted for yield of mango and banana. The dependent variables in the study are yield of mango and banana, whereas independent variables include weather variables and some other socio-economic

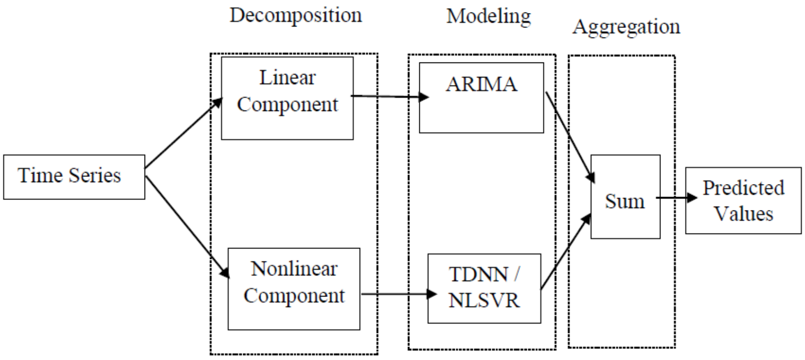

Figure 2. Schematic representation of hybrid methodology.

and agricultural variables listed in Table 1. Data on these variables were considered from 1985-1986 to 2014-2015. Data from 1985-86 to 2011-2012 were used for model building and data from 2012-2013 to 2014-2015 were used for model validation.

The summary statistics of yield of mango and banana time series is presented in Table 2, which shows that banana yield time series are highly heterogeneous as Coefficient of Variation (CV) is very high. To begin with regression analysis, multiple linear regression analysis was carried out for all the four data set separately. As explained in methodology section, weather indices have been developed and these indices are considered as independent variables in MLR model. For one independent variable, two weather indices *viz*., weighted and unweighted indices has been developed as explained in methodology section. (weighted and un-weighted), therefore, total number of independent variables becomes 22 in this study.

Table 2. Descriptive Statistics of time series under consideration.

Statistics	Mango yield	Banana yield
Mean	9.60	14.87
Median	9.55	8.60
Mode	-	32.95
Standard deviation	0.58	10.71
Minimum	8.39	2.98
Maximum	10.84	33.03
Skewness	0.12	0.52
Kurtosis	0.48	-1.44
Coefficient of variation (%)	6.00	72.01

The Multiple Linear Regression (MLR) analysis was carried out by considering all the independent variables (Table 1). It was found that the coefficient of determination (R^2) of MLR model is very high (Table 3). Although the R^2 value of MLR model is very high, most of the variables in the models are non-significant and Variance Inflating Factor (VIF) is also very high. This clearly indicates that there is a multi-collinearity problem among the independent variables. To overcome the multi-collinearity problem, one of the measures is to drop the unimportant variables which are explaining less variation in dependent variables in the model. The dropping of variable can be done through the stepwise regression analysis (Gujarati *et al.*, 2013).

Hence, stepwise regression analysis was carried out to fit the model (Table 4 and 5). In stepwise regression analysis, the value of R^2 was decreased and multicollinearity problem was also reduced as Variance Inflating Factor (VIF< 10) was less and most importantly number of significant variables were increased. For stepwise regression analysis of time series on mango yield (Table 4) infer that as area and production increases, the mango yield also increases. As the variables *viz.* net irrigated area and number of agricultural credit cooperative

Table 3. MLR Model information.

Regression model/Dependent variables	Adj R^2	Number of significant variables	Number of variables having *VIF*> 10
Mango yield	0.941	Nil	19
Banana yield	0.938	1	22

Table 4. Stepwise regression analysis of mango yield time series.

Variable	Coefficient	Std error	*t* Test	Probability	*VIF*	R^2
Intercept	6.24	0.914	6.25	0.0003	-	0.845
Area	5.88	0.001	5.88	<.0001	4.51	
Net area irrigated	5.35	1.4E-7	5.35	<.0001	3.56	
Production	4.22	0.002	4.22	0.0091	1.29	
Agricultural credit cooperative societies	2.37	0.000	2.37	0.0271	4.63	
Relative humidity	-2.42	1.310	-2.42	0.0244	1.66	

societies increased, the mango yield also increased. However, it has negative relation with relative humidity.

As discussed earlier, the data set from 1985-1986 to 2011-2012 were used for model building and data from 2012-2013 to 2014-2015 were used for validation. Performance of stepwise regression model in both training and testing data set is given in Tables 12 and 13, respectively.

Results of ARIMA Model

An ARIMA model was built using SAS 9.4 software available at ICAR-Indian Agricultural Statistics Research Institute, New Delhi. The ARIMA model for both mango and banana yield time series was built separately. The Box-Jenkins methodology of model building was followed. After identification of candidate model order, maximum likelihood method was used for parameter estimation (Tables 6 and 7). Based on the probability of residuals obtained, one can say that the residuals are non-correlated. Since, the model satisfies Box-Jenkins methodology of model building *viz.*, model identification, stationarity, parameter estimation and diagnostic checking, then, one can go for forecasting task. The forecasting performance of mango and banana yield time series in both training and testing data sets are given in Tables 12 and 13.

Table 5. Stepwise regression analysis of Banana Yield time series.

Variable	Coefficient	Std error	*t* Test	Probability	*VIF*	Adj R^2
Constant	16.7	9.37	1.79	0.0891	-	0.942
Production	0.25	0.002	15.56	0.0005	5.93	
Area	0.48	0.045	10.85	0.0003	5.32	
Solar radiation	-1.19	0.390	3.06	0.0064	2.85	
Relative humidity	0.23	0.071	3.28	0.0154	2.20	
No of regulated markets	0.49	0.017	2.94	0.0086	3.46	

Table 6. Parameter estimation of ARIMA (0 1 1) for mango yield time series.

Parameter	Estimate	Standard error	*t* Value	Approx *Pr*> |*t*|	Lag	P(Resi.) at 6 Lag
Constant	0.033	0.038	0.87	0.382	0	0.240
MA 1	0.581	0.161	3.64	0.003	1	

Results of TDNN Model

A feed forward time delay neural network was fitted for both mango and banana yield time series separately using R software with the help of package ‘forecast’ (Hyndman, 2017). The Levenberg-Marquardt (LM) back propagation algorithm was used for TDNN model building and based on repetitive experimentation, the learning rate and momentum term for all TDNN models were set as 0.02 and 0.001, respectively. Sigmoidal and linear functions were used as activation function in hidden and output layers, respectively. More than 80 percent of the observations in data set were used for model training and remaining observations were used for testing or validation. Different numbers of neural network models were tried before arriving at the final structure of the model and, finally, the adequate model given in Table 8 was obtained. After model building, forecasting the time series was done both for training and testing data set. The forecast values of TDNN model are given in Tables 12 and 13.

Results of NLSVR Model

The nonlinear support vector regression model was employed for all four time series separately using R software with the help of package ‘e1071’ (David, M. (2017). The parameter specifications, cross validation error are given in Table 9. The performance of NLSVR model strongly depends on the kernel function and set of hyper-parameters. The RBF kernel function in NLSVR requires optimization of two hyper-parameters, i.e. the regularization parameter *C*, which balances the complexity and approximation accuracy of the model and the kernel bandwidth parameter γ, which defines the variance of RBF kernel function. These tuning parameters *viz.*, *C* and γ, are user defined parameters. NLSVR tuning parameters (Table 9) were defined to minimize the training testing error. As in TDNN, the two lag delay was used as model input variables. Also, the training and testing data ratio was followed the same as in TDNN i.e. more than 80 percent of data set was used for training the model and remaining for model testing. The forecasting performance of NLSVR model in both training and testing data set are given in Tables 12 and 13.

Results of Hybrid Time Series Models

As discussed in methodology section, the first step in hybrid time series modeling was to test the nonlinearity of the residuls. The BDS test was applied to test the nonlinearity of the residulas. The BDS test result (Tables 10 and 11) shows that the residuals obtained from regression models were liner and non-significant and residuals of ARIMA models are nonlinear and significant. Therfore, hybrid models were built only with ARIMA model. Once the residual series is found to be nonlinear, it can be modelled and

Table 7. Parameter estimation of ARIMA (0 1 2) for banana yield time series.

Parameter	Estimate	Standard error	*t* Value	Approx *Pr*> |*t*|	Lag	P (Resi.) at 6 Lag
MU	0.285	0.471	0.60	0.54	0	0.824
MA 1	-0.020	0.001	11.11	< 0.0001	1	
MA 2	-0.232	0.051	4.54	< 0.0001	2	

Table 8. TDNN Model Specifications.

Time series	Activation function		Time delay	No of hidden nodes	Total No of parameters
	Hidden layer	Output layer			
Mango yield	Sigmoidal	Linear	2	4	17
Banana yield	Sigmoidal	Linear	2	10	41

Table 9. NLSVR Model specifications.

Time series	Kernel function	No of *SVs*	*C*	γ	ε	K Fold cross validation (K)	Cross validation error
Mango yield	RBF	26	1.10	1.00	0.01	10	0.037
Banana yield	RBF	18	1.50	1.62	0.10	10	0.044

Table 10. Nonlinearity testing of Regression residuals by BDS test.

Time series	Parameter	Dimension (m= 2)		Dimension (m= 3)	
		Statistic	Probability	Statistic	Probability
Mango yield	2.78	1.14	0.25	1.17	0.09
Banana yield	5.56	1.54	0.12	0.61	0.54

Table 11. Nonlinearity testing of ARIMA residuals by BDS test.

Time series	Parameter	Dimension (m= 2)		Dimension (m= 3)	
		Statistic	Probability	Statistic	Probability
Mango yield	0.34	2.82	0.004	3.58	< 0.001
Banana yield	0.78	7.60	< 0.001	8.34	< 0.001

predicted using nonlinear models. The nonlinear models, namely, TDNN and NLSVR were used for modeling and forecasting of ARIMA residuals in this study. After the confirmation of nonlinearity of ARIMA residuls, the same residuls were modelled and forecasted using TDNN and NLSVR models. Further, the forecasted residuals were combined with the forecasts obtained from original ARIMA model.

ARIMA-TDNN Hybrid Time Series Model

After the confirmation of nonlinearity of ARIMA residuls, the same residuls were modelled and forecasted using TDNN model. Further, the forecasted residuals were combined with the forecasts obtained from original ARIMA model. These modeling procedure is called ARIMA-TDNN Hybrid time series model. Finally, the forecasting performance of ARIMA-TDNN hybrid model in both training and testing data sets are given in Tables 12 and 13, respectively.

ARIMA-NLSVR Hybrid Time Series Model

The same procedure was follwed as described in ARIMA-TDNN hybrid time series model. After the confirmation of nonlinearity of ARIMA residuls, the same residuls were modelled and forecasted using NLSVR model. Further, the forecasted residuals were combined with the forecasts obtained from original ARIMA model. These modeling procedure is called ARIMA-NLSVR Hybrid time series model. Finally, the forecasting performance of

Table 12. Comparison of forecasting performance of all models in training data set.

Criteria	Stepwise regression	ARIMA	TDNN	NLSVR	ARIMA-TDNN	ARIMA-NLSVR
			Mango yield			
MAPE	18.85	3.83	2.89	2.81	1.98	1.73
			Banana yield			
MAPE	13.12	12.10	7.58	6.93	5.12	4.73

Table 13. Comparison of forecasting performance of all models in testing data set.

Year	Actual	Forecast					
		Stepwise regression	ARIMA	TDNN	NLSVR	ARIMA-TDNN	ARIMA-NLSVR
Mango yield							
2012	10.84	13.85	11.75	9.68	10.71	10.12	10.59
2013	10.04	13.69	11.15	10.14	10.73	10.62	10.44
2014	9.93	13.94	8.67	10.37	9.25	10.01	10.12
MAPE		34.83	10.71	5.37	4.97	4.40	2.73
Banana yield							
2012	25.57	28.11	23.57	28.04	27.73	27.52	26.71
2013	27.50	31.62	22.81	29.09	28.91	29.25	28.19
2014	27.59	34.89	23.10	28.92	28.83	26.25	29.88
MAPE		17.14	13.69	6.76	6.03	5.45	5.09

ARIMA-NLSVR hybrid model in both training and testing data sets are given in Tables 12 and 13, respectively.

Forecasting Performance of Models under Consideration

The models *viz.* stepwise regression, ARIMA, TDNN, NLSVR, ARIMA-TDNN and ARIMA-SVR were studied for forecasting mango and banana yield time series of Karnataka, India. Forecasting performance of each model under training and testing data set was compared. Even though we considered many exogenous variables in regression model, the univariate ARIMA, TDNN and NLSVR performed better as compared to regression models in both mango and banana yield data.

The performance of machine intelligence techniques like TDNN and NLSVR is better as compared to linear time series models under both training and testing data set. Even though we considered many exogenous variables in stepwise regression model, then also the result of univariate models, specially the machine learning models, performed better compared to regression model. Even the coefficient of variation time series were very high, the TDNN, NLSVR and hybrid models performed better. The reason could be the nonlinear machine learning techniques which can capture the heterogeneous trend in the data set and, therefore, performed well as compared to regression and ARIMA model.

The TDNN and NLSVR models performed well over linear models like stepwise regression model and ARIMA model due to their superior predictive ability in nonlinear and heterogonous data set. Among the machines, intelligence techniques like TDNN and NLSVR, the NLSVR performed better in both training and testing the data set.

As discussed in hybrid methodology section, the hybrid models have their own advantage over single models. Based on the lowest MAPE values of all models obtained for both training (Table 12) and testing data set (Table 13) considered, one can infer that hybrid model consisting of ARIMA and NLSVR i.e. ARIMA-NLSVR model, outperformed all remaining models. Both hybrid models viz., ARIMA-TDNN and ARIMA-NLSVR, outperformed the single model viz., ARIMA, TDNN and NLSVR. Finally, among all models under study, ARIMA-NLSVR model’s performance was the best. Hybrid methodology considers both linearity and nonlinearity of the data set, hence, the performance of ARIMA-NLSVR model is superior as compared to all other models under both training and testing data set for modeling and forecasting mango and banana yield time series of Karnataka.

CONCLUSIONS

Based on the results obtained in this work, one can conclude that machine intelligence techniques like time delay neural network and nonlinear support vector regression perform better as compared to classical time series models under heteroscedastic and noisy time series data. The main finding of this study is the performance of hybrid time series model which is better as compared to single models. In this study, since the data set consisted of both linear and nonlinear pattern, the hybrid model performed better as compared to single time series or machine learning techniques for modeling and forecasting mango and banana yield time series of Karnataka. Among the hybrid models, the ARIMA with Nonlinear Support Vector Regression i.e. ARIMA-NLSVR model performed superior as compared to all other models under both training and testing data set. The stepwise regression analysis shows that some variables which strongly influence the yield of mango and banana, the government or policy makers should emphasize focus on such factors, for overall development of cropping pattern under consideration. Based on the results obtained, one can conclude that the farmers or policy makers involved in mango and banana crop production can plan well in advance to further increase the productivity of crops by suitable management of the inputs and weather variable which obtained significant in this study.

The hybrid approach can be further extended using some other machine learning techniques for varying autoregressive and moving average orders so that practical validity of the model can be well established. The validity of hybrid models can be generalized by applying this approach to other horticultural and agricultural data.

REFERENCES

1. Agrawal, R., Jain, R. C. and Mehta, S. C. 2001. Yield Forecast Based on Weather Variables and Agricultural Inputs on Agro-Climatic Zone Basis. *Ind. J. Agric. Sci.*, **71 (7)**: 487-490.
2. Anonymous. 2015a. *Horticultural Statistics at a Glance*. Horticulture Statistics Division Department of Agriculture, Cooperation and Farmers Welfare Ministry of Agriculture and Farmers Welfare Government of India.
3. Anonymous. 2015b. *Karnataka at a Glance*. Department of Economics and Statistics, Government of Karnataka, India.
4. Brock, W. A., Dechert, W. D., Scheinkman, J. A. and Lebaron, B. 1996. A Test for Independence Based on the Correlation Dimension. *Econ. Rev.*, **15**: 197-235.
5. Chen, K. Y. and Wang, C. H. 2007. Support Vector Regression with Genetic Algorithm in Forecasting Tourism Demand. *Tour. Manage.*, **28**: 215-226.
6. David, M. 2017. E1071: Misc Functions of the Department of Statistics, Probability Theory Group. R Package Version **1.6-8,** https://cran.r-project.org/web/packages/e1071/index.html
7. Diebold, F. X. and Lopez, J. A. 1996. *Forecast Evaluation and Combination*: *Handbook of Statistics **14***. Elsevier Science, Amsterdam.
8. Gujarati, D. N., Porter, D. C. and Gunasekar, S. 2013. *Basic Econometrics.* Fifth Edition, Tata McGraw-Hill Education Pvt. Ltd, ISBN **10**: 0071333452/ISBN 13: 9780071333450.
9. Haykin, S. 1999. Neural Networks: A Comprehensive Foundation. New York. Macmillan, ISBN 0-02-352781-7.
10. Hyndman, R. J. 2017. Forecast: Forecasting Functions for Time Series and Linear Models. R Package Version **8.1**., https://cran.r-project.org/web/packages/forecast/index.html
11. Jha, G. K. and Sinha, K. 2014. Time-Delay Neural Networks for Time Series Prediction: An Application to the Monthly Wholesale Price of Oilseeds in India. *Neural Comput. Appl.*, **24(3)**: 563-571
12. Khan, M., Mustafa, K., Shah, M., Khan, N. and Khan, J. Z. 2008. Forecasting Mango Production in Pakistan an Econometric Model Approach. *Sarhad J. Agri.*, **24(2)**: 363-370.
13. Kumar, T. L. M. and Prajneshu, 2015. Development of Hybrid Models for Forecasting Time-Series Data Using Nonlinear SVR Enhanced by PSO. *J. Stat. Theor. Pract.*, **9(4)**: 699-711.

14. Mayer, D. G. and Stephenson, R. A. 2016. Statistical Forecasting of the Australian Macadamia Crop. *Acta Hortic.*, **1109**: 265-270. doi: 10.17660/ActaHortic.2016.1109.43
15. Montgomery, D. C., Peck, E. A. and Vining, G. 2003. *Introduction to Linear Regression Analysis*. 3rd Edition, John Wiley and Sons (Asia) Pte. Ltd.
16. Narayanaswamy, T., Surendra, H. S. and Rathod, S. 2012a. Multiple Stepwise Regression Analysis to Estimate Root Length, Seed Yield per Plant and Number of Capsules per Plant in Sesame (*Seasamum indicum* L.). *Mysore J. Agricu. Sci*., **46 (3)**: 581-587.
17. Narayanaswamy, T., Surendra, H. S and Rathod, S. 2012b. Fitting of Statistical Models for Growth Patterns of Root and Shoot Morphological Traits in Sesame (*Seasamum indicum* L.). *Environ. Ecol*., **30(4)**: 1362-1365.
18. National Horticultural Board (NHB) Data Base. 2014-2015. *Current Scenario of Horticulture in India.* http://nhb.gov.in/area-pro/NHB_Database_2015.pdf
19. Naveena, K., Rathod, S., Shukla, G. and Yogish, K. J. 2014. Forecasting of Cocnonut Production in India: A Suitable Time Series Model. *Int. J. Agric. Eng.*, **7(1)**: 190-193.
20. Naveena, K., Singh, S., Rathod, S. and Singh, A. 2017a. Hybrid ARIMA-ANN Modelling for Forecasting the Price of Robusta Coffee in India. *Int. J. Curr. Microbiol. Appl. Sci.*, **6(7)**: 1721-1726.
21. Naveena, K., Singh, S., Rathod, S., and Singh, A. 2017b. Hybrid Time Series Modelling for Forecasting the Price of Washed Coffee (Arabica Plantation Coffee) in India. *Int. J. Agric. Sci.*, **9(10)**: 4004-4007.
22. Olsen, J. and Goodwin, J. 2005. The Methods and Results of the Oregon Agricultural Statistics Service: Annual Objective Yield Survey of Oregon Hazelnut Production. *Acta Hortic.*, **686:** 533-537.
23. Omar, M. I., Dewan, M. F. and Hoq, M. S. 2014. Analysis of Price Forecasting and Spatial Co-Integration of Banana in Bangladesh, *Eur. J. Business Manage.*, **6(7)**: 244-255.
24. Pardhi, R., Singh, R., Rathod, S. and Singh, P. K. 2016. Effect of Price of Other Seasonal Fruits on Mango Price in Uttar Pradesh. *Econ. Affairs*, **61(4)**:1-5.
25. Peiris, T. S. G., Hansen, J. W. and Zubair, L. 2008. Use of Seasonal Climate Information to Predict Coconut Production in Sri Lanka. *Int. J. Climatol.,* **28:** 103–110. doi: 10.1002/joc.1517.
26. Qureshi, M. N. 2014. Modelling on Mango Production in Pakistan. *Sci. Int., (Lahore),* **26(3)**: 1227-1231.
27. Radha, T. and Mathew, L. 2007. *Fruit Crops.* New India Publ. Agency.
28. Rathod, S. Surendra, H. S., Munirajappa, R. and Chandrashekar. H. 2011. Statistical Assessment on the Factor Influencing Agricultural Diversification in Different Districts of Karnataka. *Environ. Ecol.,* **30 (3A)**: 790-794.
29. Rathod, S., Singh, K, N., Paul, R. K., Meher, R. K., Mishra, G. C., Gurung, B., Ray, M. and Sinha, K. 2017. An Improved ARFIMA Model using Maximum Overlap Discrete Wavelet Transform (MODWT) and ANN for Forecasting Agricultural Commodity Price. *J. Ind. Soc. Agric. Stat*., **71(2)**: 103–111.
30. Ray, M., Rai, A., Ramasubramanian, V. and Singh, K. N. 2016. ARIMA-WNN Hybrid Model for Forecasting Wheat Yield Time-Series Data. *J. Ind. Soc. Agric. Stat.,* **70(1)**: 63-70.
31. Soares, J. D. R., Pasqual, M., Lacerda, W. S., Silva, S. O. and Donato, S. L. R. 2014. Comparison of Techniques Used in the Prediction of Yield in Banana Plants. *Scientia Hortic*., **167**: 84-90.
32. Vapnik, V., Golowich, S. and Smola, A. 1997. 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, **9**:281-287,
33. Yadav, A. S. and Pandey, D. C. 2016. Geographical Perspectives of Mango Production in India. *Imperial J. Interdisciplinary Res.*, **2(4)**: 257-265.
34. Zhang, G. P. 2003. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing*, **50**: 159-175.

مدل های آماری برای پیش بینی عملکرد موز و مانگو در ایالت کارناتاکا هندوستان

س. رتود، و ج. س. میشرا

چکیده

بخش باغبانی نقش بارزی در رشد اقتصادی بیشتر کشورهای در حال توسعه بازی می کند. هندوستان بعد از چین بزرگترین تولید کننده میوه و سبزی در جهان است. در میان گیاهان باغبانی، بیشتر مساحت به درختان میوه اختصاص دارد. درختان میوه در توسعه اقتصادی، امنیت غذایی، ایجاد شغل، و رشد عمومی کشور نقش عمده ای دارد. در میان میوه ها، موز و مانگو بیشترین میوه های تولیدی هند هستند. به طور کلی، کارناتاکا ایالت باغبانی هند شناخته می شود. در این ایالت، موز و مانگو بیشترین تولید کننده میوه میباشند. با این تصویر، ، عملکرد مانگو و موز کارناتاکا به عنوان متغیر های مطالعه حاضر انتخاب شدند. گفتنی است که پیش بینی کردن یک جنبه اساسی در اقتصاد های در حال توسعه است تا بتوان برنامه ریزی برای رشد پایدار کشور را به گونه ای مناسب انجام داد. در این پژوهش، برای پیش بینی عملکرد موز و مانگو در کارناتاکا مدل های آماری در گروه های خطی، غیر خطی، پارامتریک، و غیر پارامتری به کار رفت. عمده ترین ایراد مدل های خطی همین فرض خطی بودن مدل است زیرا در بیشتر موارد، سری های زمانی به طور خالص(کامل) خطی یا غیر خطی نیستند چون آن ها هر دو جزء خطی و غیر خطی را دارند. برای رفع این مسله، یک مدل دو رگه (هیبرید) پیشنهاد شده که حاوی مدل های خطی و غیر خطی است. مدل هیبریدی ترکیبی از مدل Autoregressive Integrated Moving Average (ARIMA) و Support Vector Regression بود و در مقایسه با دیگر مدل ها در مرحله ساخت مدل و مرحله راستی آزمایی آن نتیجه بهتری داشت.