Indian Journal of Agricultural Sciences **88** (1): 22–7, January 2018/Article
https://doi.org/10.56093/ijas.v88i1.79546

Modeling and forecasting of oilseed production of India through artificial intelligence techniques

SANTOSHA RATHOD[1], K N SINGH[2], S G PATIL[3], RAVINDRAKUMAR H NAIK[4], MRINMOY RAY[5] and VIKRAM SINGH MEENA[6]

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 28 February 2017; Accepted: 21 August 2017

ABSTRACT

Indian agriculture has made considerable progress in respect of principle food crops, but the performance in case of oilseed crops is so far not good as compared to food grains. Production of oilseeds and oils are not meeting the increasing demand for edible oils and this widening demand-supply gap has necessitated imports of edible oils. India is world's largest importer of oil seeds as it imports more than 50 percent of its total production. Forecasting is used to analyze the past and current behavior to forecasts the future oilseeds production which intern provide an aid to decision-making and in planning for the future effectively and efficiently. Autoregressive integrated moving average (ARIMA) model is the most widely used model for forecasting time series. One of the main drawback of this model is the presumption of linearity. To model the series which contains nonlinear patterns, the artificial intelligence techniques like time delay neural network (TDNN) and non-linear support vector regression (NLSVR) model are commonly employed. In this paper an attempt has been made to forecast the oilseed production of India using ARIMA, TDNN and NLSVR models. Empirical results clearly reveal that the artificial intelligence techniques outperformed the ARIMA model.

Key words: ARIMA, NLSVR, Oilseed, TDNN, Time series

The oilseed production is also one of the important sector in Indian agriculture, which covers an area about 26.13 million ha and total production of over 32.52 million tonnes which accounts for 14.5% of gross cropped area and about 10% of the total value of output from agriculture. Oilseeds crops are grown in different agro-climatic regions of the country. The oilseeds area and production are mainly concentrated in the central and southern parts of India, mainly in the states of Madhya Pradesh, Gujarat, Rajasthan, Andhra Pradesh and Karnataka. Among different oilseeds, groundnut (6.77 MT), rapeseed and mustard (6.82 MT) and soybean (8.59 MT) together account for about 84.15 per cent of oilseeds area and 87.66 per cent of oilseeds production in the country (2015-16). Soybean is the most important oilseed crop mainly grown in Madhya Pradesh, Maharashtra, and Rajasthan accounting for more than 95 per cent of total production (RBI statistics 2016).

Edible oil is the largest imported (30%) commodity in India next only to petroleum products despite the fact that India had the world's second largest area under oilseeds. Government decided to achieve self-sufficiency in edible oilseeds through various policy and technological interventions. Government has announced many schemes and policy to increase the oil seed productivity. Integrated Scheme on Oilseeds, Pulses, Oil Palm and Maize (ISOPOM) was implemented in 2010. Currently pulses component of ISOPOM has been merged with Natural Food Security Mission (NFSM) to intensify efforts for production of pulses.

Production of oilseeds and oils does not meet the increasing demand for edible oils and this widening demand-supply gap has necessitated imports of edible oils. With competing demands on agricultural land from various crops and enterprises, the production of oilseeds can be increased only if productivity is improved significantly and farmers get remunerative and attractive prices and assured market access. However, farmers face various constraints in oilseeds production. Most of the oilseeds are grown under rainfed conditions, and only 25 percent of area under oilseeds is irrigated. Several biotic, abiotic, technological, institutional, and socio-economic constraints also inhibit exploitation of the yield potential of crops and need to be addressed. By considering facts, viz. changing policy environment, increasing demand, slow growth in domestic production and rising imports, an attempt has been made to develop the statistical and artificial intelligence models to analyze

[1,2,5]e mail: santosha.rathod@icar.gov.in, ICAR-IASRI, New Delhi. [3]ACRI, Tamil Nadu Agricultural University, Killikulam 628 252. [4]Department of Economics, VMSR Vastrad Arts, Science and VS Bellihal Commerce College, Hungund, Karnataka 587 118, [6]Crop Science Division, ICAR Headquarters, Krishi Bhawan, New Delhi 110 001.

the past performance, its trend and future forecast of oil seed production of India.

Forecasting is used to provide an aid to decision-making and in planning for the future effectively and efficiently.It is an important aspect for a developing economy so that adequate planning is undertaken for sustainable growth, overall development and poverty alleviation. Statistical forecasting models are used to develop an appropriate forecast methodology by using the past data to predict the future with the help of the trends and patterns within the data. One of the most important and widely used time series models is the autoregressive integrated moving average (ARIMA) model. The popularity of the ARIMA model is due to its statistical properties as well as the well-known Box–Jenkins methodology (Box and Jenkins 1970) in the model building process. Sarika *et al.* (2011) used ARIMA model for modeling and forecasting India's pigeon-pea production. Suresh *et al.* (2011) applied this model for forecasting sugarcane area, production and productivity in Tamilnadu state of India. Naveena *et al.* (2014) forecasted coconut production of India using ARIMA methodology. Vishwajith *et al.* (2014) forecasted pulses production in India using time series models.

The major drawback of ARIMA model is presumption of linearity, hence, no nonlinear patterns can be recognized by ARIMA model. Sometimes, the time series often contain nonlinear components, under such condition the ARIMA models are not adequate in modeling and forecasting. To overcome this difficulty, many parametric nonlinear models are available in the literature to capture the nonlinear component. These parametric nonlinear models some time fails if the data generating process is highly heterogeneous, complex and nonlinear in nature, to model such data artificial intelligence techniques are the only way to model and forecast such phenomenon. The Artificial Neural Network (ANN) and Support Vector Machine (SVM) are most widely used Artificial Intelligence (AI) techniques to model and forecast the time series data.

The major advantage of neural network is their flexible nonlinear modeling capability to model complex undefined data and no need to specify a particular model specification. Rather, the model is adaptively formed based on the structure and pattern of data (Zhang 1998, Zhang 2003, Jha and Sinha 2014, Ray *et al.* 2016, Naveena *et al* 2017, Naveena *et al* 2017a, Rathod *et al*. 2017). Sreelakhmi and Ramakanthkumar (2008), developed a neural network model for predicting short term wind speed using weather parameters. Another most important and popularly used AI technique is SVM, the SVM was originally developed for classification problems, by introducing the ε-insensitive error loss function, support vector machine (SVM) which was initially proposed for classification problems, has been successfully extended to regression problems by Vapnik in 1997, and it is called as Support Vector Regression (SVR). Many findings in literature shows that SVR methodology can be used to model the series, which contains nonlinear pattern (Alonso *et al.* 2013, Kumar and Prajneshu 2015, Zhang *et al.* 2016).Therefore, in view of the above discussion an attempt has been made to model and forecast the oilseeds production of India using ARIMA, ANN and SVM.

MATERIALS AND METHODS

Yearly data on total oilseed production (MT) of India from 1950-51 to 2015-16 were collected from agricultural statistics released by Reserve Bank of India (RBI), Government of India (RBI statistics 2016). The data from 1950-51 to 2010-11 were used for model building and 2011-12 to 2015-16 were used for model validation. The statistical software's, viz. SAS and R were used for modeling and forecasting of oil seed production of India.

ARIMA is one of the most popular and widely used model for time series modeling given by Box and Jenkins in 1970. In contrast to the regression models, the ARIMA model allows to explain by its past, or lagged values and stochastic error terms. These models are often referred to as "mixed models" because they are combination of autoregressive (AR), integration (I) referring to the reverse process of differencing to produce a stationary series and moving average (MA) operations. The Box Jenkins ARIMA (*p,d,q*) model (Box and Jenkins 1970), is expressed as follows

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

where,

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

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

ε_t = white noise or error term, *d*= differencing term and B=backshift operator, i.e. $B^a Y_t = Y_{t-a}$.

There are three steps in ARIMA model building process, viz. identification, estimation and diagnostic checking. Parameters of this model are selected at the identification stage. Identification of *d* is necessary to convert a non-stationary series into stationary. Augmented Dickey Fuller (ADF) test is use to test the stationarity. Parameters of the models are estimated by employing iterative least square or maximum likelihood techniques. The efficacy of the selected model is then tested by employing Ljung-Box test. If the model is found to be under or over fitted then, the three stages are repeated until satisfactory ARIMA orders are selected. To check the closeness of the fit, Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were calculated.

The Artificial Neural Network for time series modeling and analysis is termed as Time Delay Neural Network (TDNN) because the network contains time lags or delays in input layer. The time series phenomenon can be mathematically modelled using neural network with implicit functional representation of time, whereas in static neural network like multilayer perceptron is presented with dynamic properties. The general expression for the final output *Yt* of a multi-layer feed forward time delay neural network is

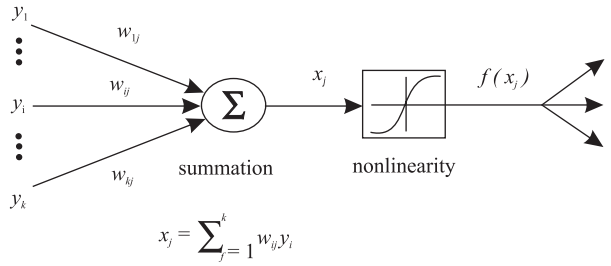

Fig 1 Neural network structure

expressed as follows.

$$Y_t = \alpha_0 + \sum^{q}_{(j=1)} \alpha_j \, g(\beta_{0j} + \sum^{p}_{(i=1)} \beta_{ij} \, Y_{t-p}) + \varepsilon_t \quad (4)$$

where, α_j *(j=0,1,2,...,q)* and β_{ij} *(i=0,1,2,...,p,* $_j$*=0,1,2,...,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 depicted in Fig 1.

The selection of appropriate parameters, i.e. number of hidden nodes as well as optimum number of lagged observation *p* for input vector is important in ANN modeling for determination of the autocorrelation structure present in a time series. Though there are no established theories available for the selection of *p* and *q*, various training algorithms have been used for the determination of the optimal values of *p* and *q*. The objective of training is to minimize the error function that measures the miss-fitting between the predicted value and the actual value. The error function which is widely used is mean squared error which can be written as:

$$E = \frac{1}{N}\sum_{t=1}^{N}(e_t)^2 = \frac{1}{N}\sum_{t=1}^{N}\{X_t - (w_0 + \sum_{J=1}^{Q} w_J g(w_{0j} + \sum_{i=1}^{P} w_{ij} X_{t-i}))\}^2 \quad (5)$$

where η is the total number of error terms. The parameters of the neural network are changed by an amount of changes in Δw_{ij} as

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \quad (6)$$

where, η is the learning rate. The error surface of multilayer feed forward neural network with non-linear activation function is complex in nature and believed to have many local and global minima.

Support vector machine (SVM) is one of the supervised machine learning technique, which was originally developed for classification tasks. With the introduction of ε-insensitive loss function (Vapnik 1997), the SVM has been extended to the nonlinear regression estimation problems and 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 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;

$$f(x) = W^T \phi(x) + b \quad (7)$$

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. The performance of NLSVR model strongly depends on the kernel function and set of hyper-parameters. The most commonly used kernel function (Table 1) is Radial Basis Function (RBF) which is given as follows;

$$k(x_i, x_j) = exp\{-y||x - x_i||^2\} \quad (8)$$

The performance of NLSVR model is strongly depends on the kernel function and set of hyper-parameters.

The performance of RBF kernel function requires optimization of two hyper-parameters; Regularization parameter C, which balances the complexity and approximation accuracy of the model and Kernel bandwidth parameters, which represents variance of RBF kernel function.γ represents the degree of kernel function to define the nonlinearity.

RESULTS AND DISCUSSION

An ARIMA model was built using SAS 9.4 software available at ICAR-Indian Agricultural Statistics Research Institute, New Delhi. The summary statistics of oilseed production time series is presented in Table 2 depict that the series is highly heterogeneous as CV is very high. The time series plot of oilseed production of India is plotted in Fig 2.

The ARIMA model has been built for oilseed production of India. The original time series was found to be non-stationary, so first differencing was done to make the stationary series time series (Fig 3).

The adequate model, i.e. ARIMA (110) has been identified based on Autocorrelation and Partial Autocorrelation Function (ACF and PACF) plots (Fig 3).

Table 1 Commonly used kernel functions in Support Vector Machine problems

Kernel type	Expression
Linear SVM	$K(x, x_i) = x_i^T x$
Polynomial of degree d	$K(x, x_i) = (x_i^T x + k)^d$
Radial Basis Function (RBF)	$K(x, x_i) = \exp\left\{ = \frac{\| x - x_i \|^2}{2\delta^2} \right\}$ Equivalently $K(x, x_i) = \exp\{-y \| x - x_i \|^2\}$
Multi-Layer Perceptron (MLP)	$K(x, x_i) = \tanh(k_1 x_i^T x + k_2)$

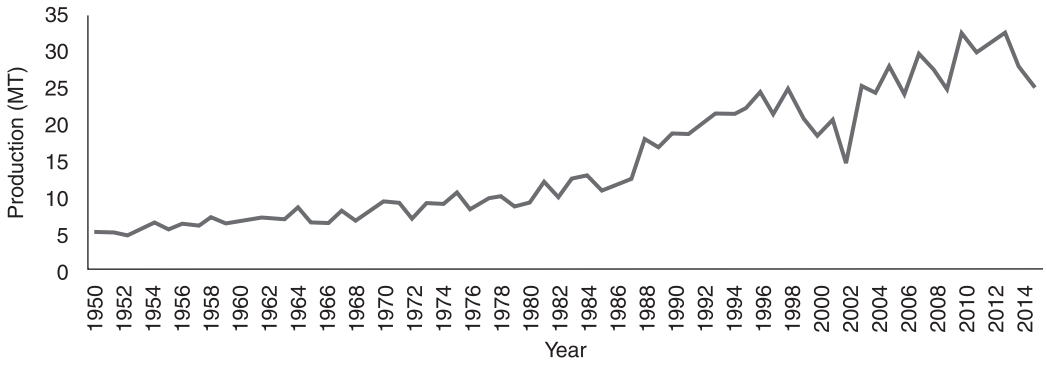

Fig 2 Time series plot of oilseed production of India

Table 2 Summary statistics of oilseed production time series

Statistic	Oilseed production	Statistic	Oilseed production
Observation	66	Maximum	32.75
Mean	14.86	Standard Deviation	8.47
Median	11.05	Skewness	0.6
Mode	6.4	Kurtosis	-1.04
Minimum	4.73	Coefficient of Variation (%)	56.98

The parameters of ARIMA models are estimated using maximum likelihood methods are given in Table 3. Auto correlation check for residuals obtained from ARIMA model of mango production time series indicates the residuals found to be non-autocorrelated as probability of chi-square is 0.45. Further the model performance in training set and testing data set is given in Tables 6 and 7.

As discussed in methodolgy section time delay neural network has been built for oilseed production of India time series using R software with the help of package 'forecast' (Hyndman 2017). Prior to select the model 2:10s:1l, many combination of time lag, hidden nodes has been tried and based on the lowest training error, a TDNN model with two tapped delay, ten hidden nodes with sigmoidal activation function and one output layer with linear identity function was selected. Based on repetitive experimentation, the learning rate and momentum term was fixed as 0.03 and 0.01 respectively. The model has been cross validated ten folds to minimize the error. Parameter specification of TDNN model has been depicted in Table 4. Further the model performance in training set and testing data set is given in Tables 6 and 7.

The nonlinear support vector regression model for oil seed production time series was analyzed using R software with the help of package 'e1071' (David 2017).The nonlinear support vector regression (NLSVR) model for oilseed

Table 3 Parameter estimation of ARIMA (1, 1, 0) by Maximum Likelihood Estimation method for Oilseed Production time series

Parameter	Estimate	Standard Error	t Value	Approx. Pr > |t|	Lag
MU	0.32	0.19	1.64	0.1012	0
AR1,1	-0.48	0.11	-4.39	<0.0001	1

Table 4 Parameter specification of TDNN model

Particulars	ANN parameter
Cross validation	10 fold
Optimum lag	2
Optimum hidden node	10
Network type	(2,10,1): Feed forward network
Activation function	Linear Sigmoidal
Learning rate	0.003
Momentum	0.001
Total no. of parameters	33

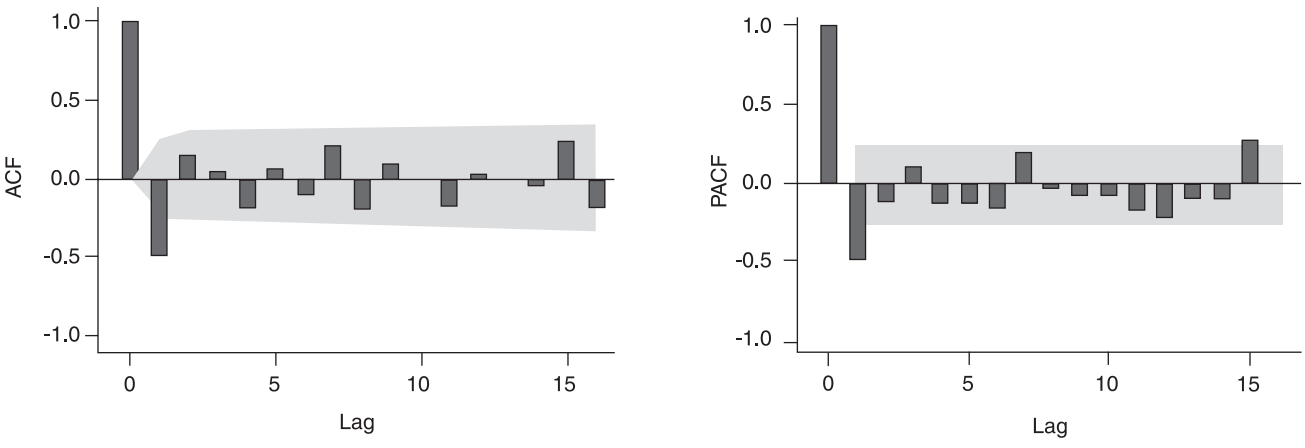

Fig 3 ACF and PACF time series oilseed production of India

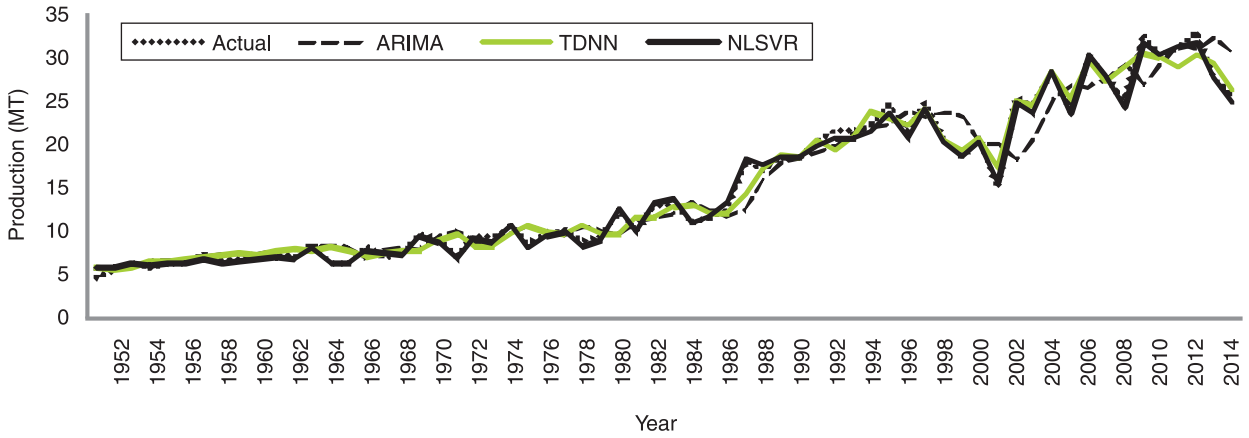

Fig 4 Actual v/s Fitted plot by different models

Table 5 Model specification of SVR for oilseed production time series

Kernel function	No. of SVs	C	γ	ε	K fold cross validation (K)	Cross Validation Error
RBF	7	8.19	3.06	0.15	10	0.035

production time series was built with following parameter specifications (Table 5). Cross validation was carried out for the considered time series and the lowest cross validation error obtained was 0.035. Further the model performance in training set and testing data set is given in % in Tables 6 and 7. The forecasting performance of different models is depicted in Fig 3.

Based on the lowest mean square error (MSE), root mean square error (RMSE) and mean absolute percentage (MAPE) values of all models obtained for both training (Table 6) and testing (Validation) data set (Table 7) considered, one can infer that both machine intelligence techniques, viz. TDNN and NLSVR outperformed over ARIMA model. Further, the value of MSE was reduced up to one third in NLSVR in comparison to that of ARIMA model which indicates that the performance of NLSVR model was superior as compared to ARIMA model. The graphical representation of forecasting performance of different models (Fig 4) clearly shows that NLSVR models better performed as compared to ARIMA and TDNN models.

The TDNN and NLSVR models performed well over ARIMA model due to their superior predictive ability in nonlinear and heterogonous data set. Among the machine intelligence techniques like TDNN and NLSVR, the NLSVR performed better in both training and testing data set. Even though, the coefficient of variation (Table 2) is very high then also the artificial intelligent techniques like TDNN and NLSVR performed better.The reason could be the nonlinear machine learning techniques can capture the heterogeneous trend in the data set and performed better as compared to regression model.

Conclusion

Table 6 Model performance of Oilseed Production time series for training data set

Criteria	ARIMA	TDNN	NLSVR
MSE	5.33	1.76	0.26
RMSE	2.31	1.33	0.51
MAPE	11.06	8.09	3.85

Table 7 Model performance of Oilseed Production time series for testing data set

Year	Actual	Forecast		
		ARIMA	TDNN	NLSVR
2011	29.80	28.84	30.82	30.28
2012	30.94	31.54	29.08	31.42
2013	32.75	30.67	31.94	33.64
2014	27.51	31.81	30.28	30.15
2015	25.30	31.82	28.06	26.77
Criteria	MSE	13.31	4.08	2.06
	RMSE	3.64	2.02	1.43
	MAPE	10.58	6.57	4.24

ARIMA models are not always adequate for the time series that contains non-linear structures. In this context, a nonlinear artificial intelligence technique like neural networks and support vector machines can be an effective way to improve forecasting performance. Based on the results obtained in this work one can infer that application of artificial intelligence techniques like time delay neural networks and nonlinear support vector regression techniques in modeling and forecasting of time series can increase the forecasting accuracy, in particular, the nonlinear support vector regression model performed better for forecasting oilseed production of India as compared to other models. This approach can be further extended by using some other machine learning techniques for varying autoregressive and moving average orders in other agricultural crops.

REFERENCES

Alonso J, Castaon A R and Bahamonde A. 2013. Support Vector Regression to predict carcass weight in beef cattle in advance of the slaughter. *Computers and Electronics in Agriculture* **91**: 116 –20.

Box G E P and Jenkins G. 1970. Time series analysis, forecasting and control. Holden-Day, San Francisco, CA.

Chen K Y and Wang C H. 2007. Support vector regression with genetic algorithm in forecasting tourism demand. *Tourism Management* **28**: 215–26.

David M. 2017. E1071: Misc Functions of the Department of Statistics, Probability Theory Group. R package version **1**: 6–8.

GOI. 2016. Agricultural statistics (RBI stat), Reserve Bank of India (RBI), Government of India.

Hyndman R J. 2017. forecast: Forecasting functions for time series and linear models. R package version **8.1.**

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 Computing and Applications* **24**(3): 563–71.

Kumar T L M and Prajneshu. 2015. Development of Hybrid Models for Forecasting Time-Series Data Using Nonlinear SVR Enhanced by PSO. *Journal of Statistical Theory and Practice* **9**(4): 699–711.

Naveena K, Rathod S, Shukla G and Yogish K J. 2014. Forecasting of coconut production in India: A suitable time series model, *International Journal of Agricultural Engineering* **7**(1):190–3.

Naveena K, Singh S, Rathod S and Singh A. 2017. Hybrid ARIMA-ANN modelling for forecasting the price of Robusta coffee in India. *International Journal of Current Microbiology and Applied Sciences* **6**(7):17–26.

Naveena K, Singh S, Rathod S and Singh A. 2017a. Hybrid time series modelling for forecasting the price of Washed coffee (Arabica Plantation Coffee) in India. *International Journal of Agriculture Sciences* **9**(10): 40–7.

NHB. 2015. National Horticultural Board Data Base 2014-15. Current scenario of Horticulture in India.

Rathod S, Singh K N, Paul R K, Meher S 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. *Journal of the Indian Society of Sgricultural Statistics* **71**(2): 103–11.

Ray M, Rai A, Ramasubramanian V and Singh K N. 2016. ARIMA-WNN hybrid model for forecasting wheat yield time-series data. *Journal of the Indian Society of Agricultural Statistics* **70**(1): 63–70.

Sarika, Iquebal M A and Chattopadhyay C. 2011. Modelling and forecasting of pigeonpea (*Cajanus cajan*) production using autoregressive integrated moving average methodology. *Indian Journal of Agricultural Sciences* **81**(6): 520–3.

Sreelakshmi K and Ramakanthkumar P. 2008. Neural networks for short term wind speed prediction. *World Academy of Science, Engineering and Technology* **32**: 721–5.

Suresh K K and Priya S R K. 2011. Forecasting sugarcane yield of Tamilnadu Using ARIMA models. *Sugar Technology* **13**(1): 23–6.

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* 9, pp 281–7. Mozer M, Jordan M and Petsche T (Eds). MIT Press, Cambridge, MA.

Vishwajith K P, Dhekale B S, Sahu P K, Mishra P and Noman M D. 2014. Time series modeling and forecasting of pulses production in India. *Journal of Crop and Weed* **10**(2): 147–154.

Zhang F, Deb C, Lee S, Yang J and Shah K. 2016. Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique. *Energy and Buildings* **126**: 94–103.

Zhang G, Patuwo B E and Hu M Y. 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* **14**(1): 35–62.