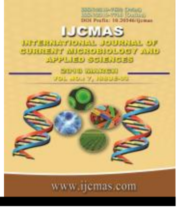
Int.J.Curr.Microbiol.App.Sci (2018) *7*(3): 1698-1707

EXCELLENT PUBLISHERS

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 Volume 7 Number 03 (2018)
Journal homepage: http://www.ijcmas.com



Original Research Article https://doi.org/10.20546/ijcmas.2018.703.201

Statistical Modeling and Forecasting of Total Fish Production of India: A Time Series Perspective

Mahalingaraya[1], Santosha Rathod[1], Kanchan Sinha[1], R.S. Shekhawat[1] and Shashikala Chavan[2]

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012, India
[2]Department of Computer Science, Jamia Millia Islamia, New Delhi-110025, India

**Corresponding author*

ABSTRACT

Indian fisheries and aquaculture is an important sector of food production, providing nutritional security to the food basket, contributing to the agricultural exports and engaging about fourteen million people in different activities. Fish production in India has increased at a higher rate compared to food grains, milk, egg and other food items. Constituting of about 6.3% of the global fish production, the Indian fisheries sector contributes to 1.1% of the GDP and 5.15% of the agricultural GDP. Forecasting is used to analyze the past and current behavior to forecasts the future fish 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 Artificial Neural Network (ANN) model commonly employed. In this paper an attempt has been made to forecast the raw jute productivity of India using ARIMA and ANN models. Empirical results clearly reveal that the machine learning techniques out performed the ARIMA model.

Keywords

ARIMA, India, Fish production, ANN, Time series

Article Info

Accepted:
16 February 2018
Available Online:
10 March 2018

Introduction

Indian fisheries and aquaculture is an important sector of food production, providing nutritional security to the food basket, contributing to the agricultural exports and engaging about fourteen million people in different activities. With diverse resources ranging from deep seas to lakes in the mountains and more than 10% of the global biodiversity in terms of fish and shellfish species, the country has shown continuous and sustained increments in fish production since independence. The 8,000 km coastline from both inland and marine resources, 3 million hectares of reservoirs, 1.4 million hectares of brackish water, 50,600 sq. km of continental shelf area and 2.2 million sq. km of exclusive economic zone supplement India's vast potential for fishery. Constituting of about 6.3% of the global fish production, the Indian fisheries sector contributes to 1.1% of the GDP and 5.15% of the agricultural GDP. The total fish production of 10.07 million metric tonnes presently has nearly 65% contribution from the inland sector and nearly the same from culture fisheries. Paradigm shifts in terms of increasing contributions from inland sector and further from aquaculture are significations over the years. With high growth rates, the different facets of marine fisheries, coastal aquaculture, inland fisheries, freshwater aquaculture, cold-water fisheries to food, health, economy, exports, employment and tourism of the country. Fish and fish products have presently emerged as the largest group in agricultural exports of India, with 10.51 lakh tonnes in terms of quantity and Rs. 33,442 crores in value. This accounts for around 10% of the total exports of the country and nearly 20% of the agricultural exports. More than 50 different types of fish and shellfish products are exported to 75 countries around the world.

Time series forecasting is a very useful technique for forecasting prices of agricultural commodities. Generally agricultural data contain both linear and nonlinear patterns, no single model is capable to identify all the characteristics of time series data on agriculture. Consequently, various types of parametric and nonparametric, linear and nonlinear time series models are used for forecasting. In contrast to the regression models, the ARIMA model allows a time series to be explained by its past, or lagged values and stochastic error terms.

ARIMA is one of the most traditional methods of non-stationary time series analysis. In contrast to the regression models, the ARIMA model allows a time series to be explained by its past, or lagged values and stochastic error terms. 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. Rathod *et al.,* (2017) forecasted oilseed production of India through artificial intelligence techniques**.** Jha *et al.*, (2014) studied monthly wholesale price of oilseeds in India using Time-delay neural networks for time series prediction. Paul *et al.*, (2010) forecasted inland fish production of India using ARIMA methodology.

The major drawback of ARIMA model is presumption of linearity, hence, no non-linear patterns cannot be recognized by ARIMA model. Sometimes, the time series often contain non-linear components, under such condition the ARIMA models are not adequate in modeling and forecasting. To overcome this difficulty, many parametric nonlinear models are developed 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. Hence artificial intelligence techniques are the only way to model such data and forecast such phenomenon. The Artificial Neural Network (ANN) is most widely used machine learning techniques to model and forecast the time series data.

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.

Materials and Methods

Data description

The raw jute productivity data from 1978 to 2015 were collected from www.india.stat.com. The data from 1978 to 2011 were used for model building i.e. training data set and data from 2012 to 2015were used for model validation i.e. testing data set.

Statistical Models

ARIMA model

An ARIMA model is usually stated as ARIMA (p, d, and q). 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. The general form of the ARIMA model expressed as ARIMA (p, d, q) (P, D, Q) $_S$, where, p = order of non-seasonal auto regressive (AR), d = order of non-seasonal difference, q = order of non-seasonal moving average (MA), P = order of seasonal auto regressive (SAR), D = order of seasonal difference, Q = order of seasonal moving average (MA), s = length of the season.

ARIMA model fitting

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

Where,

$\phi(B)=1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p$ (Autoregressive process)

$\theta(B)=1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q$ (Moving average process)

ε_t = white noise or error term

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

Diagnostic checking of the model

ARIMA model building is carried out in three stages, viz. Identification, estimation and diagnostic checking. Parameters of this model are experimentally selected at the identification stage. Identification of *d* is necessary to make the non-stationary time series to stationary. A statistical test can be employed to check the existence of stationarity, known as the test of the unit-root hypothesis. Popularly Augmented Dickey Fuller (ADF) test is utilized to test the stationarity (Dickey and Fuller, 1979). At the estimation stage, the parameters are estimated by employing iterative least square or maximum likelihood techniques.

Estimation of parameters for ARIMA model is generally done through nonlinear least squares method or maximum likelihood techniques. Several software packages are available for fitting of ARIMA models. To this end, in this paper, R software package is used. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) values for ARIMA model are computed by:

$$AIC = T'\log(\sigma^2)+2(p+q+1) \quad \ldots(2)$$

$$BIC = T'\log(\sigma^2)+(p+q+1)\log(T') \quad \ldots(3)$$

Where, T' denotes the number of observations used for estimation of parameters and σ^2 denotes the Mean square error.

The efficacy of the selected model is then tested by diagnostic checking stage where testing is done to see if the estimated model is statistically adequate i.e. whether the error terms are white noise which means error terms are uncorrelated with zero mean and constant variance. This is done by employing Ljung-Box test to the original series or to the residuals after fitting a model. A good account on Ljung-Box test can be found in Box *et al.*, (1994). If the model is found to be insufficient, the three stages are repeated until satisfactory ARIMA model is selected for the time-series under consideration.

Artificial Neural Network (ANN) Model

Neural Networks are simulated networks with interconnected simple processing neurons which analogous to the function of the biological neurons in an animal braincentral nervous system (McCulloch and Pitts, 1943). An ANN is based on a collection of connected units or nodes called artificial neurons. Each connection (analogous to a synapse) between artificial neurons can transmit a signal from one to another. The artificial neuron that receives the signal can process it and then signal artificial neurons connected to it. The ANN structure for a particular problem in time series prediction includes determination of number of layers and total number of nodes in each layer. It is usually determined through experimentation as there is no theoretical basis for determining these parameters.

Artificial neural networks (ANNs) model are considered as a class of generalized nonlinear model that are able to capture various nonlinear structures present in the data set. The main advantage of this model is that it does not require prior assumption of the data generating process, instead it is largely depend on characteristics of the data popularly known as data-driven approach (Fig. 1).

Single hidden layer feed forward network is the most popular for time series modeling and forecasting. This model is characterized by a network of three layers of simple processing units, and thus termed as multilayer ANNs. The first layer is input layer, the middle layer is the hidden layer and the last layer is output layer.

The general expression for the final output Y_t of a multi-layer feed forward time delay neural network is expressed as follows

$$y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} y_{t-p} \right) + \varepsilon_t \quad \ldots(4)$$

where, α_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 and g denotes the activation function at hidden and output layer respectively.

Activation function defines the relationship between inputs and outputs of a network in terms of degree of the non-linearity. Most commonly used activation function is logistic function which is often used as the hidden layer transfer function, i.e.

$$g(x) = \frac{1}{1+\exp(-y)} \quad \ldots(5)$$

Thus ANN model performs a nonlinear functional mapping between the input and output which characterized by a network of three layers of simple processing units connected by acyclic links

$$y_t = f\left(y_{t-1} + Xy_{t-2} + \ldots + Xy_{t-p}, w\right) + \varepsilon_t \quad \ldots(6)$$

Where, w is a vector of all parameters and f *is* a function of network structure and connection weights. Therefore, the time delay neural network resembles a nonlinear autoregressive model.

Results and Discussion

The summary statistics of Fish 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 Fish production of India is plotted in Figure 2. The ARIMA model has been built for Fish 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).

Table.1 Summary statistics of total fish production time series

Observation	38
Mean	5386.71
Standard error	393.95
Median	5323.00
Mode	#N/A
Standard deviation	2428.49
Sample variance	5897551.89
Kurtosis	-0.58
Skewness	0.52
Range	8490.00
Minimum	2306.00
Maximum	10796.00
Sum	204695.00
Coefficient of variation, CV (%)	7.31

Table.2 ADF values before and after differencing

ADF	Differencing	p-value
-1.80	0	0.6503
-2.56	1	0.3578
-4.11	2	0.01768

Table.3 Parameter estimation of ARIMA (0, 2, 1) by Maximum Likelihood Estimation method for Fish production time series

Parameter	Estimate	Std. Error	z value	Pr (>|z|)
MA1	-0.79	0.13	-6.14	<0.0001

Table.4 Parameter specification of ANN 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

Table.5 Forecasting performance of TDNN model for fish production time series

Model	Parameters	RMSE
		Training
2:2S:1L	9	131.80
2:4S:1L	17	123.94
2:6S:1L	25	120.15
2:8S:1L	33	119.97
2:10S:1L	41	116.70
3:2S:1L	11	118.57
3:4S:1L	21	104.90
3:6S:1L	31	98.41
3:8S:1L	41	87.66
3:10S:1L	51	95.45
4:2S:1L	13	113.29
4:4S:1L	25	88.99
4:6S:1L	37	74.42
4:8S:1L	49	78.80
4:10S:1L	61	71.82
5:2S:1L	15	99.66
5:4S:1L	29	73.28
5:6S:1L	43	61.87
5:8S:1L	57	54.50
5:10S:1L	71	36.64
6:2S:1L	17	71.80
6:4S:1L	33	62.03
6:6S:1L	49	50.30
6:8S:1L	65	55.66
6:10S:1L	81	36.52

Table.6 Model performance of total fish production time series for training data set

Criteria	ARIMA	ANN (NNAR(2,10)
MSE	25599.55	13935.8
RMSE	159.99	118.05
MAPE	2.74	2.16

Table.7 Model performance of total fish production time series for testing data set

YEAR	ACTUAL	FORECAST	
		ARIMA	ANN
2012	9040	9282.61	2335.27
2013	9572	9590.92	2365.84
2014	10431	9899.22	2396.41
2015	10796	10207.53	2426.97
	MSE	172077.55	7060.11
	RMSE	414.83	84.02
	MAPE	3.35	3.06

Fig.1 Neural network structure

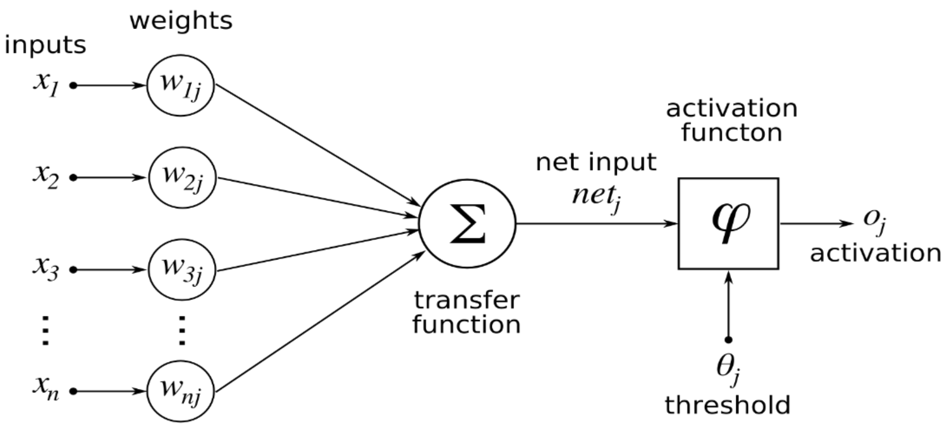


Fig.2 Time series plot of total fish production of India

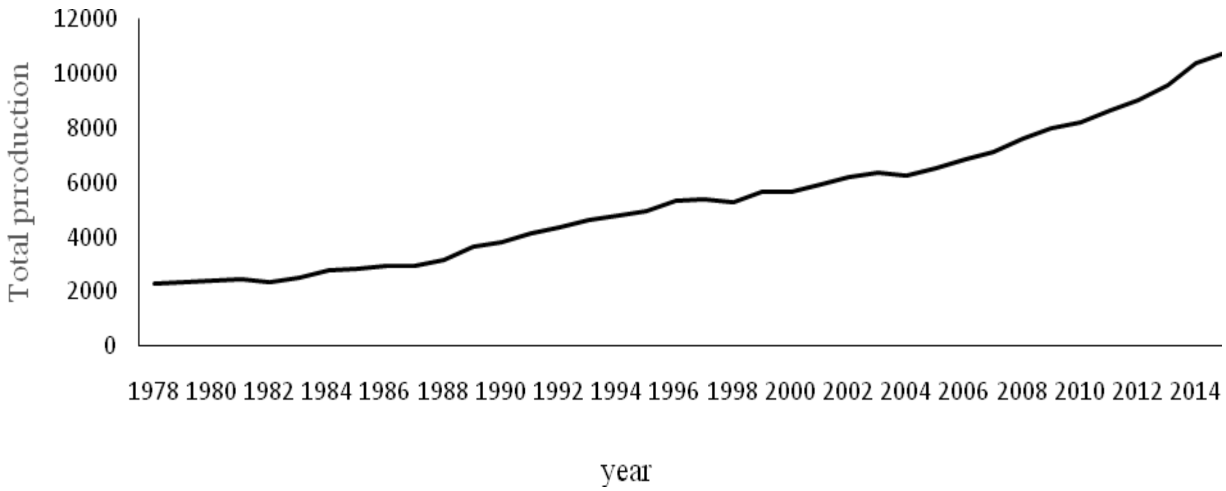


Fig.3 ACF plot of original total fish production time series

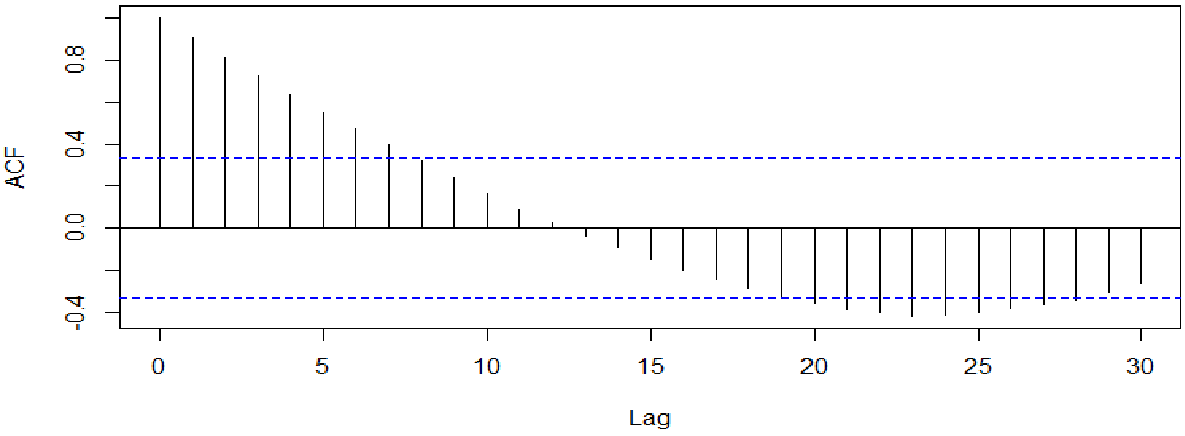


Fig.4 PACF plot of total fish production time series

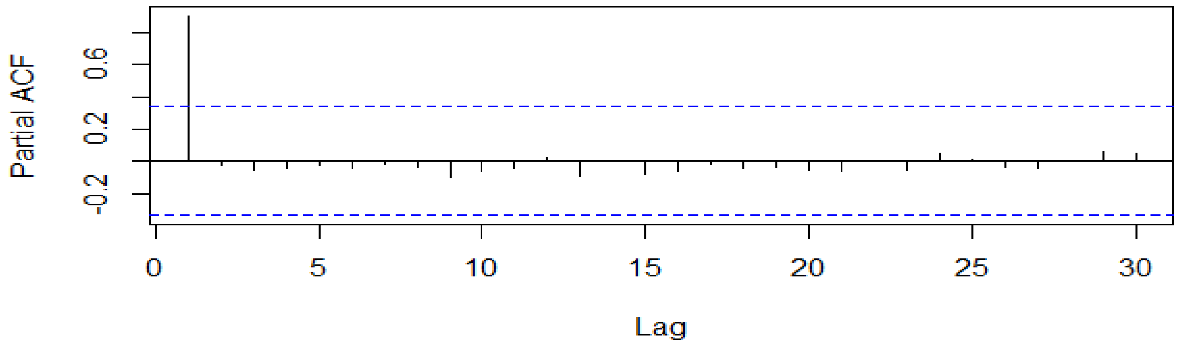


Fig.5 ACF plot of total fish production after differencing

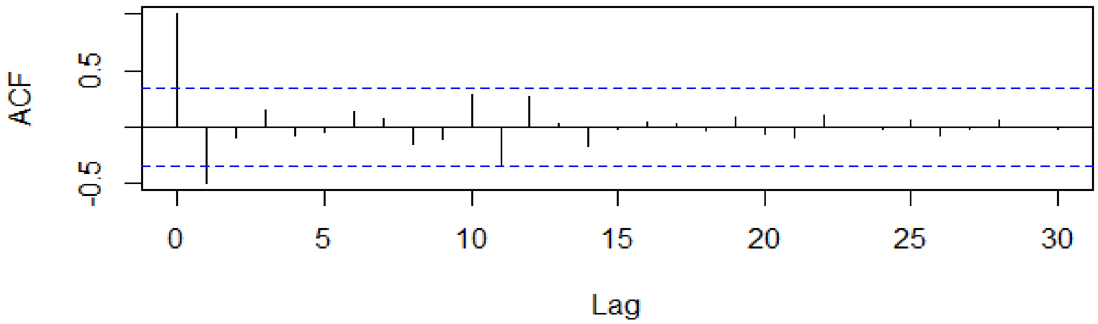
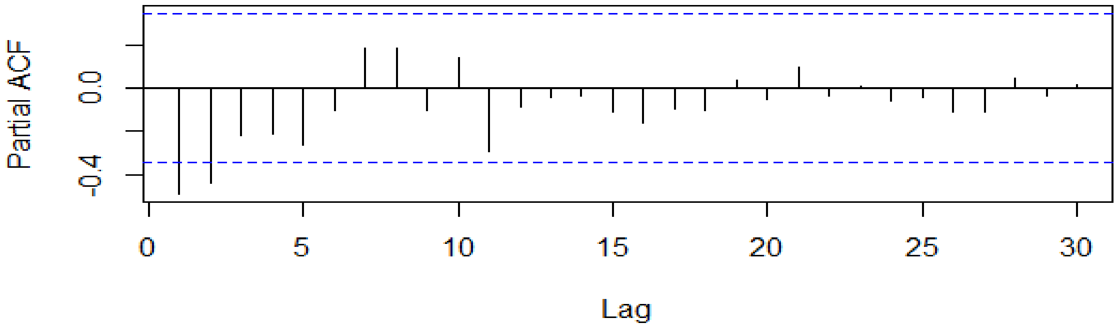


Fig.6 PACF plot of total fish production after differencing



Autocorrelation and Partial Autocorrelation Function (ACF and PACF) plots of original total fish production series is given in figure 3 and 4. The ACF and PACF plots of original series depicts that the original series is non stationary, the ADF test (Table 2) also support the same. Since, the original series is non stationary, differencing of original series was made two times to make the series stationary (Table 2). The ACF and PACF plots of last differenced series were depicted in figure 5 and 6.

After differencing the adequate model, i.e. ARIMA (0, 2, 1) has been identified based on Autocorrelation and Partial Autocorrelation Function (ACF and PACF) plots of the original series is given in figure 5 and 6.

Based on the significant ACF and PACF spikes and lowest AIC (422.56) and BIC (425.49) values the adequate model, i.e. ARIMA (0, 2, 1) has been identified. Further, the parameters of identified ARIMA (0, 2, 1) models are estimated using maximum likelihood methods (Table 3). Further the model performance in training set and testing data set is given in Tables 6 and 7.

As discussed in methodology section artificial neural network has been built for jute productivity 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 and hidden nodes has been tried (Table 5) and based on the lowest training error, a ANN 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 ANN 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.

Based on the lowest mean square error (MSE), root mean square error (RMSE) and mean absolute percentage (MAPE) values of both the models obtained for training (Table 6) and testing (Validation) data set (Table 7) considered, one can infer that both machine learning techniques, viz. ANN outperformed over ARIMA model. Further, the value of MSE was reduced TDNN in comparison to that of ARIMA model which indicates that the performance of TDNN model was superior as compared to ARIMA model.

Based on the lowest training RMSE (Table 5) the five ANN models *viz.* 2:8s:1l, 2:10s:1l, 3:10S:1l, 4:8s:1l and 4:10s:1l, are selected. These five models were further assessed based on their hold out sampling (testing set) forecasting performance. Out of total 29 neural network structures, an ANN model with two tapped delay and ten hidden nodes (2:10s:1l), was selected for forecasting banana Production of Karnataka. Based on repetitive experimentation, the learning rate and momentum term for all ANN model (Table 5) is set as 0.03 and 0.01 respectively. The forecasting performance of ANN model in training and testing data set is given in Table 6 and 7.

On the basis of the results obtained in this work one can conclude that 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 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 in modeling and forecasting of time series can increase the forecasting accuracy, in particular, the artificial neural network model performed better for forecasting total fish 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. This approach can be further extended by using some other machine learning techniques for varying autoregressive and moving average orders.

References

Box, G.E.P. and Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco, CA.

Dickey, D. A. and Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Stat. Assoc.*, 74: 427-431.

Fish statistics (2016). www.india.stat.com.

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.

Ljung and Box, G.E.P. (1978).On a measure of lack of fit in time series models. *Biometrica*, 65: 297-304.

Mcculloch, W.S. and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophy*.5:115-133.

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):1903.

Paul, R.K and Das, M. K. (2010). Statistical modelling of inland fish production in India. *Journal Inland Fish. Soc. India*, 42(2): 1-7.

Rathod, S, Mishra, G.C and Singh, K.N. (2017) Hybrid Time Series Models for Forecasting Banana Production in Karnataka State, India. *Journal of the Indian Society of Agricultural Statistics*, 71(3)193–200.

Rathod, S, Singh, K.N, Patil, S.G., Ravindrakumar, Naik, H, Ray Mand Singh, V and Meena. (2018). Modeling and forecasting of oilseed production of India through artificial intelligence techniques. *Indian Journal of Agricultural Sciences* 88 (1): 22–27.

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

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.

How to cite this article:

Mahalingaraya, Santosha Rathod, Kanchan Sinha, R.S. Shekhawat and Shashikala Chavan. 2018. Statistical Modeling and Forecasting of Total Fish Production of India: A Time Series Perspective. *Int.J.Curr.Microbiol.App.Sci.* 7(03): 1698-1707.
doi: https://doi.org/10.20546/ijcmas.2018.703.201