Journal of Pharmacognosy and Phytochemistry
Available online at www.phytojournal.com
J P P
Journal of Pharmacognosy and Phytochemistry

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2019; 8(3): 476-482
Received: 04-03-2019
Accepted: 07-04-2019

Puneet Dheer
SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

Time series modelling for forecasting of food grain production and productivity of India

Puneet Dheer

Abstract

The present study aimed for forecasting of total food grain production and productivity from 2018-2019 to 2025-2026 based on past history from 1950-51 to 2017-2018. Time series modelling and related forecasting were performed using Auto Regressive Integrated Moving Average (ARIMA), Auto Regressive Neural Network (ARNN) and ARIMA-ARNN hybrid models. ARIMA (0, 1, 1) were found suitable for the production and yield data based on the least value of Schwarz-Bayesian Criterion (SBC). Secondly, Auto Regressive Neural Network (ARNN) of order ARNN (3, 4) and ARNN (4, 3) was selected for both the dataset respectively. Lastly, ARIMA (0, 1, 1) - ARNN (3, 3) and ARIMA (0, 1, 1) - ARNN (3, 6) were found suitable for both production and yield. All the three models were tested for their forecast accuracy using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). Accordingly, the ARNN model was found to be best as compared to the individual ARIMA and hybrid ARIMA-ARNN model. Based on the ARNN model, the forecasting of total food grain production and productivity calculated which would be 356.95 million tonnes with yield of 3183.67 kg/ha by 2025-26.

Keywords: forecasting, time series modelling, ARIMA, ARIMA-ARNN

Introduction

Food is basic need for every living thing. Agriculture is particularly important being the heart of our daily life, vital to economy and society. Since independence, India has made immense progress towards food security. Indian population has tripled, and food grain production more than quadrupled. Thus, there has been a substantial increase in the availability of food grain per capita.

The expected per capita consumption levels of various commodities are worked out on the basis expected population growth rate, growth rate in per-capita consumption expenditure and the elasticity of demand with respect to per-capita consumption expenditure. (Planning commission.nic.in.). The food grain production is projected by plotting the average production figures for every 5 years from 1960-65 to current year and drawing a linear forecast trend line up to interventional year. Such an exercise gives a probable food grain production. However, with the advent of advance science and technology, several time series modelling techniques were developed for analyzing and forecasting the series, depends upon the characteristics of the time series. If the time series is linear, then Auto-Regressive Integrated Moving Average (ARIMA) can be employed most of the times. ARIMA models (Box *et al.*, 1994) [1] have been utilized for crop yield or any other agricultural production. Different time series modelling was conducted in agricultural domain. Sarika *et al.* (2011) [6] applied ARIMA model for modelling and forecasting India's pigeon pea production data. Suresh *et al.* (2011) [7] applied ARIMA model for forecasting sugarcane area, production and productivity of Tamil Nadu state of India. Earlier finding reported that combining different models enhance the accuracy of forecasting as compared to individual model. The hybrid methodology given by Zhang (2003) [8] is one of the most applied hybrid techniques which combine ARIMA and ARNN models. The ARIMA-ARNN hybrid model has been used to forecast the price of washed coffee (Naveena *et al.*, 2017) [4] and production and yield of pulses (Dheer and Yadav, 2018) [2]. Keeping these in view, three models viz., ARIMA, ARNN and ARIMA-ARNN were tested for their forecast accuracy and subsequently forecasting of total food grain production (million tonnes) and productivity (kg/ha) accommodating 68 years (1950-51 to 2017-18) respective data in India.

Materials and Methods

The time series data of production and yield of food grains for the period of 1950-51 to 2017-18 of India were analyzed (Agricultural Statistics at Glance, 2016 & business.standard.com). Out of the 68 years data, for training the model - first 48 years data were used and for model

Correspondence
Puneet Dheer
SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

validation - the last 20 years data are used. All the analysis was performed using Math Works MATLAB R2018b and R package.

Box-Jenkins ARIMA model: The most common time series model used in order to predict future outcomes based on a linear function of past data points and past errors terms is the Autoregressive Integrated Moving Average (ARIMA) also known as Box–Jenkins model. In theory and practice, ARIMA model is commonly utilized for forecasting a time series, either to predict future points or to get a better insight about the data. To satisfy the ARIMA assumptions, a sequence of steps is performed on the raw data in order to maintain a statistical stationarity property such as mean, variance and autocorrelation do not change over time. The stationarity property of a time series can be confirmed by using unit root test such as Augmented Dickey-Fuller test (ADF) and stationarity test such as Kwiatkowski–Phillips–Schmidt–Shin test (KPSS). If a series found to be a non-stationary based on these tests, differencing is performed until the data are finally made stationary. In general, an ARIMA model represented as ARIMA (*p*,*d*,*q*), consists of three parameters: (I) *p*, the order of Auto-Regression (AR), (II) *d*, the order of integration (differencing) to achieve stationarity, and (III) *q*, the order of Moving Average (MA).

$$X_{(t)} = \beta_o + \beta_1 X_{(t-1)} + \cdots + \beta_p X_{(t-p)} + \varepsilon_{(t)} + \theta_1 \varepsilon_{(t-1)} + \cdots + \theta_q \varepsilon_{(t-q)} \ (1)$$

Where$X_{(t)}$ and $\varepsilon_{(t)}$ represent the actual value and random error at time period *t* respectively, $\beta_i (i=1, 2,..., p)$ and $\theta_j (j=1, 2,..., q)$ are model parameters, and p and *q* are lagged values. Random errors $\varepsilon_{(t)}$, are assumed to be independently and identically distributed with a mean zero and a constant variance, σ^2.

After satisfying the stationarity property of the time series, the Box-Jenkins approach follows four steps:

a) **Model identification:** Examine the data by ACF (MA (*q*) term) and PACF (AR (*p*) term) to identify the potential models.
b) **Parameter estimation:** Estimate the parameters using least square for potential models and select the best model using Akaike Information Criterion (AIC) or Schwarz- Bayesian Criterion (SBC).
c) **Diagnostic checking:** Check the ACF/PACF and Ljung Box Test of residuals. Do the residuals follows random distribution? If yes go to (iv), otherwise go to (i) and repeat the same.
d) **Final model:** Generate the required forecasts by using the selected model.

Artificial Neural Network Approach for Time Series Modelling: On the other hand, Artificial Neural Networks (ANNs), is a family of statistical learning algorithms inspired from biological neural networks of the brain. For the first time, Mcculloch and Pitts (1943) [3], proposed the idea of the artificial neural network but due of the lack of computing facilities, they were not in much use until the back-propagation algorithm was discovered by Rumelhart *et al.* (1986) [5]. An ANNs is generally represented from finite numbers of artificial neurons that are associated with weights, which leads to the neural architecture and are organized in layers (input, hidden and output layer). ANNs are advantageous compared with ARIMA in many applications because ANNs do not assume linearity. ANN is a non-linear mathematical model and its ability to model a complex non-linear process that build a relationship between inputs and outputs of a system. The Autoregressive Neural Network (ARNN) model performs a nonlinear functional mapping from the past observations $[X_{(t-1)}, X_{(t-2)}, \cdots, X_{(t-p)}]$ to the future value$X_{(t)}$, i.e.,

$$X_{(t)} = f\left(g(X_{(t-1)}, X_{(t-2)}, \cdots, X_{(t-p)}, \mathrm{w})\right) + \varepsilon_{(t)} \ (2)$$

Where *'f'* is a non-linear activation function determined by the network structure (such as sigmoid, TanH, ReLU, etc.), 'g' is linear function, p is the lagged value, 'w' is a vector of connection weights with bias and $\varepsilon_{(t)}$ is a noise or error terms. Thus, the ANN is equivalent to a nonlinear autoregressive model (ARNN).

The important task of ARNN(*p*,*q*) modelling for a time series is to select an appropriate number of hidden nodes *q*, as well as to select the dimension of input vector (aka, the lagged observations), *p*. It is difficult to determine *p* and *q* values atfirst place. Hence, in practice, experiments are often conducted to select the appropriate values for *p* and *q*.

Hybrid Based Model - Autoregressive Integrated Moving Average and Artificial Neural Network

ARIMA–ARNN hybrid model was proposed (Zhang, 2003) for time series forecasting. Any time series sequence is assumed to be the sum of two components, linear and nonlinear.

$$X_{(t)} = L_{(t)} + N_{(t)} \ (3)$$

where $L_{(t)}$ and $N_{(t)}$denote the linear and non-linear components, respectively.

First, an ARIMA model is fit to the given time series sequence. Then the error sequence from ARIMA is assumed to be the nonlinear component and is modeled using an ARNN. The predictions obtained from both the ARIMA model and the ARNN model are combined to obtain the final forecast.

Let $\varepsilon_{(t)}$ denote the residual at time *t* from the linear model, then $X_{(t)}$ is actual value and $\hat{L}_{(t)}$ is forecast value:

$$\varepsilon_{(t)} = X_{(t)} - \hat{L}_{(t)} \ (4)$$

By modelling residuals $\varepsilon_{(t)}$ series using ARNNs, nonlinear relationships can be discovered. With *n* input nodes, the ARNN model for the residuals will be:

$$E_{(t)} = f\left(\varepsilon_{(t-1)}, \varepsilon_{(t-2)}, \ldots, \varepsilon_{(t-n)}\right) + e_{(t)} \ (5)$$

where *f* is a nonlinear function determined by the neural network and $e_{(t)}$ is the random error. Let's denote the $E_{(t)}$as$\hat{N}_{(t)}$, the combined forecast will be:

$$X_{(t)} = \hat{L}_{(t)} + \hat{N}_{(t)} \ (6)$$

Forecast Evaluation Criteria

There are many measurements to evaluate the residuals. We used the Root Mean Square Error (RMSE) and Mean Absolute Percentage error (MAPE).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(X_{(t)} - \hat{X}_{(t)}\right)^2}$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{(X_{(t)} - \hat{X}_{(t)})}{X_{(t)}}\right| * 100$$

Where n is the number of data points, $X_{(t)}$ is actual value at time t and$\hat{X}_{(t)}$ is predicted value. The lesser value of RMSE and MAPE, makes the better model for forecasting.

Results and Discussion

The data of production and yield of food grains were used in order to forecast for the year 2018-2019 to 2025-2026 using the models as described earlier.

First 48 years data from production and yield dataset as a training set were used to analyze the time series regarding its stationarity property and model building with an objective to forecast. The Fig. 1 shows the time series of all 68 data points for both the dataset. In order to apply the ARIMA model on the given training set, it is necessary to check the stationarity property by investigating the ACF plots in Fig. 2 and p-value of ADF test. It was observed that the dataset is non-stationary because the autocorrelation is decreasing very slowly and remains well above the significance level. This is indicative of a non-stationary series and confirmed by ADF test with p-value > 0.05 supporting the null hypothesis that the series is non-stationary. The time series was differentiated (order of 1) and again performed the ADF test and investigated the ACF plots in Fig 3 shows no significant autocorrelation. ADF test p-value <= 0.05 confirms the alternative hypothesis about the time series is stationary

In the Fig. 3, the ACF appears to cutoff to zero after lag 1 indicating MA (1) behavior and the PACF also appears to cutoff to zero after lag 1 indicating AR (1) behavior for both the production and yield. Then the ARIMA (1,1,0) and ARIMA (0,1,1) models were tried for the first differenced time series data on production and yield in India. After running the experiments for different values of *p* and *q*, it was found that AR (*p*) and MA (*q*) order identified by least SBC criterion are 0 and 1 respectively. It is partially confirmed that ARIMA (0, 1, 1) may be the best suited model for both the production and yield dataset. For final confirmation, diagnostic checking by Ljung box test was conducted on fitted residuals of ARIMA (0,1,1) for both the production and yield with p-value > 0.05 supporting the null hypothesis that the residuals follow the white noise.

Different ARNN were tried for the same datasets. The Fig. 4 and Fig. 6 describe how the prediction residuals/errors are related in time. For a good prediction model, autocorrelation should only be one nonzero value at zero lag. This suggests that the prediction residuals/errors were completely uncorrelated with each other (white noise). If there was significant correlation in the prediction residuals/errors, there is a possibility to improve the prediction may be by increasing the number of delays. Here, the autocorrelations fall within the 95% confidence limits around zero, except for zero lag, so the model seems to be adequate. If more accurate results were required, then retraining the network will change the initial weights and biases of the network, and may generate an improved network. Here, ARNN (3, 4) and ARNN (4, 3) model was found to be best for modelling both the production and yield respectively.

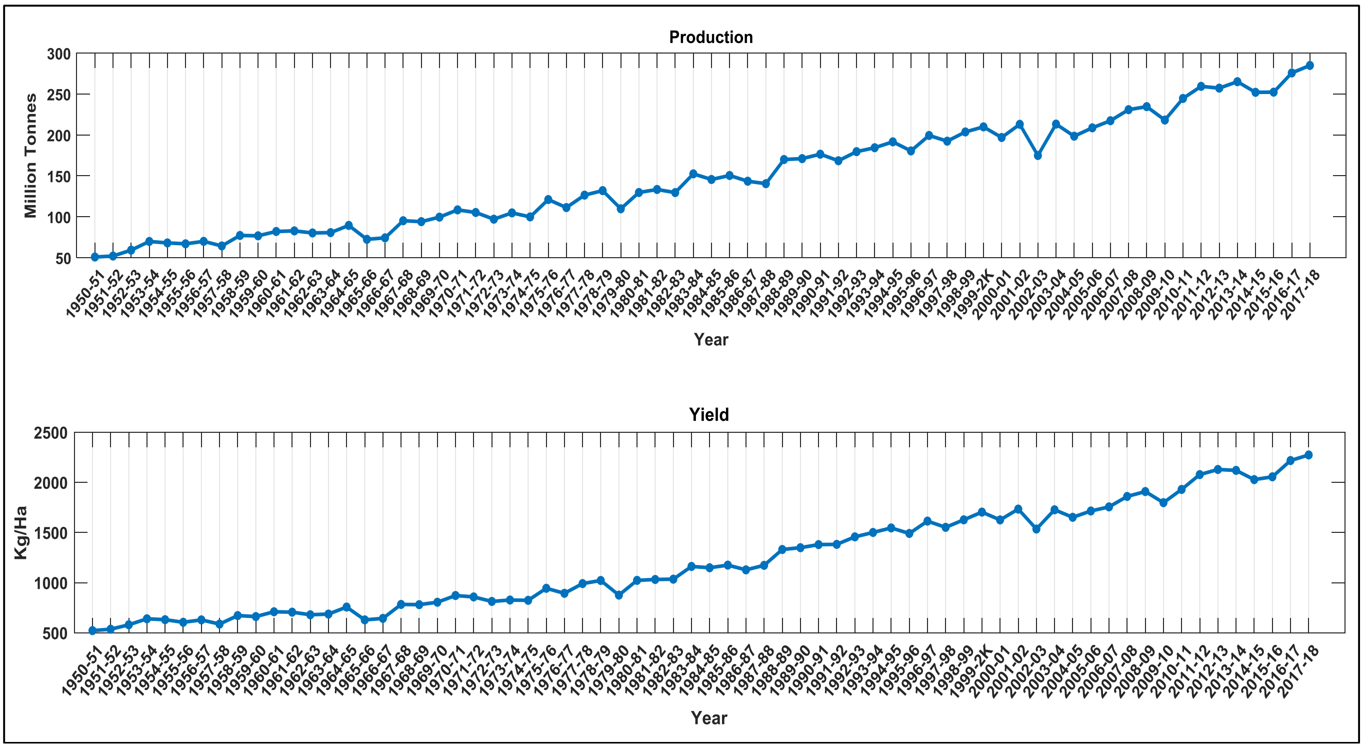

Fig 1: Time series of production and yield of food grain over years in India

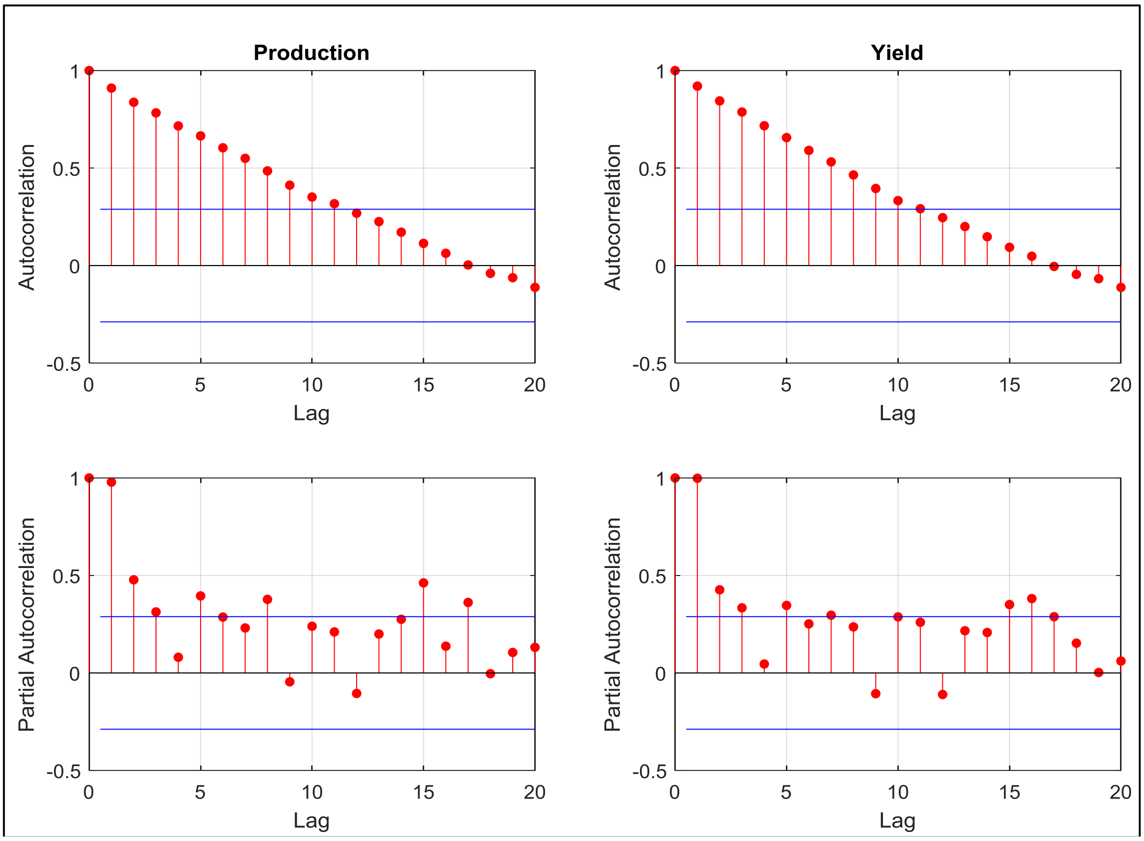

Fig 2: Autocorrelation and Partial Autocorrelation plot

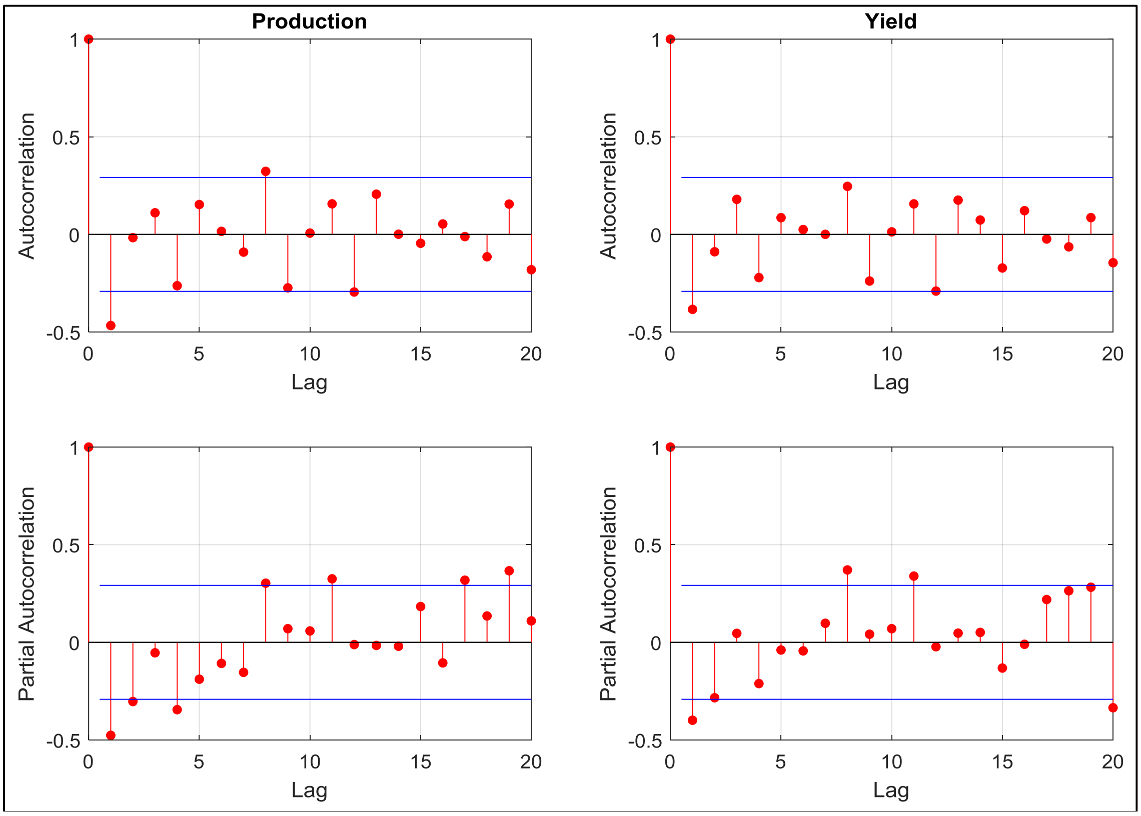

Fig 3: Autocorrelation and Partial Autocorrelation plot of differentiated time series (order of 1)

Table 1: Forecast evaluation of models on training data

Models	Production		Models	Yield	
	Rmse	Mape		Rmse	Mape
ARIMA (0,1,1)	8.39	6.12	ARIMA (0,1,1)	56.55	5.07
ARNN (3,4)	8.29	0.11	ARNN (4,3)	64.44	0.74
ARIMA-ARNN (3,3)	5.48	3.80	ARIMA-ARNN (3,6)	16.46	1.38

Table 2: Forecast evaluation of models on validation data

Models	Production		Models	Yield	
	Rmse	Mape		Rmse	Mape
ARIMA (0,1,1)	15.75	5.57	ARIMA (0,1,1)	122.633	5.04
ARNN (3,4)	14.18	1.28	ARNN (4,3)	90.16	0.82
ARIMA-ARNN (3,3)	16.18	5.48	ARIMA-ARNN (3,6)	125.40	4.83

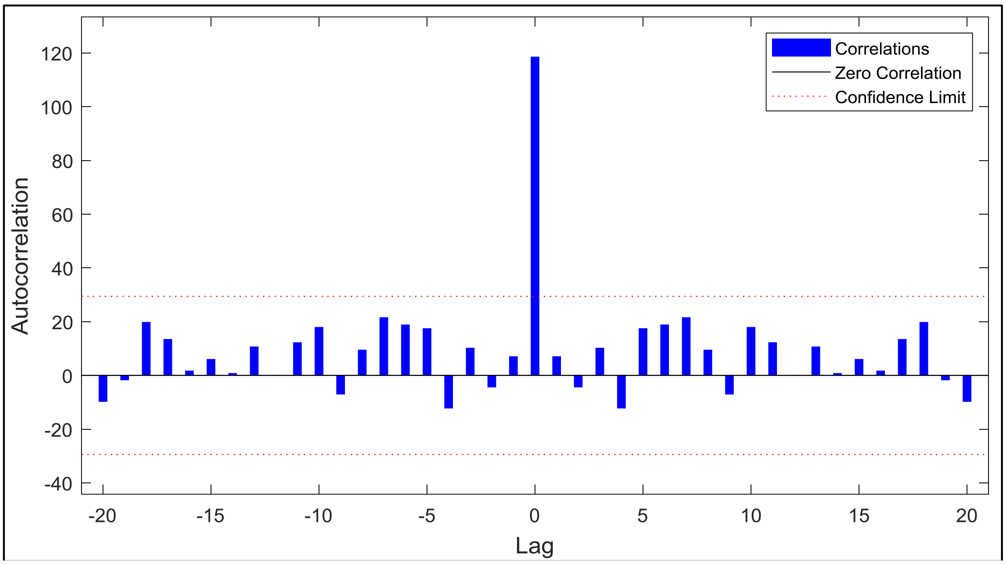

Fig 4: Autocorrelation of Residuals of ARNN (3,4) for Total Food Grain Production

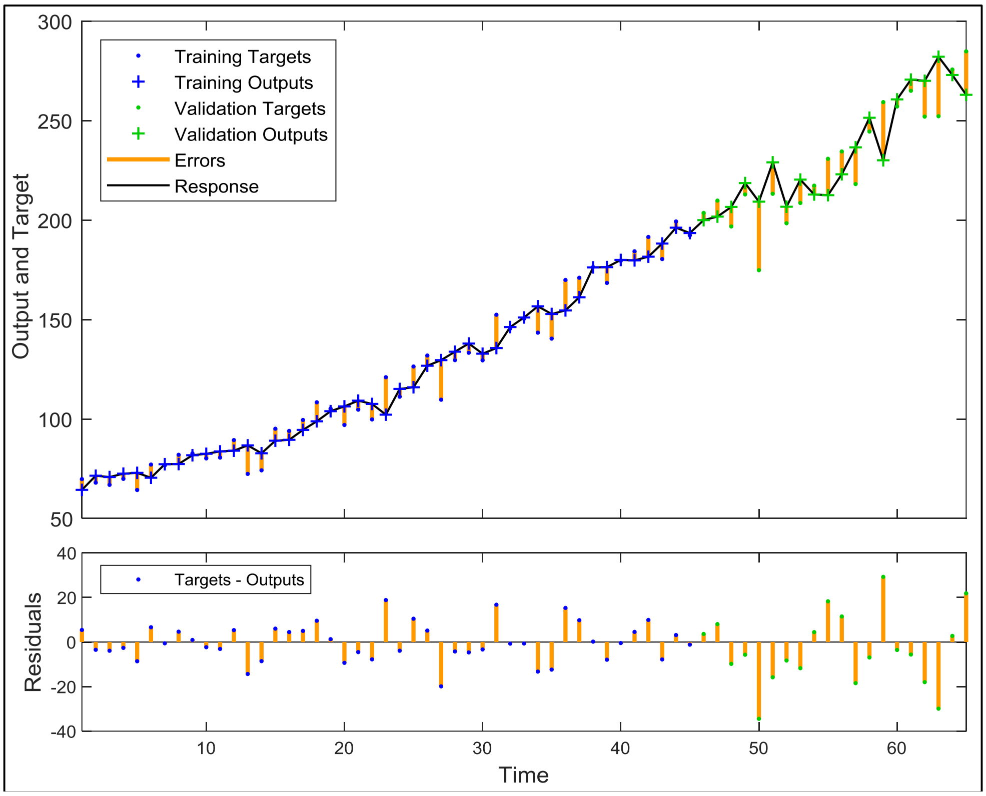

Fig 5: Predicted response from ARNN (3,4) (above) and Residuals (below) for Food Grain Production

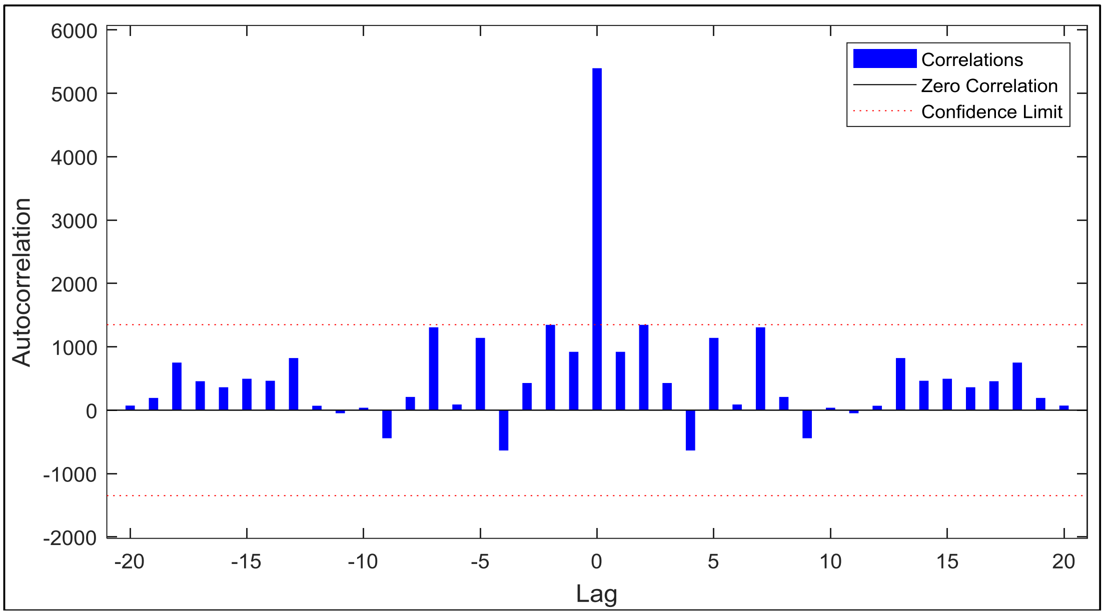

Fig 6: Autocorrelation of Residuals of ARNN (4,3) for Food Grain Yield

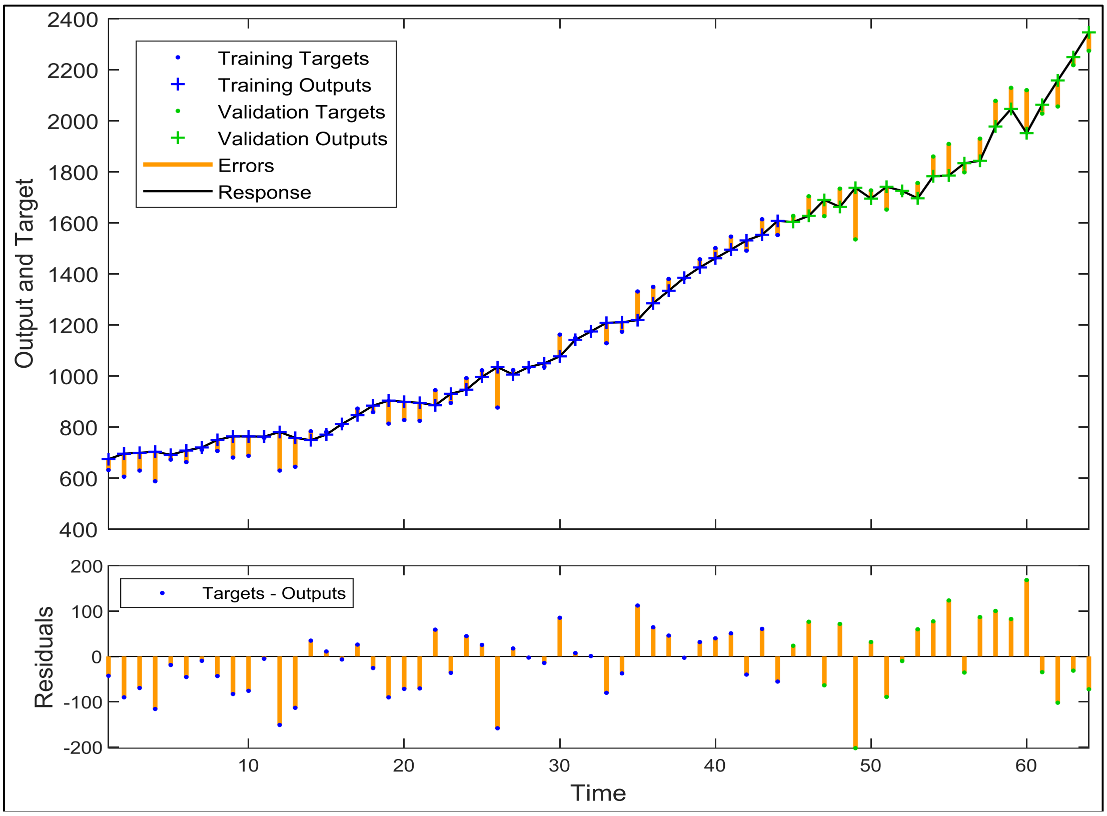

Fig 7: Predicted response from ARNN (4,3) (above) and Residuals (below) for Food Grain Yield

Table 3: Estimation of Food Grain Production (million tonnes) and Yield (kg/ha) for 2018 to 2025 in India

Year	Production (ARNN (3,4))			Yield (ARNN (4,3))		
	Estimates	Prediction Interval		Estimates	Prediction Interval	
2018-2019	293.89	271.65	316.14	2204.99	2055.41	2354.56
2019-2020	300.83	278.12	323.53	2189.95	2037.28	2342.61
2020-2021	311.07	287.90	334.24	2387.14	2231.36	2542.91
2021-2022	318.68	295.04	342.31	2638.87	2479.96	2797.77
2022-2023	330.06	305.96	354.16	2652.11	2490.05	2814.17
2023-2024	337.99	313.42	362.57	2483.47	2318.25	2648.69
2024-2025	349.65	324.61	374.70	2618.04	2449.63	2786.44
2025-2026	356.95	331.43	382.48	3183.07	3011.46	3354.68

Subsequently, to capture the linear and non-linearity in the time series, hybrid ARIMA-ARNN was selected for forecasting. The residuals of fitted data obtained from ARIMA once again fitted using ARNN. Later, outputs from both the components of the hybrid model are combined for final forecasting.

Forecasting is an important function of managing the food supply demand. In this context, we set up the ARIMA, ARNN, ARIMA-ARNN models to forecast the total food grain production and productivity in the terms of yield. The historical time series data were used to develop the above models and the adequate one was selected according to performance criteria: SBC, RMSE and MAPE. The best model was selected based on minimized performance criteria on validation dataset is ARNN (3, 4) and ARNN (4, 3), refer Table 1 and Table 2. Utilizing the best selected model, Fig. 5 for production and Fig. 7 for yield shows the fitted responses for training(in-sample) where blue plus sign represents predicted values whereas blue dot sign is the actual values and validation(out-sample) where green plus sign represents predicted values whereas green dot sign is the actual values. Finally, estimating the production and yield was shown in Table 3 from 2018-2019 to 2025-2026 with their 95% prediction interval which denotes a range of possible values for each new estimation using ARNN (3,4) and ARNN (4,3), respectively. It is visualized that the forecasting of total food grain production and productivity both are in increasing trend up to 356.95 million tonnes with a productivity of 3183.07 kg/ha for the year 2025-26, which is a shine of further spectacular improvement in self-sufficiency and sustainability for food grain production in India.

Conclusion

The results obtained proves that this model can be used for modelling and forecasting the future demand of total food production and yield, but each time we need to feed the historical data with the new data to reinforce it in order to improve the new model and forecasting. However, these results will provide to agriculture domain expert to make decisions about demands in future. Once we obtain a forecast, it will be much easier and very clear to make the right production and yield planning, thus to reduce import costs. It will help us to take right decisions related to demand and supplying of the required food. Lastly, based on the forecasting and validation results, it may be concluded that ARNN model could be successfully used for forecasting production and productivity (yield) of food grains of states as well as India for the subsequent years.

References

1. Box GEP, Jenkins GM, Reinsel GC. Time Series Analysis, Forecasting and Control. 3rd edition, Prentice Hall, Englewood Clifs, 1994.
2. Dheer, Puneet, Yadav, Pradeep. Estimation of production and yield of pulses using ARIMA-ARNN modelling. Journal of Food Legume. 2018; 31(4).
3. Mcculloch WS, Pitts W. A Logical Calculus of the ideas immanent in Nervous Activity. Bulletin of Mathematical Biophysics. 1943; 5:115-133.
4. Naveena K, Singh Subedar, Rathod, Santosha, Singh, Abhishek. Hybrid ARIMA-ANN modelling for forecasting the price of Robusta coffee in India. International Journal of Current Microbiology and Applied Sciences. 2017; 6(7):1721-1726.
5. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by backpropagating errors. Nature. 1986; 323:533-536
6. Sarika, Iquebal MA, Chattopadhyay. Modelling and forecasting of pigeonpea (*Cajanus cajan*) production using autoregressive integrated moving average methodology. Indian Journal of Agricultural Sciences. 2011; 81(6):520-523.
7. Suresh KK, Krishna, Priya SR. Forecasting sugarcane yield of Tamil Nadu using ARIMA Models. Sugar Tech. 2011; 13(1):23-26.
8. Zhang G. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003; 50:159-175.