Time series prediction Using Hybrid ARIMA -ANN Models for sugarcane

Abstract

Recently Hybrid model approach has led to a tremendous surge in many domains of science and engineering. In this study, we present the advantage of ANN to improve time series forecasting precision. The Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) models are used to separately recognize the linear and nonlinear components in the data set respectively. In this manner, the proposed approach tactically utilizes the unique strengths ARIMA and ANN to improve the forecasting accuracy. Our hybrid method is tested on two Yamunanagar and Panipat sugarcane time series of Haryana. Results clearly indicate that Hybrid ARIMA-ANN model was better perform than ARIMA models with smaller values of RMSE and MAPE for both districts.

Keywords: ARIMA, ANN, HYBRID Model, RMSE and MAPE

INTRODUCTION

Crop yield prediction is crucial for farmers, decision-makers, and commercial companies in order to improve business planning and obtain a competitive edge. Prediction is essential in many aspects of our lives. Different business sectors have created and used a variety of forecasting models, including regression models, exponential smoothing, Box-Jenkins models, neural networks, and fuzzy system models. A time series approach, such as the autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models, has been used successfully by many researchers to provide accurate prediction. Previous investigations stabilized the trends using statistical techniques. The ARMA was determined to fit the data with the greatest degree of certainty. The primary drawback is that the models of AR, MA, and ARIMA are considered to have a linear structure. In other words, the time series values are considered to have a linear correlation structure; the ARIMA model cannot capture nonlinear patterns. Time series are frequently acknowledged to be nonstationary and/or nonlinear. Due to these two characteristics, the system ought to be sensitive enough to recognize a time series' uniqueness and replicate it in prediction. Although some data are nonlinear, traditional time series approaches are constrained by the assumption of linearity. Many scientists are employing soft computing methods including fuzzy logic, neural networks, fuzzy neural networks, and

evolutionary algorithms to get beyond the limits of the conventional approaches. It is said that one such method is neural networks. Artificial neural networks (ANNs) can learn complicated systems from distorted and partial data because they have a higher noise tolerance than traditional regression-based empirical modeling. Additionally, they are more adaptable, with the ability to learn dynamic systems through retraining using new data patterns and a suitable training data set. The forecasting error performance using statistical and graphical approaches was used to express the forecasting dependability of the suggested neural networks. The "back propagation" method is the most typical learning technique. Suitable for prediction, the Back Propagation Neural Network (BPNN) is a supervised learning network. Srivastava and Brahmaprakash (1994) predicted sugarcane productivity using time series data spanning 50 years, from 1940–1941 to 1989–1990, using ARIMA models. Their results demonstrated that the time series data on sugarcane productivity for the state of Bihar may be accurately captured by an ARIMA (0, 1, 1) model. Amin and Razzaque (2000) used ARIMA model to describe how a time series variable relates to its own historical value. Varmora et al. (2004) studied the effect of climatic conditions and temporal trends on the wheat crop in the Junagadh district for Gujarat state. A regression equation to predict wheat yield has to be constructed. The actual yield in the Junagadh district deviated from the projected yield by 5.85 to 9.05 percent. Satya Pal et al. (2007) attempted to predict milk output using statistical time series modelling techniques as double exponential smoothing and Auto-Regressive Moving Average (ARIMA) using 25-year study period. Ali et al. (2013) discussed the ARIMA model to forecast Bangladesh's demand for food grain and its supply. After applying multiple ARIMA models for food grain requirements, the results showed that the forecasting model provides higher forecasting accuracy as compared to the Bangladesh Bureau of Statistics (BBS) and also aids in forecasting for the following five and ten years. Babu and Reddy (2014) showed that a balanced combination of linear and nonlinear models provides a more accurate prediction model than either linear or nonlinear models used alone when forecasting time series data from several applications. To forecast time series data, a new hybrid ARIMA-ANN model is created utilising linear autoregressive integrated moving average (ARIMA) and nonlinear artificial neural network (ANN) models. The results from each of these data sets show that the suggested hybrid model has a greater prediction accuracy for both single-step and multistep forecasts. Elwasify, A. I. (2015) investigated that the best model combination between neural networks and ARIMA models to predict the stock

market index EGX30 and it offers more accurate results than ARIMA and ANN each separately. The combination between these models combines the flexibility of the time series and the power of artificial neural networks, where one of these models makes up for the deficiency of the other model. Athira, T. (2017) used artificial neural networks and a hybrid artificial neural network to measure and forecast the particulate matter concentration of PM10. The factors that affect particulate matter concentration, such as temperature maximum and minimum, rainfall, relative humidity, and station level pressure, were taken into account as input parameters for the simulation. ANN and hybrid ANN models were created for the Trivandrum district, and statistical analysis was utilized to compare the models' performances. Bholanath et al. (2019) employed the ARIMA model to forecast wheat production in India. according to the ARIMA (1,1,0) model Wheat output is expected to rise by an average of 4% per year which was judged to be the most reliable for investigation. Gjika et al. (2019) developed a group of hybrid models and propose modifications to increase the accuracy in prediction. Ayub and Jafri (2020) investigated the excellence of hybrid ARIMA-ANN model over ANN-ARIMA in forecasting Karachi stock price. Md. M.H. Khan et al. (2020) used Wavelet transformation, Autoregressive Integrated Moving Average (ARIMA), and Artificial Neural Network (ANN) strengths were combined to evaluate a new approach of a hybrid model's capability to correctly predict upcoming droughts. Unnikrishnan and Jothiprakash (2020) proposed the combined Singular Spectrum Analysis (SSA), Auto Regressive Integrated Moving Average (ARIMA), and Artificial Neural Network (ANN) to create a hybrid model (SSA-ARIMAANN), which may produce accurate daily rainfall forecasts in a river catchment.

Materials and Methods

The Statistical Abstracts of Haryana were used to compile the time series data for sugarcane yield over a 50-year span, from 1972–1973 to 2020–2021. Data for the time period 44 years (1972-2014) have been used for model structure, and the time period 6 years (2015-2020) have been used for model validation in order to create the best model for predicting the series for the future year.

ARIMA model

The most well-liked and successful statistical models for time series forecasting are the ARIMA models, developed by Box and Jenkins. These are founded on the fundamental idea that a time

series' future values are produced by a linear function of its historical observations and white noise terms. The mathematical formulation of an ARIMA (p, d, q) model is as follows:

$$\varphi_p(B)(1-B)^d y_t = \xi_q(B)a_t$$

Here, *yt*, *εt*are the actual observation and white noise at time *t*, respectively

$$\varphi_p = \varphi_1 B - \varphi_2 B^2 .....\varphi_p B^p$$

$$\xi_q = \xi_1 B - \xi_2 B^2 .....\xi_p B^q$$

are the lag polynomials, and *B* is the lag operator. The constants *p*, *q* are the model orders, whereas, φ_i, ξ_j i=1, 2….p; j= 1,2….qare the model parameters. The term *d* represents the degree of ordinary differencing, applied to make the series stationary. The Box-Jenkins model construction process is typically used to establish the appropriate orders of the ARIMA (p, d, q) model.

Artificial Neural Network (ANN) Model

An artificial neural network is a type of computing network that mimics and models the human brain using artificial intelligence and biological inspiration. Similar to how neurons in the human brain are connected to one another, nerve cells in artificial neural networks are linked together in multiple layers of networks. Nodes are used to describe the neurons. In order to estimate complex correlations without a lot of data or prior information, ANN, a non-linear non-parametric data driven self-adaptive statistical method, is used for time series forecasting. ANN's functional form can be defined as, if the input nodes have lagged values:

$$y_t = t(y_{t-1}, y_{t-2}.....y_{t-p}) + \delta_t$$

or above functional form can be expressed mathematically.

$$y_t = w_o + \sum_{j=1}^{q} w_j . g(w_{oj} + \sum_{i=1}^{p} w_{ij} . y_{t-i}) + \Phi_t$$

y_{t-1} transformed data (input), w_{ij} weights associated with input nodes, w_j weights associated with hidden nodes, is called bias of input, *p* is the number of input nodes and *q* is the number of hidden nodes.

Hybrid approach

Neither ARIMA nor ANN are universally suitable for all kinds of time series when the data under consideration comprises both linear and nonlinear components. Models with both linear

and nonlinear components are advised in these circumstances. In order to describe the linear and nonlinear components of a time series, Zhang, 2003 devised a hybrid technique that applies ARIMA and ANN separately. According to Zhang, we have:

$$y_t = L_t + N_t$$

Where y_t L_t and N_t denote observation, linear and nonlinear components respectively at time *t*.

First, the linear component is fitted with ARIMA, and the related forecast for time t is generated. so that at time t, the residual is given by

$$e_t = y_t - L_t$$

Zhang states that the residuals dataset following ARIMA fitting only contains nonlinear components and can therefore be effectively modelled by an ANN, then the final hybrid forecast at time t is obtained as:

$$y_t = \hat{L}_t + \hat{N}_t$$

Results and discussion

In the identification stage is to identify of appropriate order of integration (d), autoregressive (p) and moving average terms (q). The order of d is identified from the data series are stationary or not. If data series ware stationary then d=0. If the data series ware non-stationary, first order differencing was done in order to transform the non-stationary series to stationary series i.e. d=1. The orders of p and q are identified from the graph of acf and pacf of differenced series. The upward trend in the sugarcane yield for selected districts, namelyYamunanagar and Panipat districts, ware observed from the time series plots shown in figure 1 and 2 respectively. These indicate existence of non- stationarity in the time series. KPSS test is also used for to test the stationarity, is shown in table1, it is found to be non-stationary at level series and stationary at first order differenced series. The plots of ACF and PACF for non-stationary and stationary series are presented in figure 4 and 5 for Yamunanagar and Panipat districts respectively. From the examination of the ACF and PACF plots of stationary series, the tentative ARIMA (0,1,1), ARIMA (1,1,0), ARIMA (1,1,1) ARIMA (2,1,0) and ARIMA (1,1,2) models ware identified, which ware shown in table 2. The ARIMA (1,1,0) and ARIMA (0,1,1) models was selected for Yamunanagar and Panipat districts respectively, as best among all the tentative models on the basis of least value of RMSE and BIC, and maximum value of log-likelihood were shown in table 2.

Figure 1: Time series plots of observed sugarcane yield for Yumananagar district

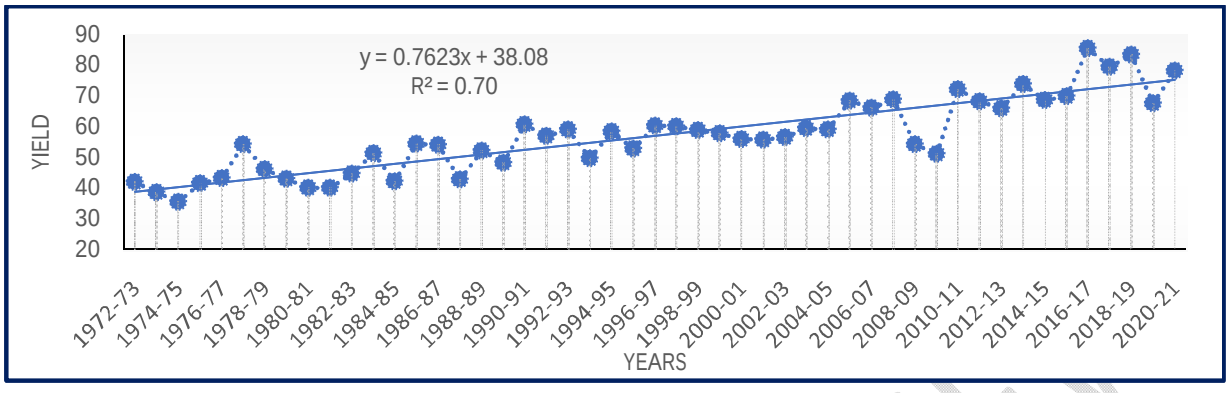


Figure 2: Time series plots of observed sugarcane yield for Panipat district

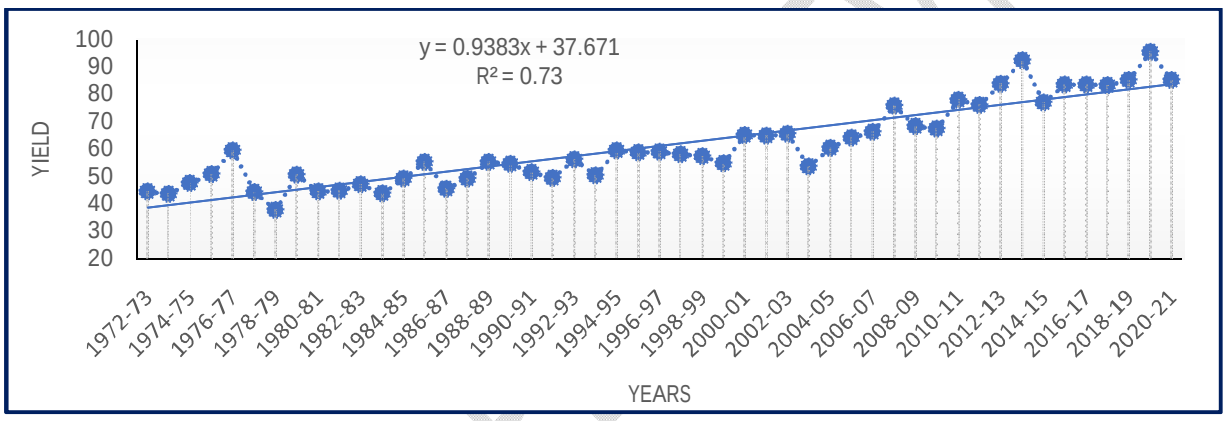


Table 1: KPSS test

Districts	Level		1st differencing	
	Statistic	p-value	Statistic	p-value
Yamunanagar	0.82	0.01	0.06	0.1
Panipat	1.05	0.01	0.11	0.1

Figure 3: ACF and PACF plots for Yamunanagar district

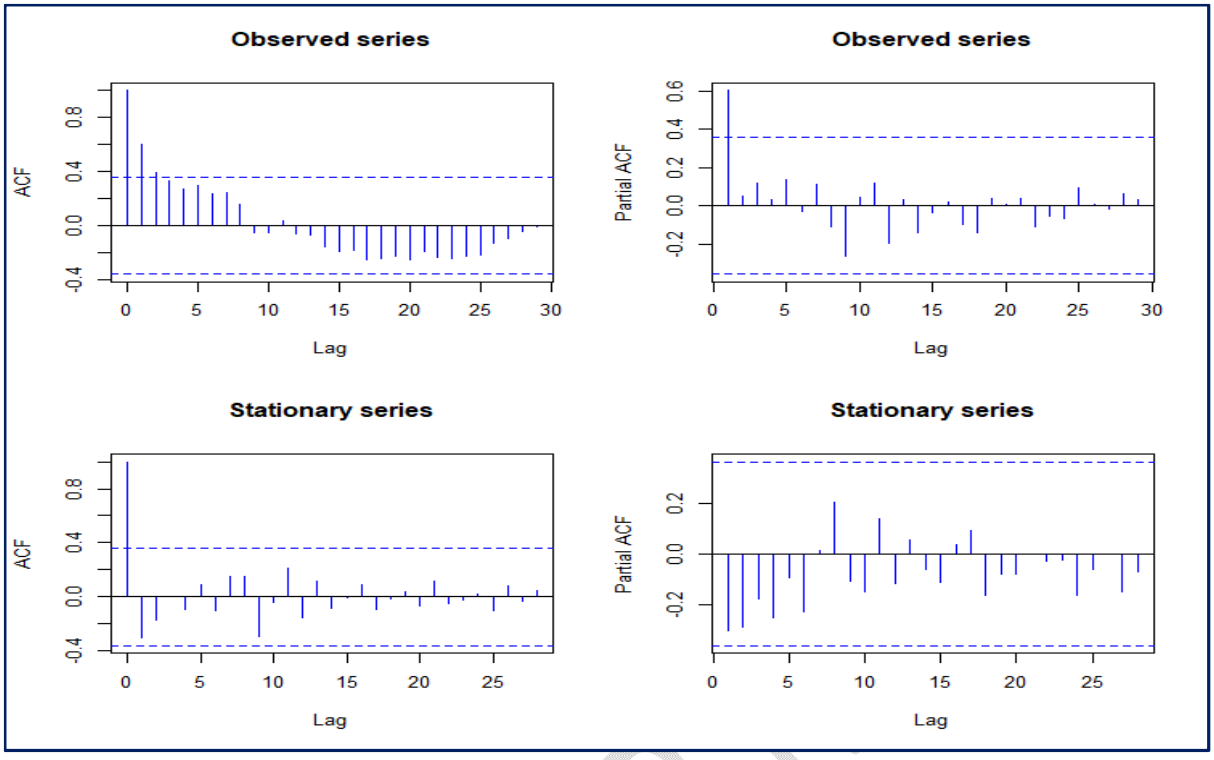


Figure 4: ACF and PACF plots for Panipat district

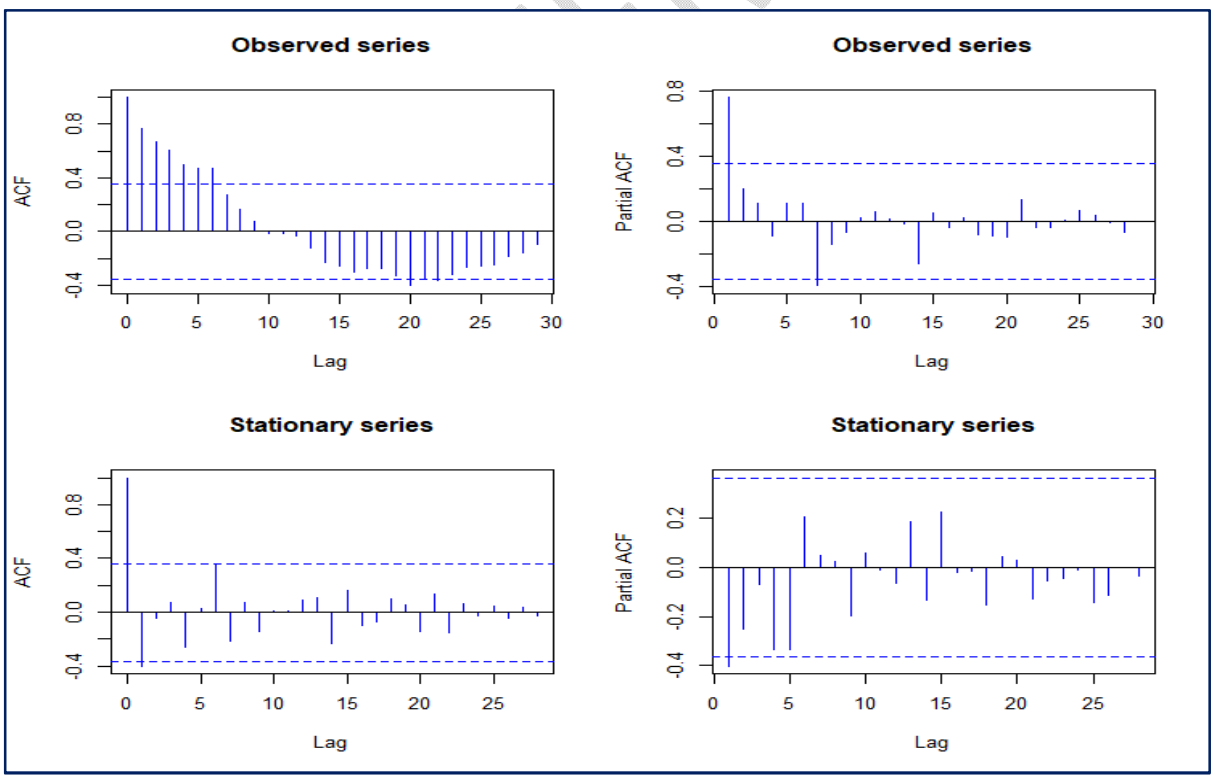


Table 2: Selection criteria values for ARIMA models.

Districts	Models	RMSE	Log-liklihood	BIC
Yamunanagar	ARIMA (0,1,1)	8.00	-146.50	296.87
	ARIMA (1,1,0)	**7.10**	**-141.78**	**291.03**
	ARIMA (2,1,0)	7.36	-142.65	296.52
Panipat	ARIMA (1,1,0)	8.03	-149.70	300.87
	ARIMA (0,1,1)	**7.24**	**-141.31**	**293.53**
	ARIMA (2,1,0)	7.26	-142.16	294.09

Parameter estimation

At the identification step, the ARIMA (0,1,1), ARIMA (1,1,0) and ARIMA (2,1,0) models were taken into consideration. Parameter estimation of ARIMA was done using a non-linear least square approach. Parameter estimation of identified ARIMA models in identification stage were shown in Table 3, all parameter were significant.

Table 3: Parameter estimates of tentatively selected ARIMA models for sugarcane yield in all the districts.

Districts	Models		Estimate	Standard error	T-ratio	Approx. prob.
Yamunanagar	ARIMA (0,1,1)	MA (1)	-0.70	0.09	-7.53	<0.01
	ARIMA (1,1,0)	AR (1)	-0.35	0.14	-2.93	<0.01
	ARIMA (2,1,0)	AR (1)	-0.49	0.13	-3.56	<0.01
		AR (2)	0.13	0.13	-2.93	<0.01
Panipat	ARIMA (1,1,0)	AR (1)	-0.36	0.13	-2.64	<0.01
	ARIMA (0,1,1)	MA (1)	-0.58	0.09	-5.89	<0.01
	ARIMA (2,1,0)	AR (1)	-0.52	0.13	-3.92	<0.01
		AR (2)	-0.42	0.13	-3.92	<0.01

For Yamunanagar ARIMA (1,1,0)

$$\widehat{Y_t} = Y_{t-1} + \emptyset_1(Y_{t-1} - Y_{t-2})$$

$$\widehat{Y_t} = Y_{t-1} - 0.35(Y_{t-1} - Y_{t-2})$$

For Panipat ARIMA (0,1,1)

$$\widehat{Y_t} = Y_{t-1} - \theta_1 e_{t-1}$$

$$\widehat{Y_t} = Y_{t-1} + 0.58 e_{t-1}$$

Table 4: Results on Stationarity and Invertibility conditions for AR and MA coefficients.

Districts	Model	Stationarity	Invertibility
Yamunanagar	**Arima (1,1,0)**	**-0.35**	**#**
Panipat	**Arima (0,1,1)**	**##**	**-0.58**

The MA model does not apply the stationary condition

The AR model does not apply the invertibility condition

The stationary and invertibility condition are fulfilled because absolute value of AR and MA coefficients are less than one for both districts shown in table 4. At parameter estimations step, ARIMA (1,1,0) for Yamunanagar and ARIMA (0,1,1) for Panipat, models were selected for diagnostic checking.

Diagnostic checking

The diagnostic check required determining the residuals from selected models were not autocorrelated and normally distributed simply white noise. An evaluation of the selected Models based on a residual plot for autocorrelations functions was shown in Figure 5 and 6 for Yamunanagar and Panipat districts, clearly indicate that all autocorrelations ware lies in the 95 percent confidence interval, indicate the residuals behavior like white noise and histogram plots indicate the normal behavior in residuals.

Table 5: Diagnostic checking of residual autocorrelations

Districts	Models	Ljung-BoxQ statistic(s)	
		Statistic	**Sig.**
Yamunanagar	ARIMA (1,1,0)	7.28	0.42
Panipat	ARIMA (0,1,1)	10.62	0.15

Figure 5: Plots of residuals from ARIMA (1,1,0) model for Yamunanagar district

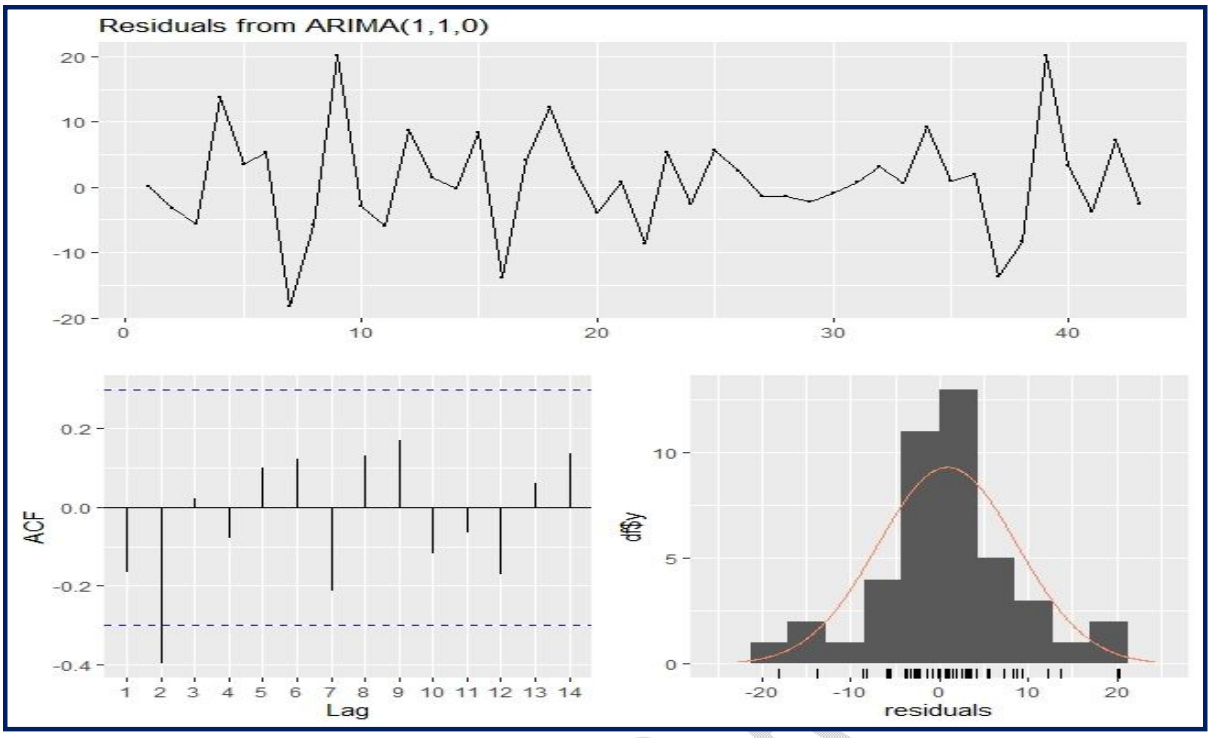
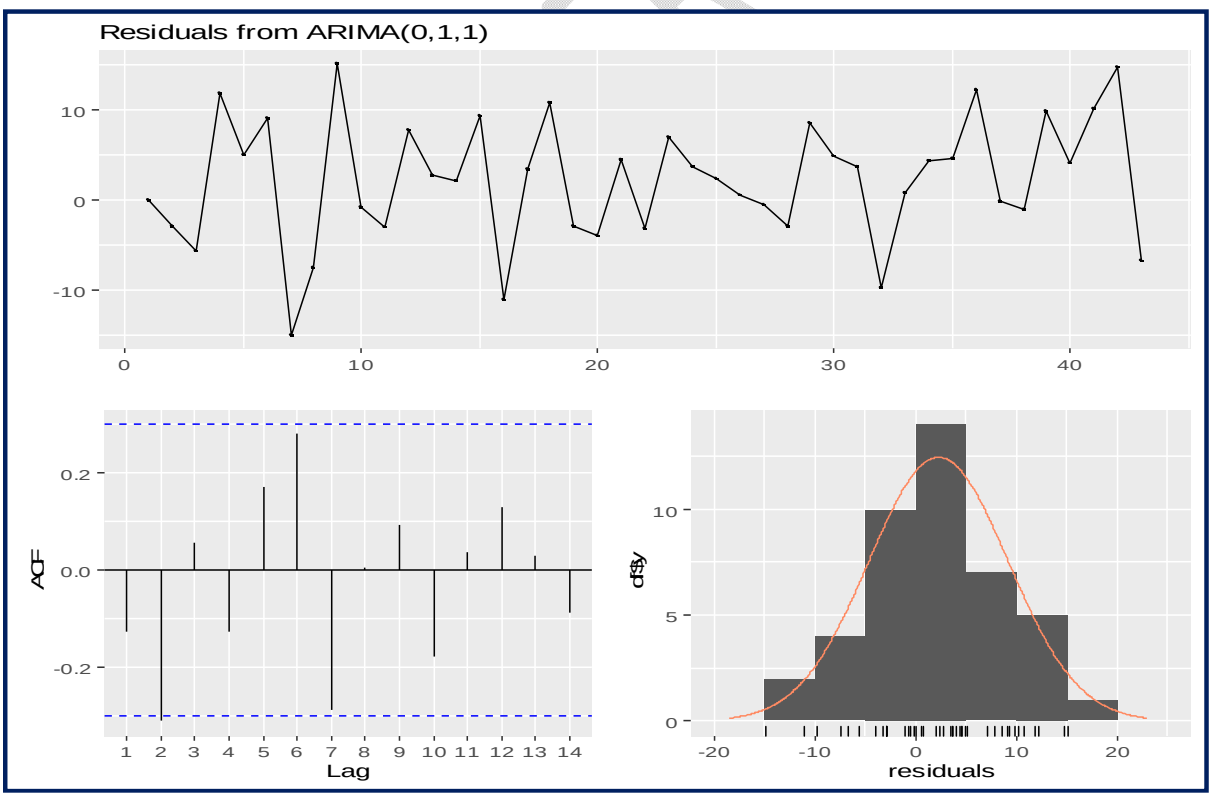


Figure 6: Plots of residual from ARIMA (0,1,1) model for Panipat district



The Box-Ljung test also used to examine the residuals because the coefficients were found to be significant. The value of the Ljung-Box "Q" statistic for all selected models were found to be non-significant shown in table 5, indicate residuals were white noise. Thus, these tests suggest that ARIMA (1,1,0)modelsforYamunanagar, and ARIMA (0,1,1) models for Panipat districts ware adequate for forecasting the sugarcane yield. Figure 7 and 8 shows the observed and forecasting value of selected ARIMA models for Yamunanagar and Panipat districts respectively.

Figure 7: Plot of observed and predicted sugarcane yield by ARIMA (1,1,0) model for Yamunanagar district

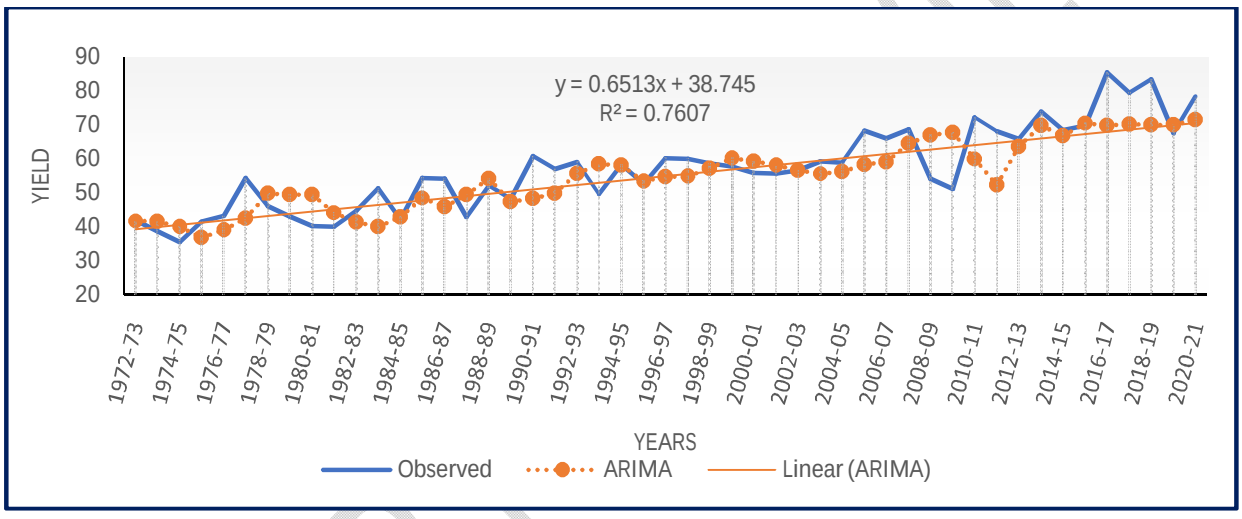
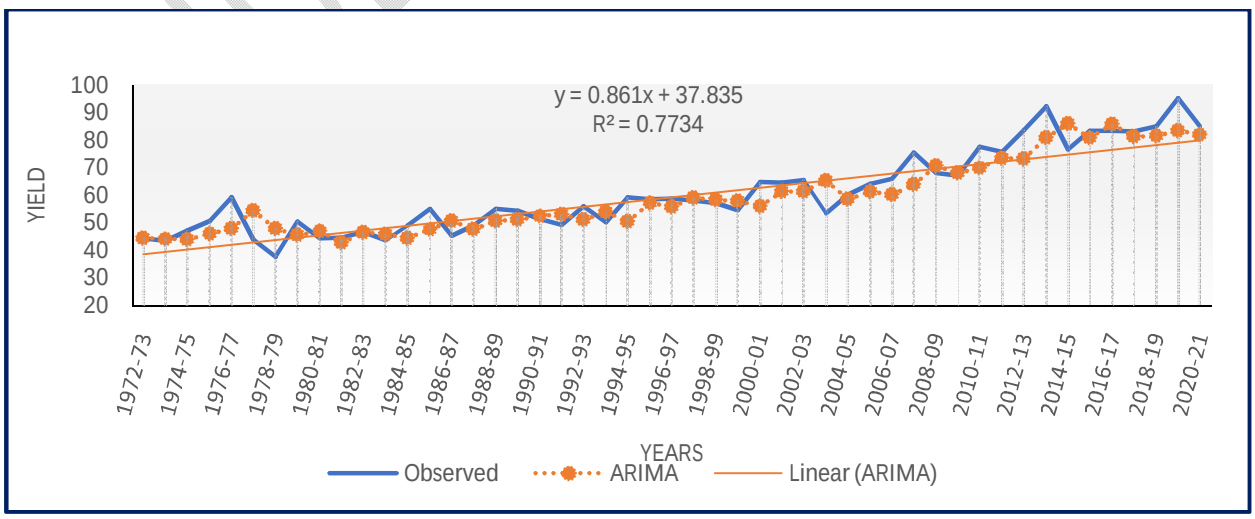


Figure 8: Plot of observed and predicted sugarcane yield by ARIMA (0,1,1) model for Panipat district



ARIMA-ANN Models

ANN models are applied on the residuals of selected as best fitted ARIMA models on all selected districts. Residuals series is divided into two data set, namely, training and testing set, RMSE and MAPE of these data set are shown in table 6. ANN (5-2-2-1) i.e., one input layer with five lag input variables, two hidden layers with two and two neurons, and one output layer with single node selected as best for Yamunanagar. Similarly, ANN (5-3-1)models selected as best for Panipat district, residuals series obtained from ARIMA (1,1,0) and ARIMA (0,1,1) models respectively on the basis of minimum value of RMSE and MAPE for Testing set.

Table 6: Statistical Performance of ANN model on residuals from best fitted ARIMA models for all selected districts

district	**Neural Network structure**				**Training set**		**Testing set**	
	Act. Fun.	Input node	Hidden node	Output node	RMSE	MAPE	RMSE	MAPE
Yamunanagar	Logistic	5	3,2	1	0.85	8.75	2.92	12.67
		5	1,2	1	1.10	6.86	3.71	9.61
	Tanh	5	2,3	1	1.19	8.35	2.75	9.36
		5	**2,2**	**1**	**0.68**	**6.27**	**1.13**	**8.32**
Panipat	Logistic	5	1,2	1	2.65	4.83	1.70	8.33
		5	6	1	0.91	8.25	0.74	5.77
	Tanh	5	5,1	1	0.88	6.08	0.69	5.04
		5	**3**	**1**	**0.65**	**1.78**	**0.66**	**4.05**

Hybrid models i.e. ARIMA (1,1,0)-ANN (5-2-2-1) and ARIMA (0,1,1) – ANN (5-3-1) were selected as best fitted for Yamunanagar and Panipat districts respectively. Figure 9 and 10 shows the observed and forecasting value of selected as best fitted Hybrid (ARIMA-ANN) models for Karnal, Ambala, Kurukshetra, Yamunanagar and Panipat district respectively

Figure 9: Plot of observed and predicted sugarcane yield by ARIMA (1,1,0) - ANN (5-4-1-1) model for Yamunanagar district

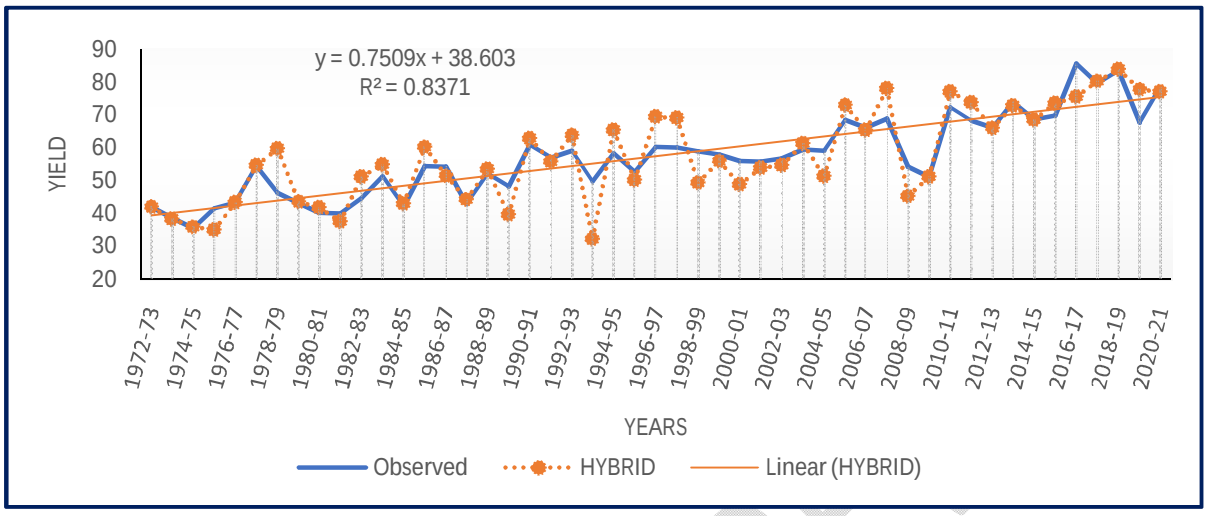
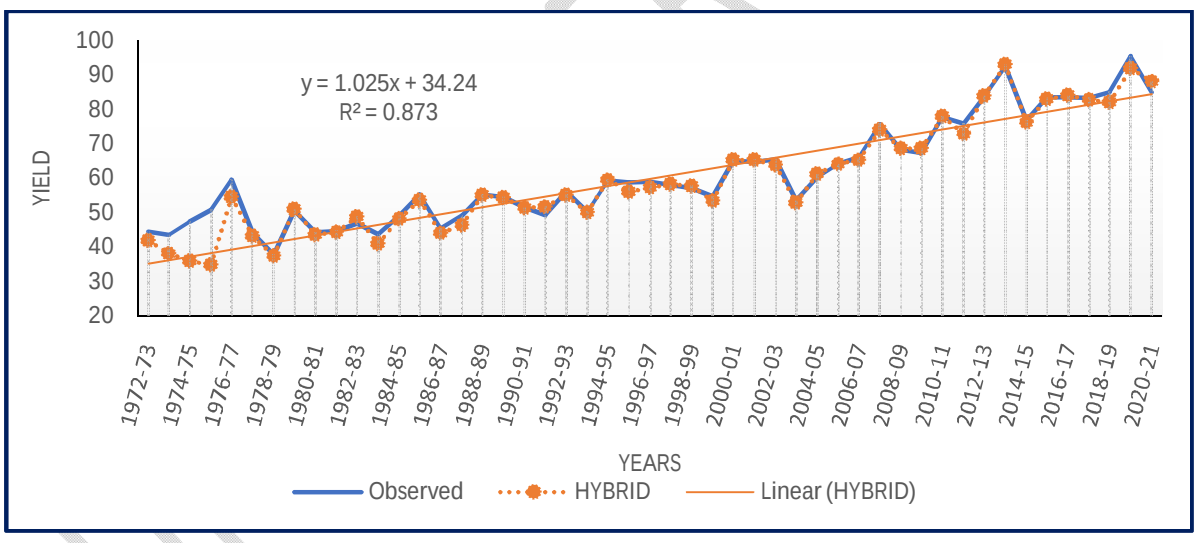


Figure 10: Plot of observed and predicted sugarcane yield by ARIMA (0,1,1) - ANN (5-2-4-1) model for Panipat district



Comparison of fitted models

For comparison purpose, the period from 2015-16 to 2020-21 of sugarcane yield was used as the validation set. The four statistical measures such as percent relative deviation (RD (%)), root mean squared error (RMSE) and mean absolute percentage error (MAPE) was used to compare the performance of ARIMA and Hybrid (ARIMA-ANN) models. Table 7 displays the observed,

forecast and percent relative deviation from the all-different models were selected as the best fitted models on validity set of sugarcane yield for Yamunanagar and Panipat district.

Table 7: Observed and forecast sugarcane yield from the selected models for Yamunanagar district and their relative percentage deviations

Yamunanagar					
Year	**Observed**	**ARIMA**		**ARIMA-ANN**	
		Forecast	**RD (%)**	**Forecast**	**RD (%)**
2015-16	69.9	70.54	-0.92	73.45	-5.08
2016-17	85.57	69.88	18.34	71.53	16.41
2017-18	79.66	70.11	11.99	80.42	-0.95
2018-19	83.53	70.03	16.16	83.86	-0.40
2019-20	67.69	70.06	-3.50	77.88	-15.05
2020-21	78.40	71.56	8.72	76.94	1.86
RMSE		10.64		6.06	
MAPE		9.96		5.46	
Panipat					
2015-16	83.35	82.6	1.15	83	0.42
2016-17	83.55	80.5	3.65	84.03	-0.57
2017-18	83.3	81.27	2.44	82.82	0.58
2018-19	85.1	80.99	4.83	82	3.64
2019-20	95.46	81.09	15.05	91.87	3.76
2020-21	85.12	82.23	3.39	88.28	-3.71
RMSE		5.94		2.58	
MAPE		4.61		2.16	

For Yamunanagar district, RMSE (6.06) and MAPE (5.46) of Hybrid ARIMA-ANN model are smaller as compare to RMSE (10.64) and MAPE (9.96) of ARIMA model. Similarly, Panipat district was shown in table 7, clearly indicate that Hybrid ARIMA-ANN model was better perform than ARIMA models with smaller values of RMSE and MAPE for both districts.

References

Ali, L. E., Islam, M., Kabir, M. R. And Ahmed, F. (2013) Forecasting production of food grain using ARIMA model and its requirement in Bangladesh. *J. Mech. Cont. & Math. Sci.* **7**(2): 1056-1066.

Amin, R.M. D., and Razzaque, M. A., (2000) Autoregressive Integrated Moving Average Modelling for Monthly Potato Prices in Bangladesh. *Journal on Financial Management and Analysis*, **13** (1): 74-80.

Athira, T. (2017) Artificial neural network and hybrid ann models for the prediction of Particulate matter (pm10) concentration. *International Journal of Current Research,* 9(6):53027-53031.

Ayub, S. and Jafri, Y.Z. (2020) Comparative Study of an ANN-ARIMA hybrid model for predicting Karachi stock price. *AmericanJournalofMathematicsandStatistics,* **10**:1-9. Doi: 10.5923/j.ajms.20201001.01.

Babu, N.C. and Reddy, B. E. (2014) A moving-average-filter-based hybrid ARIMA–ANN model forq1forecasting time series data. *Applied Soft Computing*, 1-12.

Bholanath, Dhakre, D.S. and Bhattacharya, D. (2019) Forecasting wheat production in India: An ARIMA modelling approach. *Journal of Pharmacognosy and Phytochemistry*. **8**(1): 2158-2165.

Elwasify, A. I. (2015) A Combined Model between Artificial Neural Networks and ARIMA Models. *International Journal of Recent Research in Commerce Economics and Management,*2(2):134-140.

Gjika, E., Ferrja, A. and Kamberi, A. (2019) A Study on the Efficiency of Hybrid Models in Forecasting Precipitations and Water Inflow Albania Case Study, *Advances in Science, Technology and Engineering Systems Journal*, **4**(1): 302-310.

Md. M.H. Khan, N.S. Muhammad and A. El-Shafie (2020) Wavelet based hybrid ANN-ARIMA models for meteorological droughtforecasting, *Journal of Hydrology*, 590: 1-9.

Satyapal., Ramasubramanian, V. And Mehta, S. C. (2007) Statistical Models for Forecasting Milk Production in India. *Journal of the IndianSociety of Agricultural Statistics*, **61** (2): 80-83.

Srivastava, S. And Brahmaprakash (1994) Pattern of Marketing Arrivals and Prices of Gram in Uttar Pradesh. *Indian journal of agriculturalmarketing*, **8** (1): 13-16.

Unnikrishnan, P. and Jothiprakash, V. (2020)Hybrid SSA-ARIMA-ANN Model for Forecasting Daily Rainfall, *Water Resources Management*, 34:3609–3623.

Varmora, S. L., Dixit, S. K., Patel, J. S. And Bhatt, H. M. (2004) Forecasting of wheat yield on the basis of weather variables. *journal of Agrometeorol.*, **6(2):** 223-228.