Journal of Applied and Natural Science
13(3), 907 - 912 (2021)
ISSN : 0974-9411 (Print), 2231-5209 (Online)
journals.ansfoundation.org

Research Article

Forecast and error analysis of vegetable production in Haryana by various modeling techniques

Manoj Kumar*
Department of Mathematics & Statistics, CCS Haryana Agriculture University (HAU), Hisar (Haryana), India
P. K. Muhammed Jaslam
Department of Mathematics & Statistics, CCS Haryana Agriculture University (HAU), Hisar (Haryana), India
Sunil Kumar
Pulses Section, Department of Plant Breeding, CCS Haryana Agriculture University (HAU), Hisar (Haryana), India
Ashok Dhllion
DES, Krishi Vigyan Kendra, Mahendragargh (Haryana), India

*Corresponding author. Email: m25424553@gmail.com

Article Info
https://doi.org/10.31018/jans.v13i3.2629
Received: March 20, 2021
Revised: August 2, 2021
Accepted: August 8, 2021

How to Cite
Kumar, Manoj *et al.* (2021). Forecast and error analysis of vegetable production in Haryana by various modeling techniques. *Journal of Applied and Natural Science*, 13(3), 907 - 912. https://doi.org/10.31018/jans.v13i3.2629

Abstract
Crop forecasting is a formidable challenge for every nation. The Government of India has developed a number of forecasting systems. The national and state governments need such pre-harvest forecasts for various policy decisions on storage, distribution, pricing, marketing, import-export and many more. In this paper, univariate forecasting models such as random walk, random walk with drift, moving average, simple exponential smoothing and Autoregressive Integrated Moving Average (ARIMA) models are considered and analyzed for their efficiency for forecasting vegetable production in the Haryana state. The State annual data on vegetable production were divided into the training data set from 1966-67 to 2013-14 and the test data set from 2014-15 to 2018-19. Suitable models were selected on the basis of error analysis on the training data and a percent error deviation test on the test data. Model diagnostic checking was carried out on ACF and PACF in residual terms through runs above and below the median, runs up and down and Ljung-Box tests. It is inferred that ARIMA (2,1,1) was found to be optimal and that the forecast values for the years 2019-20 to 2023-24 were estimated on the basis of this model, which were 7.82,8.23,8.72,9.2 and 9.72 million tonnes for the year 2019-20 to 2023-24, respectively. The significance of the mode is that we can forecast the values using this best fit model and forecast values are very important for the policymakers and other government agencies for proper policy decision regarding food security.

Keywords: ARIMA, Autocorrelations function, Forecasting models, Time series, Vegetable production, Ljung-Box tests

INTRODUCTION

India is the second largest producer of fruit and vegetable products in the world next to China. Diverse agro-climate zones with distinct seasons make it possible to grow a wide range of vegetables in India. The total area under vegetables was 4.44lakh hectares in 2018-19 with the production of 7.31 million tons (http://hortharyana.gov.in/en). Vegetables are the greatest sources of nutrients, dietary fiber, phytochemicals and vitamins. Short duration, higher productivity of vegetables has resulted in greater economic returns to farmers. Among various states in India, West Bengal, Uttar Pradesh and Madhya Pradesh are the leader vegetable producers contributing nearly 40% to the total production in the country (2nd Advance Estimate, 2016-17). Horticulture crops cover 5.28 lakh hectares area, which is 8.17 % of the gross cropped area of the Haryana state. Production of horticultural crops in the state was 80.85 lakh MT during the year 2017-18 (Horticultural Statistics at a glance (2017). The State of Haryana is blessed with a favourable climate for the production of high-quality fruit and vegetables, exclusive good soil for fruit and vegetables with high production potential and proximity to major markets such as Delhi and the tri-city of Chandigarh. Horticulture crops can become one of

the key components of doubling farmer's income. Keeping in mind the emerging challenges in the field of horticulture crops and providing nutritional protection for the masses, the state department is starting to work on a vision to make Haryana as modern fruit and vegetable cultivation state, a pioneer in the domestic and export markets.

Forecasting is the method that enables to make predictions of the future on the basis of past and present data and analysis of trends. Crop production forecast is an essential parameter for founding a support policy decision regarding food security, effective land-use allocation, technological and environmental issues. Verma *et al*. (2015); Kumar *et al.* (2016, 2017 a b and 2019) made a number of studies for better forecasting using various pre-harvest forecasting techniques. Fildes and Lusk (1984) advise that forecasters should consider a range of methods and analyze their comparative performance over a random selection of series. In this context, the present study was an attempt to forecast vegetable production in Haryana, which will help the public, researchers and decision-makers with longitudinal data on state vegetable production in the future.

MATERIALS AND METHODS

Vegetable production data from 1966-67 to 2018-19 of Haryana state (Horticultural Department, Government of Haryana) have been used in this study. Complete data is split into training and testing, where data from 1966-67 to 2013-14 is considered as training and the rest period as testing data. five-time series models *viz.* random walk, random walk with drift, moving average, simple exponential smoothing and ARIMA model have been tried to fit for forecasting vegetable production. Five tests run on the residuals of training data i.e. test for excessive runs up and down, test for excessive runs above and below media, Ljung-Box test for excessive autocorrelation, test for difference in mean 1st half to 2nd half, test for difference in variance 1st half to 2nd half to determine whether each model is adequate for the data. Similarly, model diagnostic checking can also be done through a minimum of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), minimum of Akaike Information Criterion (AIC), Hannan-Quinn Criterion (HQC) and Schwarz Bayesian Criterion (SBC). The accuracy of the estimate was evaluated by computing relative deviation (RD %) on the test data set.

Time series forecast models:

A brief description of different time series models are given by various authors like Hyndman and Koehlers (2006), Hyndman and Athanasopoulos (2018), Hanke and Wichern (2008), (Box *et. al* (1976) and Fathony *et al*. (2008).

Random walk model

It's a non-stationary stochastic time series model also denote as I (1) model. Suppose $a_t$is a white noise error term with mean 0 and variance σ^2 . Then the series Y_t is said to be random walk if

$$Y_t = Y_{t-1} + a_t \quad (1)$$

It means the value of *Y* (production) at time *t* is equal to the sum of its value at (*t*–1) and a random shock.

The above equation can be re-written as:

$$Y_t - Y_{t-1} = a_t = \Delta Y_t \quad (2)$$

Where Δ denotes the differencing operator.

Random walk with drift

Modifying the equation (1), as follows:

$$Y_t = \delta + Y_{t-1} + a_t \quad (3)$$

Where δ is known as the drift parameter. The name drift comes from the fact that if one writes the preceding equation as

$$(Y_t - Y_{t-1}) = \delta + a_t = \Delta Y_t \quad (4)$$

It shows that Y_t drifts upward or downward, depending on δ being positive or negative. However, the model in equation (4) is also an I(1) model. For I(1) model with drift, the mean, as well as the variance, increases over time, again violating the conditions of (weak) stationary. In short, random walk model, with or without drift, is a non-stationary stochastic process.

Simple moving average

This technique uses a projection from the last few years, say T. The new average value is determined by eliminating and replacing the oldest value with the newest. The technique is ideal for data that is stationary and does not contain trend or seasonal components.

The moving average forecast can be computed using the following equation:

$$F_t = \frac{\sum_{i=1}^{n} y_{t-i}}{n} \quad (5)$$

where, *i* = an index that corresponds to time periods, *n* = number of periods (data points) in the moving average, y_{t-i} = actual value in period *t-i* andF_t = forecast for time period t.

Simple exponential smoothing

It is a process that continually repeats enumeration through the use of the newest data. This approach can be used if trend and seasonal factor do not significantly affect the results. A parameter called the smoothing constant (*α*) is required to smooth out the data with single exponential smoothing. A convinced weighting is given for each data point, *α* for the newest data and (1–

α) for older data etc. The value of α must be 0 to 1. The following is a smoothed-value equation:

$$S_n = \alpha\left[Y_n + (1-\alpha)Y_{n-1} + (1-\alpha)^2 Y_{n-2} + ...\right] \quad (6)$$

Forecasting value with single exponential smoothing can be done by substituting this equation:

$$\hat{Y}_{n+1} = \alpha Y_n + (1-\alpha)\hat{Y}_n \quad (7)$$

The initial value S_0can be calculated from the average of several observations. The first several observations can be chosen to determine S_0.

ARIMA technique

Univariate Box-Jenkins ARIMA forecasts are based only on past values of the variable being forecast. They are not based on any other data series, and uniquely suited to short-term forecasting. The Box-Jenkins procedure for finding a good forecasting model consists of the following three stages *i.e*., identification, estimation and diagnostic checking stage (Kumar *et al*. 2019). It is a generalization of ARMA (Autoregressive moving average) model denoted by ARMA (*p*, *q*) can be written as

$$\phi_1 Y_{n-1} + \phi_2 Y_{n-2} + ... + \phi_p Y_{n-p} + e_n - \theta_1 e_{n-1} - \theta_2 e_{n-2} - ... - \theta_q e_{n-q} \quad (8)$$

This technique affords a model with the smallest number of parameters for explaining the available data. The initial differencing step is done to lessen the non-stationary. They are denoted by ARIMA (*p*,*d*,*q*), where *p* denotes the order of autoregressive processed denotes the degree of differencing,*q* denotes the order of moving-average process

Diagnostics checking and error analysis

The models that are estimated are acceptable only when the residuals are random. For this purpose, several alternative models that may be appropriate were to be fitted. The ACF and PACF of the residuals of these models are then estimated. If the plot of these ACF and PACF exhibit a non-significant pattern, then the corresponding model is valid and can be considered for forecasting. Three standard tests to test the randomness of residuals based on ACF and PACF are: (1) Runs above and below median (2) Runs up and down and (3) Ljung-Box tests.

To measure the adequacy of the fitted model, the error analysis is useful which compares the results of the fitting of various models. Smaller values of these accuracy measures indicate a good fitted model with minimum forecasting error (Karim *et al*. 2010). The most pertinent accuracy measures can be calculated using the following equations:

$$RMSE = \sqrt{\frac{1}{N}\sum e_n^2} \qquad MAE = \frac{\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|}{n}$$

$$AIC = 2k - 2\ln(\hat{L})$$

$$HQC = -2L_{\max} + 2k\ln(\ln(n))$$

$$SBC = k\ln(n) - 2\ln(\hat{L})$$

where, *k* is the number of estimated model parameters, *n* is the number of observations, $L_{\max}$ is the log-likelihood, $\hat{L}$ is the maximized value of the likelihood function and e_n is the residual term of n^{th} observation.

Percent relative deviation (RD %)

This measures the deviation (in percentage) of forecast yield from the observed yield and is measured as:

Percent deviation= {(observed yield – forecasted yield)/ observed yield} * 100 (9)

RESULTS AND DISCUSSION

The present study observed that ARIMA (1,1,0) and ARIMA (0,1,1) model was found to be the best fit model for the forecasting of soybean and cotton yield as reported by Kumar *et al.* (2017 a, b). Tripathi *et al*. (2014) used the ARIMA model to forecast the rice area, production, and productivity of Odisha and India. They observed that ARIMA (1, 1, 1) was best fitted for forecasting rice productivity and production in Odisha whereas ARIMA (2,1,0) model was the best fit for rice productivity and production for all of India. Sharma et al. (2018) used the ARIMA model to forecast the maize production in India and found that that ARIMA (2,1,0) was the most suitable model for forecasting maize production in India for the years 2018 to 2022. Monika *et al.* (2021) studied the behaviour of production of the wheat forecast using the hybrid model approach and found that ARIMA. (1,1,0) with drift was selected on the basis of the lowest AIC and BIC values. So from the above discussion, different authors used ARIMA model techniques and tried to find the best fit model for forecasting purposes. In the present study, ARIMA (2,1,1) model was the best fit for forecasting purposes. By using these best fitted models, crop yield forecasting can be done for ensuring food security, managing import/export and implementing price policy. All over the world, Scientists applied different types of models to obtain accurate forecasts for the area, production and productivity of different field crops.

All five models discussed in the materials and methods have been developed. The method of constructing the ARIMA model is defined hereby briefly. At the identification stages, the appropriate order of the AR and MA polynomials, *i.e*. the values of *p* and *q* must be calculated with the aid of the ACFs and PACFs of the stationary time series. The graphical presentation of vegetable production in the state of Haryana in Fig. 1 clearly

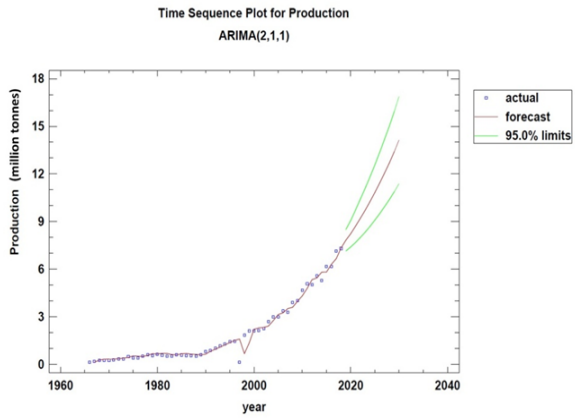

Fig. 1. *Forecast plot of vegetable production.*

shows that the data series is non-stationary. The plotting of the ACFs in Figure 2 also shows that the decline of the ACFS gradually suggests non-stationarity, with most of all the autocorrelation up to the 16th lags substantially different from zero, indicating the same non-stationarity state. Thus, the series considered here were transformed into stationary series by differencing of order one of the original ones. The PACFs in Fig. 2 show a large spike at lag 1, only indicating that the series might have an auto-regressive portion of order one. ARIMA (1,1,0), ARIMA (0,1,1), ARIMA (1,1,1)and ARIMA (2,1,1) were considered at the identification level. ARIMA estimation was rendered using the least square non-linear method.

The error analysis table compared the results of the fitting of various models to the data (Table 1). The model with minimal RMSE, MAE, AIC, HQC and SBC values was chosen, *i.e*. ARIMA (2, 1, 1) and was used to generate the forecast values.

The currently selected model, ARIMA (2, 1, 1), passed 4 tests out of 5 tests run on the residuals, i.e. this model is adequate for the data. The residual normal probability plot of the fitted model is shown in Fig. 3, which can evaluate residual normality. An approximately straight line should be generated if the points derive from the normal distribution, and here also most of the residual points lie close to the straight line.

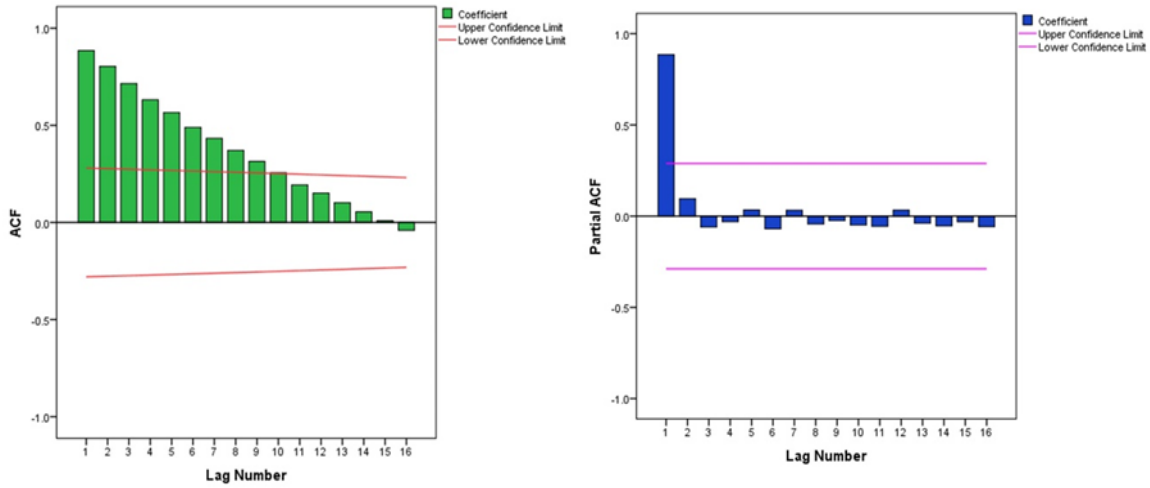

Fig. 2. *ACF and PACF plot for vegetable production.*

Table 1. Error analysis for model comparison.

Model	RMSE	MAE	AIC	HQC	SBIC
Random walk	377.63	195.57	11.87	11.87	11.87
Random walk with drift = 115.543	363.41	190.04	11.83	11.85	11.87
Simple moving average of 2 terms	392.80	240.15	11.99	12.00	12.03
Simple exponential smoothing with alpha = 0.8365	371.03	203.29	11.87	11.89	11.91
ARIMA(1,1,0)	373.45	208.30	11.89	11.90	11.93
ARIMA(0,1,1)	375.03	207.38	11.90	11.91	11.93
ARIMA(1, 1,1)	377.34	206.28	11.95	11.98	12.03
ARIMA(2,1,1)	336.39	180.01	11.76	11.81	11.88

The selected ARIMA model summary is given in Table 2. This model assumes that the best forecast for future data is given by a parametric model relating the most recent data value to previous data values and previous noise. The output summarizes the statistical significance of the terms in the forecasting model. Terms with *p*-values less than 0.05 are statistically significantly different from zero at the 95.0% confidence level. The P-value for the AR (1), AR (2) and MA (1) terms are less than 0.05, so it is significantly different from 0. The estimated standard deviation of the input white noise equals to 338.607.

None of the 24 autocorrelations coefficients and partial autocorrelations coefficients were statistically significant in this study, implying that the time series may well be completely random (white noise). The residual ACF and PCF plot is shown in Figure 4.

The results of the comparison between actual and ARIMA vegetable production estimates for the test data set in terms of RD percent are shown in Table 4. The future forecast of production in 000' tons for the next five forecast years (2019-2020 to 2023-2024) along with 95.0 percent forecast limits for the forecast is also given in Table 4. These limits show where the true data value at a selected future time is likely to be with 95.0% confidence, assuming the fitted model is appropriate for the data. Also, Fig. 1 displays the actual and forecast production of vegetables with a 95% confidence limit.

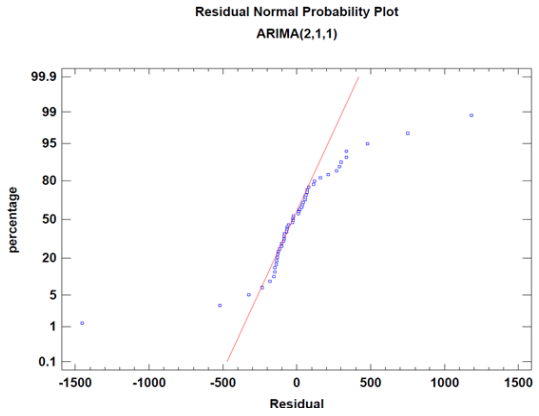

Fig. 3. *Residual normal probability plot.*

Conclusion

In this study, on the basis of error analysis, ARIMA (2, 1, 1) model is best fit for forecasting vegetable produc-

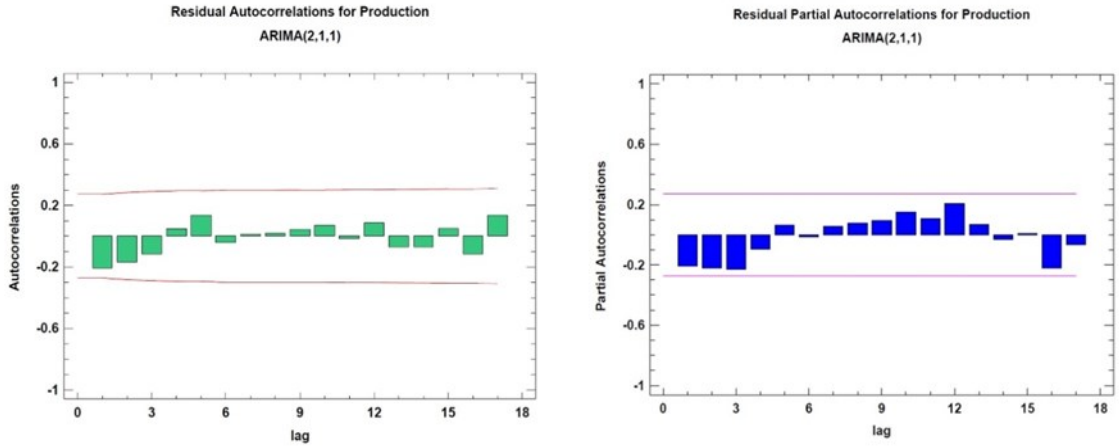

Fig. 4. *Residual ACF and PACF plot for vegetable production.*

Table 2. ARIMA model summary.

Parameter	Estimate	Standard Error	*t* cal.	P>*t*
AR(1)	0.676	0.140	4.836	0.001
AR(2)	0.392	0.150	2.613	0.012
MA(1)	0.970	0.041	23.664	0.001

Table 3. RD % for test data and forecast value of vegetable production.

Test Year	Actual Production (000'tonnes)	Predicted Production (000' tonnes)	RD (%)	Forecast year	Forecasted Production (000' tonnes)	Lower 95% Limit	Upper 95% Limit
2014-15	5286	5808	–9.878	2019-20	7822	7141	8502
2015-16	6157	5820	5.474	2020-21	8235	7402	9068
2016-17	6180	6309	–2.085	2021-22	8717	7684	9750
2017-18	7141	6663	6.696	2023-23	9205	7999	10411
2018-19	7305	7335	–0.415	2023-24	9724	8339	11108

tion in Haryana state. This model provided a forecasted production estimate of 7.82, 8.23, 8.72, 9.2 and 9.72 million tonnes for the forecast year 2019-20 to 2023-24, respectively. These forecasted estimates will be helpful to the government, agro-based industries, traders and agriculturists alike.

Conflict of interest

The authors declare that they have no conflict of interest.

REFERENCES

1. Box, G. E. P. & Jenkins, G. M. (1976). Time series analysis: Forecasting and control. Holden-Day, University of Michigan (ISBN: 0816211043, 9780816211043).
2. Fathony, R.Z.A., Wiboowo, S.H. & Amelia, L. (2008). *Zaitun Time Series 0.1.4. Software*.
3. Fildes, R. & Lusk, E.J. (1984). The choice of a forecasting model. *Omega,* 12(5), 427-435. doi: 10.1016/0305-0483 (84)90042-2
4. Hanke, J. E. & Wichern, D. W. (2008). *Business Forecasting.* Pearson Education, 9th edition, New Delhi (ISBN: 978 -0132301206).
5. Horticultural Statistics at a glance (2017). Horticultural Statistics Division, Ministry of Agricultural and Farmers Welfare, Government of Haryana. http:// hortharyana.gov.in/en/statistical-data.
6. Hyndman, R.J. & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*, Second Edition. OTexts, Melbourne, *Australia* (ISBN: 978-0987507112).
7. Hyndman, R.J. & Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688. https://doi.org/10.1016/ j.ijforecast.2006.03.001
8. Karim, M.R., Awal, M.A. & Akter,M.(2010). Forecasting of wheat production in Bangladesh. *Bangladesh Journal of Agricultural Research*, 35(1), 17-28.doi.org/10.3329/bjar.v 35i1.5863
9. Kumar, M., Battan, K.R. & Sheoran, O.P. (2019). Pre-harvest forecast model for rice yield using principal component regression based on biometrical character with R-software. *International Journal of Agricultural and Statistical Sciences*, 15(1), 323-326. http://www.connectjournals.com/toc2.p...
10. Kumar, M., Raman, R.K. & Kumar, S. (2017a). Forecasting of soybean yield in India through ARIMA model .*International Journal of Pure and Applied Bioscience*, 5 (5), 1538-1546. doi: http://dx.doi.org/10.18782/2320-7051.5834
11. Kumar, Manoj, Rajendra & Hasija, R.C. (2017b). Arima modelling and forecasting of cotton productivity in India. *Environment & Ecology* 35(1A), 224 - 228, January - March 2017. https://www.cabdirect.org/cabdirect/abstract/ 20173068119
12. Kumar, Manoj., Paul, R.K. & Singh, B.K. (2016). Estimating area, production and productivity of cotton crop in Haryana state. *Journal of Cotton Research and Development*, 30(2), 317-323. doi https://doi.org/10.31018/jans.v11i4.2175
13. Monika Devi, Joginder Kumar, D.P.Malik & PradeepMishra (2021). Forecast of wheat production in Haryana using hybrid time series model. *Journal of Agriculture and Food Research,* 5,-5. doi.org/10.1016/j.jafr.2021.100175
14. Tripathi Rahul, Nayak, A.K. Raja, R., Shahid Mohammaad, Kumar, Anjani, Mohanty, Sangita, Panda, B. B., Lal, B. & Gautam, Priyanka (2014). Forecasting Rice Productivity and Production of Odisha, India, Using Autoregressive Integrated Moving Average Models. Advances in Agriculture Volume 2014, 9 pages. doi.org/10.1155/2014/621313.
15. Sharma Pawan Kumar, Dwivedi Sudhakar, Ali, Lyaquat & Arora, R.K. (2018). Forecasting Maize Production in India using ARIMA Model. *Agro Economist - An International Journal,* 5(1), 01-06. doi: 10.30954/2394-8159.01.2018.1
16. Verma, U., Piepho, H.P., Hartung, K., Oggutu, J.O. & Goyal, A. (2015). Linear mixed modeling for mustard yield prediction in Haryana state (India). *Journal of Mathematics and Statistical Science,*1(3), 96-105. http://www.sspub.org/jmss/linear-mixed-modeling-for-mustard-yield-prediction-in-haryana-state-india/