Journal of Pharmacognosy and Phytochemistry

Available online at www.phytojournal.com

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2017; 6(6): 1126-1132
Received: 25-09-2017
Accepted: 27-10-2017

A Dash
College of Agriculture, Chiplima (OUAT), Odisha, India

DS Dhakre
Institute of Agriculture, PSB, Visva-bharati, Santiniketan, West Bengal, India

D Bhattacharya
Institute of Agriculture, PSB, Visva-bharati, Santiniketan, West Bengal, India,

Correspondence
A Dash
College of Agriculture, Chiplima (OUAT), Odisha, India

Forecasting of food grain production in Odisha by fitting ARIMA model

A Dash, DS Dhakre and D Bhattacharya

Abstract
Agricultural scenario of a state is reflected by the analysis of its food grain production status. In odisha, food grains share a major portion of the total cropped area in kharif season. and rabi. In kharif season, cereals form the major share of food whereas, in rabi the major share is by pulses. A time series modeling approach (Box-Jenkins' ARIMA model) has been used in this study to forecast food grain production in Odisha. The order of the best ARIMA model was found to be 2,1,0 (without constant) for kharif food grain production and 1,1,0 (without constant) and for rabi food grain production. The selected best fit models are also validated by using the data which were held up and not used for model building. Further, efforts were made to forecast, as accurate as possible, the future food grain production for kharif and rabi season for a period upto three years by using the best fit model. The forecast results have shown that the food grain production will show a positive growth from 2014-15 to 2016-17 for both kharif and rabi season.

Keywords: Forecasting, Production, ARIMA, autocorrelation function, partial autocorrelation function.

1. Introduction

Agriculture is the backbone of rural economy and livelihood of Odisha. It provides employment both directly and indirectly to about 64 % of the total workforce. It is the largest private enterprise of the State as almost two-thirds of the population of the state is dependent upon agriculture. So, the development of the state is mainly dependent upon the growth in agriculture sector. The agricultural scenario of the state can be best reflected from the analysis of food grains status. This is evident from the fact that, food grain shares almost 86 % and 64 % of the total cropped area in the state, in kharif and rabi season respectively. The forecasting of food grains is of utmost importance for the sake of proper policy framing. The objective of the study is to develop appropriate ARIMA models for the time series of paddy area and production in Odisha and to make three year forecasts with appropriate prediction interval.

2. Methodology

ARIMA (Auto Regressive Integrated Moving Average) model describe the present behaviour of the variable under study in terms of linear relationships with its past values. It is an extrapolation method that requires only historical time series data on the variable under study. ARIMA models are developed basically to forecast the corresponding variable. These models have been developed to forecast the cultivable area, production, and productivity of various crops of Tamil Nadu by (Balanagammal, *et al*. 2000) [1].

These models are also called Box-Jenkins Models on the basis of these authors' pioneering work regarding time series forecasting techniques (Box *et al*. 1994) [2]. The main objective in fitting ARIMA model is to identify the stochastic process of the time series and predict the future values accurately.

The main steps for setting of Box-Jenkins forecasting model are:

(i) Identification of appropriate ARIMA model (ii) Estimation of the parameters
(iii) Diagnostic checking (iv) Forecasting.

During identification, more than one model may be found suitable for being fitted. In this case, the estimation of parameters is done for each of the selected model. Basing on significance of the estimated parameters, model fit statistics and diagnostic checking of the selected models, the best fit ARIMA model is selected. Then, cross validation of the best fit model is done. After successful cross validation of the best fit model, it is used for forecasting. Usually short term forecasting is done.

Description of different steps involved in fitting the appropriate ARIMA model by in the present study is as folows:

2.1 Identification of the model

Identification of an ARIMA model involves two steps (Box *et al*. 2007) [3]. The first (and most important) step in identifying an ARIMA model is the determination of the order of differencing needed to stationarise the series. Normally, the correct amount of differencing is the lowest order of differencing that yields a time series which fluctuates around a well-defined mean value. The stationarity of the series is determined by the plot of ACF (Auto Correlation Function). The ACF plot is a bar chart of the correlation coefficient between the time series and lags of itself. If the plot of ACF of the time series values decays fairly rapidly to zero, either from above or below, then the time series values is considered to be stationary. If the graph of ACF dies down extremely slowly, then the time series values are considered to be non-stationary. The non-stationary series is converted to stationary series by differencing. The no. of differences used to stationarise the series is denoted as 'd'. Usually one or two differences are sufficient to make the series stationary.

After a time series has been stationarised by differencing, the next step in selection of an ARIMA model consists of ARMA (Auto Regressive Moving Average) model for the stationary series achieved through differentiation. The ARMA component is further decomposed into Autoregressive (AR) and Moving Average (MA) components (Pankratz, A. (1983) [6]. The Auto Regressive (AR) components shows the correlation between the observations of current values of the time series with its past values. For example, AR (1) means that the correlation of the observations of any period with its immediate past value. The moving Average (MA) component represents the duration of the influence of a random (unexplained) shocks. For example, MA (1) means that a shock on the value of the series at time t is correlated with the shock at time t +1. By looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of the differenced series, the numbers of AR and/or MA terms that are needed can be tentatively identified. The ACF plot is a bar chart of the coefficients of correlation between a time series and lags of itself, whereas, the PACF plot is a plot of the partial correlation coefficients between the series and lags of itself. The lag beyond which the PACF cuts off is the indicated number of AR terms. The lag beyond which the ACF cuts off is the indicated number of MA terms. The no. of AR terms is denoted as 'p' and no. of MA terms is denoted as 'q'. If there is a unit root in the AR part of the model i.e., if the sum of the AR coefficients is almost exactly one, then the number of AR terms must be reduced by one and the order of differencing must be increased by one. If there is a unit root in the MA part of the model i.e., if the sum of the MA coefficients is almost exactly one, then the number of MA terms must be reduced by one and the order of differencing must be reduced by one.

Let Y_t represents the value of the time series for t = 1, 2, 3,……,n

If the order of differencing, d=1 then $y_t = Y_t - Y_{t-1}$;

If the order of differencing, d=2, then $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} - Y_{t-2}$

The forecasting equation of ARIMA model with p no. of AR terms and q no. of MA terms and order of differencing d is expressed as:

$$Y_t = \mu + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \ldots\ldots + \theta_p Y_{t-p} - \Phi_1 \varepsilon_{t-1} - \Phi_2 \varepsilon_{t-2} - \ldots\ldots - \Phi_q \varepsilon_{t-q}$$

Where, Y_t, Y_{t-1}, Y_{t-2},…..Y_{t-p} are the stationarised values which may be the original values of the series or the values obtained after first or second order differencing.

μ - Constant term; θ_1, θ_2,…, θ_p are the AR coefficients; Φ_1,Φ_2,….,Φ_q are the MA coefficients.

2.2 Estimation of the parameters of the model

After identifying the p (no. of AR terms) and q (no. of MA terms) values, the parameters of the autoregressive and moving average terms are estimated using simple least squares. Sometimes it is not possible to choose a particular value of p and q from the ACF and PACF plot. In such cases more than one combination are tried and accordingly more than one ARIMA models are fitted using different possible combinations of the values of p and q. In the present study, the parameters of the AR and MA terms are obtained by help of the forecasting tool of SPSS 20.0.

2.3 Selection of appropriate ARIMA model

First the estimated parameters i.e. the constant and the coefficients of the AR and/or MA terms are tested for significance by t-test. If the constant is not significant, then ARIMA model without constant is fitted. The coefficients (usually the higher order) of AR and/or MA terms should be significant. After testing the significance of the parameters, the residual diagnostic test for selected model (s) is/are done. The residual ACF and PACF plots were obtained by using the forecasting tool of SPSS 20.0. If none of the residual ACF and PACF are significant, then the model can be considered to be adequate.

A formal test of the fitness of the model is also done by using Box-Ljung statistic which is used for testing independency of the errors.

The Box-Ljung test of the residuals (G. M. Ljung and G. E. P. Box 1978) [5]. is done in following manner:

Null hypothesis, H_0: The errors are random.

Alternate hypothesis: The errors are non-random.

Test statistic, $$Q = n(n+2)\sum_{k=1}^{m} \frac{r_k^2}{n-k}$$

Where, r_k is the estimated autocorrelation of the series at lag k;

m is the no. of lags being tested.

The null hypothesis is rejected i.e., the errors are not independent if $Q \geq \chi^2_{1-\alpha, h}$

The null hypothesis is accepted i.e., the errors are independent if $Q < \chi^2_{1-\alpha, h}$

Where, $\chi^2_{1-\alpha, h}$ is the chi-square distribution table value with h degrees of freedom and α level of significance.

Here, degrees of freedom, h = m – p - q; p and q are the no. of AR and MA terms respectively.

The Box-Ljung test is done by help of forecasting tool of SPSS 20.0

The normality and heteroscedasticity of residuals are tested by Shapiro-Wilk's test and Park's test respectively, as described earlier in the section of this chapter.

Evaluation of the models

Among the models satisfying the residual diagnostics test, the best fit model is chosen from the model fit statistics such as, Mean Absolute Percentage Prediction Error (MAPPE) and Normalised BIC (Etebong P. Clement, 2014) [4].

Mean Absolute Percentage Prediction Error (MAPPE): The model having lowest value of these measures is considered to be the best fit ARIMA model for the given data.

$$MAPPE = \sum_{i=1}^{n} \frac{|P_i - O_i|}{O_i} X100$$

Normalised BIC - The Bayesian Information Criterion is a criterion for model selection among a finite set of models.

BIC = -2 ln L + k{ln(n)};

Where, L is the Likelihood f function, k is the no. of parameters involved in the model, n is the no. of observations (Singh, *et al*. 2015) [8]

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over fitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model.

After exploring the best fit ARIMA model, it is cross validated by obtaining the forecast values from the model for the last three years of our data. The actual values of the last three years data are known and were not used for model building and were held up for cross validation of the selected best fit model. From the actual and forecasted values of the last three years data, the MAE and MAPE are obtained. Low values of these measures ensure the appropriateness of the selected model for forecasting.

2.4 Forecasting: Using the selected best fit model, the forecasted values are obtained for next three years along with the confidence intervals by forecasting tool of SPSS 20.0. The process might continue to obtain forecast for any further period, the standard error associated with the prediction increases. But it is advisable to use ARIMA for short-term forecast since uncertainty increases as prediction is made for farther periods. (Sarika, *et al*. 2011) [7].

3. Result and Discussion

For forecasting production of kharif and rabi foodgrains, appropriate ARIMA model is fitted to the respective data. Though the available data in the present study is from 1970-71 to 2013-14, the data used for model building is from the year 1970-71 to 2010-11. The data from 2011-12 to 2013-14 is held up for cross validation of the selected ARIMA model.

From the plot of the original data regarding production of kharif food grains in Odisha, it is found that the data is non-stationary as seen in fig 1(a). The plot of first difference of the data is found to be stationary as seen in fig 1(b). From the plot of the original data regarding production of rabi food grains in Odisha, it is found that the data is non-stationary as seen in fig 2(a). The plot of first difference of the data is found to be stationary as seen in fig 2(b).

From the ACF and PACF plot of the first difference values of production of kharif food grains as given in figure 3, the tentative values of q and p that are found suitable respectively are q = 0, p = 2. Thus the ARIMA model that is found suitable for the production of kharif food grains is ARIMA (2,1,0). From the ACF and PACF plot of the first difference values of production of rabi food grains as given in Figure 4, the tentative values of q and p that are found suitable respectively are q = 0, p = 2. Thus the ARIMA models that are found suitable for the production of kharif food grains is ARIMA (1,1,0). The ARIMA models are fitted by using SPSS 20.0.

Table 1 provides the estimate of the coefficients of AR of the fitted ARIMA model for production of kharif and rabi food grains. From the table, it is found that in case of production of kharif and rabi food grains, though the coefficient estimates are significant but the estimated constant is not significant in both the cases. So ARIMA (2,1,0) without constant is also fitted for the production of kharif food grains in which the coefficient estimate of the AR(1) and AR(2) components are significant. Also, ARIMA (1,1,0) without constant is fitted for the production of rabi food grains in which the coefficient estimate of the AR(1) component is significant. Thus ARIMA (2,1,0) without constant and ARIMA (1,1,0) without constant are found to be more suitable for production of kharif and rabi food grains respectively.

Table 2 provides the residual diagnostic and model fit statistics of the fitted models. From the table, it is evident that the Box-Ljung statistic, Shapiro-Wilk's Statistic and coefficient of ln(t) in Park's test of heteroscedasticity are all non-significant for the ARIMA models fitted for production of kharif and rabi food grains. This shows that the residuals are white noise, normally distributed and homoscedastic for the ARIMA models fitted for production of kharif and rabi food grains. From the model fit statistics, it is found that the RMSE, MAPE and Normalised BIC are lowest for ARIMA (2,1,0) without constant model in case of production of kharif food grains. and for ARIMA(1,1,0) without constant model in case of production of rabi food grains

As evident from the figure 5 which shows the ACF and PACF of the residuals obtained from fitting ARIMA (2,1,0) without constant model to production of kharif food grains, none of the aurocorrelation and partial autocorrelations of residuals at all lags are significant. Also from figure 6 which shows the ACF and PACF of the residuals obtained from fitting ARIMA (1,1,0) without constant model to production of rabi food grains, none of the autocorrelations and partial autocorrelations of residuals at all lags are significant.

Table 3 shows the cross validation of the best fit models for production of kharif and rabi food grains. For cross validation, the forecasted values of the period 2011-12 to 2013-14 (retained for cross validation) are obtained for production of kharif and rabi food grains by using the respective best fit ARIMA model. Since the actual data regarding production of kharif and rabi food grains for the period is known, Mean Absolute Percentage Error (MAPPE) is obtained for each variable. From the cross validation study it is found that MAPPE for the best fit models is quite low. So the selected best fit models can be used for future forecasting.

Table 4 shows the forecasting values of production of kharif and rabi food grains with 95 per cent confidence limits for the year 2014-15, 2015-16 and 2016-17 obtained by using the respective best fit ARIMA model.

Figure 7 shows the observed, fitted and forecasted values of production of kharif food grains obtained from the ARIMA (2,1,0) without constant model. Figure 8 shows the observed, fitted and forecasted values of production of rabi food grains obtained from the ARIMA (1,1,0) without constant model.

Table 1: Coefficients of the AR and MA components of the fitted ARIMA model considered for forecasting production of kharif and rabi food grains in Odisha

	Best fit ARIMA model	Constant (μ)	Coefficient of autoregressive components		Coefficient of moving average components	
			α_1	α_2	θ_1	θ_2
Kharif	2,1,0	44.89 (144.987)	-0.934** (0.149)	-0.436** (0.148)	-	-
	2,1,0 (without constant)	-	-0.932** (0.147)	-0.434** (0.147)	-	-
Rabi	1,1,0	28.56 (40.291)	-0.547** (0.136)	-	-	-
	1,1,0 (without constant)	-	-0.540** (0.135)	-	-	-

(Figures in the parentheses indicates the standard error)
** Significant at 0.01 level of significance * Significant at 0.05 level of significance

Table 2: Model fit statistics and residual diagnostics of fitted ARIMA models for production of kharif and rabi food grains in Odisha

	Model	Model fit statistics			Residual diagnostics		
		RMSE	MAPPE	Normalised BIC	Ljung – Box Q Statistic	Shapiro-Wilk's Statistic	Coefficient of lnt in Park's test
Kharif	2,1,0	980.64	14.680	14.145	12.517 (8.142)	0.983 (0.786)	0.307 (0.316
	2,1,0 (without constant)	968.26	14.624	14.028	12.061 (8.622)	0.980 (0.887)	0.504 (0.361)
Rabi	1,1,0	182.21	13.591	10.687	18.987 (14.244)	0.912 (0.822)	0.405 (0.312)
	1,1,0 (without constant)	181.48	13.443	10.582	19.143 (18.264)	0.917 (0.711)	0.411 (0.324)

(Figures in the parentheses indicates the standard error)
** Significant at 0.01 level of significance * Significant at 0.05 level of significance

Table 3: Cross validation of the selected best fit ARIMA model for production of kharif and rabi food grains in Odisha

	Model		Actual value (Y)	Forecasted value ($\hat{Y}$)	Error	MAPPE
Kharif	2,1,0 (without constant)	2011-12	6245.11	7517.85	-1272.74	14.898
		2012-13	9817.65	7574.32	2243.33	
		2013-14	7845.23	7730.45	114.78	
Rabi	1,1,0 (without constant)	2011-12	1371.02	1590.79	-219.77	8.709
		2012-13	1581.15	1634.61	-53.66	
		2013-14	1787.23	1667.21	120.02	

Table 4: Forecasted values for production of kharif and rabi food grains in Odisha by using the selected best fit ARIMA model

		Forecasted value	Lower Confidence Limit (95%)	Upper Confidence Limit (95%)
Kharif	2014-15	7915.52	5510.75	10320.28
	2015-16	8038.51	5556.45	10520.57
	2016-17	8214.84	5554.21	10875.48
Rabi	2014-15	1707.17	1172.48	2241.85
	2015-16	1744.44	1154.68	2334.21
	2016-17	1784.46	1150.71	2418.20

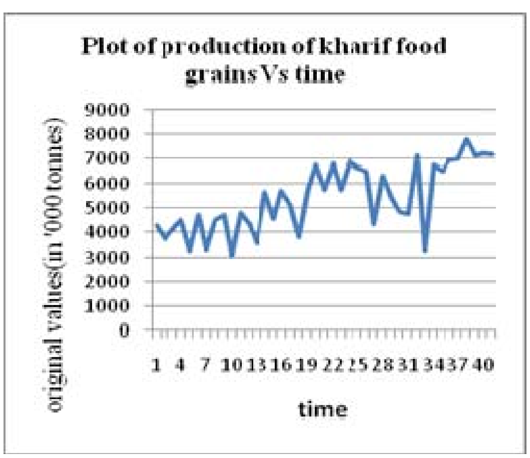

Fig 1(a): Plot of original values of production of kharif food grains Vs time

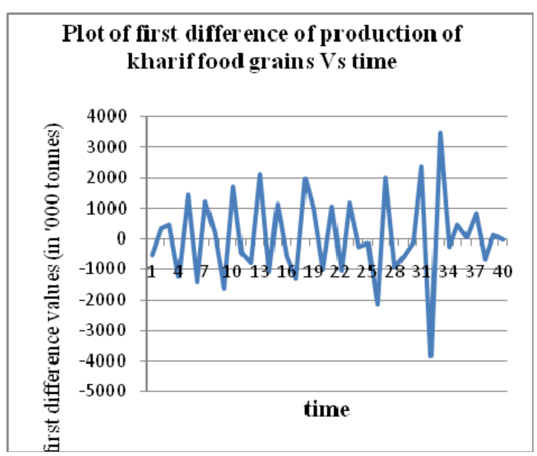

Fig 1(b): Plot of first difference values of production of kharif food gruins Vs time

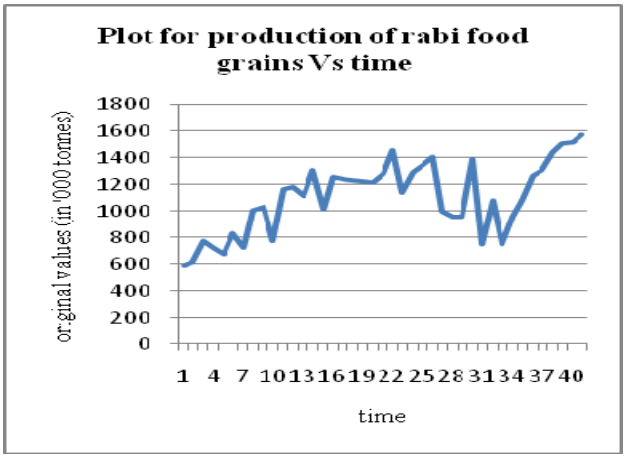

Fig 2(a): Plot of original values of production of rabi food grains Vs time

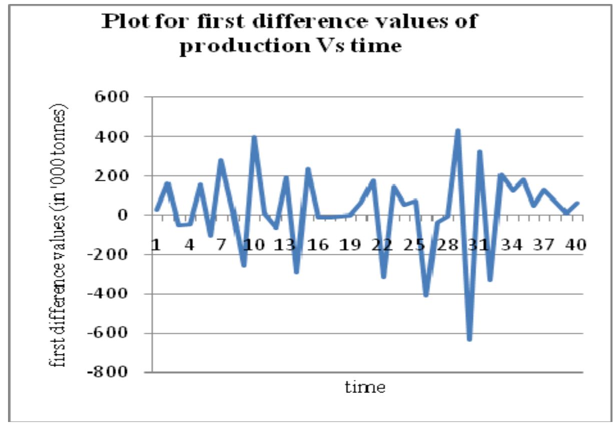

Fig 2(b): Plot of first difference values of production of rabi food gruins Vs time

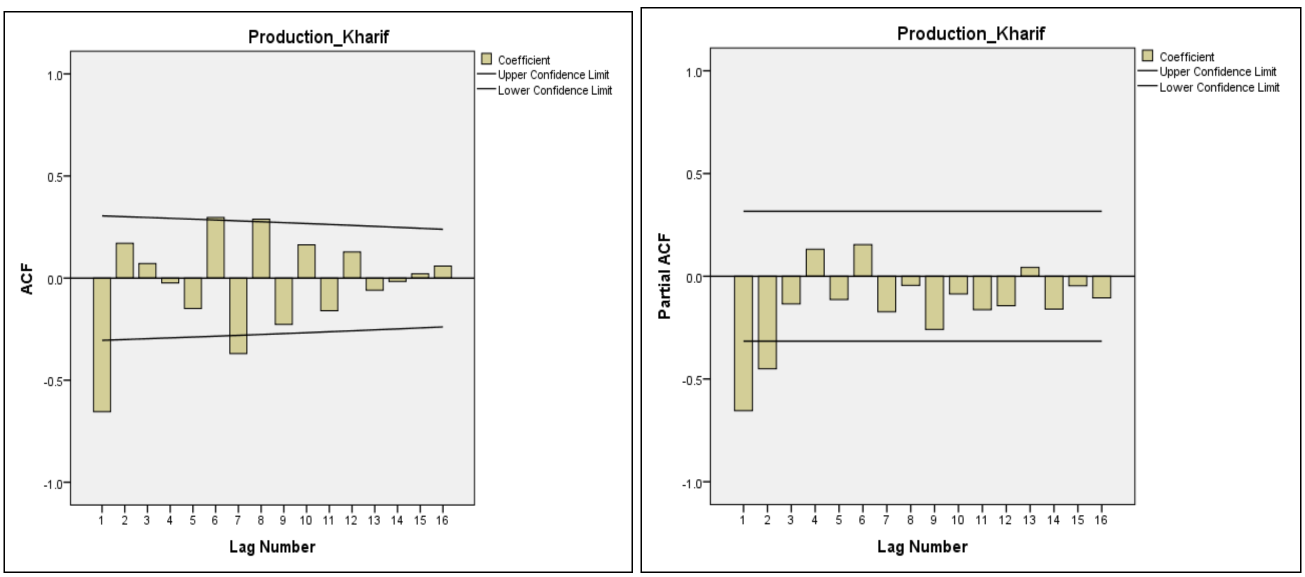

Fig 3: ACF and PACF plot of first difference values of production of kharif food grains

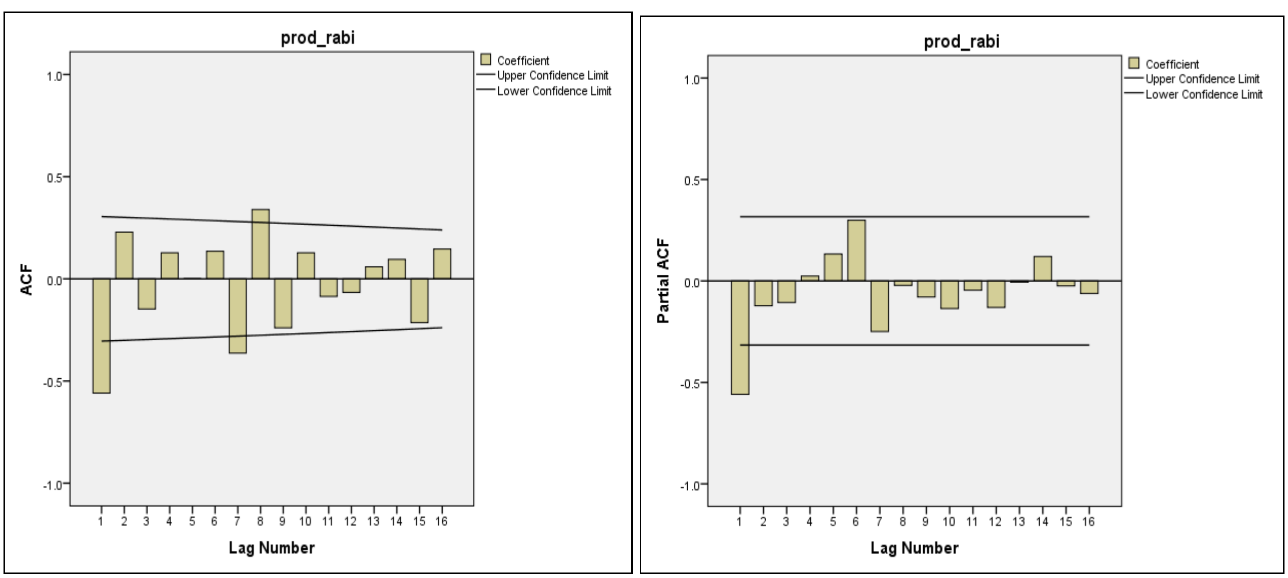

Fig 4: ACF and PACF plot of first difference values of production of rabi food grains

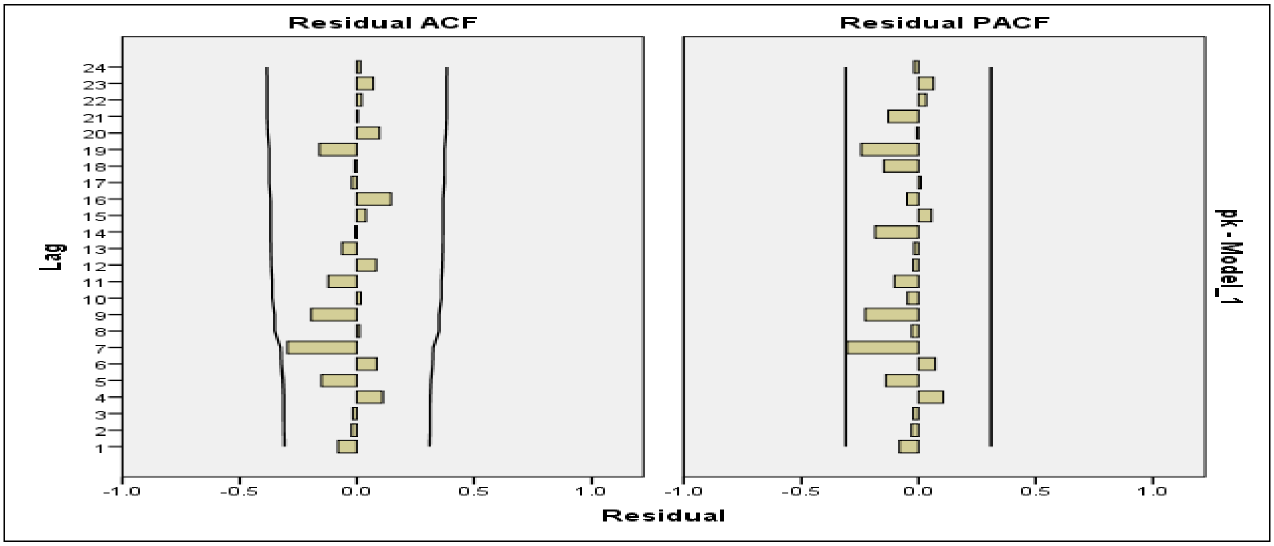

Fig 5: Residual ACF and PACF obtained by fitting ARIMA (2,1,0 without constant) to production of kharif food grains

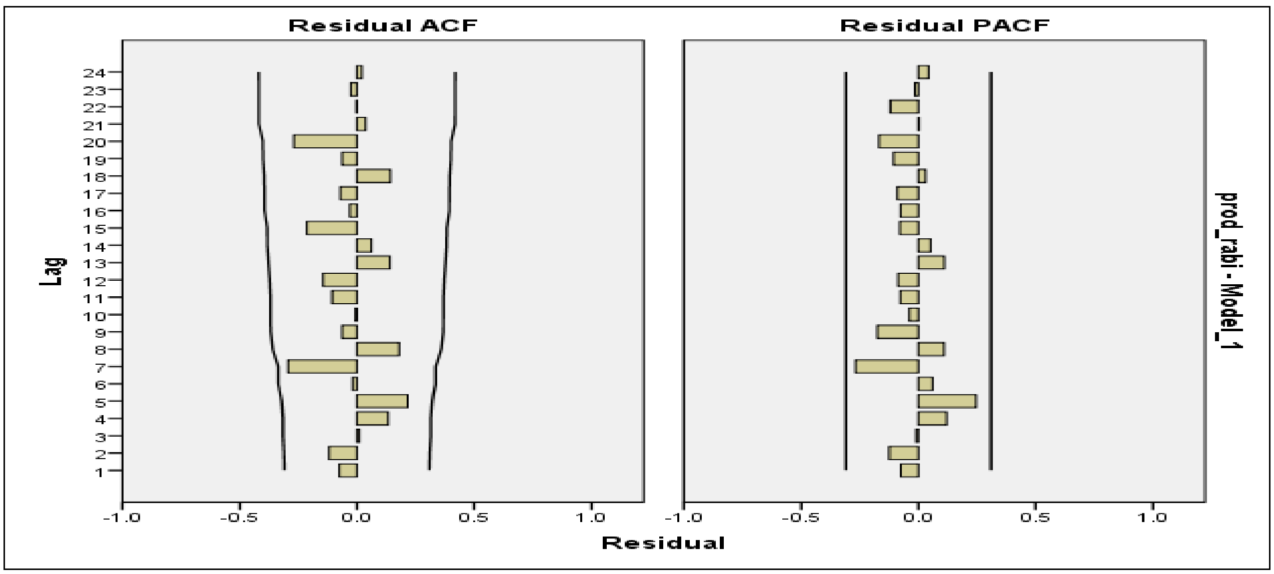

Fig 6: Residual ACF and PACF obtained by fitting ARIMA (1,1,0 without constant) to production of rabi food grains

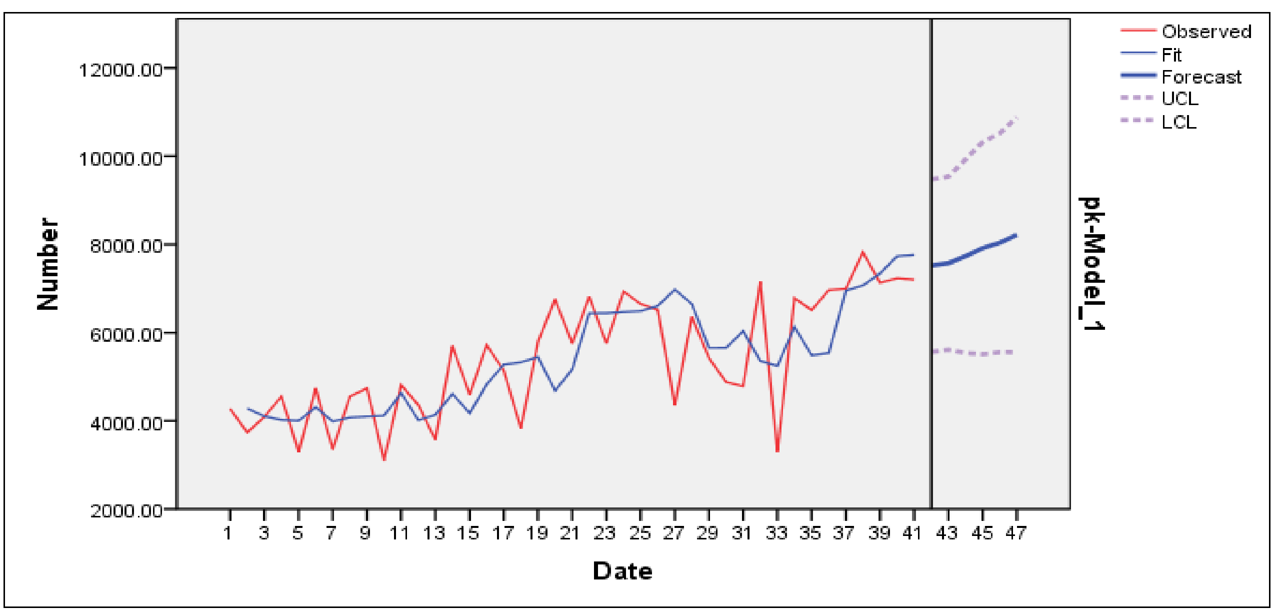

Fig 7: Graph of observed, fitted and forecasted values of production of kharif food grains obtained by fitting ARIMA (2,1,0) without constant model

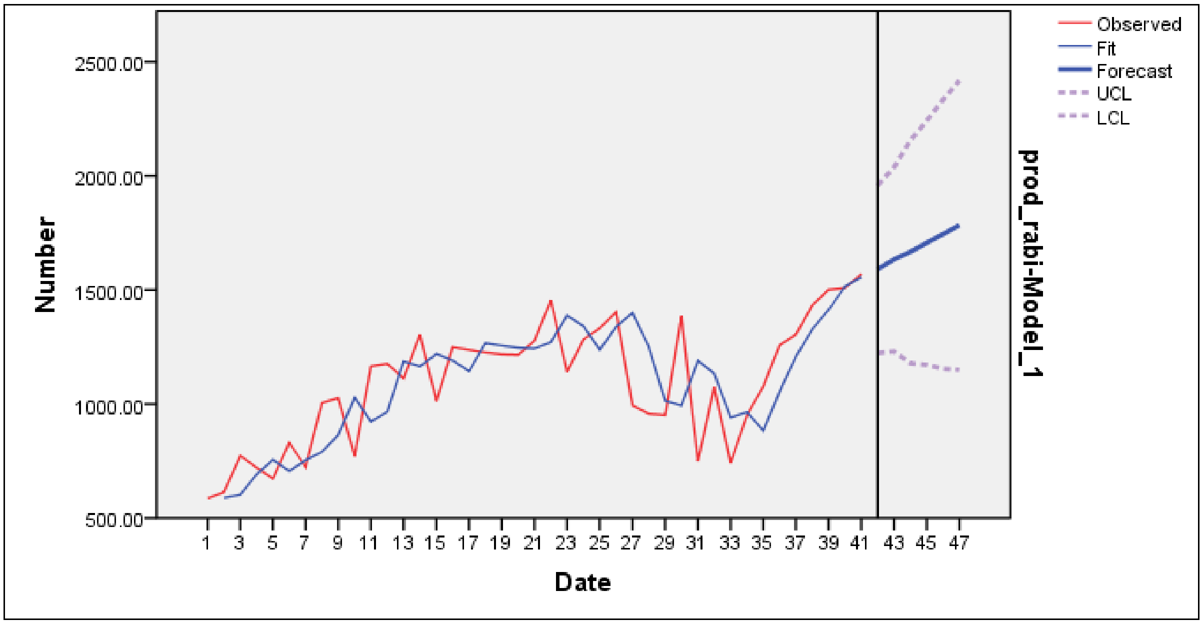

Fig 8: Graph of observed, fitted and forecasted values of production of rabi food grains obtained by fitting ARIMA (1,1,0) without constant model.

4. Conclusion

In the present study ARIMA (2,1,0) without constant and ARIMA (1,1,0) without constant are the models are found to be the most appropriate ARIMA models for forecasting production of kharif and rabi food grains respectively. From the forecast available by using the model developed, it can be seen that forecasted food grain production during kharif season and rabi season expected to increase in the next three years from the study period i.e. 2014-15 to 2016-17. The validity of the forecasted values can be checked when the data for the lead periods become available. The model can be used by researchers for forecasting cultivated areas, production and yield in India. However data need to be updated from time to time with incorporation of current values.

5. References

1. Balanagammal D, Ranganathan CR, Sundaresan K. Forecasting of agricultural scenario in Tamilnadu: A time series analysis. Journal of Indian Society of Agricultural Statistics, 2000; 53(3):273-286.
2. Box GEP, Jenkins GM, Reinsel GC. Prentice Hall, Englewood Cliff, N.J., 3rd edition, 1994.
3. Box GEP, Jenkins GM, Reinsel GC. John Wiley & Sons, Hoboken, N. J., 4th Edition, 2007.
4. Clement EP. Using Normalized Bayesian Information Criterion (BIC) to Improve Box - Jenkins Model Building. American Journal of Mathematics and Statistics. 2014; 4(5):214-221.
5. Ljung GM, Box GEP. On a Measure of a Lack of Fit in Time Series Models. Biometrika. 1978; 65(2):297-303.
6. Pankratz A. Forecasting with univariate Box-Jenkins Models: Concepts and Cases, John Wiley and Sons, 1983.
7. Sarika, Iquebal MA, Chattopadhyay C. Modelling and forecasting of pigeonpea (Cajanus cajan) production using autoregressive integrated moving average methodology. Indian Journal of Agricultural Sciences. 2011; 81(6):520-523.
8. Singh DP, Dhakre DS, Shrivastava A, Sharma S. Application of ARIMA Model for Forecasting Paddy Production in Bastar Division of Chhattisgarh. Environment & Ecology. 2015; 33(3):1083-1088.