International Journal of Chemical Studies

P-ISSN: 2349–8528
E-ISSN: 2321–4902
IJCS 2019; 7(3): 3899-3904

Received: 10-03-2019
Accepted: 12-04-2019

SK Mahapatra
SRF, ICAR-NASF Funded ICT Project, Directorate of Extension Education, OUAT, Bhubaneswar, Odisha, India

A Dash
Assistant Professor (Agril. Statistics), College of Agriculture, Chiplima, OUAT, Bhubaneswar, Odisha, India

Correspondence
SK Mahapatra
SRF, ICAR-NASF Funded ICT Project, Directorate of Extension Education, OUAT, Bhubaneswar, Odisha, India

Forecasting area and production of green gram in Odisha using ARIMA model

SK Mahapatra and A Dash

Abstract

A study was conducted on forecasting of area and production of green gram in Odisha. Box-Jenkins Autoregressive integrated moving average (ARIMA) time-series methodology was considered for forecasting of area and production of green gram. The different ARIMA models are judged on the basis of Autocorrelation Function (ACF) and Partial autocorrelation Function (PACF) at various lags The data from 1971-72 to 2006-07 are used for model building and from 2007-08 to 2015-16 used for successful cross-validation of the selected model on the basis of the absolute percentage error. The ARIMA models are fitted to the original time series data as well as the first difference data to check the stationarity. The possible ARIMA models are identified on the basis of significant coefficient of autoregressive and moving average components. The best fitted models are selected on the basis of low value of Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). ARIMA (1,1,0) without constant found to be best fitted for area under green gram having absolute percentage error ranging from 0.34% to 10.63% in cross-validation of model. ARIMA (2,1,0) is the best fitted model for production of green gram having absolute percentage error ranges from 6.02% to 26.11% in cross-validation of the model. The best fitted ARIMA model has been used to forecast the area, production for the year 2016-17 to 2018-19. The model showed area forecast for the year 2018-19 to be about 895.75 thousand hectare with lower and upper limit 462.18 and 1329.31 thousand hectare respectively. The model also showed the forecasting in production of green gram for the year 2018-19 to be about 313.66 thousand tones with lower and upper limit 134.27 and 493.04 thousand tonnes respectively.

Keywords: ARIMA, Green gram, area, Production, ACF, PACF, RMSE, MAPE

Introduction

Pulses are an important commodity group of crops that provide high quality protein complementing cereal proteins for pre-dominantly substantial vegetarian population of the country. Pulses are the major sources of protein in the diet. Green gram is the important pulse crop in the state of Odisha. Green gram or mung bean (Vigna radiata L.) is the major pulse crop of the state Odisha covering total area of 8.36 lakh ha with average productivity 434 kg per ha. The contribution of green gram to the total pulse area in Odisha is 41.23%, during the year 2015-16. The share of green gram towards total production is 38.57%, during 2015-16.

Crop area estimation, forecasting of production and crop yield are an essential procedure in supporting policy decision regarding land use allocation, food security and environmental issues. Various approaches have been used for forecasting such agricultural systems. Several studies have been carried out on the univariate time series models known as Auto regressive integrated moving average (ARIMA) models. The popularity of ARIMA model is due to its statistical properties as well as known Box and Jenkins methodology. This study helps to the policy makers to get an idea about the future requirements, enabling to take appropriate measures like selection of high yielding varieties, conducting trainings to farmers to improve cultural practices, adequate supply of inputs and use of latest technologies. Import and export of these crops can also be planned.

Materials and Methods

The secondary data on area, production green gram for the period from 1971-72 to 2015-16 were collected from Directorate of Agriculture and Food Production, Government of Odisha.

In time series analysis, the Box–Jenkins method applies autoregressive moving average ARMA or ARIMA models to find the best fit of a time-series model to past values of a time series.

To overcome the difficulty in describing the dynamic structure of the data by fitting Autoregressive (AR) and Moving average (MA) models, the autoregressive moving average (ARMA) models was introduced. The ARMA models, which includes the order of differencing (which is done to stationarise the data) is known as Autoregressive integrated moving average (ARIMA) models. The ARIMA model with parameter (p,d,q) is fitted by univariate Box-jenkins techniques (Box and Jenkins, 1976) [1]. This model includes Autoregressive of order *p*, differencing to make stationary series of degree *d* and moving average of order *q*.

The time series data is said to be stationary, if it has constant mean and variances over the time. The original data plotted first and verified for stationarity, if the data is found to be non-stationary from the graph, then the first difference of the data is plotted and checked for stationarity (Dash, *et al*. (2017)) [2] In this manner we proceed till the data become stationary. The maximum order of differencing (d) is usually 2.

The values of *p* (order of Autoregression) and *q* (order of moving average) can be determined by examining the plots of the Autocorrelation and partial autocorrelation of the stationeries value of the variables. The autocorrelation of *y* at lag *k* is the correlation between *y* and itself lagged by *k* periods, i.e., it is the correlation between y_t and y_{t-k}.The partial autocorrelations of *y* at lag *k* is the coefficient of *y*_LAG*k* in a regression of y on *y*_LAG1,*y*_LAG2,….., up to *y*_LAG*k*. Thus, the partial autocorrelation of y at lag 1 is the same as the autocorrelations of y at lag1.The partial autocorrelation of *y* at lag2 is the coefficient of *y*_LAG2 in a regression of *y* on *y*_LAG1 and y_LAG2, and so on. The way to interpret the partial autocorrelation at lag k is that it is the amount of correlation between *y* and *y*_LAG*k* that is not explained by lower-order autocorrelations.

The adequacy of the selected model is checked using Box-Ljung test. A formal test of the fitness of the model is also done by using Box-Ljung test of the residuals (Ljung and Box (1978)) [3] is done in following manner:

Null hypothesis is set as H_0: The errors are distributed randomly.

And the alternate hypothesis H_1: The errors are non-random.

The Box-Ljung test statistic,

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

Where n is the number of observations,

$r_k$is the estimated autocorrelation of the series at lag k = 1,2,…,m and *m* is the number of lags being considered. Here 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 level of significance α such that $P(\chi^2_h > \chi^2_{1-\alpha, h=}) = 1-\alpha$

Here the degrees of freedom, h = (m – p – q); p and q are the numbers of AR and MA terms, respectively. The Box-Ljung test is done by help of forecasting tool of SPSS 20.0.

The following model fit statistics are used to select the best fit model:

1. Root mean square error (RMSE)

$$\text{RMSE} = \frac{\sum_t e_t^2}{n-2}$$

2. Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{\frac{\sum_t |Y_t - \hat{Y}_t|}{yt}}{n} \times 100$$

The models among the selected ARIMA models have lowest value of RMSE and MAPE is considered to be the best-fit model from the respective data set.

Nearly 20% of the data at the end period is not used for model building and left out for cross-validation of the selected model. The actual value of the left out period and the forecasted value of the left out period from the selected model are used for cross-validation. Here the data from 1971-72 to 2006-07 are used for model building and data from 2007-08 to 2015-16 are used for cross-validation.

For this, the percentage error is calculated.

$$\%\text{ of Forecasting Error} = \left(\frac{Y-\hat{Y}}{Y}\right) \times 100$$

Where, Y is the observed value of remaining eight years

$\hat{Y}$is the forecast values of remaining eight years

Lower the value of forecasting error percentage, better is the prediction by the selected model.

After successful cross-validation of the selected best fit ARIMA model, it is used for forecasting; generally short term forecasting is used in case of ARIMA techniques. This is because if we go on predicting for longer periods than the error associated with the prediction will increase. So, ARIMA should be used for short term forecasting (Sarika *et al.* 2011) [5].

Results and Discussion

Appropriate ARIMA model is fitted to the data on area and production of Green gram for the purpose of forecasting. Data used for model building is from the year 1971-72 to 2006-07. The data from 2007-08 to 2015-16 is used for cross validation of the selected model and the forecasting is done for the years 2016-17,2017-18 and 2018-19 by using the selected best fit model.

The original plot of data on area under green gram as shown in Fig. 1(a) reveals that the data is non-stationary i.e.do not have constant mean and variance. Thus, the first difference of the data is plotted and shown in Fig. 1(b). This plot reveals that the first difference of data is found to be stationary i.e. have constant mean and constant variance.

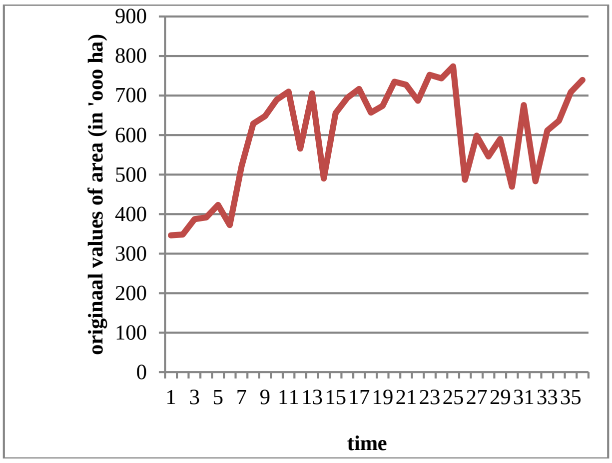

Fig 1(a): Plot of original values of area under green gram vs time

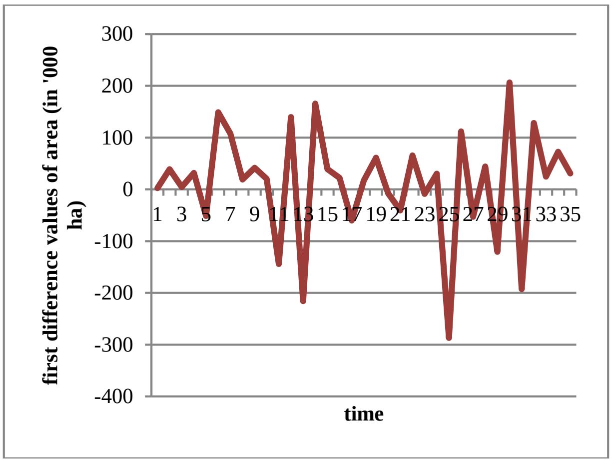

Fig 1(b): Plot of first difference values of area under green gram vs time

The ACF and PACF plot of first difference values of area under green gram shown in Fig. 2, which suggests that the tentative value of *q* and *p* that would be suitable for area under green gram is *q*=0 and *p*=1. Thus the ARIMA model that is found suitable for area under green gram is ARIMA (1,1,0).

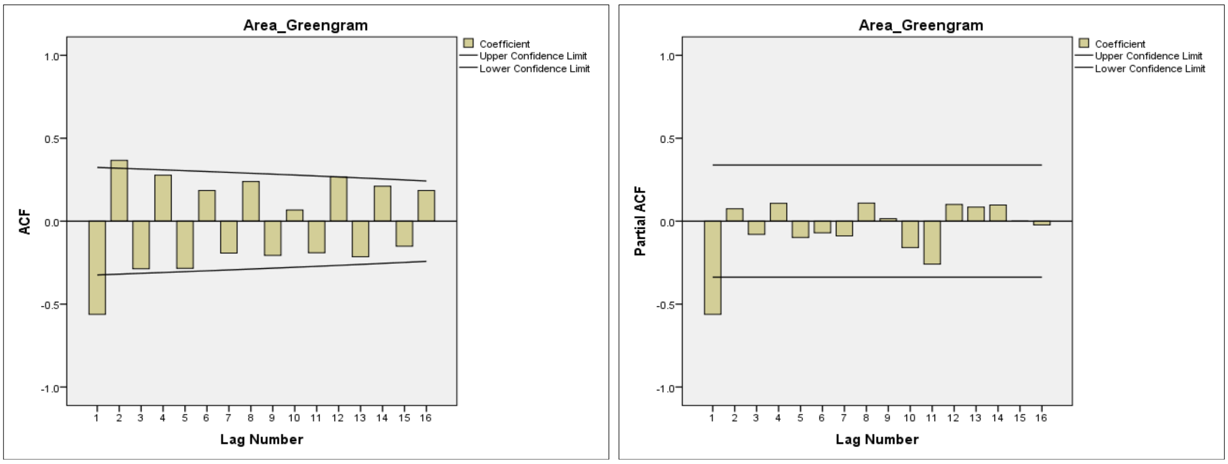

Fig 2: ACF and PACF plot of first difference values of area under green gram

The original plot of data on production of green gram as shown in Fig 3(a) reveals that the data is non-stationary. Thus, the first difference of the data on production is plotted and shown in Fig 3(b). This plot shows that the first difference of data is found to be stationary.

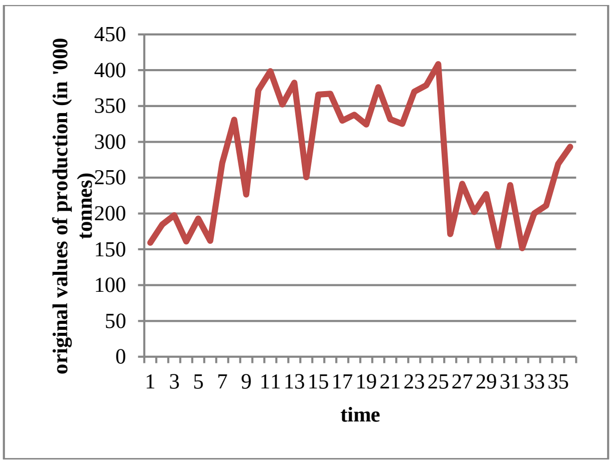

Fig 3(a): Plot of original values of production of green gram vs time

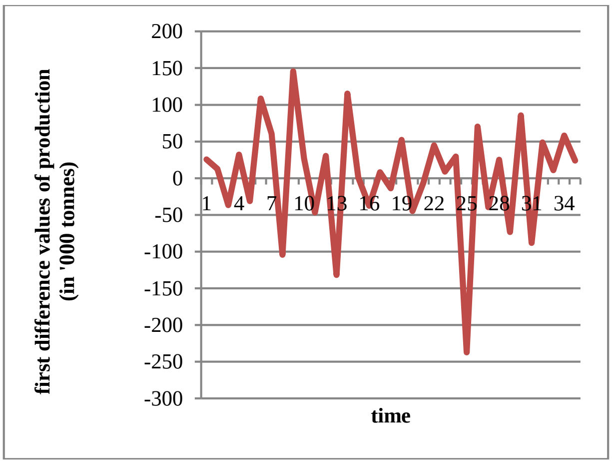

Fig 3(b): Plot of first difference values of area under green gram vs time

The ACF and PACF plot of first difference values of green gram production is shown in Fig. 4., which suggests that the tentative value of *q* and *p* that would be suitable for production of green gram is *q*=0 and *p*=1 and *q*=0 and *p*=2, Thus the ARIMA model that found to be best fitted for production of green gram is ARIMA(2,1,0).

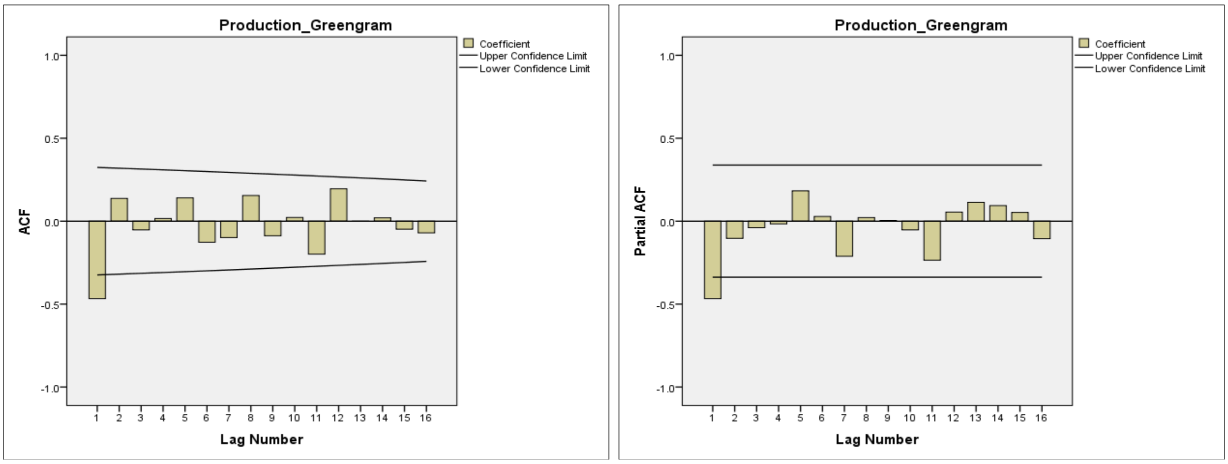

Fig 4: ACF and PACF plot of first difference values of production of green gram

The study of the table 1 shows that when ARIMA (1,1,0) is fitted to the data on area under green gram, the constant is not significant. So, ARIMA (1,1,0) without constant is also fitted. The estimated coefficient of AR (1) is found to be significant. Thus, the selected ARIMA model for area under green gram is ARIMA (1,1,0) without constant. In case of production of green gram the constant and all the estimated coefficient for both the fitted ARIMA model i.e. ARIMA(1,1,0) and ARIMA(2,1,0) are found to be significant. So, both of them are found to be suitable for production of green gram. In case of productivity, when ARIMA (0,1,1) is fitted to the data on productivity of green gram, the constant is not significant. So, ARIMA (0,1,1) without constant is also fitted. The estimated coefficient of MA (1) is found to be significant. Thus, the selected ARIMA model for green gram productivity is ARIMA (0,1,1) without constant.

Table 1: Coefficients of the AR and MA components of the fitted ARIMA model considered for forecasting area and production of green gram 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
Area	110	23.747 (21.448)	-0.554** (0.146)	-	-	-
	110 (without constant)	-	-0.537** (0.145)	-	-	-
Production	110	254.655** (54.816)	0.619** (0.134)	-	--	-
	210	235.561** (68.508)	0.413* (0.165)	0.338* (0.166)	-	-

(Figures in the parentheses indicate the standard error)
* Significant at 5% level of significance; ** Significant at 1% level of significance

The study of table 2 shows that all the fitted model satisfy the assumptions of normality of error as they all have non-significant S-W Statistic and also all the models are found to be adequate due to non-significant Ljung-Box Q Statistic. For area under green gram, the ARIMA (1,1,0) without constant has low value of RMSE and MAPE. So the best fit model is ARIMA (1,1,0) without constant. In case of production of green gram, the ARIMA (2,1,0) with constant has low value of RMSE and MAPE than the ARIMA(1,1,0). So the best fit model is ARIMA (2,1,0) with constant. For productivity of green gram, the ARIMA (0,1,1) without constant has low value of RMSE and MAPE. So the best fit model is ARIMA (0,1,1) without constant.

Table 2: Model Fit Statistics and Residual Diagnostics of the ARIMA models fitted for area, production and productivity of green gram in Odisha

	Model	Model fit statistics		Residual diagnostics	
		RMSE	MAPE	Ljung – Box Q Statistic	Shapiro-Wilk's Statistic
Area	110	90.971	10.852	11.755	0.929
	110 (without constant)	90.472	10.462	10.698	0.928
Production	110	69.045	21.772	14.590	0.912
	210	65.670	19.317	13.705	0.917

(Figures in the parentheses indicate the standard error) (Models highlighted as bold are the best fit models)

The figures 5 and 6 which show the residual ACF and PACF plots of the best fitted ARIMA model for area, production and productivity of green gram respectively confirms the adequacy of the model.

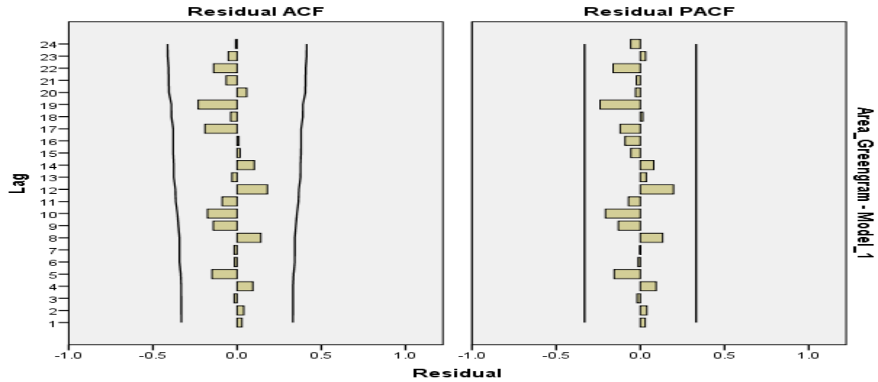

Fig 5: Residual ACF and PACF plot for best fit ARIMA (1,1,0) without constant model for area under green gram

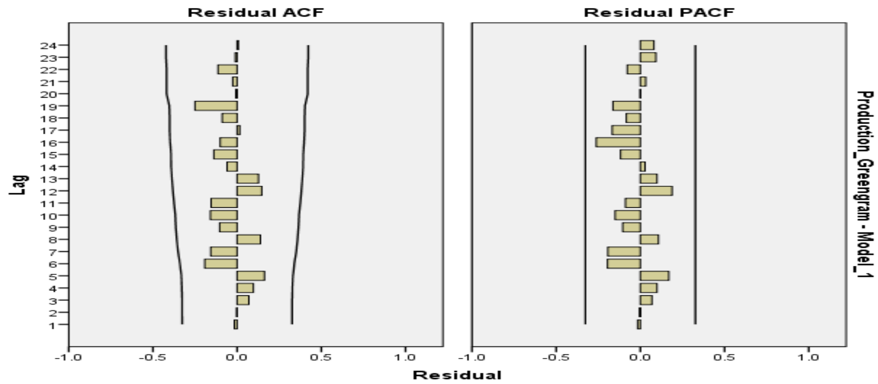

Fig 6: Residual ACF and PACF plot for best fit ARIMA (2,1,0) with constant model for production of green gram

The cross validation of the selected best fit ARIMA(1,1,0) without constant model for area under green gram presented on the table 4.3 shows that the absolute percentage error are quite low, thus the selected model is successfully cross validated.

Table 3: Cross validation of the selected best fit ARIMA (1,1,0) without constant model for area under green gram in Odisha

Year	Actual value (in '000 ha) (Y)	Forecasted value (in '000 ha) $(\hat{Y})$	Error $(Y-\hat{Y})$	Absolute Percentage Error $\left(\frac{\lvert Y-\hat{Y}\rvert}{Y}\right)$ X 100
2007-08	749.1	741.02	8.08	1.078628
2008-09	756.09	758.71	-2.62	0.34652
2009-10	832.18	768.19	63.99	7.689442
2010-11	836.03	782.57	53.46	6.394507
2011-12	799.68	794.81	4.87	0.608994
2012-13	833.11	808.69	24.42	2.931186
2013-14	857.07	822.18	34.89	4.070846
2014-15	869.15	836.37	32.78	3.771501
2015-16	768.89	850.67	81.78	10.63611

The cross validation of the selected best fit ARIMA(2,1,0) with constant model for production of green gram presented on the table 4.4 shows that the absolute percentage error are quite low, thus the selected model is successfully cross validated.

Table 4: Cross validation of the selected best fit ARIMA (2,1,0) model for production of green gram in Odisha

Year	Actual value (in '000 ton) (Y)	Forecasted value (in '000 ton) $(\hat{Y})$	Error $(Y-\hat{Y})$	Absolute Percentage Error $\left(\frac{\lvert Y-\hat{Y}\rvert}{Y}\right)$ X 100
2007-08	306.05	287.6	18.45	6.028427
2008-09	341.79	293.84	47.95	14.0298
2009-10	341.69	294.96	46.73	13.67614
2010-11	362.79	297.94	64.85	17.87535
2011-12	331.43	299.95	31.48	9.498235
2012-13	396.93	302.2	94.73	23.86567
2013-14	407.99	304.22	103.77	25.43445
2014-15	414.48	306.23	108.25	26.11706
2015-16	361.07	308.14	52.93	14.65921

The forecasted values for the area and production of green gram are obtained from respective best fit ARIMA model are presented in table.6 shows that both area and production shows an increase in their forecasted values but there is decrease in the value of productivity from 2016-17 to 2018-19.

Table 5: Forecasted values (with 95% confidence limits) for area and production of green gram in Odisha by using the selected ARIMA model

	Year	Forecasted value	Lower Confidence Limit (95%)	Upper Confidence Limit (95%)
Area (in '000 ha)	2016-17	865.41	466.32	1264.49
	2017-18	880.40	463.69	1297.11
	2018-19	895.75	462.18	1329.31
Production (in'000 tonnes)	2016-17	310.02	131.39	488.65
	2017-18	311.86	132.78	490.94
	2018-19	313.66	134.27	493.04

Figure 7 and 8 shows that the observed values and fit values of area and production of green gram along with their upper and lower limit as obtained from their last best fit ARIMA model.

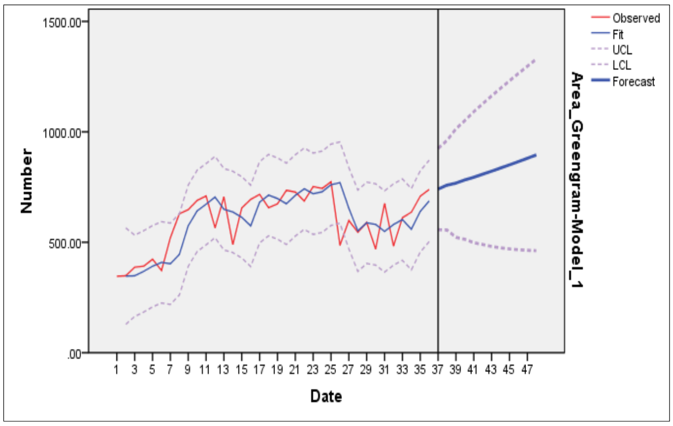

Fig 7: Observed and fit values of area along with upper and lower limit by using best fit ARIMA (1,1,0) without constant model

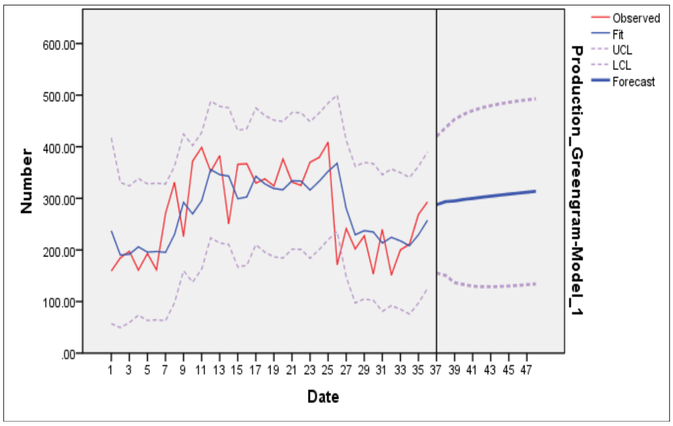

Fig 8: Observed and fit values of production along with upper and lower limit by using best fit ARIMA (2,1,0) with constant model

Conclusion

ARIMA (1,1,0) without constant is selected as the best fitted model for area under green gram due to its low RMSE and MAPE. In case of production of green gram due to the significant constant and also the significant value of AR (1) and AR (2), low MAPE and RMSE, ARIMA (2,1,0) is selected as the best fitted model.

After successful cross validation of the best fitted model on area under the green gram, it is found that the absolute percentage error is lowest (0.34%) during the year 2008-09 and the absolute percentage error is highest (10.63%) during the year 2015-16.Similarly in case of production of green gram after successful cross validation it is found that the absolute percentage error is lowest (6.02%) during the year 2007-08 and the highest absolute percentage error (26.11%) is seen in 2014-15.

After successful forecasting it is found that the area under green gram is expected to be increase for the future year i.e. from 2016-17 to 2018-19. Whereas the production of green gram for the future year is expected to be increase at a slower rate due to the increase in area.

References

1. Box GEP, Jenkins GM. Time series analysis: forecasting and control, second edition, Holden day, 1976.
2. Dash A, Dhakre DS, Bhattacharya D. Forecasting of food grain production in Odisha by fitting appropriate ARIMA model. Journal of Pharmacognosy and Phytochemistry. 2017; 6(6):1126-1132.
3. Ljung GM, Box GEP. On a Measure of a Lack of Fit in Time Series Models. Biometrika, 1978; 65(2):297-303.
4. Odisha Agriculture statistics, Directorate of Agriculture and Food Production, Government of Odisha, Bhubaneswar
5. Sarika Iquebal MA, Chattopadhyay C. Modeling and forecasting of pigeon pea (*Cajanus cajan*) production using autoregressive integrated moving average methodology. Indian Journal of Agricultural Sciences. 2011; 81(6):520-523.