Environment and Ecology 37 (4B) : 1501—1509, October—December 2019
ISSN 0970-0420

Using ARIMA Model to Forecast Mustard Production in Odisha

Madhu Chhanda Kishan, Abhiram Dash

Received 21 August 2019 ; Accepted 12 October 2019 ; Published on 9 November 2019

ABSTRACT

Odisha is one of the agriculture dependent states of India. The oil seeds cover 0.3% of the total cultivated area. This thesis attempted to describe the status of forecasting that is done on one of the important oilseed crops of Odisha i.e. mustard on it's area, yield and production for the future years from 2016-2017 to 2018–2019 by the help of ARIMA (Auto-Regressive Integrated Moving Average) models. The secondary data regarding the oilseeds are collected for the years from 1975-1976 to 2015-2016 from various volumes of Odisha Agricultural Statistics. The data for the year from 1975-1976 to 2006-2007 are used for building of the ARIMA model and for the year from 2007-2008 to 2015-2016 are kept for cross validation of the selected ARIMA model on the basis of Absolute Percentage Error (APE). The different ARIMA models are judged on the basis of Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) at various lags. The possible ARIMA models are identified on the basis of significant coefficient of auto-regressive and moving average components. The best fitted models for different variables under study are selected on the basis of low value of Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). ARIMA (1, 2, 0) without constant is selected as the best-fitted model for the area for mustard having absolute percentage error less than 10% in most cases during cross - validation of the model. For the yield of mustard (0, 1, 1) without constant is selected for best-fit ARIMA model having absolute percentage error ranges from within 10% during cross - validation of the model. By using the forecasted values of area and yield the production valued for future years are calculated. In mustard the production forecast is in accordance with yield forecast.

Keywords ARIMA, Production, Auto-correlation function, Partial auto-correlation function, Forecasting.

INTRODUCTION

India has the largest area and production of oilseeds in the world. Out of the major oil seed crops mustard stands out as an important commercial oil seed crop in India. India is the fourth largest mustard producer in the world, 11% of world's total production. This crop accounts nearly one-third of the oil production in India. In India most of the farmers are small and marginal and mustard being a rain-fed crop adds

Madhu Chhanda Kishan
PG Scholar in Department of Agricultural, Statistics College of Agriculture, Bhubaneswar (OUAT) 751003, Odisha, India

Abhiram Dash*
Assistant Professor (Agricultural Statistics), College of Agriculture, Chiplima (OUAT) 751003, Odisha, India
email: abhidash2stat@gmail.com
*Corresponding author

security to the livelihood of the farmers. Mustard is a multi-purpose crop and it's grown with different pattern all over the country. Beside the oil-value of the crop it's seeds are used as condiments in preparation of pickle and flavorings food items. The oil is mostly used for human consumption throughout the country for cooking as well as frying. The leaves of the young plants are also used as green leafy vegetable as they adds sulfur minerals in the diet. The oil cakes are used as manure and cattle feed. Increase trend has been experienced by the production of India due to increasing used of mustard and there is an urgent need for undertaking the basic and strategic research for stabilizing and increasing the mustard status.

Odisha is one of the agriculture dependent states of India. The oil seeds covers the total area of 6.30 lakh ha from the net sown area of 54.24 lakh ha and the grossed crop area is 90.54 lakh ha according to the 2015-16 data. It's 0.3% of the total area. Odisha has total oilseed yield rate of 928 kg/ha which is less than the total oilseed yield of India i.e. 1153 kg/ha according to the records of 2015-2016. That implies there is enough room for oilseed production development in Odisha. Mustard is one of the important oilseed crop grown in Odisha. Mustard contributes the total area, yield and production of 1.45 lakh ha, 424 kg/ha and 6.16 lakh tones.

Forecasting plays a pivotal role in agriculture. It is an important and necessary aid when it comes to crop planning and planning is the backbone of effective operations. Forecasting in agriculture sector comprises of forecasting of production/yield/area of the crop, also the forewarning of incidence of pest and diseases to the crop. The forecasting based on the time series data are prime importance in policy forming decisions on agricultural activities.

ARIMA stands for Auto-Regressive Integrated Moving Average Model, one of the time-series data forecasting statistical model. ARIMA model is also known as Box-Jenkins model. The main application focuses on the area of short term forecasting which requires atleast 40 historical data points. It works best when the data exihibits a stable pattern over time with a minimum amount of outliers.

Objectives of the study were as follows: To select and fit possible ARIMA models to data on area and yield of mustard for the year 1975–1976 to 2006–2007, To select the best fit ARIMA models on basis of significance of coefficients, model diagnosis test and model selection criteria, To cross-validate the selected best fit model by using the available data for the year 2007-2008 to 2015-2016, To use the best fit model for forecasting the area and yield of mustard for the years 2016-2017, 2017-2018, 2018-2019 after cross validation of the selected model, To forecast the production of mustard by using the forecast of area and yield.

MATERIALS AND METHODS

The study period consists of 41 years of data from the year 1976-1977 to 2015-2016. The data collected for the above time period covers the area and yield of mustard.

Secondary data relating to the area and yield of mustard of Odisha for the period from 1976-1977 to 2015-2016 has been collected from Odisha Agricultural Statistics published by the Directorate Agriculture and Food Production, Government of Odisha. The area and yield are expressed in '000 ha and kg/ha respectively (1 ha = 10000 m).

ARIMA is a statistical analysis model that uses the time series data to forecast the future trends. It retains a form of regression analysis seeking to predict future movements and the random walks by examining the differences between values in the series instead of using the actual data values. The differenced series have lags referred as auto-regressive and forecasted data lags are referred as moving average.

This model is represented as ARIMA (p, d, q), where p represents order of auto-regression shows degree of differencing, q shows the order of moving

average. Trends, seasonality, cycles, errors and non-stationary aspects of data set while making the forecasts are taken accounted by the ARIMA modelling.

Fitting of the Box-Jenkins ARIMA model

The Auto-Regressive Moving Average (ARMA) models was introduced to overcome the difficulty in describing the dynamic structure of the data by fitting Auto-Regressive (AR) and Moving Average (MA) models. Auto-Regressive Integrated Moving Average (ARIMA) models are the ARMA models that includes the order of differencing (which is done to stationaries the data). The ARIMA model with parameter (*p, d, q*) is fitted by univariate Box-Jenkins techniques (Box and Jenkins 1976). This model includes Auto-Regressive of order *p,* differencing to make stationary series of degree *d* and moving average of order *q*.

Test for stationarity

If the time series data have constant mean and variance over time then it is stationary. After the original data are plotted, it is verified for stationarity, if the data are non-stationary from the graph, then the first difference of the data are plotted and checked for stationarity. This process is repeated till the data becomes stationary. The maximum order of differencing (*d*) is usually 2.

To determine the order of AR (i. e. *p*) and MA (i.e. *q*)

By examining the plots of the auto-correlation and partial auto-correlation of the stationaries values of the variables the value of *p* (order of auto-regression) and *q* (order of moving average). The auto-correlation 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 auto-correlations of *y* at lag *k* is the coefficient of *y*–lag lag *k* in a regression of *y* on *y* –lag 1, *y*–lag 2....., up to *y*–lag *k*. Thus, the partial auto-correlation of *y* at lag 1 and the auto-correlations of *y* at lag 1 are same. The partial auto-correlation of *y* on *y*–lag 1 and *y* – lag 2 and so on. The amount of correlation between *y* and *y*–lag *k* is the way to interpret the partial auto-correlation at lag *k*, it is not explained by the lower-order auto-correlation.

To determine the *p* and *q* from the plots of ACF and PACF the rules are

If the ACF plots cuts off sharply at lag *k* (i.e., if the auto-correlation is significantly different from zero at lag *k* and extremely low in significance at the next higher lag and the ones that follow), while there is a more gradual decay in the PACF plot (i.e., if the drop off in significance beyond lag *k* is more gradual), then set *q*-k and *p*=0. This is a so-called MA (*q*) signature. On the other side, when there is a more gradual decay in the ACF plot and the PACF plot cuts off sharply at lag *k*, *p*=*k* and *q*=0 are set. *p*=1 and *q*=0 is set when there is a single spike at lag 1 in both the ACF and PACF plots. If it is positive (this is an AR (1) signature) and set *p*=0 and *q*=1 if it is negative (this is a MA (1) signature).

The highest order AR or MA coefficient should be significant after they are identified by all the correct form of the model. If the highest order coefficient is not significant but the ACF and PACF plot look good then the value of *p* and *q* should be reduced by 1, as the case may be. When there are some significant residual auto-corrections or partial auto-correlations at the few lag, the rules that to be followed are: There is an increase in *q* by 1 when there is a spike at a low-order lag in the residual ACF plot and the model is refitted. Conversely, there is an increase in *q* by 1 when there is a spike at a low-order lag in the PACF plot and the model is refitted.

By using Box-L jung test the adequacy of the selected model is checked. A formal test of the fitness of the model is also done by using Box-L jung test of the residuals (L jung and Box 1978) is done in following manner:

Null hypothesis: H_0 : The errors are distributed randomly.

Alternate hypothesis: H_1 : The errors are non-random.

The Box -L jung test statistic,

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

Where, n is the number of observations,

r_k is the estimated auto-correlation of the series at lag k =1, 2,...., m.

m=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, p = Number of AR,
Q = Number of MA.

The degree of freedom, h= (m–p–q) (Dash et al. 2017). By the help of forecasting tool of SPSS 20.0. The Box-L jung test is done. The model fit statistics used to select best fit model are : Root Mean Square Error (RMSE).

$$RMSE = \frac{\Sigma_t e^2_t}{n-2}$$

Mean absolute percentage error (MAPE)

$$MAPE = \frac{\Sigma_t \left| Y_t - \hat{Y}_t \right| / y_t}{n} * 100$$

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

Cross-validation of the selected model

The cross-validation of the selected model is worked on the 20% of the data that is not used for model building at the end period. For the cross-validation of the model the actual value of the left out period and the forecasted value of the left out period of the selected model are used. Here the data from 1976-77 to 2006-07 are used for model building and data from 2007-08 to 2015-16 are used for cross-validation.

The percentage error is calculated as follows:

$$\% \text{ of forecasting error} = \left[\frac{Y - \hat{Y}}{Y} \right] \times 100$$

Where, Y = observed value of remaining 8 years,

$\hat{Y}$ = The forecasted value of remaining 8 years.

The selected best-fit ARIMA model is used for forecasting after the cross-validation. ARIMA techniques are generally used in case of short term forecasting because the prediction for longer periods will have more errors associated with it. So, ARIMA should be used for short term forecasting (Biswas et al. 2014, Debnath et al. 2013).

RESULTS AND DISCUSSION

Data are fitted to the data by the help of ARIMA model on area and yield of mustard for forecasting. Basically the data used for model building is from the year 1976-1977 to 2006-2007. The data from 2007-2008 to 2015-2016 is used for the cross-validation of the selected model. By using the best fit model the forecasting is done for the years 2016-2017, 2017-2018 and 2018-2019.

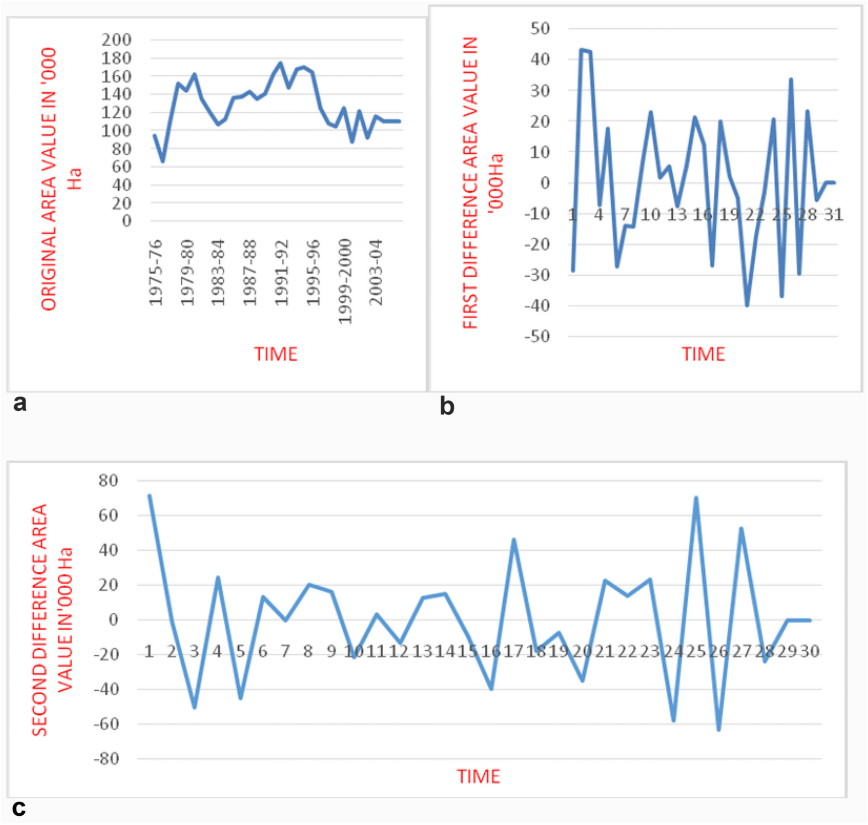

Fig. 1 (a). Plot of original value of area of mustard vs time. **Fig. 1 (b).** Plot of first difference area value of mustard vs time. **Fig. 1 (c).** Plot of second difference area value of area of mustard vs time.

Forecasting of area, yield and production of mustard by fitting appropriate ARIMA model

The original plot of data on area under mustard as shown in Fig. 1 (a) explains that the data is non-stationary that says it don't have constant mean and variance. Thus, the second difference of the data are plotted after plotting the first difference Fig. 1 (b) of the data which were not stationary and shown in Fig. 1 (c) this plot shows the second difference of data are found to be stationary which have constant

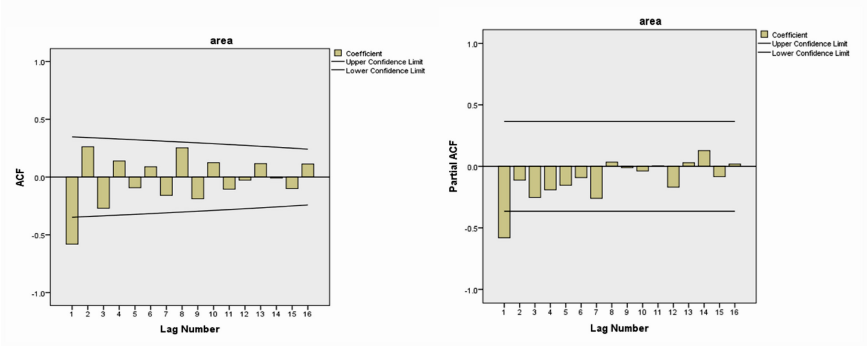

Fig. 2. ACF and PACF plot of first difference values of area of mustard.

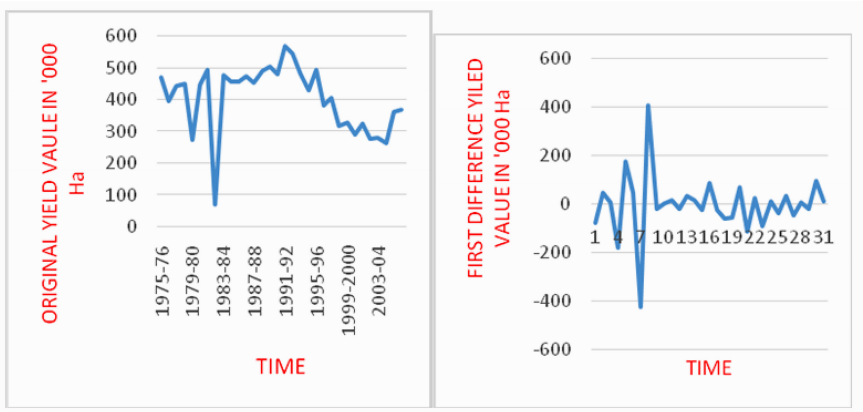

Fig. 3 (a). Plot of original value of yield of mustard vs time. **Fig. 3 (b).** Plot first difference yield value of mustard vs time.

mean and variance.

The ACF and PACF plots of the second difference value of mustard area is shown in the Fig. 2. Which shows that the provisional value of q and p that would be satisfactory of mustard are $q = 0$ and $p = 1$. Thus the ARIMA model found fitted for area of mustard is ARIMA (1, 2, 0).

The original plot of data on yield under mustard as shown in Fig. 3 (a) explains that the data are non-stationary that says it don't have constant mean and variance. Thus, the first difference of the data are plotted and shown in Fig. 3 (b), this plot shows the first difference of data are found to be stationary which have constant mean and variance.

The ACF and PACF plots of the first difference value of mustard yield is shown in the Fig. 4. Which shows that the provisional value of q and p that would be satisfactory of mustard are $q = 1$ and $p = 0$. Thus the ARIMA model found fitted for yield of mustard is ARIMA (0, 1, 1).

The study of Table 1 shows that when ARIMA (1, 2, 0) is fitted to the data on area under mustard, the

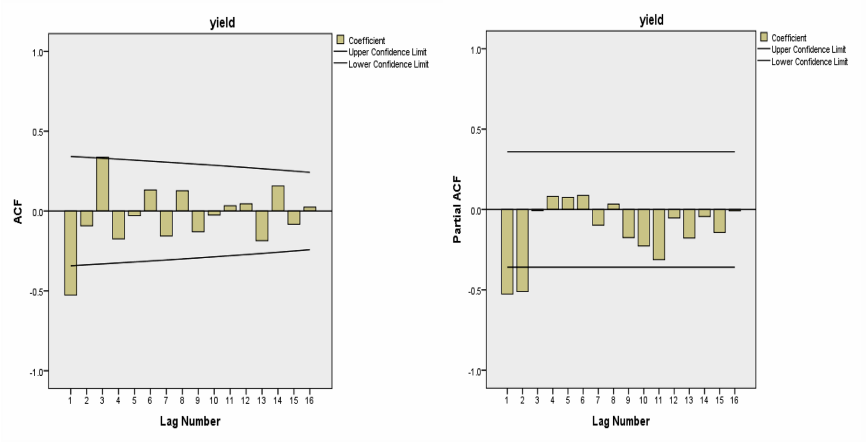

Fig. 4. ACF and PACF plot of first difference values of yield of mustard.

Table 1. Coefficient of AR and MA components of the fitted ARIMA model considered for forecasting area and yield of mustard in Odisha. Figures in the parentheses indicate the standard error. *Significant at 5% level of significance, **Significant at 1% level of significance.

	Best fit ARIMA model	Constant (μ)	Coefficient of auto-regressive components α_1	α_2	Coefficient of moving average components α_1
Area	(1, 2, 0)	0.014 (0.047)	– 0.719** (0.113)	–	– 0.719** (0.113)
	(1, 2, 0) without constant	–	– 0.722** (0.111)	–	– 0.722** (0.111)
Yield	(0, 1, 1)	– 0.009 (0.033)	–	–	–
	(0, 1, 1) without constant	–	–	–	–

Table 1. Continued.

	Best fit ARIMA model	Constant (μ)	Coefficient of auto-regressive components α_1	α_2	Coefficient of moving average components θ_1	θ_2
Area	(1, 2, 0)	0.014 (0.047)	– 0.719** (0.113)	–	–	–
	(1, 2, 0) without constant	–	– 0.722** (0.111)	–	–	–
Yield	(0, 1, 1)	– 0.009 (0.033)	–	–	0.801** (0.110)	–
	(0, 1, 1) without constant	–	–	–	0.809** (0.108)	–

constant is not significant. So ARIMA (1, 2, 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 mustard is ARIMA (1, 2, 0) without constant. In case of yield of mustard ARIMA (0, 1, 1) is fitted to the data, the constant is

Table 2. Model fit statistics and residual diagnostics of the ARIMA models fitted for area and yield of mustard in Odisha. Models highlighted as bold are the best fit models.

	Model	Model fit statistics RMSE	MAPE	Residual diagnostics Ljung–Box Q statistic	Shapiro-Wilk's statistic
Area	120	31.026	17.190	9.687	0.921
	120 (without constant)	30.032	17.028	9.582	0.914
Yield	011	105.224	30.515	12.302	0.923
	011 (without constant)	104.254	30.855	12.818	0.917

Table 3. Cross validation of the selected best fit ARIMA (1, 2, 0) without constant model for area of mustard in Odisha.

Year	Actual value (in '000 ha) (Y)	Forecasted value (in '000 ha)	Error	Absolute percentage error
2007-08	110.29	110.87	−0.58	0.5258863
2008-09	109.93	110.78	−0.85	0.773219321
2009-10	112.19	110.16	2.03	1.809430431
2010-11	112.45	112.84	−0.39	0.346820809
2011-12	126.67	114.4	12.27	9.686587195
2012-13	116.37	131.32	−14.95	12.84695368
2013-14	145.36	123.94	21.42	14.73582829
2014-15	121.98	145.46	−23.48	19.24905722
2015-16	99.69	136.33	−36.64	36.75393721

Table 4. Cross validation of the selected best fit ARIMA (0, 1, 1) without constant model for yield of mustard in Odisha.

Year	Actual value (in kg/ ha) (Y)	Forecasted value (in kg / ha)	Error	Absolute percentage error
2007-08	375.1	358.38	16.72	4.457478006
2008-09	382.97	366.79	16.18	4.224874011
2009-10	369.37	375.26	−5.89	1.594607034
2010-11	375.1	379.63	−4.53	1.207677953
2011-12	415.73	384.36	31.37	7.545762875
2012-13	422.02	395.97	26.05	6.172693237
2013-14	423.98	406.81	17.17	4.049719326
2014-15	424	416.19	7.81	1.841981132
2015-16	421.01	423.96	−2.95	0.700695945

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 mustard yield is ARIMA (0, 1, 1).

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 L jung-Box Q statistic. For area under mustard ARIMA (1, 2, 0) without constant has low value of RMSE and MAPE, so the best fit model is ARIMA (1, 2, 0) without constant. For yield of mustard ARIMA (0, 1, 1) without constant has low value of RMSE and MAPE, so the bestfit model ARIMA (0, 1, 1) without constant.

The cross validation of the selected best fit ARIMA (1, 2, 0) without constant model for area under mustard presented on the Table 3 shows that the absolute percentage error are quite low, thus the selected model is successfully cross validated.

The cross validation of the selected best fit ARIMA (0, 1, 1) without constant model for yield under mustard presented on the Table 4 shows that the absolute percentage error are quite low, thus the selected model is successfully cross validated.

In Table 5, the forecasted values for area and yield of mustard are obtained from the respective best fit ARIMA model. It shows that there is a decrease in the forecasted values of area and yield from 2016-2017 to 2018-2019.

Figures 5 and 6 that the observed values and the fit values of area and yield of mustard along with their upper and lower limit as obtained from their last best fit ARIMA model.

In Table 6 by using the forecasted values for area and yield of mustard are obtained from the respective best fit ARIMA model the production is calculated. It shows that there is a decrease in the calculated values of production from 2016-2017 to 2018-2019.

CONCLUSION

ARIMA (1, 2, 0) and ARIMA (1, 2, 0) without constant are selected model for the forecasting of area under mustard. The constant is not significant in ARIMA (1, 2, 0) model, so ARIMA (1, 2, 0) without constant is fitted. It is found that AR (1) is significant. Due to low value of RMSE and MAPE, ARIMA (1, 2, 0) without constant is selected for the best-fit model. In case of yield of mustard ARIMA (0, 1, 1) and ARIMA (0, 1, 1) without constant are selected

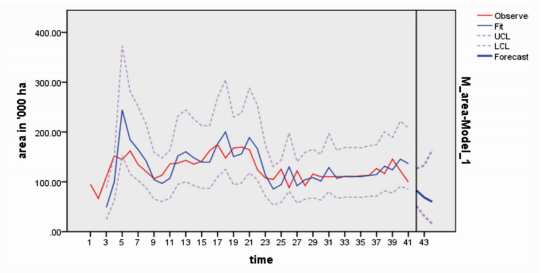

Fig. 5. Observed and fit values of area along with upper and lower limit by using best fit ARIMA (1, 2, 0) without constant model of mustard in Odisha.

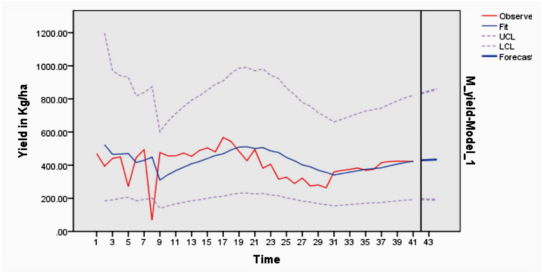

Fig. 6. Observed and fit value of yield along with upper and lower limit by using best fit ARIMA (0, 1, 1) without constant of mustard in Odisha.

model for forecasting. As the constant is found to be not significant in ARIMA (0, 1, 1), ARIMA (0, 1, 1) without constant is fitted. It is found that MA (1) is significant. Due to the low value of RMSE and MAPE, ARIMA (0, 1, 1) without constant is selected for best-fit model. After the successful cross validation of best fitted model on area under the area of mustard, it is found that the lowest absolute error percentage (0.35%) is seen during the year 2014-2015. Similarly in case of the yield of mustard after successful cross validation it is found that the lowest absolute percentage error (0.70%) seen in 2015-16 and the absolute percentage error (7.54%) is seen in 2011-12. After successful forecast of area and yield, the predicted production value is calculated and it is found that the forecasted values of area of mustard is found to decrease over the future years. The forecasted values of yield is found to increase over the future years. The forecasted values of production is found to decrease over the future period. This shows that the production forecast is in accordance with yield forecast in case of mustard.

Table 5. Forecasted values (with 95% confidence limits) for area and yield of mustard 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	82.94	51.80	126.52
	2017-18	68.59	31.17	132.81
	2018-19	59.87	15.58	164.11
Yield (kg/ha)	2016-17	429.79	195.03	832.88
	2017-18	432.00	193.01	846.26
	2018-19	434.24	192.06	859.72

Table 6. Value of production forecast by using the forecasted value of area and yield of mustard in Odisha.

	Year	Calculated value
Production in '000 tonnes	2016-17	34.65
	2017-18	29.63
	2018-19	26.00

REFERENCES

Biswas B, Dhaliwal LK, Singh SP, Sandhu SK (2014) Forecasting wheat production using ARIMA model in Punjab. Int J Agric Sci 10 : 1.

Dash A, Dhakre DS, Bhattacharya D (2017) Forecasting of food grain production in Odisha by fitting ARIMA model. J Pharmac Phytochem 6 (6) : 1126—1132.

Debnath MK, Bera K, Mishra P (2013) Forecasting area, production and yield of cotton in India using ARIMA model. STM J JoPC 2013 : 16—20.