



Application of ARIMA model for forecasting Paddy production in Bastar division of Chhattisgarh

D.P. Singh*, Prafull Kumar and K. Prabakaran***

* Assistant Professor (Agricultural Statistics), S.G. College of Agriculture & Research Station, Jagdalpur (C.G.)

** Assistant Professor (Plant Breeding & Genetics), S.G. College of Agriculture & Research Station, Jagdalpur (C.G.), INDIA.

*** Assistant Professor (Agricultural Statistics), Department of Agricultural Economics, Agricultural College & Research Institute, Madurai, INDIA.

Abstract: Paddy area and production in Bastar division of Chhattisgarh data for the period of 1974-75 to 2010-2011 were analyzed by time series methods. Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) were calculated for the data. Appropriate Box-Jenkins Auto Regressive Integrated Moving Average (ARIMA) model was fitted. Validity of the model was tested using standard statistical techniques. ARIMA (2, 1, 2) and ARIMA (2, 1, 0) model were used to forecast paddy area and production in Bastar division of Chhattisgarh for four leading years. The results also shows paddy area forecast for the year 2015 to be about 598.22 thousand hectare with upper and lower limit 665.39 and 531.06 thousand hectares respectively. The model also shows paddy production forecast for the year 2015 to be about 1126.61 thousand tones with upper and lower limit 1430.16 and 823.05 thousand tonnes respectively.

Keywords: - Forecasting, Area, Production, Paddy, ARIMA, ACF, PACF.

I. INTRODUCTION

Paddy is one of the chief grains of India. Moreover, this country has the biggest area under rice cultivation, as it is one of the principal food crops. It is in fact the dominant crop of the country. India is one of the leading producers of this crop. Paddy is the basic food crop and being a tropical plant, it flourishes comfortably in hot and humid climate. Rice is mainly grown in rain fed areas that receive heavy annual rainfall. That is why it is fundamentally a *Kharif* crop in India. It demands temperature of around 25 degree Celsius and above and rainfall of more than 100 cm. Rice is also grown through irrigation in those areas that receives comparatively less rainfall.

The paddy production in Chhattisgarh in the *Kharif* season 2012 recorded a new height. Chhattisgarh Known as rice-bowl of the country, the paddy production in Chhattisgarh touched the figure of 11.7 million tonne (MT). The state government had set the target to produce 9 MT of paddy last year. The final result was however surprisingly for them with about 25 per cent higher production.

The objective of the study is to develop appropriate ARIMA models for the time series of paddy area and production in Bastar division of Chhattisgarh and to make four year forecasts with appropriate prediction interval.

II. MATERIALS AND METHODS

The annual data on paddy cultivated area and production for the period from 1974-75 to 2010-11, Department of Agriculture, Govt. of Chhattisgarh, were used for forecasting the future values using ARIMA models. Originally ARIMA models have been studied extensively by George Box and Gwilym Jenkins during 1968 and their names have frequently been used synonymously with general ARIMA process applied to time series analysis, forecasting and control. The ARIMA methodology is also called as Box-Jenkins methodology. The Box-Jenkins procedure is concerned with fitting a mixed Auto Regressive Integrated Moving Average (ARIMA) model to a given set of data. The main objective in fitting this ARIMA model is to identify the stochastic process of the time series and predict the future values accurately. These methods have also been useful in many types of situation which involve the building of models for discrete time series and dynamic systems. But, this method was not good for lead times or for seasonal series with a large random component (Granger and Newbold, 1970).

However, the optimal forecast of future values of a time-series are determined by the stochastic model for that series. A stochastic process is either stationary or non-stationary. The first thing to note is that most time series are non-stationary and the ARIMA model refer only to a stationary time series. Since the ARIMA models refer only to a stationary time series, the first stage of Box-Jenkins model is reducing non-stationary series to a stationary series by taking first order differences.

The main stages in setting up a Box-Jenkins forecasting model are as follows.

1. Identification
2. Estimating the parameters
3. Diagnostic checking and
4. Forecasting

Step 1: Identification: Appropriate values of p , d and q are found first. The tools used for identification are the Autocorrelation Function (ACF), the Partial Autocorrelation Function (PACF) and the resulting correlograms and partial correlograms.

The general ARIMA (p, d, q) model is presented in simple form as

$$\varphi(B) \nabla^d X_t = \theta(B) U_t$$

whereas B is the backshift operator defined by

$$B^m X_t = X_{t-m} \quad (m = 0, 1, 2, \dots, p)$$

$\varphi(B)$ is autoregressive operator of order ' p ' defined by

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p$$

∇ is the backward difference operator defined by

$$\nabla X_t = X_t - X_{t-1} = (1-B) X_t$$

∇^d means the d^{th} difference of the series values X_t , $\theta(B)$ is the moving average operator of order ' q ' defined by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$

U_t is white noise process having a normal probability distribution with mean zero and variance σ^2 .

An example of ARIMA model is given below to clarify the general representation of the ARIMA (1,1,1) in explaining some features of the general ARIMA (p, d, q) model.

As could be seen in ARIMA(1,1,1) model where $p=1, d=1, q=1$

$$\varphi(B) = 1 - \varphi_1 B$$

$$\nabla^1 = (1-B)^1 = 1 - B \text{ and}$$

$$\theta(B) = 1 - \theta_1 B^1 = 1 - \theta_1 B$$

Thus the model becomes

$$(1-\varphi_1 B)(1-B) X_t = (1-\theta_1 B) U_t$$

$$\text{i.e. } (1-\varphi_1 B^1)(X_t - X_{t-1}) = U_t - \theta_1 U_{t-1}$$

$$\text{i.e. } X_t - \varphi_1 X_{t-1} - X_{t-1} + \varphi_1 X_{t-2} = U_t - \theta_1 U_{t-1}$$

$$\text{i.e. } (X_t - X_{t-1}) - \varphi_1 (X_{t-1} - X_{t-2}) = U_t - \theta_1 U_{t-1}$$

$$\text{i.e. } W_t - \varphi_1 W_{t-1} = U_t - \theta_1 U_{t-1}$$

φ_1 and θ_1 are the parameters of the model ARIMA (1, 1, 1).

Similarly φ_i ($i = 1, 2, \dots, p$) and θ_j ($j = 1, 2, \dots, q$) are the parameters of the general ARIMA (p, d, q) model. φ_i ($i = 1, 2, \dots, p$) are the Autoregressive (AR) parameters and θ_j ($j = 1, 2, \dots, q$) are the moving average (MA) parameters

The Autoregressive (AR) operator $\varphi(B)$ is assumed to be stationary and $\theta(B)$, the moving average (MA) operator is assumed to be invertible, $\{a_t\}$ is sequence of independent and identically distributed random variables with mean zero and variance σ^2 , and the ARIMA(p, d, q) process becomes ARMA (p, q) process by suitable transformation of the variables and is given by $\varphi(B) w_t = \theta(B) a_t$

$$\text{i.e., } w_t = \sum_{i=1}^p \varphi_i w_{t-i} - \sum_{j=1}^q \theta_j a_{t-j} + a_t$$

where, w_t consists of $(n - d)$ observations

Step 2: Estimation: Having identified p and q values estimation of parameters of the autoregressive and moving average terms are estimated using simple least squares. The least square criterion for AR(1) is furnished below.

$$(1-\varphi_1 B) X_t = U_t$$

$$(1-\varphi_1 B) (X_t - \mu) = U_t \quad (\text{Since } (X_t = X_t - \mu))$$

$$X_t - \varphi_1 X_{t-1} - \mu + \varphi_1 \mu = U_t$$

$$X_t = \varphi_1 X_{t-1} + \mu - \varphi_1 \mu + U_t$$

$$X_t = \mu (1 - \varphi_1) + \varphi_1 X_{t-1} + U_t$$

Where $\mu (1 - \varphi_1)$ is the constant term

Step 3: Diagnostic checking: Having chosen a particular ARIMA model and having estimated its parameters the fitness of the model is verified. One simple test is to see if the residuals estimated from the model are white noise, if not we must start with other ARIMA model. The residuals were analyzed using Box-Ljung Statistic.

Step 4: Forecasting: One of the reasons for the popularity of the ARIMA modeling is its success in forecasting. In many cases, the forecasts obtained by this method are more reliable than those obtained from the traditional

econometric modeling, particularly for short-term forecasts. An Autoregressive Integrated Moving Average Process model is a way of describing how a time series variable is related to its own past value. Mainly an ARIMA model is used to produce the best weighted average forecasts for a single time series (Rahulamin and Razzaque 2000). The accuracy of forecasts for both Ex-ante and Ex-post were tested using the following tests (Markidakis and Hibbon, 1979) such as Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE).

III. RESULTS AND DISCUSSION

In this study, we used the data for paddy cultivated areas and production for the period 1974-75 to 2010-11. As we have earlier stated that development of ARIMA model for any variable involves four steps: Identification, Estimation, Verification and Forecasting. Each of these four steps is now explained for paddy cultivated areas and production as follows.

Model Identification

For forecasting paddy area and production, ARIMA model estimated only after transforming the variable under forecasting into a stationary series. The stationary series is the one whose values vary over time only around a constant mean and constant variance. There are several ways to ascertain this. The most common method is to check stationarity through examining the graph or time plot of the data is non-stationary. Non-stationarity in mean is connected through appropriate differencing of the data. In this case difference of order 1 was sufficient to achieve stationarity in mean.

The newly constructed variable X_t can now be examined for stationarity. The graph of X_t was stationary in mean. The next step is to identify the values of p and q . For this, the autocorrelation and partial auto correlation coefficients of various orders of X_t are computed (Table 1). The Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) (Fig 1 and Fig 2) show that the order of p and q can at most be 1. We entertained three tentative ARIMA models and chose that model which has minimum AIC (Akaike Information Criterion) and SBC (Schwartz Bayesian Criterion). The models and corresponding AIC and SBC values are

	ARIMA (p, d, q)	AIC	SBC
1. Paddy Area	210	339.12	344.00
	211	341.21	347.54
	212	336.55	344.00
2. Paddy Production	210	451.67	456.42
	211	453.40	459.74
	212	455.16	463.08

So the most suitable model is ARIMA (2,1,2) for Paddy area and ARIMA (2, 1, 0) for Paddy production has the lowest AIC and SBC values.

Model Estimation and Verification

Paddy cultivated areas and production model parameters were estimated using SPSS package. Results of estimation are reported in Table 2 and Table 3. The model verification is concerned with checking the residual of the model to see if they contain any systematic pattern which still can be removed to improve on the chosen ARIMA. This is done through examining the auto correlations and partial auto correlations of the residuals of various orders. For this purpose, the various correlations up to 24 lags were computed and the same along with their significance which is tested by Box-Ljung test provided in Table 4 and Table 5. As the results indicated majority of correlations are not significantly different from zero at a reasonable level for paddy area and none of these correlations significantly different from zero at a reasonable level for paddy production also. This proves that the selected ARIMA model is an appropriate model. The ACF and PACF of the residual (Fig 3 and Fig 4) also indicate "good fit" of the model.

Forecasting with ARIMA model

ARIMA models are developed basically to forecast the corresponding variable. To judge the forecasting ability of the fitted ARIMA model important measure of the sample period forecasts accuracy was computed. The Mean Absolute Percentage Error (MAPE) for paddy cultivated area turns out to be 2.097 and paddy production turns out to be 16.32. This measure indicates that the forecasting inaccuracy is low. The forecasts for paddy area and production during 2011-12 to 2014-15 showing increasing trend are given in Table 6.

IV. CONCLUSIONS

In our study the developed model for paddy cultivated areas and production was found to be ARIMA (2, 1, 2) and ARIMA (2, 1, 0) respectively. From the forecast available by using the developed model, it can be seen that forecasted paddy cultivated areas and production increases the next four years. The validity of the forecasted value can be checked when the data for the lead periods become available. The model can be used by researchers for forecasting paddy cultivated areas and production in Bastar division of Chhattisgarh. However, it should be updated from time to time with incorporation of current data.

References

- [1] Box, G.E.P. and G.M. Jenkins, 1976 Time Series of Analysis, Forecasting and Control, San Francisco, Holden-Day, California USA.
- [2] Ganger, C.W.J. and Newbold, P., 1970, Special methods. *Econometrica* 30: 424-438.
- [3] Makridakis, S. and Hibon, M., 1979, Accuracy of forecasting: An empirical investigation. *J. Roy. Statist. Soc. A* 41(2): 97-145.
- [4] Rahulamin, M.D. and Razaque, M.A., "Autoregressive Integrated Moving Average Modeling for Monthly Potato Prices in Bangladesh", *Journal of Financial Management and Analysis*, 13(1): 74-80, 2000.
- [5] Pindyck, S.R. and Rubinfeld, D.L., *Econometric Models and Economic Forecasts*, New York: McGraw Hill, Inc, 1991.

Tables

Table1. Auto Correlations and Partial Auto Correlations for rice area and production

Paddy area					Paddy production				
Lags	ACF	SE	PACF	SE	Lags	ACF	SE	PACF	SE
1	-0.368	0.160	-0.368	0.167	1	-0.734	0.160	-0.734	0.167
2	-0.218	0.158	-0.409	0.167	2	0.305	0.158	-0.504	0.167
3	0.358	0.155	0.130	0.167	3	0.053	0.155	0.057	0.167
4	-0.084	0.153	0.079	0.167	4	-0.309	0.153	-0.220	0.167
5	0.082	0.151	0.313	0.167	5	0.366	0.151	-0.123	0.167
6	-0.068	0.148	0.019	0.167	6	-0.178	0.148	0.241	0.167
7	-0.027	0.146	-0.250	0.167	7	-0.052	0.146	0.073	0.167
8	0.04	0.143	-0.189	0.167	8	0.145	0.143	-0.166	0.167
9	0.005	0.140	-0.047	0.167	9	-0.185	0.140	-0.165	0.167
10	-0.009	0.138	-0.019	0.167	10	0.146	0.138	-0.049	0.167
11	0.066	0.135	0.224	0.167	11	-0.021	0.135	-0.008	0.167
12	-0.068	0.132	0.063	0.167	12	-0.098	0.132	-0.082	0.167
13	0.023	0.130	0.070	0.167	13	0.139	0.130	0.026	0.167
14	0.064	0.127	-0.071	0.167	14	-0.139	0.127	0.005	0.167
15	0.093	0.124	-0.128	0.167	15	0.082	0.124	-0.097	0.167
16	0.023	0.121	0.139	0.167	16	0.068	0.121	0.121	0.167

Table 2. Estimates of the fitted ARIMA model for rice area

Model Fit statistics								Box- Ljung Q	
St. R-sq.	R-sq.	RMSE	MAPE	MAE	Max APE	Max AE	Normalized BIC	Statistics	Sig.
0.457	0.861	23.453	2.097	12.60	11.788	85.38	6.808	4.149	0.995

Table 3. Estimates of the fitted ARIMA model for rice production

Model Fit statistics								Box- Ljung Q	
St. R-sq.	R-sq.	RMSE	MAPE	MAE	Max APE	Max AE	Normalized BIC	Statistics	Sig.
0.645	0.757	121.74	16.322	84.257	86.107	306.846	9.902	13.848	0.610

St. R-sq-Stationary R-Square, R-Sq.-R-Square, RAME-Root Mean Square Error, MAPE-Mean Absolute Percentage Error, MAE-Mean Absolute Error, Max APE-Maximum Absolute Percentage Error, Max AE-Maximum Absolute Error.

Table 4. ACF and PACF of residuals for rice area

Lag	ACF	Std. Error(a)	Box-Ljung Statistic			Lag	PACF	Std. Error
			Value	df	Sig.(b)			
1	-0.368	0.160	5.30	1	0.021	1	-0.368	0.167
2	-0.218	0.158	7.21	2	0.027	2	-0.409	0.167
3	0.358	0.155	12.51	3	0.006	3	0.130	0.167
4	-0.084	0.153	12.81	4	0.012	4	0.079	0.167
5	0.082	0.151	13.11	5	0.022	5	0.313	0.167
6	-0.068	0.148	13.32	6	0.038	6	0.019	0.167
7	-0.027	0.146	13.36	7	0.064	7	-0.250	0.167
8	0.04	0.143	13.44	8	0.098	8	-0.189	0.167
9	0.005	0.140	13.44	9	0.144	9	-0.047	0.167
10	-0.009	0.138	13.44	10	0.200	10	-0.019	0.167
11	0.066	0.135	13.68	11	0.251	11	0.224	0.167
12	-0.068	0.132	13.94	12	0.305	12	0.063	0.167
13	0.023	0.130	13.97	13	0.375	13	0.070	0.167
14	0.064	0.127	14.23	14	0.432	14	-0.071	0.167
15	0.093	0.124	14.79	15	0.466	15	-0.128	0.167
16	0.023	0.121	14.83	16	0.537	16	0.139	0.167

Table 5. ACF and PACF of residuals for rice production

Lag	ACF	Std. Error(a)	Box-Ljung Statistic			Lag	PACF	Std. Error
			Value	df	Sig.(b)			
1	-0.734	0.160	21.04	1	0.000	1	-0.734	0.167
2	0.305	0.158	24.79	2	0.000	2	-0.504	0.167
3	0.053	0.155	24.09	3	0.000	3	0.057	0.167
4	-0.309	0.153	28.99	4	0.000	4	-0.220	0.167
5	0.366	0.151	34.91	5	0.000	5	-0.123	0.167
6	-0.178	0.148	36.36	6	0.000	6	0.241	0.167
7	-0.052	0.146	36.48	7	0.000	7	0.073	0.167
8	0.145	0.143	37.51	8	0.000	8	-0.166	0.167
9	-0.185	0.140	39.24	9	0.000	9	-0.165	0.167
10	0.146	0.138	40.35	10	0.000	10	-0.049	0.167
11	-0.021	0.135	40.37	11	0.000	11	-0.008	0.167
12	-0.098	0.132	42.07	12	0.000	12	-0.082	0.167
13	0.139	0.130	42.07	13	0.000	13	0.026	0.167
14	-0.139	0.127	42.28	14	0.000	14	0.005	0.167
15	0.082	0.124	43.28	15	0.000	15	-0.097	0.167
16	0.068	0.121	44.03	16	0.000	16	0.121	0.167

Table 6. Forecasted values of Rice Cultivated Areas and Production with 95% Confidence Level (CL)

Year	Paddy Area (000'ha)	LCL	UCL	Paddy Production (000' tonnes)	LCL	UCL
2011-12	593.97	546.71	641.23	1048.76	802.55	1294.98
2012-13	599.55	550.94	648.15	1100.00	853.00	1346.99
2013-14	597.47	544.29	650.65	1093.27	805.11	1381.43
2014-15	598.22	531.06	665.39	1126.61	823.05	1430.16

LCL – Lower Confidence Level UCL – Upper Confidence Level

Figures

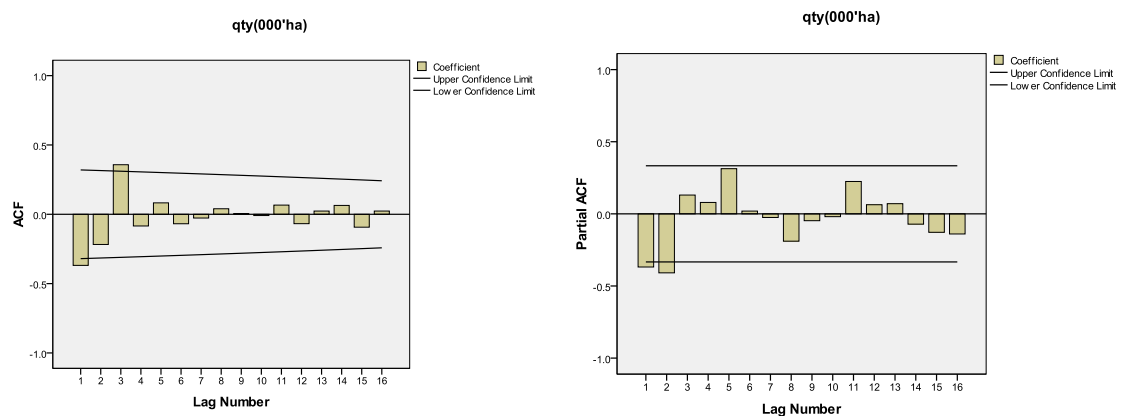


Fig.1 Paddy area ACF and PACF differenced data

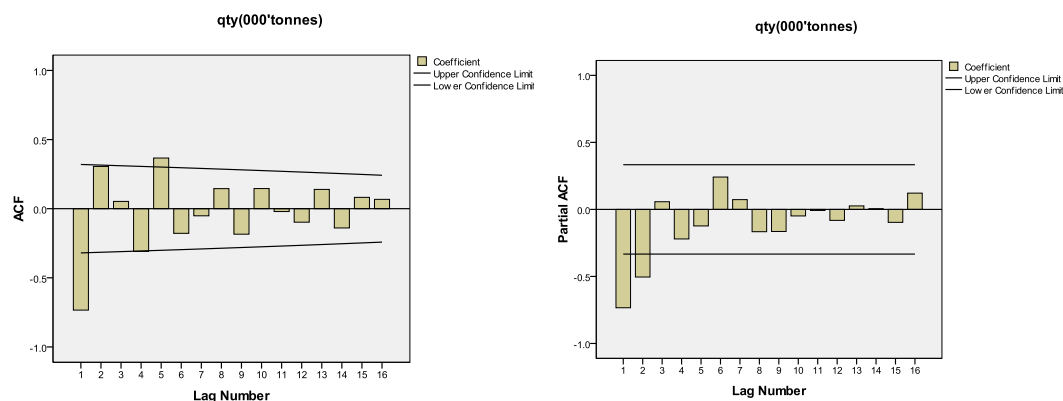


Fig.2 Paddy Production ACF and PCF differenced data

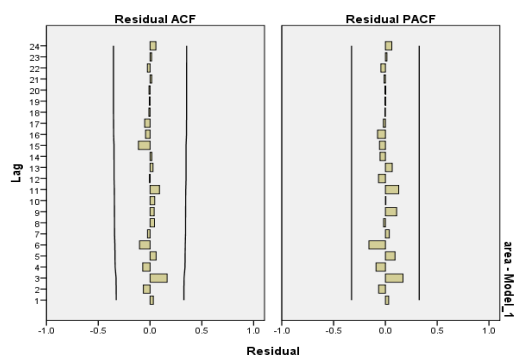


Fig. 3 ACF and PACF of residuals of fitted ARIMA model for Rice area.

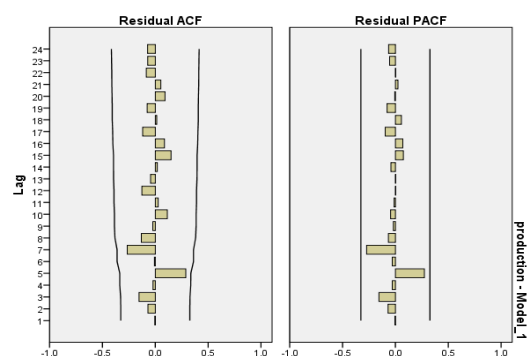


Fig. 4 ACF and PACF of residuals of fitted ARIMA model for Rice production.