Journal of Crop and Weed, 9(2):26-31(2013)

ARIMA modeling to forecast area and production of rice in West Bengal

R. BISWAS AND B. BHATTACHARYYA

Department of Agricultural Statistics
Bidhan Chandra Krishi Viswavidyalaya,
Mohanpur-741252, Nadia, West Bengal

Received: 18-09-2013, Revised: 29-10-2013, Accepted: 15-11-2013

ABSTRACT

Crop area estimation and forecasting of crop yield are an essential procedure in supporting policy decision regarding land use allocation, food security and environmental issues. The present paper intended to give a comprehensive picture of current status of rice production in West Bengal. Time series modeling using ARIMA model was developed for individual univariate series of both area and production of rice in West Bengal since independence. Box and Jenkins linear time series model, which involves autoregression, moving average, and integration, termed as ARIMA (p, d, q) model was applied.

Keywords: ACF, AIC, ARIMA, PACF and SBC

Rice is India's preeminent crop, and is the staple food of 65% of India's population of the Eastern and southern parts of the country. Its cultivation is a major source of employment in South Asia. India, Bangladesh and Pakistan supply almost 30% of the world's paddy (http://en.wikipedia.org/ wiki/, 2013). India is the world's second largest producer of rice followed by China, accounting for 20% of all world rice production (http://en.wikipedia.org/ wiki/, 2013). Forecasting has an important role in the management of agriculture. Univariate time series modeling have been useful in developing the forecasting for production of the crops. During the last few decades many sophisticated statistical forecasting models have been developed due to the availability of advanced computers. One of such models includes the Autoregressive Integrated Moving Average (ARIMA) models (Box-Jenkins, 1976). These ARIMA models have been widely used in attempts to forecast the area and production of rice in West Bengal. Autocorrelation function, partial autocorrelation function and spectral density function indicate nonstationary of the series.

The ARIMA approach in modeling time series with trend is to filter and then fit a stationary model of the class. Recent relevant references include Diebold (2001), Patterson (2000) and Gujarati (2003)**.** Forecast would be obtained based on early indicators. The residuals from the model have been used to detect the hidden periodicity by the spectral density function.

In this study attempts have been made to examine the class of ARIMA models that may best fit the area and production of rice in West Bengal and the forecast skill of the best fitted model have been also studied. Search for hidden periodicity has also been investigated in this study.

MATERIALS AND METHODS

Separate time series data of area and production of rice in West Bengal were collected over the period of 1947-1948 to 2007-2008. The entire time series data of 60 years have been divided into two parts. Data for 1947 to 2004 were utilized for model development and last three years were considered as test period. Let Z_t {t=1, 2,…} and t∈ T be a univariate time series where t is a subset of the observations and T is an index set. The observations are at equally spaced time points. A non-seasonal series can often be represented as a process whose differences an autoregressive are moving average as prescribed by Box-Jenkins (1976). Differencing a series is likely to be appropriate when there is finite autocorrelation between adjacent observations. Differencing is considered as a filter in forming the series stationary.

The general form of ARIMA (p, d, q) model is represented by

$$Z_t = Ø_1 Z_{t-1} + Ø_2 Z_{t-2} + \ldots\ldots\ldots + Ø_p Z_{t-p} - \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots\ldots\ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t ;$$

Where, Z_t stands for the value of a stationary time series at time t and ε_t's represent random error being independently and normally distributed with zero mean and constant variance for t = 1, 2…..n; d the degree of differencing and Ø's and θ's are coefficients to be estimated.

The Box-Jenkins method consists of the following steps:

i) **Identification**

Identification of the model for ARIMA (p,d,q) is based on the concepts of time-domain and frequency-domain analysis i.e. autocorrelation function (ACF), partial autocorrelation function (PACF) and spectral density function. Once the order of differencing has been diagnosed and the differenced univariate time series can be analysed by the method of both time-domain and frequency-domain approach (Cressie, 1988).

ii**) Estimation**

The appropriate p, d and q values of the model and their statistical significance can be judged by t-distribution. A model with minimum values of RMSE, MAPE, AIC, BIC, Q-statistics and with high R-square, may be considered as an appropriate model

Email: ria.biswas6@gmail.com

for forecasting. The model selection criteria includes Akaike Information criterion and Schwarz's Bayesian Information criterion, Mean squared error (MSE), Root Mean squared error (RMSE), Mean absolute error (MAE) and Minimum Absolute Percentage Error (MAPE).

iii**) Diagnostic checking**

Considerable skill is required to choose the actual ARIMA (p,d,q) model so that the residuals estimated from this model are white noise. So the autocorrelations of the residuals are to be estimated for the diagnostic checking of the model. These may also be judged by Ljung-Box statistic under null hypothesis that autocorrelation co-efficient is equal to zero.

iv) **Forecast**

ARIMA models are developed basically to forecast the corresponding variable. The entire data is segregated in two parts, one for sample period forecasts and the other for post-sample period forecasts. The former are used to develop confidence in the model and the latter to generate genuine forecasts for use in planning and other purposes.

The relative measure of forecast accuracy is described by Koreisha and Fung (1999), and Pankratz (1983).

One of the first analytical techniques developed is frequency domain analysis i.e., periodogram analysis. Finally it is represented by spectral density function with spectral window and autocorrelation. It may be represented as:

$$I(T) = \frac{1}{\pi}\left[\lambda_0\gamma_0 + 2\Sigma\lambda_k\gamma_k \cos k(2\pi f)\right]$$

, where λ_k is the spectral window i.e., Parzen window, γ_k is the autocorrelation at k lags, and

$f_i = \frac{i}{N}$; $i = 0,1,2,.....,N/2$, k=0,1,2,…..M+1

where M is the truncation point and It is the maximum lag used in the particular window employed. Here Parzen window is employed.

The results from the simple spectral analysis indicate the several neighbouring spectral estimates, which may also be found statistically significant. But to overcome the problem the procedure to smoothing spectrum has been used to detect the sharp peaks in the fundamental frequencies for the time series of the production of rice in West Bengal. The most popular procedure of smoothing is that of the Fourier transformation in the truncated sample autocorrelation have to be weighted by the neighbouring estimates. The Parzen window suggested by Priestley (1981) has been adopted here for spectral window. Both the time-domain and frequency-domain analysis are used here to detect the series as nonstationary in mean. For testing the presence of periodicity in time series of the production of rice in West Bengal,

model $g^* = \frac{I_{\max}}{\left\{\sum_{p=1}^{N/2} I_p\right\}}$ Under the null hypothesis that residual process have the periodicities in I_{max}.

The method of Woodroofe and Van Ness (1967) is adopted for its significance. It is given by:

$b'_m = \frac{\left(1+\sqrt{\left(\frac{2}{d}\right)}z\right)}{\pi}$; where z at 5% level of significance and d (degree of freedom) $= 3.71\frac{N}{M}$ for Parzen window. The test of the spectral ordinate i.e., detecting the hidden periodicity in residuals was described by Priestley (1981), Mukhopadhyay (1997). The Fisher's g^* statistic are compared with the calculated b'_m.

RESULTS AND DISCUSSIONS

Fig. 1 represents the gross cultivated area and yearly production of rice in West Bengal from 1947-48 to 2007-08 along with the best(adjudged through R^2 value) fitted trend. Both the curve shows increasing pattern. For gross cultivated area of rice the trend curve shows the linear trend (R^2=0.89 and significant at 95% confidence limit) whereas the production trend follows an exponential growth (R^2=0.95 and significant at 95% confidence limit).

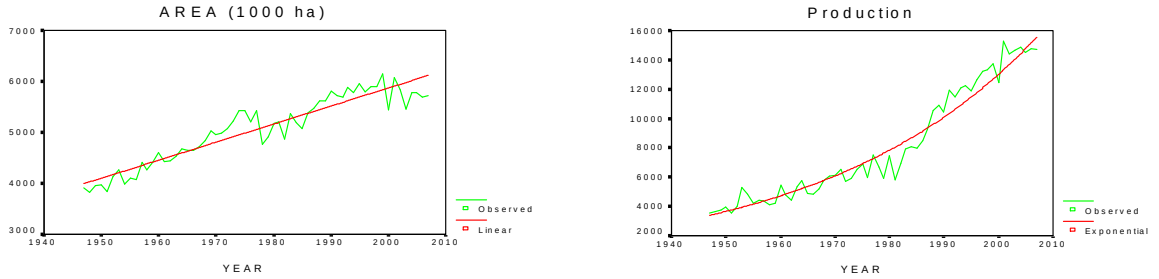

Fig.1: Trend curve for gross cultivated area and production of rice in West Bengal

Fig. 2 show the autocorrelation function and partial autocorrelation function of the historical observations of the gross cultivated area of rice in West Bengal. From the figure two facts stand out. First, the ACF declines very slowly.ACF up to 15 lags are individually statistically different from zero or, they all are outside the 95% confidence bound and secondly, after the first lag the PACF drops dramatically and all PACF after lag 1 are statistically insignificant. The correlogram represents that ACF remain close to 1.0 throughout, declining to zero gradually. So it is expected that high positive autocorrelations exists indicating increasing mean. When the variance of the realization appears fairly stable, the means are definitely increasing and this series has an upward trend. It is regarded as nonstationary in mean.

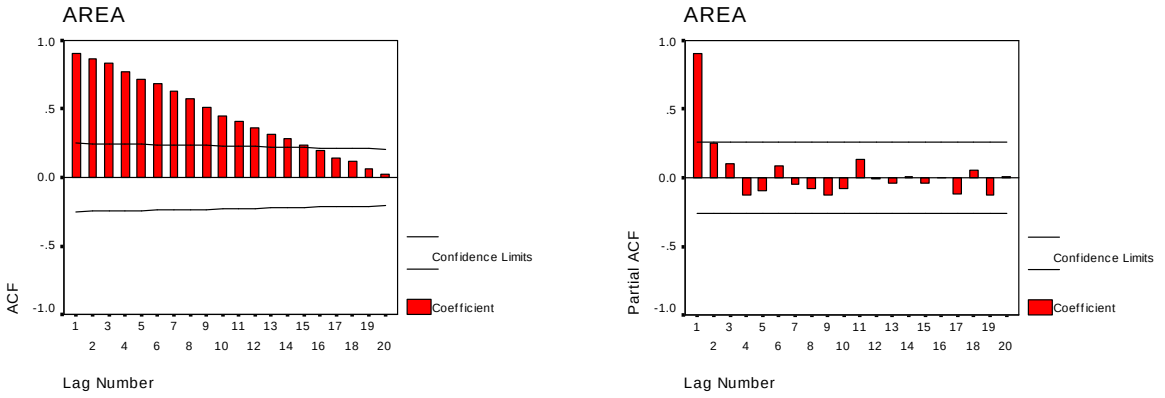

Fig. 2: Autocorrelation ,Partial Autocorrelation Correlogram of gross cultivated area of rice

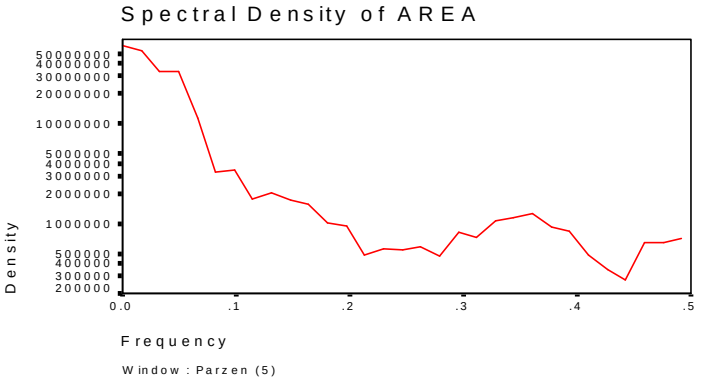

Fig. 3: Spectral Density Function of gross cultivated area of rice

Fig. 3 shows that spectral density function has the highest peak in low frequencies. This pattern indicates the nonstationary in mean of the time series with considering high positive autocorrelation in neighboring points.

Fig. 4 shows the correlogram of time series data for production of rice. Here also, the figure exhibits the similar pattern in support of nonstationarity of the time series. The ACF declines and ACF upto 15 lags are individually statistically different from zero or, they all are outside the 95% confidence limit. After the first lag the PACF drops abruptly and all PACF after lag 1 are statistically insignificant. ACF remain close to 1.0 initially and declining gradually suggesting high positive autocorrelations with increasing mean. It is regarded as nonstationary in mean.

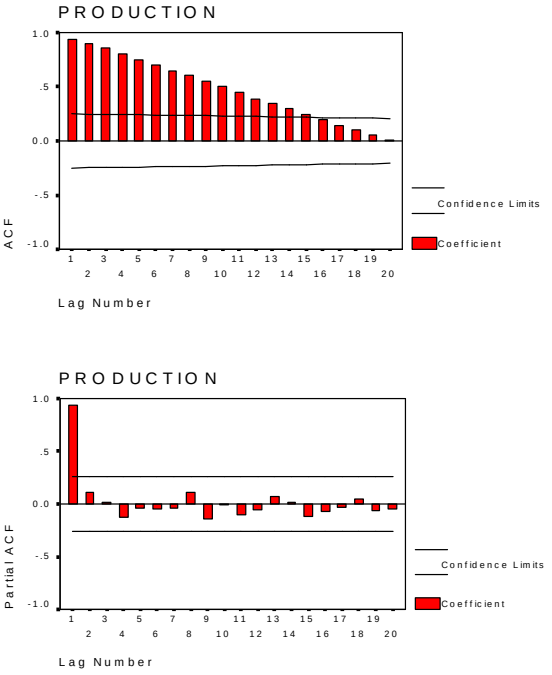

Fig. 4: Autocorrelation, Partial Autocorrelation and Correlogram of rice production

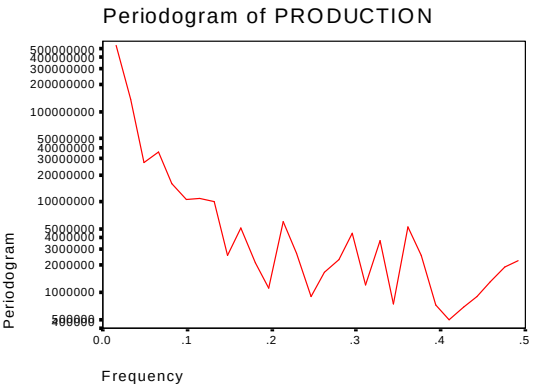

Fig. 5: Spectral Density Function of rice production

Fig. 5 shows that spectral density function has the highest peak in low frequencies. This also indicates the nonstationary in mean of the time series with considering high positive autocorrelation in neighbouring points.

For both of the univariate time series data treated separately the PACF has the 1st spike significant and the others are nonsignificant. So, both the series posses the ARIMA system as the other PACF spikes have a wave with positive and negative values. So, differencing is the procedure for filtering the series. Then differencing the series changes the variable under consideration as suggested by Chatfield (1971), Box-Jenkins (1976) and Cressie (1988). To make the series stationary for area and production, differenced series was used. First difference series of rice area and production showed stationarity for the present study. Some tentative ARIMA models were considered and the best fitted model is accepted on the basis of minimum AIC (Akaike's Information Criterion), SBC (Schwarz's Bayesian Criterion) and MAPE (Mean Absolute Percentage Error) and Maximum R^2 value as presented in Table- 1 and 2.

Table1: Comparison of different ARIMA models with model fit statistics for gross cultivated area of rice

Area-Model	**Model fit statistics**			
	R^2	**RMSE**	**MAPE**	**Normalized BIC**
ARIMA (2,1,3)	0.930	182.496	2.537	10.891
ARIMA	**AIC**	**SBC**	**MAPE**	**R^2 value**
(2,1,1)	805.56	813.93	2.670	0.919
(2,1,2)	805.90	816.381	2.570	0.922
(2,1,3)	801.85	814.42	2.520	0.929

For the series of gross cultivated area of rice it is observed from the results that ARIMA (2,1,3) provides the best fitted model. On the basis of estimated parameters the mathematical model is obtained as follows:

$$Z_t = 0.00164 + 0.02Z_{t-1} - 0.753Z_{t-2} + \varepsilon_t - 0.725\varepsilon_{t-1} + 1.120\varepsilon_{t-2} - 0.462\varepsilon_{t-3}$$

ARIMA (2, 1, 1) is found to be the best fitted model for the production of rice. Estimation of parameters suggest the model may be represented mathematically as:

$$Z_t = 188.479 - 0.661Z_{t-1} - 0.352Z_{t-2} + \varepsilon_t + .081\varepsilon_{t-1}$$

Diagnostic checking of the models are concerned with the residual plots of ACF and PACF as presented in figure-6 and 7. As all the ACF and PACF are within the confidence bound the model ensures that the errors or residuals possesses a white noise.

Table2: Comparison of different ARIMA models with model fit statistics for production of rice

Area-Model	**Model fit statistics**			
	R^2	**RMSE**	**MAPE**	**Normalized BIC**
ARIMA (2,1,1)	0.969	680.474	7.600	13.319
ARIMA	**AIC**	**SBC**	**MAPE**	**R^2 value**
(2,1,1)	957.41	965.78	7.600	0.969
2,1,2)	959.38	969.85	7.608	0.969
(2,1,3)	9591	972.55	7.619	0.970
(3,1,1)	959.25	969.72	7.578	0.969

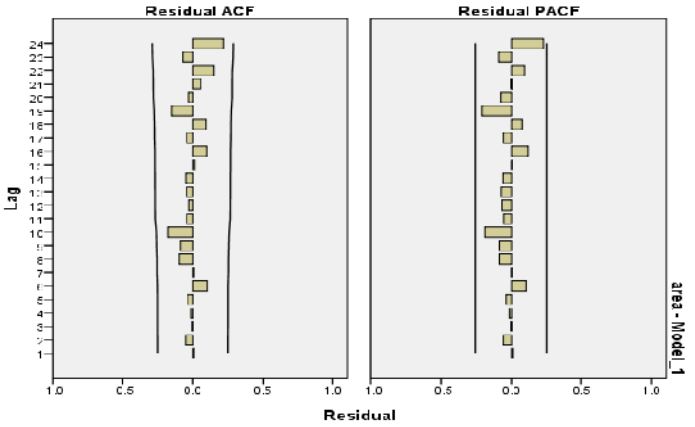

Fig. 6: Residual plots of ACF and PACF of gross cultivated area of rice

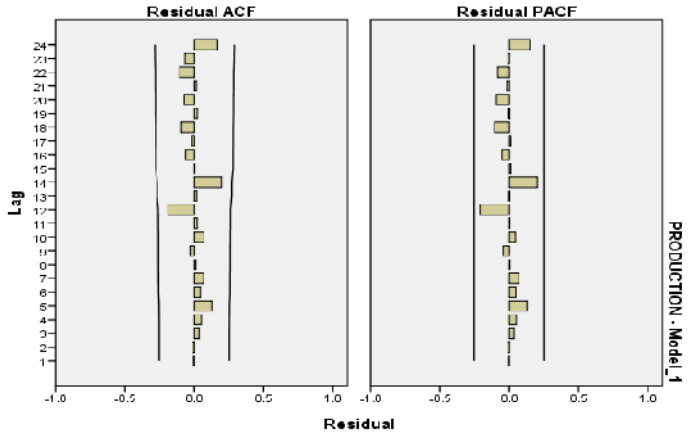

Fig. 7: Residual plots of ACF and PACF of rice production

The ARIMA models developed in the present study are finally used to forecast the corresponding variables. There are two kinds of forecasts: sample period forecasts and post-sample period forecasts. The sample period forecast are made for last four years of the data-set *i.e.*, for 2003-2007 to establish the confidence about the model and the latter *i.e.*, upto 2016, to generate genuine forecasts for use in planning and other purposes. The forecasting values thus obtained are presented in the Table - 3. The MAPE to forecast gross cultivated area is 2.53 which indicate that the accuracy level of forecasting is much higher. MAPE for forecasting production is 7.6, which also shows quite good level of forecasting accuracy.

Table 3: Forecasting table

Year	Area (`000 ha)		Production (`000 tonnes)	
	Observed	**Estimated or forecasted**	**Observed**	**Estimated or forecasted**
2003	5456.6	5754.0	14662.2	14345.9
2004	5783.6	5860.5	14884.9	15192.5
2005	5782.9	5859.5	14510.8	14996.1
2006	5687.0	5679.7	14745.9	15020.0
2007	5719.8	5691.0	14719.5	15079.7
2008	.	5868.0	.	15004.5
2009	.	5922.3	.	15204.9
2010	.	5851.9	.	15351.5
2011	.	5869.3	.	15563.5
2012	.	5978.9	.	15751.2
2013	.	6018.3	.	15931.9
2014	.	5989.6	.	16125.7
2015	.	6016.9	.	16313.4
2016	.	6094.3	.	16500.5

The study revealed some significant observations. Gross cultivated area and yearly production of rice in West Bengal from 1947-48 to 2007-08 both exhibit an increasing trend. After green revolution the area is increased only 22.98% while the production is increased 168.73%. ACF, PACF and spectral density function indicate both area and production data of rice are nonstationary in mean. So the filter of the series by differencing is adequate to form stationary. Trend in data is also detectable by the peak in the spectral density function at the low frequencies. The wide spikes in spectral density function suggest a non-deterministic cycle in rice production in West Bengal. For the series of gross cultivated area ARIMA(2,1,3) model is found to be the best fitted model whereas for the series of production ARIMA (2,1,1) is found to be the best fitted one. The model exhibits good accuracy level for future projection of area and production of rice in the state.

REFERENCES

Box, G.E. and Jenkins, G.M. 1976. *Time Series Analysis Forecasting and Control*. Holden-Day. San Fran., USA.

Chatfield, C. 1975. *The Analysis of Time Series: Theory and Practice*. Chapman and Hall, London.

Cressie, N.1988. A Graphical Procedure for Determining Non-stationary in Time Series. *JASA*. [**83**: 1108-15.

Diebold , E.X. 2001. *Elements of Forecasting*.South-Western Publishers. USA

Gujarati, D.N. 2003. *Basic Econometrics*. 4th Ed. Mc. Graw Hill. New York. USA

Koreisha, S.G. and Fang, V. 1999. The impact of management errors on *ARIMAPrediction. J. Forecasting*. **18**: 95-100.

Mukhopadhyay S.K.1997. Empirical Spectral analysis for determining hidden periodicities in monsoon rainfall. *J. Interacad*., **1:** 61-68.

Patterson, K. 2000 . *An Introduction to Applied Economics: A Time Series Approach.* St. Martin's Press. USA.

Priestley M.B. 1981. *Spectral Analysis and Time Series*. Vol.1. Academic Press.

Pankratz, A. 1983. *Forecasting with Univariate Box-Jenkin Models Concept and Case*. John. Wiley and Sons. New York.

Woodroofe, M.B. and Van Ness, J.W. 1967. *Ann. Math. Statist*., **38**: 1558-69.

http://en.wikipedia.org/wiki/Rice_production_in_India last accessed on 07.11.13