Int.J.Curr.Microbiol.App.Sci (2019) *8*(7): 1719-1730

International Journal of Current Microbiology and Applied Sciences
***ISSN: 2319-7706* Volume 8 Number 07 (2019)**
Journal homepage: http://www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2019.807.204

Statistical Study on Modeling and Forecasting of Jute Production in West Bengal, India

Soumitra Sankar Das[1*], Soumik Ray[2], Abhishek Sen[3], G. Samba Siva[4] and Shantanu Das[5]

[1]Department of Agricultural Statistics & Computer Application, BAU, RAC, Kanke 834006, India
[2]Department of Agricultural Statistics, CUTM, Paralakhemundi, Gajapati, Orisha 761211, India
[3]Department of Soil Science and Agricultural Chemistry, UBKV, Pundibari, West Bengal 736165, India
[4]ICAR-Central Research Institute for Dryland Agriculture, Hyderabad 500 059, India
[5]College of Agriculture, Tripura, Lembucherra 799210, India

**Corresponding author*

Keywords

Parametric regression model, Auto Regressive Integrated Moving Average (ARIMA), Normality test, Forecasting

Article Info

Accepted:
15 June 2019
Available Online:
10 July 2019

A B S T R A C T

Present investigation was an attempt to study the trend of jute production in West Bengal for the period starting from 1950 to 2016. For stochastic trend estimation, a number of time series parametric regression models viz. Linear model, Quadratic model, Exponential model, Logarithmic model, Power model and Auto Regressive Integrated Moving Average (ARIMA) were employed and compared for finding out an appropriate econometric model to capture the trend of jute production of the country. Based on the performance of several goodness of fit criteria viz. Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and R-squared values best fitted model was selected. The assumptions of 'Independence' and 'Normality' of error terms were examined by using the 'Run-test' and 'Kolmogorov-Smirnov (K-S) test' respectively. This study found ARIMA (1, 1, 2) as most appropriate to model the jute production of West Bengal. The forecasted value by using this model was obtained as 9149.22 (In ' 000 Bales of 180 Kgs. each) by 2021.

Introduction

Jute is a natural fibre popularly known as the golden fibre. It is one of the cheapest, strongest among all natural fibres and considered as fibre of the future. Jute occupies second position (next to cotton) in world's textile fibre production where India is the

largest producer of jute in the world. Approximately 60 percent of the total world production of jute is cultivated in India with an annual estimated production of 11494 thousand bales of jute. Jute is a bio-degradable crop grown mainly in the Ganges delta. State of West Bengal occupies tops the list of jute production and contribute alone more than 80 per cent of the jute the country produces. Generally the crop (jute) is grown through-out the state except the hilly region of the north and the plateau area of the west. Murshidabad, West Dinajpur, Cooch Behar. Hugli. 24 Parganas (north and south), Nadia, Malda, Barddhaman, Jalpaiguri, Haora and Medinipur districts are the important producers. In West Bengal *Corchorus capsularis* (white jute) is grown in lowlands (Bills). A number of studies have been considered up by several authors in different forms to analyze the production behavior of jute with different objectives. Among the studies are works of importance (Sen, 1967; Narain, 1977; Reddy, 1977; Sawant, 1983; Boyce, 1987; Chakraborty, 1987; Dey, 1999; Chattapadhyay and Das, 2000; Sarkar and Sahu, 2002).

Most of these studies have emphasized analyzing the trends and production of jute along with other crops oronly for jute. This study was devoted for the analysis of trend of jute production in West Bengal with an attempt to have some idea about the possible future behavior of jute production in these areas by using forecasting methods. Among the different methods of forecasting on the basis of past information, different parametric trend models (linear, quadratic, logarithmic, power and exponential) are important. With the publication of B-J methodology (1978), it has taken a valuable place in the process of forecasting in which "the data speak for themselves," and as such, this methodology has been utilized in this study to foresee the future of jute production. An attempt has also been made to compare the above methods with the help of the actual data for the years 1950 onward.

Materials and Methods

Data with respect to production of jute in West Bengal for the period of 1950-51 to 2016-17 has been collected from Directorate of Economics and Statistics, Department of Agriculture and Cooperation, Government of India. Before analysis, as the study is dealing with time series, present data set have been verified initially for existence of outlier and randomness. Descriptive statistics are used to explain the basic features of the data in any study. The selected descriptive measures i.e. mean, standard error, standard deviation, skewness, kurtosis along with simple growth rates have been used to explain behavior of each series in this study. Simple Growth Rate (SGR) has been calculated by using the following formula:

$$\text{SGAR (\%)} = \frac{X_t - X_0}{X_0 \times n} \times 100$$

Where X_t is the value of series for the last period, X_0 is the value of the series of first period and n is the number of the periods (Dhekale *et al.*, 2014).

For detecting outlier in time series, Grubbs test was used in the present's scenario as the test is particularly useful in case of large sample and easy to follow. SPSS software has been used for testing the outliers.

Some of the parametric and non-parametric trend models are also applied to study the behavior of the data series. The models along with their equations are given below:

Linear model

A linear model is one in which all the parameters appear linearly and it is formulated as $X_t = a + bt + e_t$.

Quadratic model

The quadratic model can be used to model a series which "takes off" or a series which "dampens". It expressed as $X_t = a + bt + ct^2 + e_t$.

Power model

It is a non-linear regression model, which is based on the following equation:

$X_t = at^b$.

Exponential model

The equation of exponential model is

$$X_t = a\left[Exp(bt)\right] + e_t$$.

Logarithmic model

The equation of logarithmic model is given by

$X_t = a + b\ln(t) + e_t$.

In order to apply these models, e_t is expressed as error term which is independently and identically normally distributed. In all the trend models, model significant was tested by F test and individual regression coefficient is testing using t test. The best fitted model is selected on the basis of maximum value of R^2 and minimum values of RMSE, MAPE and MAE.

Time series analysis

Generally, a time series, as a stochastic process, is an ordered sequence of observations made sequentially in time.

The most important feature of such data is the likely lack of independence between successive observations in time. Time series data can be univariate as the case with the jute production under consideration or multivariate (Akpanta, 2014).

The ARIMA (Auto Regressive Integrated Moving Average) class of model is only applied to a univariate time series data. This method of time series modelling is often referred to as the Box-Jenkins approach. The act of ARIMA modelling gained its credence from Box and Jenkins 1976, (Box and Jenkins, 1976). A good ARIMA model requires at least 50 observations and a reasonably large sample size is required for a seasonal time series (Pankratz, 1983). With the ARIMA models forecast are made using the past of the process and are particularly suitable for short term forecasting and also forecasting seasonally enriched series. Box-Jenkins models are only reasonable for stationary time series with equi-spaced discrete time intervals.

A time series is said to be stationary if its mean, variance and autocorrelation functions remains unchanged over time.

However, in practice many time series data are non-stationary and could be transformed to stationary by a simple differencing exercise, usually the first difference is enough to coerce a non-stationary time series into a stationary one and the second difference is seldom required. Non-stationarity implies trend and is typically induced by serial correlation.

Box-Jenkins Auto Regressive Integrated Moving Average (ARIMA) Models

Box-Jenkins methodology (Box and Jenkins of Time Series Analysis: Forecasting and Control) is used here for time series analysis which is technically known as the ARIMA methodology.

The ARIMA Model Includes:

The Autoregressive (AR) model.
The Moving Average (MA) Model.
The ARMA Model.

The Autoregressive (AR) Model

The Simplest form of the ARIMA model is called the autoregressive model. Let z_t stand for the value of a stationary time series at time t, that is, a time series that has no trend, but fluctuates about a constant value referred to as the *level* of the series. (We deal with trends below.) By autoregressive, we assume that current z_t values depend on *past* values from the same series. In symbols, at any t,

$$z_t = C+\varphi_1 z_{t-1}+\varphi_2 z_{t-2}+.....+\varphi_p z_{t-p}+\varepsilon_t$$

$$Z_t = C + \varepsilon_t + \sum_{i=1}^{p} \varphi_i z_{t-i}$$

Where C is the constant level, z_{t-1}, z_{t-2},….., z_{t-p} are past series values (lags), the ϕ's are coefficients (similar to regression coefficients) to be estimated, and ε_t is a random variable with mean zero and constant variance. The ε_t's are assumed to be independent and represent random error. Some of the ϕ's may be zero. If z_{t-p} is the furthest lag with a nonzero coefficient, the AR model is said to be of order p, denoted *AR (p).*

The Moving Average (MA) Model

z_t can also be modeled as a linear combination of white noise stochastic error terms. We call this type of model a moving average (MA) model. If z_t is considered as a weighted average of the uncorrelated ε_t's, MA(q) moving average component of order q, which relates each z_t value to the residuals of the q previous z estimates may be expressed as

$$z_t = e_t - q_1 e_{t-1} - q_2 e_{t-2} - \ldots. - q_q e_{t-q}$$

The ARMA Model

The AR and MA models for stationary series to account for both past values and past shocks may be combined. Such a model is called an *ARMA (p, q)* model with *p* order AR terms and *q* order MA terms. Thus an *ARMA (p, q)* model is written as

$$z_t = C+\varphi_1 z_{t-1}+\varphi_2 z_{t-2}+.....+\varphi_p z_{t-p}+\varepsilon_t - q_1 e_{t-1} - q_2 e_{t-2} - \ldots. - q_q e_{t-q}$$

Augmented Dickey Fuller (ADF) test (Stationarity test)

D.A. Dickey and W.A. Fuller (1979) established the Augmented Dickey Fuller test and it can be presented as

$$\Delta Y_t = \alpha + \beta_t + \delta Y_{t-1} + \sum \lambda_i \Delta Y_{t-i} + e_i$$

Where $^{\Delta Y_t}$ is the first difference of Y and α allows for a non-zero intercept or drift component i.e., constant t is included to allow for deterministic trend as the data may be trend stationary. The null hypothesis here is Y_t has a unit root (H_0: δ=0) against δ is negative. Thus the test consists of testing the negativity of δ in above equation. The test statistics is given by

$$DF_\tau = \frac{\delta}{SE(\delta)}$$

It can be compared to the relevant critical value for the Dickey-Fuller Test. If the test statistics is less than the critical value, then the null hypothesis of δ=0 is rejected and the data is stationary.

Box-Jenkins procedures

The objective of Box-Jenkins modelling approach is to find a parsimonious ARIMA model that describes the inherent generating process of the observed time series. The Box-Jenkins method consists of the following steps:

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).

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 Root Mean Square Error (RMSE), Mean Absolute Percent Error (MAPE), Q-statistics and with high R-square, may be considered as an appropriate model for forecasting. The model selection criteria include Mean squared error (MSE), RMSE, MAE and MAPE.

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.

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.

Model selection criteria using goodness of fit statistics

Among the competitive Box- Jenkins ARIMA model best model is selected on the basis of maximum R^2, RMSE, MAPE, and MAE. Any model which has fulfilled most of the above criteria is selected. This section provides definitions of the goodness-of-fit measures used in time series modeling.

R-squared

An estimate of the proportion of the total variation in the series that is explained by the model. This measure is most useful when the series is stationary. High positive values mean that the model under consideration is better than the baseline model.

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{X}_i - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

Root Mean Square Error (RMSE)

The square root of mean square error. A measure of how much a dependent series varies from its model-predicted level, expressed in the same units as the dependent series.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2}{n}}$$

Mean Absolute Percentage Error (MAPE)

A measure of how much a dependent series varies from its model-predicted level. It is

independent of the units used and can therefore be used to compare series with different units.

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{X_i - \hat{X}_i}{X_i}\right|}{n} \times 100$$

Mean absolute error (MAE)

Measures how much the series varies from its model-predicted level. MAE is reported in the original series units.

$$MAE = \frac{\sum_{i=1}^{n}\left|X_i - \hat{X}_i\right|}{n}$$

Results and Discussion

The data on the jute production on west Bengal (In ' 000 Bales of 180 Kgs. each) was analyzed using SAS and SPSS (statistical software) and the following results obtained:

The production under jute has varied between 1330 to 9325 kg with an average of 4972.319 kg registering a positive simple growth rate 6.39 percent in a year. The positive growth rate confirms that production of jute increase over the study period. Positive skewness (0.259) and negative kurtosis (-1.409) indicate that there has been increasing order during early half of the study period and its remain steady for long time.

The research work undertaken in the paper was based on forecasting jute production in West Bengal with respect to the parametric regression model along with Auto Regressive Integrated Moving Average (ARIMA) model. The best model has been selected on the application of the model performance criteria and it (*i.e.* best model) has been used to determine the forecast value for next five years.

At first, jute production data was tested for outliers by Grubbs method. It was observed that the number of extreme observations (*i.e.* outlier) in the present data was zero, which is depicted in Table 1.

Before analyzing by ARIMA, five (5) parametric regression models were also fitted on the data and the values of the precision coefficients were given in Table 2.

Table 2 reveals that all five models were almost equally precise, however, out of the five models investigated, the Quadratic model was superior to other selected regression models based on goodness of fit criteria of models. It might be due to time series data of jute production follows a quadratic growth pattern.

After consideration of these five (5) parametric regression models, ARIMA technique was employed in addition. At first, stationarity of jute production data was tested by time series plots and Augmented Dickey Fuller (ADF) test. The time series plot clearly indicated that the data was non stationary because of prominent increasing trend as shown in Figure 1.

ADF test for unit root also confirmed that the data of jute production data was non-stationary and it became stationary at first difference as the calculated values were lesser than critical values at a given levels of significance (i.e. 5%, 2% & 1%) depicted in Table 3. This was also supported by the trend of time series plot at first difference given in Figure 2.

After fixing the value of d as 1, values of p and q were determined. From correlogram of ACF and PACF, it was observed that there was only one significant spike for ACF at lag1 and two significant spikes PACF at lag 1 and lag 2, depicted in Figure 3.

Table.1 Descriptive statistics of jute production in West Bengal	
Statistics	**Production**
Mean	4972.319
Standard Error	301.4204
Standard Deviation	2467.233
Kurtosis	-1.40979
Skewness	0.259181
Minimum	1330
Maximum	9325
SGAR%	6.39
Grub test (Outliers detection)	No outliers

Table.2 Parametric regression models for estimation of Jute production				
Model	**R^2**	**RMSE**	**MAPE**	**MAE**
Linear	.888	832.590	17.569	666.259
Quadratic	.898	799.182	15.452	611.354
Exponential	.870	866.36	15.664	662.436
Power	.726	1141.3	23.601	990.183
Logarithmic	.643	1486.332	35.161	1260.861

Table.3 Result of ADF test						
Test	**ADF statistic**	**Critical value at**			**Prob.**	**Decision**
		1%	**5%**	**10%**		
ADF at level	-1.887	-3.568	-2.921	-2.599	0.629	Data Non-Stationary
ADF at first difference	-5.487	-4.152	-3.495	-3.181	< 0.0001	Data Stationary

Table.4 ARIMA Model Fit statistics				
Model	**R^2**	**RMSE**	**MAPE**	**MAE**
(1,1,1)	0.901	788.179	15.690	577.242
(0,1,1)	0.901	781.847	15.693	577.311
(1,1,0)	0.882	855.834	17.299	618.172
(1,1,2)	0.904	785.108	15.895	569.261
(2,1,1)	0.902	790.997	15.771	579.477

Table.5 ARIMA (1, 1, 2) model parameters estimation				
Model Parameter	**Estimate**	**Std. Error**	**t-value**	**Sig.**
Intercept	17.912	22.524	0.795	0.430
Autoregressive, Lag 1	-.999	0.048	-20.881	.000
Difference				
Moving Average, Lag 1	-.223	.205	-1.089	.281
Moving Average Lag 1	.768	.186	4.123	.000

Table.6 Tests of normality and randomness of residuals								
Kolmogorov-Smirnov test						Run test		
Statistic	df	Critical Value			Sig.	Z-value	No of Runs	Sig.
		5%	2%	1%				
0.087	66	0.167	0.186	0.201	0.200	-1.489	28	.137

Table.7 Forecasting of Jute production with control limits				
Year	Actual (In ' 000 Bales of 180 Kgs. each)	Predicted(In ' 000 Bales of 180 Kgs. each)	UCL(In ' 000 Bales of 180 Kgs. each)	LCL(In ' 000 Bales of 180 Kgs. each)
2011	8558.6	8870.68	6691.65	11272.36
2012	8228.2	8690.51	6535.04	11068.63
2013	8771.8	8908.95	6725.13	11315.4
2014	8341.2	8761.31	6596.74	11148.49
2015	7667.1	8992.66	6798.19	11409.73
2016	8187.7	8569.7	6430.4	10931.57
2017		8787.51	6619.74	11177.84
2018		8669.42	6468.92	11103.18
2019		8970.75	6675.14	11511.95
2020		8848.43	6523.26	11429.89
2021		9149.22	6729.74	11837.3

Fig.1 Time series plot of jute production

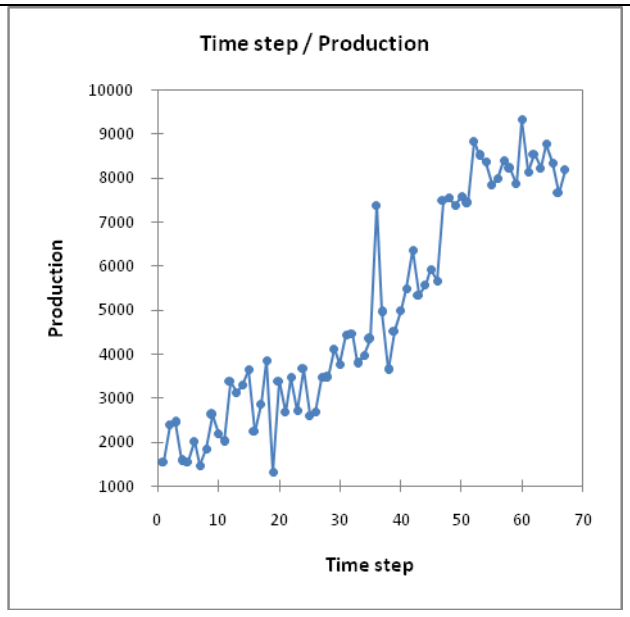


Fig.2 Time series plot for first differenced of jute production

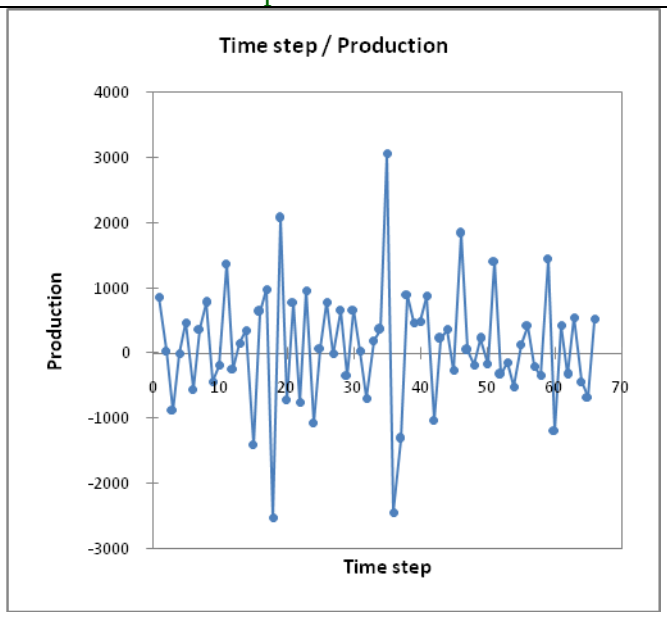
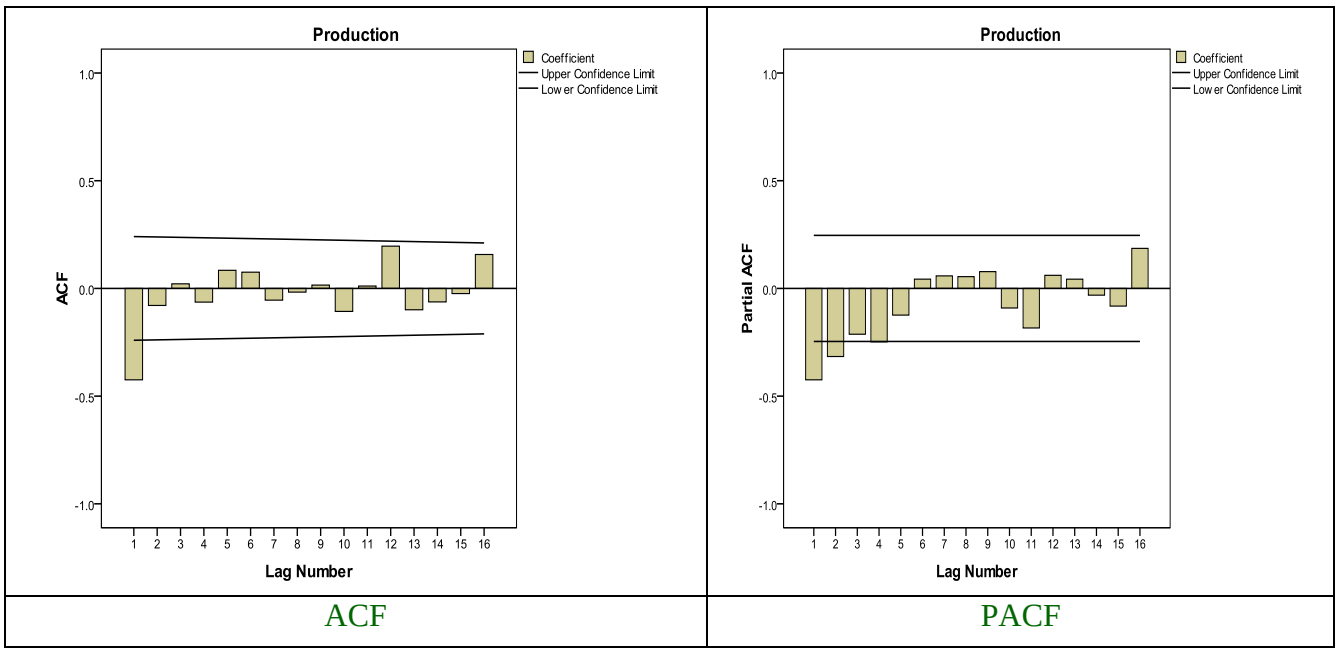


Fig.3 Correlogram of ACF and PACF for first differenced of jute production



ACF PACF

Fig.4 Residual ACF and PACF of ARIMA (1, 1, 2)

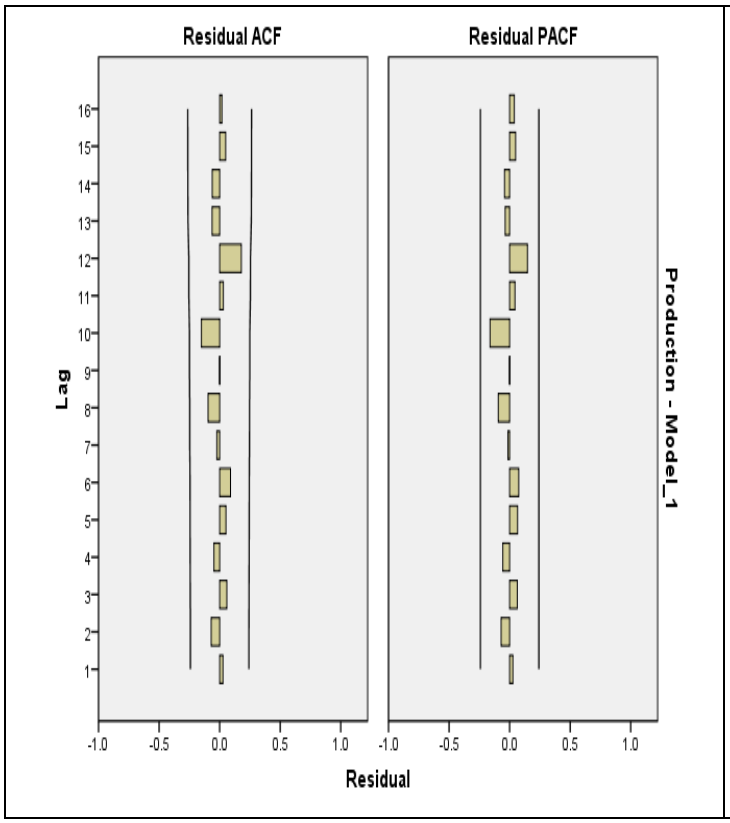


Fig.5 Histogram of residuals

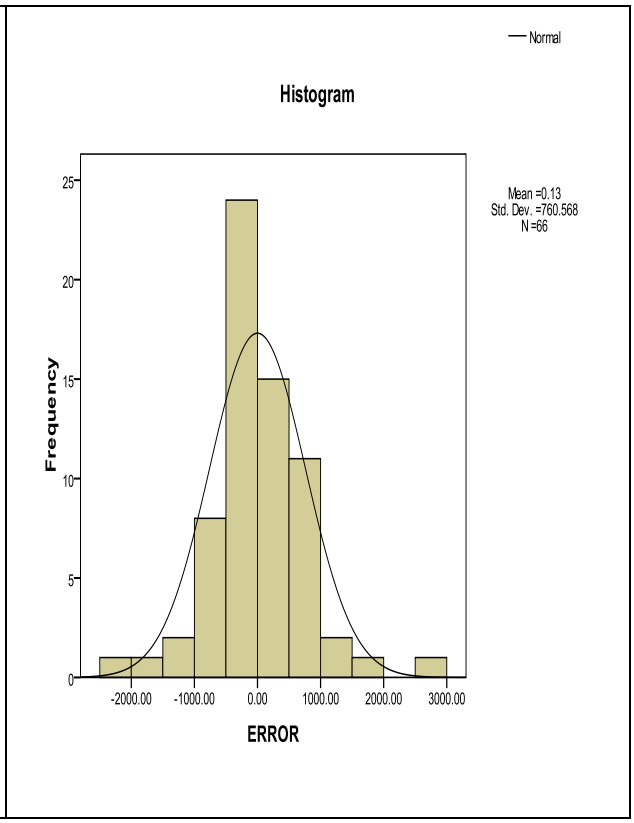
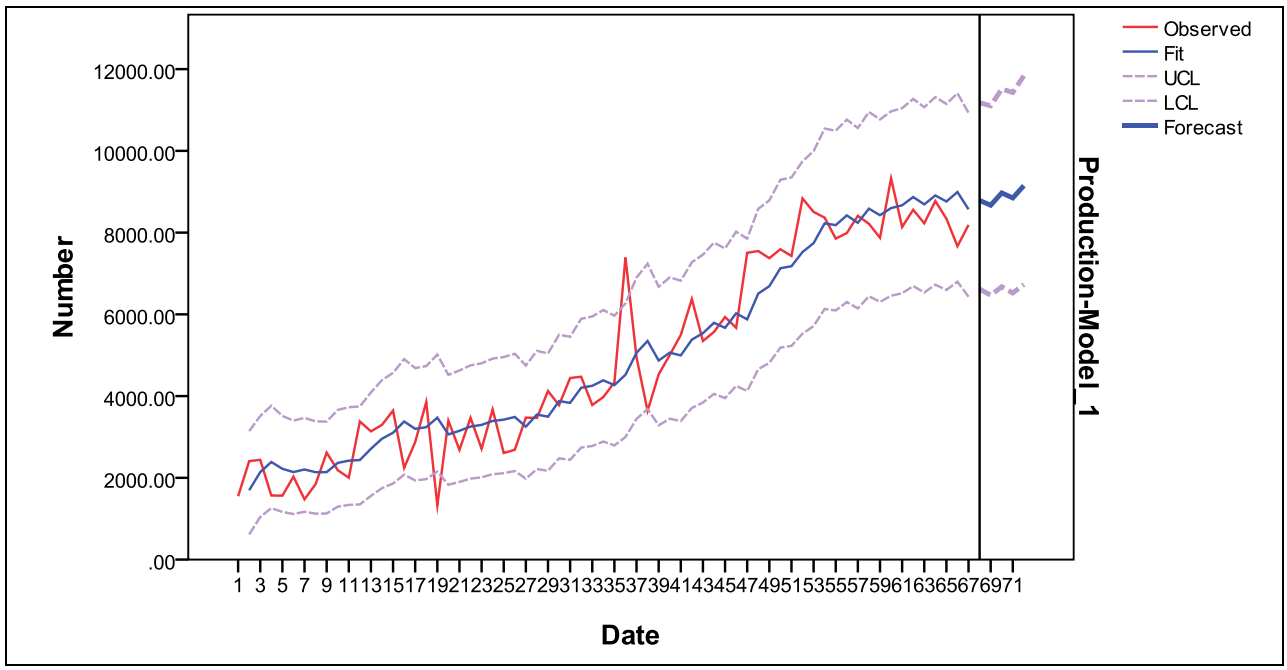


Fig.6 Forecasting of wheat production by ARIMA (1, 1, 2) model



The present study, possible ARIMA (p, d, q) models such as (1, 1, 1), (0, 1, 1), (1, 1, 0), (1, 1, 2) and (2, 1, 1) were compared to each other. Among all possible models, ARIMA (1, 1, 2) was selected as best and most appropriate model based on criteria of goodness of fit of model such as minimum values of RMSE, MAPE, MAE, MSE, and high R-squared value, which is presented in Table 4.

From Table 4, it can be concluded that ARIMA model performed better than the earlier selected models viz. Quadratic. ARIMA (1,1,2) model consider as a best model due to lower values of goodness of fit criteria's for model ***i.e.*** R^2, RMSE, MAPE and MAE. Thus ARIMA (1, 1, 2) has been selected as a best model for further analysis. The parameters were estimated for the best selected ARIMA (1, 1, 2) model as depicted in Table 5.

From the residual ACF and PACF plots of ARIMA (1, 1, 2), it was clear that all autocorrelations and partial autocorrelations lie between 95% control limits as shown in Figure 4. This also confirmed the 'good fit' of this selected model.

For checking normality of residuals, K-S test was performed. It was observed that the calculated value of the test statistic was $D_{n\ (Cal.)} = 0.087$ for ARIMA (1, 1, 2) model given in Table 6. As the calculated value of $D_{n\ (Cal.)} < D_{n\ (Tab.),\ 0.05} = 0.167$, the null hypothesis (that the observed distribution is Normal) is accepted. For checking the randomness of residuals, Run test was performed and it was observed that the probability value was greater than the 5% level of significance (*i.e.* >0.05) indicating residuals were distributed independently also. Thus it can be concluded that the residuals are independent and follow Normal distribution. Histogram of residuals was also confirmed the normality for the residuals which is depicted in Figure 4 and 5.

Finally, forecasting was done for jute production of West Bengal from 2011-12 to 2020 by using ARIMA (1, 1, 2) model where first six years data used for validation of the

model can be regarded as in sample forecast and last five years data were used for prediction purpose, which is popularly known as out sample forecast. Predicted values with 95% upper control limits (UCL) and lower control limits (LCL) were presented in Table 7.

By using ARIMA (1, 1, 2) model, it was observed that the actual and predicted values were closely related and predicted values were lies within the 95% confidence intervals as captured in Figure 6.

Time series analysis for forecasting may be regarded as a useful practice of a model to estimate future values based on previously observed values. The present dissertation work intended to establish the importance of the use of ARIMA models, made an attempt towards short term prediction of jute production in West Bengal. Box and Jenkins methodology of univariate ARIMA model has been employed to develop appropriate econometric model than traditional parametric regression model. ARIMA (1, 1, 2) model was found as most appropriate among other ARIMA models and hence, forecasting behavior has been employed for jute production of India.

From the forecasted values, it can be concluded that for a few coming years production of jute will follow an increasing trend and it has been estimated as 9149.22 (In' 000 Bales of 180 Kgs. each) for the year 2021. The analysis of best fitted ARIMA model and predicted forecasting pattern can play vital role to deal with future food security scenario and planning for policy makers in India. Finally, technological improvement, better management practice, high government policies i.e. agricultural funding, price support programmes etc. and enhancing relationship between farmers and research workers maybe important factors in sustaining this trend of production for long term.

References

Akpanta, A.C. and Okorie, I. E. 2014. Application of Box-Jenkins Techniques in Modelling and Forecasting Nigeria Crude Oil Prices. International Journal of Statistics and Applications. 4(6): 283-291.

Box, G.E.P. and Jenkins, G.M. 1976. Time series analysis, forecasting and control, 2nd ed. (Holden-Day, San Francisco, 1976).

Boyce, J.K. 1987. Agrarian impasse in Bengal: Institutional constraints to technological change. New York: Oxford University Press.

Chattapadhyaya, A.K. and Das, P.S. 2000. Estimation of growth rate: A critical analysis with reference to West Bengal agriculture. Indian Journal of Agricultural Economics 55(2).

Cressie, N.1988. A Graphical Procedure for Determining Non-stationary in Time Series. JASA. 83: 1108-1115.

Dekhale, B. S., Sahu, P. K., Viswajith, K. P., Mishra, P. and Noman, M. D. 2014. Modeling and forecasting for tea production in West Bengal. J. Crop and Weed. 10(2): 94-103.

Dey, U.K. 1999. Nature and causes of inter district variations in yield of rice in West Bengal, 1970–71 to 1994–95. Indian Journal of Agricultural Economics 59(4)

Dickey D.A and Fuller W.A. 1979. Distribution of estimators for Autoregressive Time Series with a Unitroot. Journal of the American Statistical Association 74: 427-431. doi: 10.1080/01621459.1979.10482531

Narain, D. 1977. Growth of productivity in Indian agriculture. Indian Journal of Agricultural Economics. 32(1): 1–44.

Pankratz, A. 1983. Forecasting with Univariate Box-Jenkins Models: Concepts and Cases. Wiley series in Probability and Mathematical Statistics.

Reddy, V.N. 1977. Statistical fitting of growth curves with illustrations from data on Indian economy. Calcutta, India: Indian Institute of Management.

Sarkar, C. and Sahu, P.K. 2001. Statistical account of the growth in jute and aus paddy—Two competing crops in West Bengal agriculture. Proceedings of the 90th Indian Science Congress, January 3–7, 2001, Lucknow, India.

Sawant, S.D. 1983. Investigation of the hypothesis of deceleration in Indian agriculture. Indian Journal of Agricultural Economics 38(4): 475–496.

Sen, S.R.1967. Growth and Instability in Indian Agriculture. Journal of the Indian Society of Agricultural Statistics, 21: 832–833.

How to cite this article:

Soumitra Sankar Das, Soumik Ray, Abhishek Sen, G. Samba Siva and Shantanu Das. 2019. Statistical Study on Modeling and Forecasting of Jute Production in West Bengal, India. *Int.J.Curr.Microbiol.App.Sci.* 8(07): 1719-1730. doi: https://doi.org/10.20546/ijcmas.2019.807.204