Available online at www.ijpab.com

DOI: http://dx.doi.org/10.18782/2320-7051.5838

ISSN: 2320 – 7051

Int. J. Pure App. Biosci. **5 (6):** 1042-1051 (2017)

Research Article

International Journal of Pure & Applied Bioscience

OPEN ACCESS

Sugarcane Productivity in Bihar- A Forecast through ARIMA Model

Mahesh Kumar[1*], Rohan Kumar Raman[2] and Subhash Kumar[1]

[1]Department of Statistics, Mathematics and Computer Application

Dr. Rajendra Prasad Central Agricultural University, Pusa, Samastipur, Bihar -848125

[2]ICAR-Central Inland Fisheries Research Institute, Barrackpore, Kolkata-700120

*Corresponding Author E-mail: mahesh_smca@yahoo.co.in

Received: 5.10.2017 | Revised: 12.11.2017 | Accepted: 16.11.2017

ABSTRACT

This study is undertaken to attempt forecasting the sugarcane (Saccharum officinarum) productivity of Bihar through fitting of well-known Box Jenkins univariate Auto Regressive Integrated Moving Average (ARIMA) model. Time series data on sugarcane productivity in Bihar from 1939-40 to 2014-15 were taken from Sugarcane Research Institute (SRI) Pusa, Bihar and Indian sugar** for the study. The data on sugarcane productivity in Bihar from the year 1940 to 2010 were utilized to build an ARIMA model and validated through five-year productivity data from 2011 to 2015. Akaike information criterion (AIC) was selected for best model selection criteria. ARIMA (0, 1, 1) model found best suitable model for sugarcane productivity in Bihar based on AIC model selection criteria. The performances of models are validated by comparing with actual values of sugarcane productivity in Bihar data. Using developed ARIMA (0, 1, 1) model, two years ahead, year 2016 and 2017 sugarcane productivity in Bihar forecasted showing increasing productivity with 4.22 % and 5.15 % prediction standard error.*

Key words: *Time series, ARIMA, AIC, Sugarcane productivity in Bihar, Forecasting*

*Data source: *Sugarcane Research Institute (SRI), Pusa, Bihar and ** Indian Sugar, 2016.*

INTRODUCTION

Sugarcane (*Saccharum officinarum*) is the leading cash crop of Bihar. In India it is grown in area of 5 (five) million hectares with production of 341.20 million tonne. The average productivity is 68.25 tonne / hectare, 2012-13 (Source: Agricultural Research Data Book 2015). Although available technology was demonstrated that sugarcane productivity can easily be enhanced to the extent at of least 80 tonnes per hectares. It is now a complex scientific activity aimed at producing maximum amount of agricultural produce with minimum expenditure in terms of time, space and energy to meet the needs of a growing population and economy. In Bihar rice is cultivated over an area of about 0.252 million hectare with production of 12.60 million tonne and productivity is 50 tonne /ha, 2012-13 (Source: Indian Sugar) which is much lower than most of the rice growing states of the country.

Cite this article: Kumar, M., Raman, R.K. and Kumar, S., Sugarcane Productivity in Bihar- A Forecast through ARIMA Model, *Int. J. Pure App. Biosci.* **5(6):** 1042-1051 (2017). doi: doi: http://dx.doi.org/10.18782/2320-7051.5838

In spite of recent technological advances, the sugarcane productivity is low. Forecasting of sugarcane productivity is of immense value and plays an important role in many important decisions. The univariate Box-Jenkins[3] approach for forecasting is based on the solid foundation of classical probability theory and mathematical statistics. It is a family of models out of which one appropriate model is selected having optimal univariate forecast. For the purpose of present study, the data on yield (t/ha) of sugarcane has been collected for the period of 76 years i.e. from 1940 to 1915 for Bihar from the Sugarcane Research Institute, Dr. Rajendra Prasad Central Agricultural University, Pusa, Bihar and Indian Sugar, May 2016 for building forecast model and generating short term forecast on sugarcane productivity.

Review of Literature

Several studies have been carried out to develop suitable forecast models for various crops. Most of earlier studies are based on multiple regression technique. Multiple regression techniques violets the principle of parsimony (which is a characteristics of a good model) due to the presence of correlation among the independent variables or multicollinearity, instead of increasing the value of multiple correlation co-efficient significantly. Some review of the research work is cited here.

Agrawal *et al*[1]., have developed a forecast model for predicting rice yield of Raipur district. They have employed stepwise regression technique and explanatory variables included in the model are weekly rainfall, number of rainy days, relative humidity and maximum temperature. George and Kumar[4] have develop pre-harvest forecast of cashew yield by adopting the conventional regression techniques. All these investigations, however, have the inherent drawback due to the presence of multicollinearity and an orthogonal transformation of the explanatory variables seems to be useful to tackle this problem to a certain extent. Gupta[5] has discussed about ARIMA model and forecasts on tea production in India. He developed and applied an ARIMA forecasting model for tea production in India. The model is developed using monthly tea production data in India for the period January 1979 to July 1991 and forecasts are made for the future 12 month periods. Boran and Bora[2] have discussed about the monthly rainfall around Guwahati using a seasonal ARIMA model. The model parameters are estimated using Marquardt algorithm for nonlinear optimization the various stages of model building have been presented in a simple algorithm form. The model is used to predict rainfall for the month ahead and month wise rain fall for the year ahead. Min[7] has discussed about forecasting for the changes in number of hogs and hog's farms. This study was carried out to forecast the changes in the number of pigs and pigs' farms in the Korea, Republic by total and herd size using ARIMA models. In view of the presence of autocorrelation among observations in the three data sets, ARIMA model of various types have been developed separately for describing marine inland and total fish production of India. The identified models are then used to forecast the future fish production.

MATERIALS AND METHODS

Objective of the present study is to develop an adequate forecast model for describing the sugarcane productivity in Bihar. Univariate Box-Jenkin Autoregressive Integrated Moving Average (ARIMA) technique has been applied for obtaining the same. This approach automatically select most reliable forecast model from the family of ARIMA model by going through three iterative stages i,e. Identification stages, estimation stages and diagnostic checking stage. This technique provides a parsimonious model that is a model with smallest number of parameters for describing the available data. The secondary data covering the period from the year 1939-40 to 2014-15 for Bihar. Data for 2011-12 to 2014-15 have been used for computing the forecast error and the rest were used for building the models. Building an ARIMA (p, d, q) model basically consisted of three steps, namely; (a) Identification of the order of the model (b) Estimation of model parameters and (c) Diagnostic checking for adequacy of the fitted model[3].

Mathematically, an ARIMA (p,d,q) model is given by-

$\phi (B) \Delta^d \bar{Z}_t = \Theta (B) a_t$

Where,

Δ^d	=	$(1-B)^d$
$\phi (B)$	=	$(1- \phi B- \phi B^2- \ldots\ldots\ldots\ldots - \phi B^p)$
$\Theta (B)$	=	$(1- \Theta_1 B\ \Theta_2 B^2- \ldots\ldots\ldots\ \Theta_q B^q)$
Z	=	$Z_t - \mu$
Z_t	=	Stationary time series data
d	=	Order of differencing
a_t	=	random shock
p	=	Order of auto-regression
q	=	Order of moving average

Under identification phase, the first order differencing (r=1,2, …) of Z_t is done till a stationary time series is achieved. The order p & q is decided on the basis of ACF & PACF and the criteria led down by Box and Jenkins[3], after determining the value of p, d and q. The model parameter estimated. Diagnostic checking of the fitted model is done through some important statistics such as t-test and χ^2 (Chi-square)[6] of the residual ACF.

A brief description of various models of ARIMA family are cited here:

ARIMA model

ARIMA model is an algebraic statement telling how the observations on a variable are statistically related to past observation on the same variable. In fact, ARIMA model is a family of models consisting of three kinds of model, which are given below;

Autoregressive model; This can be represented as

$Z_t \quad = \quad C + \phi_1 Z_{t-1} + a_t$ …(1)

Where

C	=	μ (1- ϕ1) = Constant term
μ	=	Constant parameter

ϕ = Deterministic coefficient. its value determines the relationship between Zt and Z_{t-1} (Lagged observation)

a_t = Random shock having some continuous statistical distribution.

The term $\phi_1 Z_{t-1}$ is autoregressive term, and the longest lag attached to it is t-1 thus, above is autoregressive model of order 1, denoted as AR (1). The parameters of model (1) are estimated by least square method. Approximate estimates for μ and ϕ_1 can be obtained as Z (mean of the available observation) and r_1 (autocorrelation function) respectively. Similarly, second order autoregressive model denoted as AR (2) can be represented as

$Z_t \quad = \quad C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + a_t$

In this model Z_t is linearly related to the past observation Z_{t-1} and Z_{t-2}. The lease square estimate of ϕ_1 and ϕ_2 are approximated by

$\phi_1 = r_1 (1-r_2) / 1-r^2_1$ and $\phi_2 = r_2 - r^2_1) / 1-r^2_1$

Where,

r_1 & r_2 are autocorrelation function for first and second lag respectively.

In general, one can represent autoregressive model of order p denoted as AR (p) as a liner combination of p-past values and a random term i.e.

$Z_t = C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t$

Moving Average(MA) model: A moving average model of order one denoted as MA (1) can be represented as

$Z_t = C - \Theta_1 a_{t-1} + a_t$ …(2)

Where,

C = μ (1- Θ_1) = constant term

Θ_1 = Moving average coefficient determines the statistical relationship between Z_t and a_t-1 (lagged random shock)

a_t = random shock with mean '0' and variance σ^2.

Estimation of parameters of MA model: Estimation of parameters of MA model is more difficult than an AR model because efficient explicit estimators cannot be found. Instead some numerical iteration method is used. For example, to estimate μ and Θ of equation (2) i.e.

$Z_t \quad = \quad C - \Theta_1 a_{t-1} + a_t$

residual sum of square (RSS) $\sum a^2_t$ in terms of observed Z's and the parameters μ and Θ and them differentiate with respect to μ and Θ to obtain estimates μ and Θ. Unfortunately, the RSS is not a quadratic function of the parameters and so explicit least square estimates cannot be found. An iterative procedure suggested by Box-Jenkins is used in which suitable values of μ and Θ such as μ = Z and Θ given by the solution of equation (3)

$Z_t = C + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} - \Theta_1 a_t\text{-}_1 \ldots - \Theta_q a_{t-q} + a_t$ …(3)

Then the RSS may be calculated recursively from

$$a_t = Z_1 - c + \Theta_1 a_{t-1} \text{ with } a_0=0$$

This procedure then can be repeated for a grid of points in (μ, Θ) plane. We may then by inspection choose that value of (μ, Θ) as estimates which minimized RSS. The lease square estimates are also maximum likelihood estimated conditional on a fixed value of a_0 provided a_t is normally distributed.

Autoregressive Moving Average Model (ARMA): The combination of AR (p) and MA (q) models to describe a given series is known as ARMA (p, q) which can be represented as

$$Z_t = C + \phi_1 Z_{t-1} + \dots + \phi_p Z_{t-p} - \Theta_1 a_{t-1} \dots - \Theta_q a_{t-q} + a_t$$

The Box-Jenkins modeling procedure

Box-Jenkins proposed a practical three stage procedure for finding a good model. a sketch of the broad outline of the Box-Jenkins modeling procedure is summarized schematically in figure 1

Stage 1 Identification of the order of the model:

It $\bar{Z}$ be the mean of a stationary time series such that $Z_t = Z_t - \bar{Z}$ denoting the number of observations by n and the number of computable lags by k the estimated autocorrelation function (ACF) r_k of the observations separated by k time periods.

Estimation of parameters of MA model: Estimation of parameters of MA model is more difficult than an AR model because efficient explicit estimators cannot be found. Instead some numerical iteration method is used. For example, to estimate μ and Θ of equation (2) i.e.

$$Z_t = C - \Theta_1 a_{t-1} + a_t$$

residual sum of square (RSS) $\sum a^2_t$ in terms of observed Z's and the parameters μ and Θ and them differentiate with respect to μ and Θ to obtain estimates μ and Θ. Unfortunately, the RSS is not a quadratic function of the parameters and so explicit least square estimates cannot be found. An iterative procedure suggested by Box-Jenkins is used in which suitable values of μ and Θ such as μ = Z and Θ given by the solution of equation (3)

$$Z_t = C + \phi_1 Z_{t-1} + \dots + \phi_p Z_{t-p} - \Theta_1 a_{t-1} \dots - \Theta_q a_{t-q} + a_t \qquad \dots(3)$$

Then the RSS may be calculated recursively from

$a_t = Z_1 - c + \Theta_1 a_{t-1}$ with $a_0=0$

This procedure then can be repeated for a grid of points in (μ, Θ) plane. We may then by inspection choose that value of (μ, Θ) as estimates which minimized RSS. The lease square estimates are also maximum likelihood estimated conditional on a fixed value of a_0 provided a_t is normally distributed.

Autoregressive Moving Average Model (ARMA): The combination of AR (p) and MA (q) models to describe a given series is known as ARMA (p, q) which can be represented as

$$Z_t = C + \phi_1 Z_{t-1} + \dots + \phi_p Z_{t-p} - \Theta_1 a_{t-1} \dots - \Theta_q a_{t-q} + a_t$$

The Box-Jenkins modeling procedure

Box-Jenkins proposed a practical three stage procedure for finding a good model. a sketch of the broad outline of the Box-Jenkins modeling procedure is summarized schematically in figure 1

Stage 1 Identification of the order of the model:

It $\bar{Z}$ be the mean of a stationary time series such that $Z_t = Z_t - \bar{Z}$ denoting the number of observations by n and the number of computable lags by k the estimated autocorrelation function (ACF) r_k of the observations separated by k time periods.

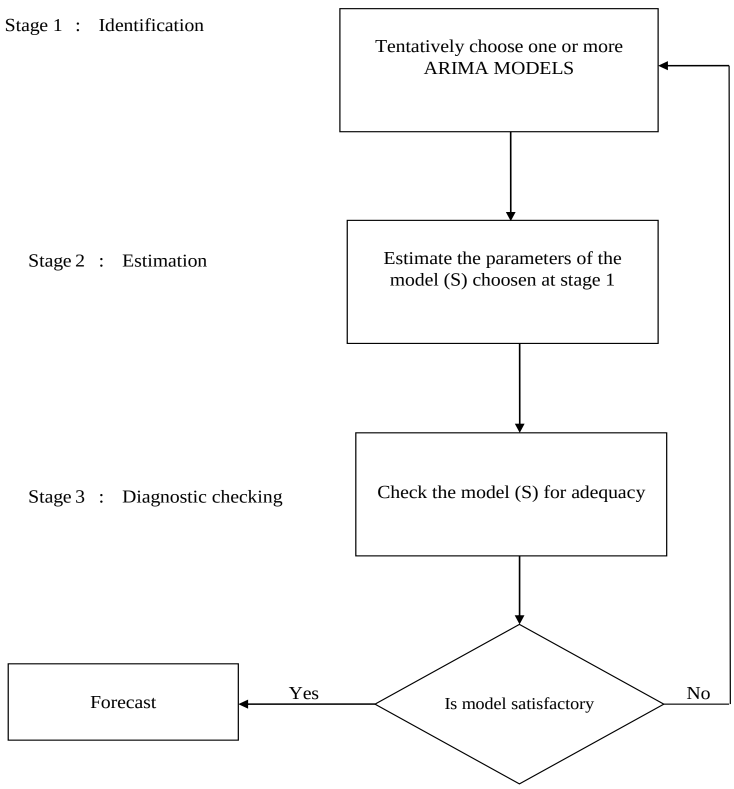

Fig. 3.1 Stages in the Box-Jenkins iterative approach to model building

Stage 2 Estimation of model parameter;

Box-Jenkins time series models written as ARIMA (p, d, q) amalgamate three type of processes namely auto-regressive (AR) or order p; differencing to make a series stationary of degree d and moving average (MA) of order q. At the parameter estimation stage the aim is to obtain estimates of the tentatively identified ARMA model parameters of Stage-I for given values of p and q. In general, ARIMA coefficients (the ϕ 's and ϴ's) must be estimated using a nonlinear least square procedure, while several nonlinear least square methods are available, the one most commonly used t estimate ARIMA models is known as "" Marquardt's compromise".

Stage 3 Diagnostic checking for the adequacy of the model[3]:

This is the third stage of model formulation. At this stage the decision about the statistical adequacy of the model is taken. Most important test of the statistical adequacy at an ARIMA model involves the assumptions that the random shocks (a_t) are independent. Meaning not autocorrelate, since in practice the random shocks cannot be observed the estimate at residual(a_t) is taken in to test the hypothesis about the independent of random schocks. This mainly performed by the examination of residual ACF, t test for the residual ACF and χ^2 –test based on L-Jung and Box for the residual autocorrelation.

RESULTS AND DISCUSSION

This study is undertaken to attempt forecasting the sugarcane productivity of Bihar through fitting of well-known Box Jenkins univariate Auto Regressive Integrated Moving Average (ARIMA) model. Time series data on sugarcane productivity in Bihar from 1939-40 to 2014-15 were taken from Sugarcane Research Institute (SRI)*Pusa , Bihar and Indian sugar** for the study. The data on sugarcane productivity in Bihar from the year 1940 to 2010 were utilized to build an ARIMA model and validated through five-year productivity data from 2011 to 2015.

ARIMA model for sugarcane productivity in Bihar:

Model Identification:

The first and foremost stage in the identification of an ARIMA model judging the Stationary behavior of the under lying process. The graph of the time series data (Fig-1 to Fig.- 3) indicates the non-stationary of the underlying process. The stationary was achieved by the first order differentiation of the original time series data (Fig- 4 to Fig. 6). The autocorrelation and partial autocorrelation up to lag 15 (fifteen) are presented in Table- 1

In addition to the time series plot of the original series Z_t (Fig. 1), the non-stationary behavior is also indicated by the failure of auto correlation value for Z_t to die out rapidly (Table 1) autocorrelation of (ΔZ_t) are however small after the first lag and showed a cut off after lag 1. This suggest first order difference are stationary and hence d = 1. For further modeling first order difference are therefore considered. The ACF of the first order difference drops off after lag 1. Since the autocorrelation function has exponentially decaying pattern, its partial autocorrelation functions cuts off to zero after lag 1. This suggest an autoregressive model of the order 1. Since in practice P is usually not larger than 2 or 3 for non-seasonal models, all models up to the order of 2 have been fitted. The ACF of first order difference drop after lag 1. Since autocorrelation function has exponentially decaying pattern and its partial autocorrelation function cuts off to zero after lag 1 which suggest ARIMA (0,1,1).

The ACF and PACF plot of the data showed the non-stationarity (Fig.- 1, 2 and 3)

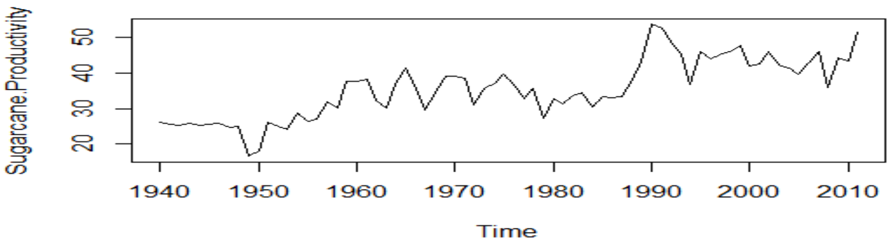

Fig. 1:

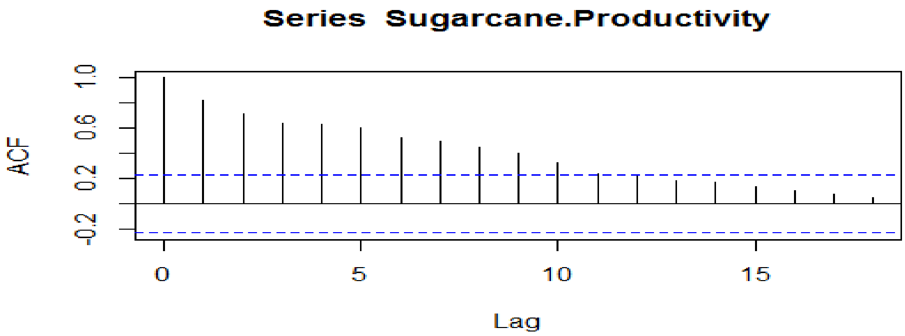

Fig. 2:

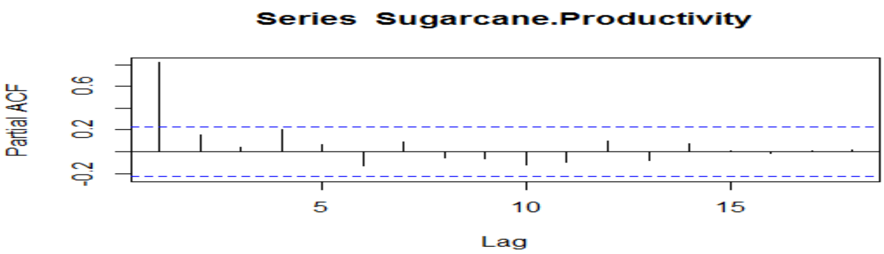

Fig. 3:

After first order differencing the data showed stationary; d=1(Fig. 4.4, 4.5 and 4.6)

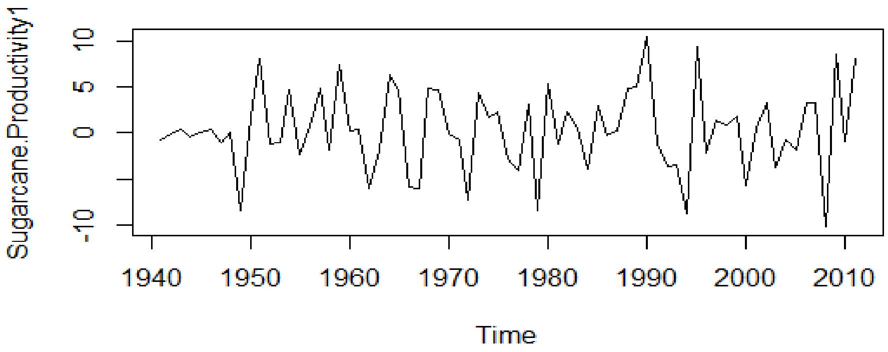

Fig. 4:

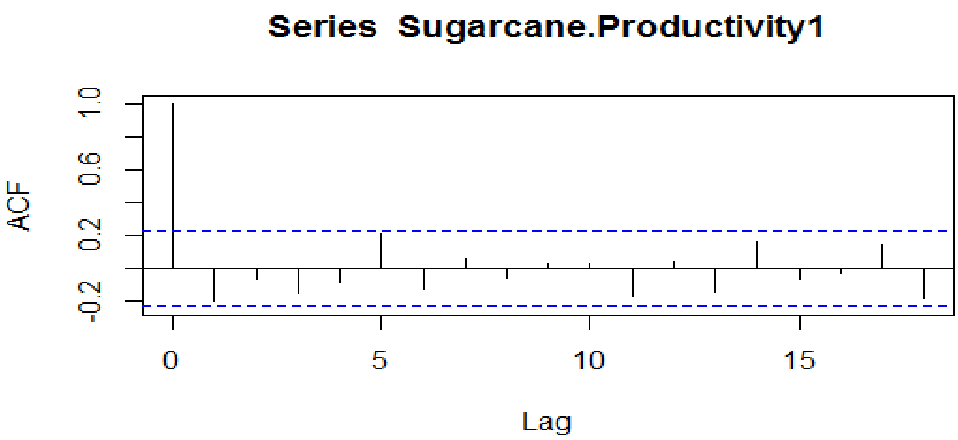

Fig. 5:

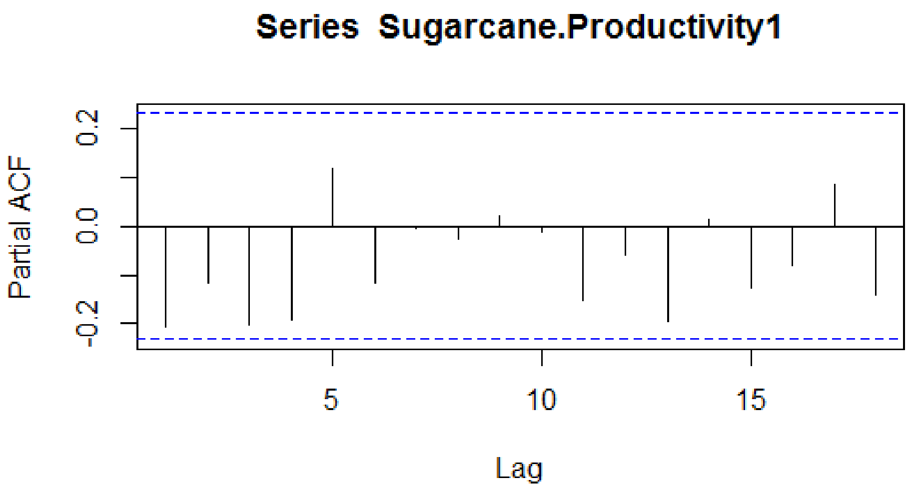

Fig. 6:

Only one line at lag 5 in ACF function is touching the significant line so q=1. PACF function, no line touching the significance level, hence p=0. Hence our ARIMA model will be ARIMA (0,1,1)In figure-7 the graph shows forecast trend likely to be constant with increasing interval trend for the year 2016 and 2017. The figure- 8, graph shows the well fitted and forecast curve are likely to be superimposed to observed. This shows that model (0,1,1) is good fitted.

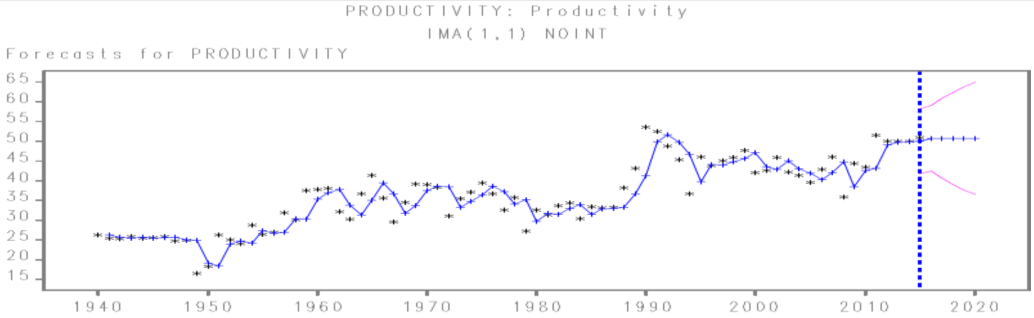

Fig. 7:

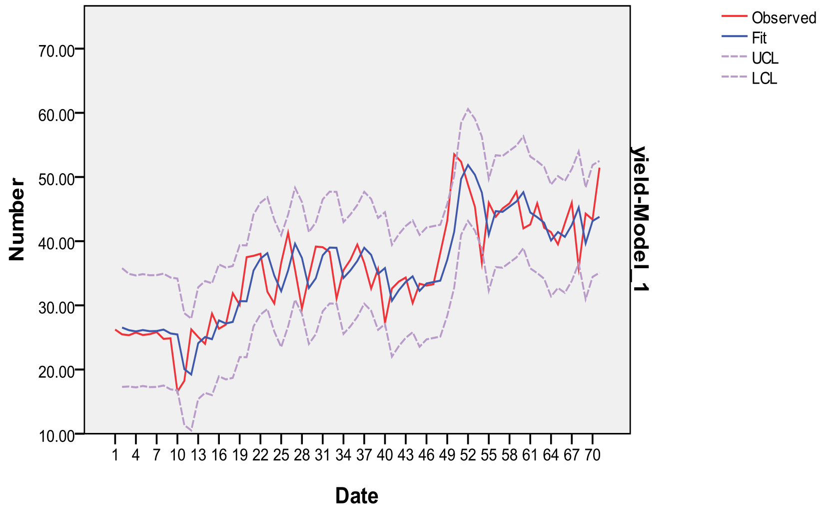

Fig. 8:

1. Estimation of model parameter: The various parameters of the different models have been presented in Table- 3. The tentative selection of model is presented as ARIMA (0,1,1) which containing only one moving average term θ_1. The output of fitted model has been presented in Table- 3. The invertibility condition have been satisfied. Since $|\theta| < 1$. The significant estimated coefficient θ_1 indicates the parsimonious of the model. Since the correlation coefficient between the estimates of the parameters is not high, the resultant model can be considered as stable.

2. Selection of good model: A good model should be parsimonious, stationary and invertible. The estimated coefficients should be of high quality and stable. The five models under comparison fulfills the stationary and invertibility condition wherever necessary. For selection of the parasimonious model, a guiding principle suggested by Box and Jenkins has been used. Based upon the significance of estimated coefficient ARIMA (1,1,0), ARIMA (0,1,1) and ARIMA (2,1,0) results in a parsimonious model. A comparison of different models suggests ARIMA (0, 1, 1) Table- 3. Since it fulfils the invertibility conditions, parsimoniousness, stability of estimates of the parameters and has lowest AIC, MAPE, RMSE, BIC and highest R^2 value in comparison to other parasimonious models.

3. Diagnostic checks: For deciding the statistical adequacy of the model diagnostic checks with respect to the independence of random shocks (a_t) has been performed. A statistical adequate model is one whose random shocks are not auto correlated. For this purpose, residual autocorrelation function (ACF) has been calculated and presented in the output Tables 2. A t-test has been performed to test the significance of null hypothesis H_0: $\hat{\rho}_k$ (a) = 0 for each residual autocorrelation coefficient. The respective standard error has been computed using Bartlettes (1946) approximation formulae. The non-significance of residual ACF indicates the independence of random shocks. The independence of the random shocks is also confirmed by a Chi- squares test suggested by L-Jung and Box- test using Q-statistics[6] which comes to be 22.25 (Non-significant Chi-square value at 17 d.f.) for the selected model ARIMA (0,1,1) Table 2. Thus the selected ARIMA model of the order (0,1,1) or MA (1) seems to be appropriate. The forecast error for the one step ahead and two step ahead has been computed as 4.2210% and 5.1509 respectively (Table 4)

Computation of forecast and their confidence intervals: After confirming the validity of the model it is used to forecast the future values of the observed time series. Thus using ARIMA (0,1,1) the forecast along the confidence intervals have been computed for the 5 periods ahead and it presented in Table 4. The table shows the increasing trend of sugarcane productivity in Bihar by the year 2016 and 2017. The sugarcane productivity of Bihar is expected to be 50.0698 tonne /hac for both the years 2016 and 2017 with a confidence limit of 58.9638 to 42.4179 and 60.7864 to 40.5953 respectively.

Table 1: Autocorrelation (ACF) and partial autocorrelation (PACF)

Lag	ACF	PACF
1	-0.196	-0.196
2	-0.169	-0.216
3	-0.074	-0.170
4	-0.022	-0.133
5	0.100	0.012
6	-0.016	-0.032
7	0.042	0.053
8	-0.102	-0.078
9	0.068	0.056
10	-0.084	-0.100
11	-0.060	-0.110
12	-0.031	-0.145
13	-0.067	-0.186
14	0.289	0.173
15	-0.273	-0.257

Table 2: Output of fitting ARIMA (0,1,1) for Bihar

Parameter Estimates

Parameter	Estimates	Standard Error	t-value
Θ_1	0.335	0.118	3.022^{**}
Constant	- 5.012	5.119	-0.979

Autoregressive factor

$$\theta(B) = 1+0.335B$$

Forecast model

$$Z_t – Z_{t-1} = - 5.012 - 0.335 (Z_{t-1} – Z_{t-2}) + a_t$$

Diagnostic Check

Lags	**Residual ACF**	**SE**
1	0.071	0.120
2	-0.126	0.120
3	- 0.233	0.122
4	-0.078	0.128
5	0.110	0.129

6	0.008	0.130
7	0.042	0.130
8	-0.053	0.130
9	-0.008	0.131
10	0.136	0.131
11	-0.140	0.133
12	-0.070	0.135
13	-0.059	0.135
14	0.257	0.136
15	-0.125	0.143

Q-statistics (L Jung Box Test) = 22.250
D.F. = 17

Table 3: Best fitting model for Bihar

Model	AIC	MAPE	RMSE	BIC	R^2
ARIMA(1,1,0)	414.14	9.37	4.421	3.155	0.722
ARIMA(0,1,1)	412.88	9.16	4.366	3.130	0.729
ARIMA (1,0,1)	422.08	9.73	4.521	3.258	0.715
ARIMA (0,1,0)	415.09	9.66	4.484	3.122	0.710
ARIMA (2,1,0)	415.27	9.17	4.443	3.225	0.723

Table 4: Forecast and their confidence intervals of ARIMA (0,1,1) for Bihar

Periods	Forecast	95% limits		Actual	Forecast standard error
		Lower	Upper		
2011	43.1260	51.3990	34.8531	51.4600	4.2210
2012	48.9547	57.2276	40.6817	50.0000	4.2210
2013	49.6857	57.9587	41.4128	50.0000	4.2210
2014	49.9055	58.1785	41.6326	50.0000	4.2210
2015	49.9716	58.2446	41.6986	51.0000	4.2210
2016	50.6908	58.9638	42.4179		4.2210
2017	50.6908	60.7864	40.5953		5.1509

REFERENCES

1. Agarwal, *et al*., Forecasting of rice yield using climatic variables. *J. Indian Society of Agricultural Statistics. Models* **30:** 145-146 (Abstract) (1978).
2. Boran, D.K. and Bora, P.K., Predicting the monthly rainfall around Guwahati using a seasonal ARIMA model. *J. Indian Society of Agricultural Statistics*, **47:** 278-287 (1995).
3. Box, G.E.P. and Jenkins, G.M., Time series analysis: Forecasting and control 2 nd Ed Holder-Day, San Francisco (1976).
4. George, M.V. and Kumar, Vijay K., Forecasting of cashew yield, *J. Indian Society of Agricultural Statistics*, **31:** 95-96 (1979).
5. Gupta, G.S., ARIMA Model forecasting on tea production in India. *The Indian Economic journal*, **41(2):** 88-110 (1993).
6. L-Jung, G.M. and Box, G.E.P., On a measure of lack of fit in time series models. *Biometrika*, **65:** 297-303 (1978).
7. Min, B.J., Forecasting for the changes in number of hogs and hog farms. *Korean Jouranal of animal* sciences, **37:** 558-566 (1995).