FORECASTING ANALYSIS FOR SUGARCANE PRODUCTIONIN TAMILNADU USING ARIMA MODEL

C. Devaki[1], Dr. A. Kachi Mohideen[2].

[1]Assistant Professor, Department of Statistics,
Bharathiyar Arts & Science College for Women, Attur, Selam – 636 112.Tamilnadu, India.

[2]Assistant Professor, Department of Statistics,
Periyar EVR College (Autonomous), Trichy – 620 023. Tamil Nadu, India.

Abstract

The paper describes an Empirical study of modeling and forecasting time series data of sugarcane production in Tamil Nadu. Box-Jenkins ARIMA methodology has been used for forecasting. The data for a period of 1961-2015 was to develop the forecast model. The order of the best ARIMA model was found to be (2,1,0).Further, efforts were made to forecast, as accurate as possible, the future sugarcane production for a period up to five year by fitting ARIMA (2,1,0) model to our time series data.

Key words: ARIMA, Forecasting, Sugarcane Production and Auto- regressive process.

1. Introduction

The aim of standard regression model is to explore an association between department and independent variables to identify the impact of this covariate on the response. The classical linearregression model is an important statistical tool but its use is limited because of its standard assumptions. Regression models using time series (TS) data occur quite often and the assumption of uncorrelated or independent errors for time series data is often not appropriate. The ARIMA model with special applications extends the regression with ARIMA errors classof models, Vector autoregressive with exogenous variable (VARX) and vector ARIMA (VARMA etc. Time series models have advantages in certain situations. They can be used more easily for forecasting purpose because the historical sequence of observations upon study variable are readily available at equality spaced intervals over discrete point of time. These successive observations are statistically dependent and Time Series modeling is concerned with techniques for the analysis of such dependence.

The production of sugarcane is fluctuated from year to year due to fluctuation of area under sugarcane cultivation is given by M.N.Shekh and M.M.Haque (1986). Yield and production could not be increased to the desired level due to various bottlenecks in production and marketing of sugarcane. Forecast of sugarcane production are discussed by M.D.Moyazlem Hossain and Farug Abdult(2015) S.R.Krishna priya and K.K.Suresh (2009-2010),Kumar Manoj and Anand madhu(2013), Faqur Muhammad *et.al.,* (1992). In this paper, we are forecasting the sugarcane production in Tamil Nadu.The data are collected chronologically 1961-2015.The data analysis by using ARIMA model in SPSS.

3. Model Description.

An ARIMA model is characterzed by the notation ARIMA (p,d,q); where p, d and q denote the orders of auto-regression, differencing and moving respectively. For a given TS process $\{Y_t\}$, a first order auto- regressive process denoted by ARIMA (1,0,0) or simply AR(1) is,

$$Y_t = \mu + \phi_1 Y_{t-1} + e_t$$

And a first order moving average process denoted by ARIMA (0,0,1) or simply MA(1) is given by

$$Y_t = \mu + \phi_1 e_{t-1} + e_t$$

Alternatively, the model may be mixture of these procedure and of higher orders as well.

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + C Y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} + e_t$$

This is called a mixed auto regressive moving average model of order(p,q). It contains both AR and MA terms.

Alternatively, an ARIMA (p,d,q) may be written in polynomial form as:

$$\phi_p(B)\Delta^d Y_t = c + \theta_q(B) e_t$$

Where

Y= Variable under forecasting

B=Lag operator

e =Error term

t =Time subscript

ϕ_p= non-seasonal AR, the autoregressive component represented as a polynomial in the back shift operator

$(1 - B)^d$ =Non- seasonal difference

$\phi_p(B)$ =Non – seasonal MA, the moving – average operator, represented as a polynomial in the back shift operator

$\phi's$ and θ's are the parameters to be estimated and c is a constant related to the mean of the process

3.1. Autocorrelation function

Autocorrelation refers to the way the observations in TS process are related to each other and is measured by simple correlation between current observation (Y_t) and observation from K periods before the current one is(Y_t, Y_{t-k}). Thus, for a given series Y_t, autocorrelation at lag k i.e. the correlation between (Y_t, Y_{t-k}) is given by

$$r_k = \frac{\sum_{t=1}^{n-k}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

It ranges from -1 to +1. Box and Jenkins has suggested that maximum number of useful r_k are roughly n/4 where n is the number of time- period.

3.2. Partial autocorrelation function

Partial autocorrelation function (pacfs) are used to measure the degree of association between Y_t and Y_{t-k} when the Y-effects at other time lags 1,2,3,…., k-1 are taken care of the general characteristics of theoretical acfs and pacfs as follows to AR, MA and ARMA are three stage iterative process of ARIMA model building may be described as follows.

3.3. Model identification

The foremost step in the process of ARIMA modeling is to check for stationary of theseries as the estimation procedures are available only for stationary series. There are two kinds of stationary viz., stationary in 'mean' and stationary in 'variance'. A look at the graph of TS data and structure of acfs and pacfs may provide clues for the presence of stationary. The stationary could be achieved by differencing the original series if the data series is found to be non-stationary. This is applicable for both seasonal and non-seasonal TS data. Thus, if 'Y_i' denotes the original series, the non-seasonal difference of first order is$Y_t = Y_t - Y_{t-1}$. The next step in the identification process is to find the initial values of the initial values of the orders of seasonal/non-seasonal parameter i.e., P,Q and p,q i.e., one or more models are tentatively chosen that seem to provide statistically adequate representation of the available data.

3.4. Parameter estimation

At the estimation stage, it is attempted to obtain precise estimates of a small numberof parameters of the model. Box and Jenkins is use of least-square method. Linear least-square may be used to estimate only pure AR models. All other models require a non-linear least-square method. Marquardt (1963) has designed a powerful algorithm for estimating ARIMA model through iterative improvement where some preliminary estimates are chosen and then the computer programmed refines them iteratively so as to minimize the sum of squared residuals.

Different ARIMA model can be obtained for various combinations of AR and MA orders individually/collectively. Low Akaike Information Criteria(AIC)/ Schwarz-Baysian Information Criteria (SBC) may be preferred to choose a suitable ARIMA model among the alternative ones i.e., $\mathrm{AIC}=(-2\log L + 2m), \text{where} m = p + q + P + Q$ and L is likelihood function. Since $-2\log L$ is approximately equal to $\{n(1 + \log 2\pi) + n\log\sigma^2\}$ and σ^2 is the model MSE. Thus, AIC may be written as $\{n(1 + \log 2\pi) + n\log\sigma^2 + 2m\}$.

As an alternative to AIC, sometimes SBC is also used which is given by $\{\log \sigma^2 + (m \log n)/n\}$.This stage also provides some warning signals if the estimated coefficients do not satisfy certain mathematical inequality conditions as expressed below.

3.5. Checking for stationary condition

To check for stationary, the estimated AR coefficients are to be examined if they satisfy the stationary conditions

For AR (1),

$$|\phi_1| < 1$$

For AR (2),

$$|\phi_2| < 1$$

$$\phi_2 + \phi_1 < 1$$

$$\phi_2 - \phi_1 < 1$$

……

……

For AR(p), $\phi_2 + \phi_1 + \phi_3 + \cdots + \phi_3 < 1$

Inevitability condition that ARIMA model must satisfy is called invertibility. This requirement implies that the MA coefficients must satisfy certain conditions. All pure AR processes are invertible and no further checks are required.

For MA(1),

For AR(1),

$$|\phi_1| < 1$$

For AR(2),

$$|\phi_2| < 1$$

$$\phi_2 + \phi_1 < 1$$

$$\phi_2 - \phi_1 < 1$$

For moving average models of order greater than 2, the invertibility conditions become complicated. However these models do not frequently occur in practice. For higher order models, the necessary(but not sufficient) condition for inevitability may be checked whether

$$\theta_1 + \theta_2 \ldots\ldots + \theta_p < 1$$

3.6. Diagnostic checking

After the tentative model has been fitted to the data, it is important to perform diagnostic checks to test the adequacy of the model and if need be, to suggest potential improvements. One way to examine the goodness of fit is by plotting the residual acfs of the fitted model. If most of the sample autocorrelation coefficients of the residuals are within the Limits $\pm 1.96/\sqrt{n}$ where n is the number of observations upon which the model is based then the residusls are white noise indicating that the model is a good fit. An alternative way to accomplish this through the analysis of residual. The residual acfs should ideally have autocorrelation coefficients that are all statistically zero. According to Pankratz (1991), all residual acfs should be zero but actually all of them need not be necessary zero because of sampling error.The residual acfs along with t test and chi-square test using Bartlett's approximation estimated correlations. Ljung and Box (1978) suggested a test statistic based on all the residual autocorrelations.

$$Q^* = n(n+2)\sum_{k=1}^{p}(n-k)^{-1} r_k^2 \hat{e}_t with (p-m) d.f.$$

The statistic Q^* approximatelynfollows a chi-squared distributionwith (p-m) degrees of freedom, where nis the total number of observations used to estimate the model, p is the number of residual autocorrelation and m is the number of parameters estimated in ARIMA model.

The fitted Arima (0,1,1) model for selam and Namakkal districts may be elaborated as below:

$$(1-B)Y_t = (1-\theta_1 B)e_t$$

$$Y_t - BY_t = e_t - \theta_1 B e_t$$

$$Y_t = Y_{t-1} - \theta_1 e_{t-1} + e_t \cdots (1)$$

For Namakkal, Erode and Krishnagiri districts, the fitted ARIMA (1,1,0) model may be expressed as follows.

$$(1-\phi_1)(1-B)Y_t = e_t$$

$$Y_t - (1-\phi_1)BY_t + \phi_1 B^2 Y_t = e_t$$

$$Y_t - (1-\phi_1)Y_{t-1} + \phi_1 Y_{t-2} = e_t$$

$$Y_t = (1+\phi_1)Y_{t-1} - \phi_1 Y_{t-2} = e_t \cdots (2)$$

The equations 1 and 2 are the corresponding forecast equations. The presence of lagged values of dependent variable and random show in equ.2 having the lagged values of dependent variable indicates the presence of only autoregressive component. Finally, a comparison between ARIMA based yield estimate with observed /DOA yield estimates was made in terms of percent relative deviations. The results presented in Table 5 indicate the percent deviations of the estimated yield(s) from the observed yield(s). A graphical view of observed and estimated sugarcane yield for all the districts is presented in figure 5.

4. Results and Discussion

The Box Jenkins' methodology was applied in obtaining the suitable ARIMA models for district-level sugarcane yield forecasting in Haryana. Autocorrelation functions of sugarcane yield shown in Figure1 indicated that the data series were non-stationary for all the districts under consideration. Differencing of order one was sufficient for making an appropriate stationary series.

The orders of AR and MA components were deter-mined through acfs and pacfs of the stationary series. Marquardt algorithm (1963) was used to minimize the sum of squared residuals. Log Likelihood, AIC (1969), Schwarz's Bayesian Criterion, SBC (1978) and residual variance decided the criteria for the selection/estimation of AR and MA coefficients in the model. The residual acfs along with the Chi-square test (Ljung and Box, 1978) were used to ascertain the random shocks as white noise.

After experimentation with different lags of moving average and autoregressive processes, ARIMA (0,1,1) Namakkal, Erode and Krishnagiri districts were fitted for achieving sugarcane yield forecasts.

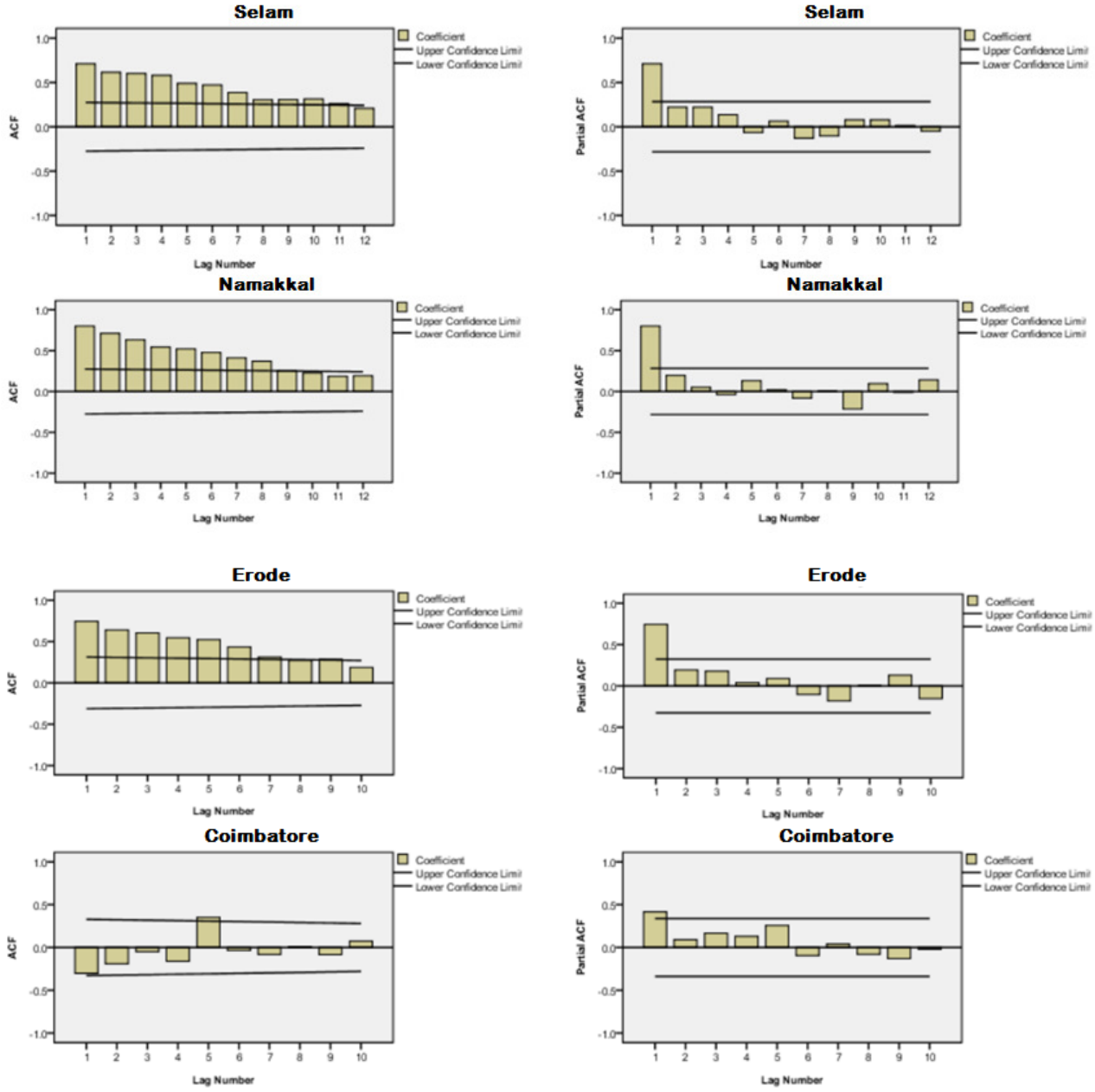
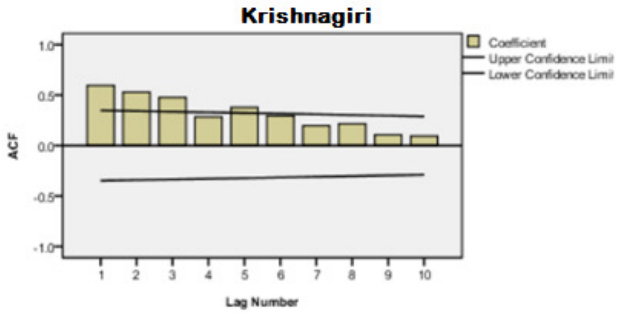


Fig-1: Autocorrelation and partial autocorrelation of sugarcane yield for all the districts.



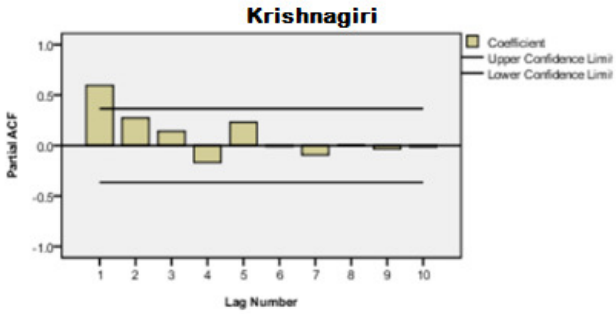


Table 1: Parameter estimates of fitted ARIMA models.

District/model	Constant Difference	Estimate	Standard error	t-value
Selam, Sug- ARIMA (0,1,1)	MA log 1	0.67	0.14	<0.01
	Constant	1		
	Difference	0.85	0.09	<0.01
Namakkal Sug- ARIMA (0,1,1)	MA log 1	0.93	0.20	<0.01
	Constant	1		
	Difference	0.74	0.10	<0.01
Erode Sug- ARIMA (1,1,0)	MA log 1	0.63	0.91	0.49
	Constant	1		
	Difference	-0.36	0.16	0.02
Coimbatore Sug- ARIMA (1,1,0)	AR log 1	0.79	1.03	0.44
	Constant	1		
	Difference	-0.36	0.18	0.06
Krishnagiri Sug- ARIMA (1,1,0)	AR log 1	0.80	0.80	0.32
	Constant	1		
	Difference	-0.48	0.17	0.01

Table -2: Youle-Walker estimates of selected AR models

Districts	Selected Autoregressive order				
	Lag-1	Lag-2	Lag-3	Lag-4	Lag-5
Selam,	-0.806	-0.709	-0.535	-0.306	-0.284
Namakkal	-0.709	-0.604	-0.402	-0.252	
Erode	-0.601	-0.408	-0.336	-0.484	
Coimbatore	-0.578	-0.293	-0.279	-0.457	-0.265
Krishnagiri	-0.359				

Table 3: Sugarcane yield estimates and Selection Criteria values for ARIMA models

District	Models	RSME	MAPE	SBC
Selam	ARIMA (1,1,1)	5.72	9.54	3.73
	ARIMA (0,1,1)	5.66	9.55	3.63
	ARIMA (1,1,0)	6.47	10.36	3.89
Namakkal	ARIMA (1,1,1)	6.4	9.62	4.01
	ARIMA (0,1,1)	6.31	9.68	3.88
	ARIMA (1,1,0)	7.54	10.91	4.23
Erode	ARIMA (1,1,1)	5.58	7.51	3.79
	ARIMA (0,1,1)	5.49	7.42	3.64
	ARIMA (1,1,0)	6.39	9.11	3.94
Coimbatore	ARIMA (1,1,1)	5.14	8.86	3.51
	ARIMA (0,1,1)	5.12	9.01	3.42
	ARIMA (1,1,0)	5.5	9.99	3.57
Krishnagiri	ARIMA (1,1,1)	6.74	9.49	4.13
	ARIMA (0,1,1)	6.67	9.54	4.01
	ARIMA (1,1,0)	8.06	10.18	4.38

Results on Stationary and Invertibility conditions for AR and MA coefficients.

Variable (Sugarcane yield)	Model	Stationarity	Invertibility
Selam	ARIMA (0,1,1)	*	0.85
Namakkal	ARIMA (1,1,0)	-0.38	**
Erode	ARIMA (1,1,0)	-0.49	**
Coimbatore	ARIMA (0,1,1)	*	0.74
Krishnagiri	ARIMA (1,1,0)	-0.36	**

*Stationarity condition is not applicable since the model is MA model,*Invertibility condition is not applicable since the model is AR model

It is clear from the Table 3 that both the Stationarity and invertibility conditions are satisfied because absolute values of AR and MA coefficient for all the districts are less than one.

The diagnostic check involved testing whether the residuals from the estimated equations were white noise. The model verification was concerned with checking the residuals to see if they contained any systematic pattern which can be removed to improve the autocorrelation coefficients using Bartlett's approximation for the standard error of the estimated autocorrelations. All chi-squared statistics in this concern.(Table 4) showed that none of the residual acfs in any of the districts were significantly different from zero at a reasonable level. This ruled out any systematic pattern in the residuals.

Table 4: Diagnostic checking of residual autocorrelations.

Districts	Models	Ljung-Box Q statistic(s)	
		Statistic	Sig.
Selam	ARIMA (0,1,1)	14.23	0.65
Namakkal	ARIMA (1,1,0)	21.26	0.21
Erode	ARIMA (1,1,0)	20.09	0.27
Coimbatore	ARIMA (0,1,1)	24.11	0.12
Krishnagiri	ARIMA (1,1,0)	16.73	0.47

After experimenting with different lags of moving average and autoregressive process; ARIMA (0,1,1) for selam and Coimbatore districts and ARIMA (1, 1, 0) for Namakkal, Erode and Krishnagiri districts were found to be best fit models for sugarcane yield production in the state.

Table - 5: Sugarcane yield estimates and their associated percent relative

Deviations based on ARIMA models

Linear mixed Modeling															
Year	Selam			Namakkal			Erode			Coimbatore			Krishnagiri		
	Observe yield (q/ha)	Estimate yield (q/ha)	RD (%)	Observe yield (q/ha)	Estimate yield (q/ha)	RD (%)	Observe yield (q/ha)	Estimate yield (q/ha)	RD (%)	Observe yield (q/ha)	Estimate yield (q/ha)	RD (%)	Observe yield (q/ha)	Estimate yield (q/ha)	RD (%)
2015	79.77	69.60	12.75	74.16	74.67	-0.69	75.99	70.40	7.36	67.22	67.15	0.10	68.29	60.25	11.77
2016	78.38	70.29	10.32	69.93	75.75	-8.32	83.72	71.29	14.84	71.58	68.09	4.88	66.02	60.52	8.34
2017	81.6	70.93	13.07	77.09	76.70	0.51	74.26	72.14	21.91	79.68	68.98	13.43	74.01	60.73	17.94
2018	78.81	71.56	9.20	75.47	77.61	-2..83	76.91	72.98	5.12	71.23	69.86	1.93	68.66	60.93	11.26
2019	85.04	72.18	15.12	81.64	78.49	3.85	83.56	73.80	11.68	70.55	70.72	-0.24	69.9	61.12	12.57
Average absolute percentage deviation			11.34			33.24			12.18			4.12			12.38

$$RD\% = 100\ X\ (\{observed\ yield - Estmated\ yield\}/\ observed\ \ yield)$$

5. Conclusion

This study is development of ARIMA model for sugarcane yield estimation in selam,Namakkal, Erode and Krishnagiri districts in Tamilnadu. The model fitted using the sugarcane yield from 1961-2015. Under ARIMA methodology involved the determination of appropriate orders of AR and MA polynomials i.e., the values of q and q. The sugarcane yield(s) data were found to be non-stationary for all the districts. Thus differencing of orders were determined from autocorrelation and partial autocorrelation functions of the stationary series with different lags of autoregressive and moving average process, ARIMA (0,1,1) for selam and Coimbatore districts and ARIMA (1, 1, 0) for Namakkal, Erode and Krishnagiri districts were fitted. Different statistics viz., RMSE, mean absolute percent deviation to select AR and MA coefficients of the model. Parameter estimations of the selected models satisfied the stationary and invertibility conditions under ARIMA structure.

References

1. Dickey, D.A. and W.A. Fuller. (1979). Distribution of the estimators for autoregressive time series with a unit-root. *Journal of the American Statistical Association,* Vol.74,pp.427-431.
2. Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* Vol, 21, pp.243-247.
3. Meinhold R.J. and Singpurwalla N.D. (1983). Understanding the kalman filter. *Amer. Statist.* Vol.37, pp.123-127.
4. Faqir Muhammed,Muhammad Siddique Javed and Mujahid Bashlr.(1992). Forecasting sugarcane production in Pakistan using Arima models. *Pakistan journal Agriculture Science,* Vol.9 (No.1).pp.31-36.
5. Box,G.E.P. and G.M.Jenkins.(1970). Time series Analysis: Forecasting and control. *Holdwn-Day, San Francisco, California,USA*.
6. M.N.Sheikh and M.M.Haque. (1986). Production consumption and demand for sugar industry in Bangladesh.Eds Shalauddinn *et.al., BSFIC, Shilpa Bhaban, Motijheel commercial Area, Bangladesh.*
7. Naidu, M.G., P.Balasiddamuni, R.Abbaiah, T.Gangaram and T.Sudha (2013). A modified Box-Jenkins methodology for forecasting, *International Journal of Agricultural Statistical Science,* Vol. 9(2), pp. 481-491.
8. Bera, M.K.K.Chakravarty, Md. Shahjahan and S.Nandi (2002). Area, Production and Productivity of rice in Major Rice Growing Districts of west Bengal during Nineteen Eighties, *Economics Affairs,* Vol, 47(2), pp.108 – 114.
9. Mohammed.A.H. (2014). Forecasting major fruit crops -productions in Bangladesh using Box-Jenkins ARIMA model, *Journal of Economics and sustainable development,* Vol. 5(7), pp.96-107.