Research Journal of Chemical and Environmental Sciences
Res J. Chem. Environ. Sci. Vol 9 [1] February 2021: 08-14
Online ISSN 2321-1040
CODEN: RJCEA2 [USA]

Website: www.aelsindia.com/rjces.htm

RJCES

ORIGINAL ARTICLE

Statistical Modeling for Prediction of Sugarcane Yield in India using ARIMA Model

Ajay Kumar[1], Veer Sain[2], P.K. Muhammed Jaslam[1], Deepankar[1]*, Nitin Bhardwaj[1], Vinay Kumar[1]
[1]Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, India
[2]Department of Agricultural Economics, CCS Haryana Agricultural University, Hisar, India
*Corresponding Author E-mail: deepankarverma7@gmail.com

ABSTRACT
Sugarcane is one of the important commercial crops in India. Sugarcane occupies about 3% of the total cultivated area and it is one of the most important non-food grain crops which contribute to about 7.5% gross value of the agricultural production in the country. Crop yield forecasts and crop production estimates are necessary for national food security including early determination of the import/export plan and price and to provide timely information for optimum management of growing crops. The present study was planned to model and predict the sugarcane yield of India usingdifferent ARIMA models. The yield data of sugarcane crop for a period of (1951-2018) have been utilized to develop the forecast models. The results of the study revealed that the ARIMA(2,1,0) model have been selected as best model among all the models for prediction of the sugarcane yield on the basis of Root mean square error (RMSE), Mean absolute error (MAE), Mean absolute percentage error (MAPE) and Adjusted R^2. The performance of the model is validated by comparing with actual values.
***Keywords**: ARIMA, Forecast, RMSE, MAPE, MAE, Sugarcane yield*

Received 21.12.2020 Revised 24.01.2021 Accepted 21.02.2021

INTRODUCTION

Forecasting is a method that can allow decision-makers to make their future decisions more correctly, whether from the economic or non-economic fields. In fact, national governments need forecasting to make different policy decisions on storage, pricing, marketing, import-export, etc. Continuous changes are usually characterized by the productivity of agricultural crops due to many factors such as variations in precipitation and economic, technological and agricultural conditions. Studying the essence and course of changes in that productivity is useful in evaluating efforts to increase agricultural production. Moreover, the productivity forecast of various agricultural crops allows accurate predictions of productivity levels to be made in the years to come. Therefore, forecasting is one of the main tools in the field of agricultural production to make effective growth policies and successful economic plans.

Time series models have advantages in certain situations. They can be used more easily for forecasting purposes because the historical sequences of observations upon study variables are readily available at equally spaced intervals over discrete point of time. These successive observations are statistically dependent and time series modelling is concerned with techniques for the analysis of such dependence. The application of Box-Jenkins [1] univariate autoregressive integrated moving average (ARIMA) models in the field of agriculture for forecasting a variety of study variables of interest for different crops / regions etc. may be of immense importance.

India has a very well-developed network for recording and aggregating crop statistics at various village levels. India is second largest sugarcane producers in the world after Brazil, producing around 350 million tonnes of cane per annum. Around 60-65 percent of total sugarcane area in the country is in the sub-tropics, and this covers U.P, Bihar, Haryana and Punjab. Sugarcane can be grown in a wide range of climate from warm tropical south to foothills of Himalayas. Under warm humid condition, it can continue its growth, unless terminated by flowering. The crop does its best in tropical region, receiving a rainfall of 750-1200mm.

Panse [7-9] in a series of papers studied the trends in yield(s) of rice and wheat with a view to compare the yield rates during the plan period(s) with that of the pre-plan period(s). Paul *et al.* [10] have conducted the time-series analysis for modeling and forecasting of spices export data in India. Suresh and

Priya [11] used ARIMA models to predict the area, production and yield of sugarcane in Tamilnadu and found that ARIMA (1, 1, 1) and ARIMA (2, 1, 2) is best for area & yield and for production respectively. In 2016, Hossain and Abdulla studied that the ARIMA (0, 2, 1) model is best for predicting the potato production in Bangladesh using Auto Regressive Integrated Moving Average (ARIMA) method and found that the fitted model is statistically significant. Kumar *et al.* [2] forecasted the productivity of sugarcane in Bihar through fitting of ARIMA model and select ARIMA (0, 1, 1) model as a best model. Kumar *et al.* [3] developed a model to forecast the yield of wheat in Haryana by using annual time series data from 1980-81 to 2009-10. They applied various methods as a random walk, random walk with drift, linear trend, moving average, simple exponential smoothing, and ARIMA models and compared each other to find out the best model to forecast the yield. Kumar and Verma [4] conducted a study to find out mustard yield forecast models for Bhiwani and Hisar districts of Haryana using autoregressive integrated moving average (ARIMA) technique. They found that ARIMA(0,1,1) and ARIMA(1,1,0) model is suitable for Bhiwani and Hisar districts respectively.

MATERIAL AND METHODS

Time series (TS) data related to observations on a variable that happens in a time interval. In every time series analysis, one fundamental assumption is that a few part or effect of the past will continue to stay in future. The most widely used technique for modeling and forecasting the TS data is Autoregressive integrated moving average (ARIMA) technique.

Data description and Statistical methodology

The present research dealt with the time series modeling for prediction of sugarcane yield in India. The sugarcane yield data (Source: Ministry of Agriculture & Farmers Welfare, Govt. of India.)from 1950-51 to 2014-15 has been used to find out suitable model and the remaining data set i.e. 2015-16 to 2017-18 used to check the validity of the developed ARIMA models.

Autoregressive Integrated Moving Average (ARIMA)

Autoregressive integrated moving average forecasts are based only on the previous value of the predicted component. This approach covers both continuous and discrete data. The ARIMA model needs a minimum sample size of approximately 30-40 observations and only applies to stationary time series data. The mean, variance and autocorrelation function of a stationary time series is generally constant over time. Mostly, non-stationary time series arising in practice which can be converted into stationary time series through differencing. Differencing is applied when the mean of a time series is fluctuating over time and log transformation is applied if the variance of a time series is fluctuating through time. The Autoregressive integrated moving average techniques have three stages 1) Identification stage 2) Estimation stage and 3) Diagnostic checking

At the identification stage, two graphical tools (autocorrelation function and partial autocorrelation function) are used to calculate the statistical relationships within a data set. These functions facilitate the selection of one or more relatively suitable ARIMA models. At the second step, we can correctly measure the coefficients of the model selected at the identification stage. This step also provides warning signals when such mathematical inequality requirements are not satisfied by the estimated coefficients. Finally, at third step, residuals are taken to test the independency of the random shock and to decide if an estimated model is statistically sufficient or not.

An estimated autocorrelation function r_k describe the correlation between ordered pairs ($\overline{Y}_t, \overline{Y}_{t+k}$) separated by various time spans (*k=1, 2, 3,.....*) and r_k is an estimate of corresponding parameter ρ_k. We can define autocorrelation function as

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

An estimated partial autocorrelation function ϕ_{kk} describe the correlation between ordered pairs ($\overline{Y}_t, \overline{Y}_{t+k}$) separated by various time spans (k = 1, 2, 3...) with the effect of intervening observations ($\overline{Y}_{t+1}, \overline{Y}_{t+2}, \ldots\ldots, \overline{Y}_{t+k-1}$) accounted for and ϕ_{kk} is an estimate of corresponding parameter ρ_{kk}.

Comparison and post-sample validity checking of fitted models

For the comparison and evaluation of the fitted models, various statistics criterion have been used which are follows as:

Percent Relative Deviation (RD %)

This calculates the difference (in percentage) of the predicted output from the observed output and is defined as:

Percent Relative = {(observed yield – forecasted yield)/observed yield}*100

Root Mean Square Error (RMSE)

It is used to compare two models and its formulation is given as

$$RMSE=\left[\left\{\frac{1}{n}\sum_{i=1}^{n}(O_i-E_i)^2\right\}\right]^{\frac{1}{2}}$$

Mean Absolute Percentage Error (MAPE)

This evaluates the accuracy of a model and can be calculated as:

$$MAPE=\frac{100}{n}\times\sum_{i=1}^{n}\left|\frac{O_i-E_i}{O_i}\right|$$

where, O_i and E_i are the values observed and predicted, and n is the number of years of forecasting.

RESULTS AND DISCUSSION

The time versus yield graph (Figure 1) is showing overall increasing trend for sugarcane crop in India.

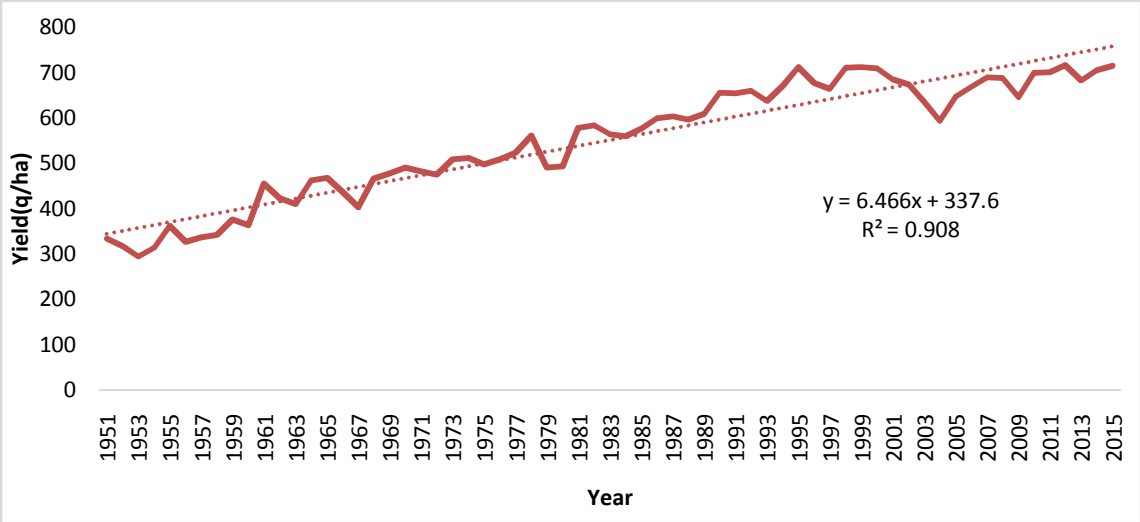


Figure 1: Time versus Yield graph of Sugarcane crop

Identification of order for AR and MA polynomials

At the identification steps, an appropriate order of AR and MA polynomials i.e. the values of p and q were determined with the help of acfs and pacfs of the stationary time series. The graphical presentation of sugarcane yield (q/ha) in Figure 1 clearly shows that the data series are non-stationary. All of the acfsupto $n/4^{th}$ lags significantly differ from zero reflecting the same non-stationarity condition (Table 1). The plotting of acfs in Figure 2 also indicates that the acfs decline gradually implying non-stationarity. Thus, the series considered here were transformed into stationary series by differencing of order one of the original ones (Figure 4). Further, pacfs in Figure 3 show a significant spike at lag 1, just suggesting that the series may have an autoregressive component of order one. The same can be observed from the parameter values and corresponding t-test as well.

Table 1: Autocorrelations of sugarcane yield

Lag	Autocorrelation	Std. Error	Box-Ljung Statistic		
			Value	df	Sig.
1	0.79	0.16	23.80	1	<0.01
2	0.62	0.16	38.72	2	<0.01
3	0.51	0.16	49.03	3	<0.01
4	0.40	0.16	55.71	4	<0.01
5	0.28	0.15	59.17	5	<0.01
6	0.12	0.15	59.85	6	<0.01
7	0.08	0.15	60.12	7	<0.01
8	-0.01	0.14	60.12	8	<0.01
9	-0.10	0.14	60.61	9	<0.01
10	-0.12	0.14	61.42	10	<0.01
11	-0.14	0.14	62.53	11	<0.01
12	-0.08	0.13	62.88	12	<0.01
13	-0.10	0.13	63.48	13	<0.01
14	-0.13	0.13	64.54	14	<0.01
15	-0.05	0.12	64.73	15	<0.01

Fig 2: Autocorrelation of sugarcane yield

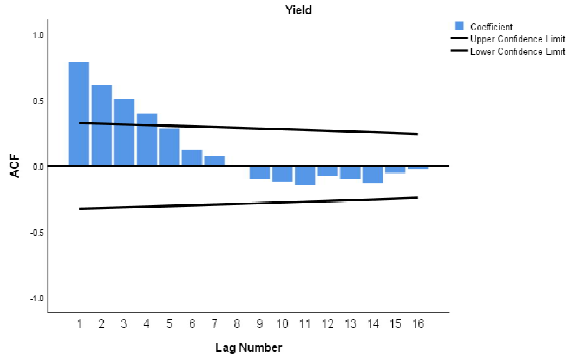


Fig 3: Partial autocorrelation of sugarcane yield

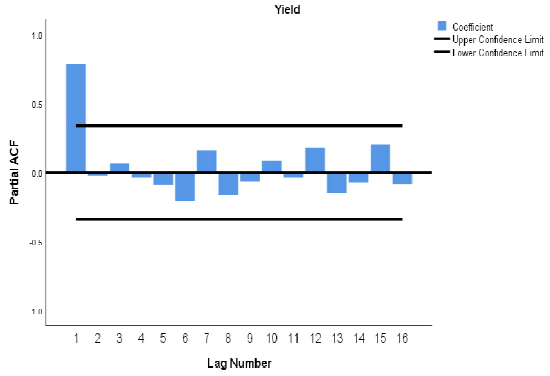


Figure 4: Autocorrelation of sugarcane yield after 1st differencing

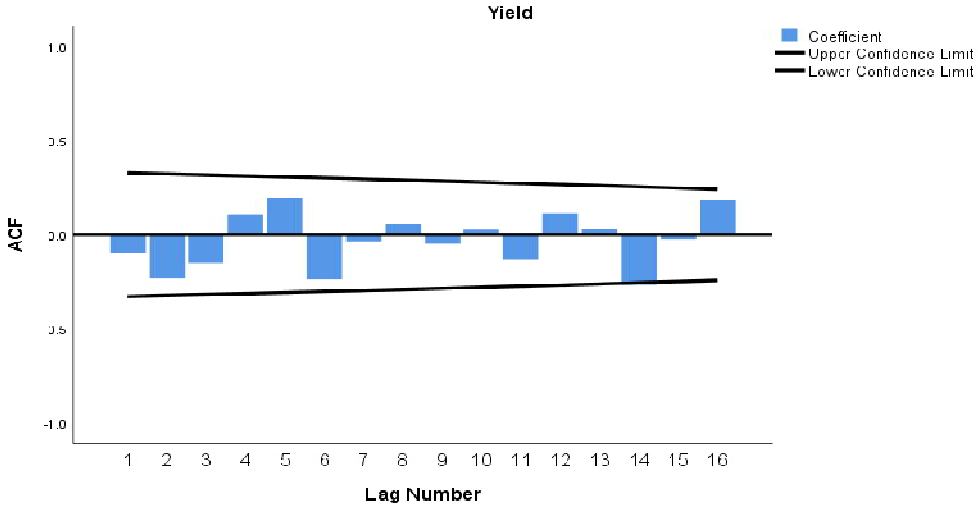
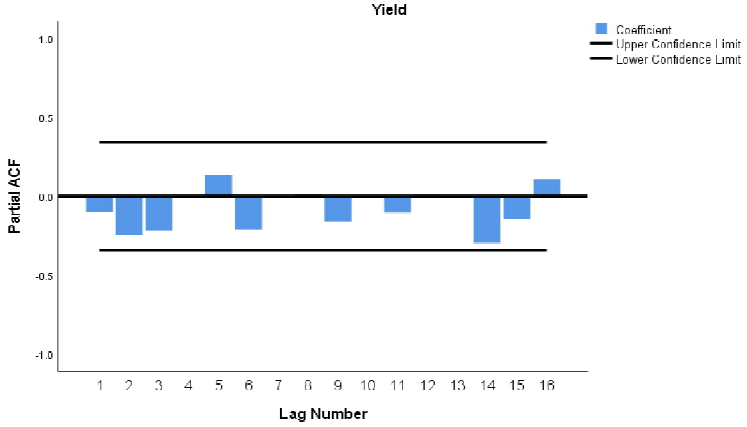


Figure 5: Partial Autocorrelation of sugarcane yield after 1st differencing



Parameter Estimation

After trying with different lags of AR and MA orders; the models ARIMA (2,1,0), ARIMA (0,1,2) and ARIMA (1,1,2), were considered at the identification stage. ARIMA estimation was carried out using non-linear least squares (NLS) approach. The relatively popular method due to Marquardt (1963) was used for the purpose. Parameter estimates of the fitted ARIMA models are given in Table 2 subsequently followed by the related results shown in Tables 3 and 4.

Table 2: Parameter estimates of ARIMA models for sugarcane yield

Models		Parameter Estimate	Standard Error	Approx. Prob.
ARIMA (2,1,0)	AR(2)	-0.27 -0.39	0.12 0.12	0.03 <0.01
ARIMA (0,1,2)	MA(2)	0.32 0.28	0.12 0.12	0.01 0.03
ARIMA (1,1,2)	AR(1) MA(2)	-0.17 0.17 0.35	0.46 0.44 0.21	0.71 0.70 0.11

Table 3: Selection criteria values for choosing ARIMA models

Models	RMSE	MAPE	MAE	BIC
ARIMA (2,1,0)	29.00	4.03	20.91	6.93
ARIMA (0,1,2)	29.08	4.19	21.81	6.94
ARIMA (1,1,2)	29.19	4.05	21.00	7.01

ARIMA (2,1,0), (0,1,2) and (1,1,2) were fitted for sugarcane yield estimation in India. These models were used to obtain sugarcane yield forecasts for the post-sample period 2015-16 to 2017-18 as has been given in Table 6.

Table 4: Results on Stationarity and Invertibility conditions for AR and MA coefficients of fitted ARIMA models

Model	Stationarity	Invertibility
ARIMA (2,1,0)	-0.27 -0.39	**

** Invertibility condition is not applicable since the model is AR model

Parameter estimates of the fitted models satisfied the stationarity condition since absolute value of AR coefficient is less than one.

Diagnostic checking

The model verification concerns with checking the residuals to see if they contained any systematic pattern which can be removed to improve the chosen ARIMA models. Approximate t-values were calculated for residual acfs using Bartlett's approximation for the standard error of the estimated autocorrelations. All Chi-Squared statistic(s) in this concern were calculated using the Ljung-Box (1978) formula as has been shown in Table 5. The graphical Figure 5 shows that none of the residual acfswas significantly different from zero at a reasonable level. This ruled out any systematic pattern in the residuals.

Table 5: Diagnostic checking of residual autocorrelations based on fitted ARIMA models

Model	Ljung-box Q Statistic		
	Statistic	df	Sig
ARIMA (2,1,0)	18.84	16	0.28

Fig 5: Residual ACFS and PACFS plots based on fitted ARIMA models

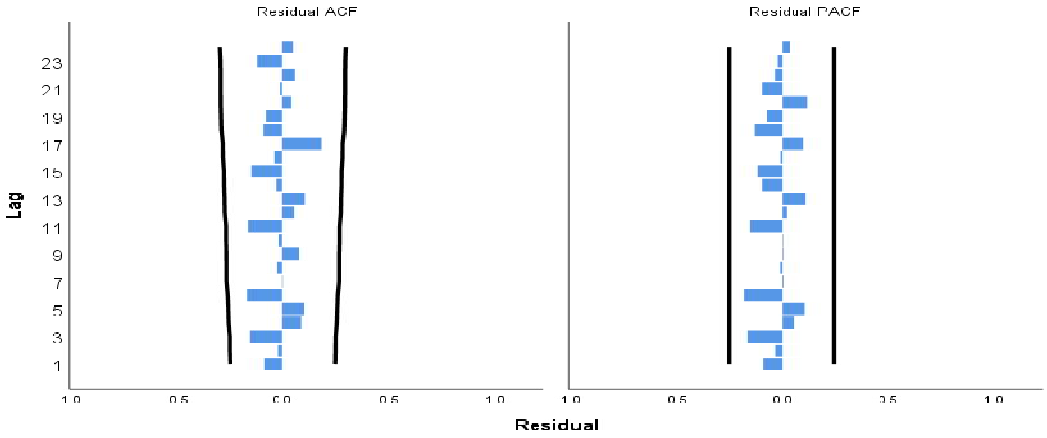


Fig. 6: Observed and Estimated sugarcane yield on fitted ARIMA models

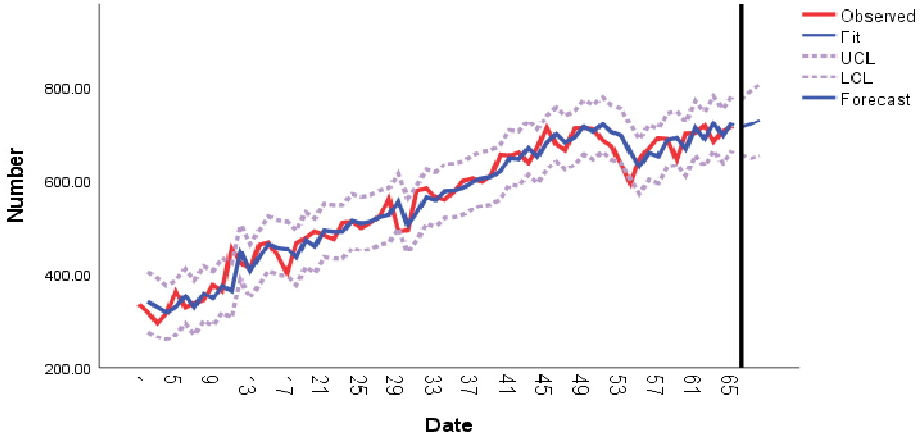


Table 6: Estimated sugarcane yield based on ARIMA models and their associated percent relative deviations (RD%) = 100×(Obs. Yield-Estimated Yield)/Obs. Yield)

Model	Forecast Year	Observed Yield (q/ha)	Estimated Yield (q/ha)	Percent Relative Deviation
ARIMA (2,1,0)	**2015-16**	707.20	713.73	-0.92
	2016-17	690.01	720.38	-4.40
	2017-18	801.98	729.28	9.07
Av. Abs. percent dev.				**4.80**

The prediction performance(s) of alternative models were observed in terms of per cent deviations of sugarcane yield forecasts about the real-time yield(s). It has been observed that ARIMA technique is appropriate to forecast the sugarcane yield. Finally, the fitted model is accomplished of providing satisfactory estimates of sugarcane yield well in advance of the crop harvest.

REFERENCES

1. Box, G. E. P. and Jenkins, G. M. (1970) Time series analysis: Forecasting and control, 1st ed., *Holden Day, San Francisco*.
2. Kumar, A. and Verma, U. (2020) Forecasting mustard yield in Haryana with ARIMA model, *The Pharma Innovation*, **9**(4), 136-140.

3. Kumar, A., Deepankar, P K., M. J. and Kumar, A. (2019) Wheat yield forecasting in Haryana: A time series approach, *Bulletin of Environment Pharmacology and Life Sciences*, **8**(3), 63-69.
4. Kumar, M., Raman, R. K. and Kumar, S. (2017) Forecasting of soybean yield in India through ARIMA model, *International Journal of Pure and Applied Bioscience*, **5**(5), 1538-1546.
5. Ljung, G. M. and Box, G. E. P. (1978) On a measure of lack of fit in time series models, *Biometrika*, **65**, 297-303.
6. Marquardt, D. W. (1963) An algorithm for least-squares estimation of non-linear parameters, *Journal of the Society for Industrial and Applied Mathematics,* **11**(2), 431-441.
7. Panse, V. G. (1952) Trends in areas and yields of principal crops in India, *Agriculture Situation in India*, **7**, 144-48.
8. Panse, V. G. (1959) Recent trends in the yield of rice and wheat in India, *Indian Journal of Agricultural Economics*, **14**, 11-38.
9. Panse, V. G. (1964) Yield trends of rice and wheat in first two five years plans in India, *Journal of Indian Society of Agricultural Statistics*, **16**(1), 1-50.
10. Paul, R. K., Prajneshu and Ghosh, H. (2009) Garch non-linear time series analysis for modelling and forcasting of India volatile spices export data. *Journal of the Indian society of Agricultural Statistics,***63**(4), 123-131.
11. Priya, S.R.K. and Suresh, K.K. (2011) Forecasting sugarcane yield of Tamilnadu using ARIMA models. *Sugar Tech*, **13**, 23–26.

CITE THIS ARTICLE

A Kumar, V Sain, P.K. Muhammed Jaslam, Deepankar, N Bhardwaj, V Kumar· Statistical Modeling for Prediction of Sugarcane Yield in India using ARIMA Model. Res. J. Chem. Env. Sci. Vol 9[1] February 2021. 08-14