A-6169
[1-6]

RESEARCH ARTICLE Indian Journal of Agricultural Research

# Development of ARIMA Model for Forecasting Sugarcane Production in Assam

Manashi Hazarika[1], Kaushik Kishore Phukon[2]

10.18805/IJARe.A-6169

## ABSTRACT

**Background:** Assam is a state of northeast India which lives on agriculture. Sugarcane is one of the important cash crops grown in the state. The cultivated area of sugarcane is 29,768 hectare in the state during 2021-22 with a total production of 11,60,025 tones. Sugarcane cultivation is not evenly distributed in all the districts of Assam. The major sugarcane producing districts in state are Karbi Anglong, Nagaon, Sonitpur, Dima Hasao and Golaghat.

**Methods:** This paper is an attempt to develop an appropriate time series model by using Box-Jenkins methodology to forecast the production of sugarcane in the state for few coming years. The study is based on secondary data collected from various publications of Directorate of Economics and Statistics, Govt. of Assam. Year wise time series data on production of sugarcane for a period of 60 years from 1961-62 to 2021-22 have been analyzed for this study.

**Result:** The results of this study establish ARIMA(1,2,1) as an suitable model for forecasting sugarcane production in the state. The proposed forecasting model indicates an increasing trend of sugarcane production in the coming years in Assam.

**Key words:** ARIMA, Forecast, Production, Sugarcane, Time series.

## INTRODUCTION

Sugarcane is one of the leading sugar plants of the world. It is a large spices of grass in the genus Saccharum and the family Poaceae. Sugarcane is a stout perennial grass that growing from 8 to 15 feet high with solid, heavy stalks. The plant is grown for the juice which the stalks contain and from which sugar is made. Ethanol, biogases, molasses, crude wax etc. are some byproducts of this crop. Sugarcane is now considered as an important energy source in terms of bio-ethanol production. However, Sugarcane is mainly used for the production of white sugar, Jaggery (Gur) and Khandsari. Sugarcane has been cultivated in India for many centuries. Soldiers of Alexander, the great, saw sugar cane during its conquest of India in 326 BC (Purseglove, 1979). Today, sugarcane is a major commercial crop of India which plays a vital role in country's nation economy. Indian economy is an agrarian economy and the country lives in villages. Around 7.5% of the country's rural population is primarily engaged in sugarcane farming (GOI, 2020). India is the second largest producer of Sugarcane after Bazil and followed by China. These three countries contributed about 62% area and 64% production of sugarcane in the world. Sugarcane is grown in more than 115 countries with an average cultivation area 26.54 million hectare and with a total production of 1878.79 million tones and productivity of 71 tones/ha. Sugar cane accounts for 21 per cent of the global crop production over the period 2000-2019 (FAO, 2019, 2021). In India, sugarcane is grown in almost all states with a variation regarding area of cultivation and production statewide. Sugarcane plays the role of raw materials for the second largest agro-based industries in the country after textile.

[1]Department of Agricultural Statistics, Sarat Chandra Sinha College of Agriculture, Assam Agricultural University, Dhubri-783 376, Assam, India.
[2]Department of Information Technology, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya, Tulungia, Bongaigaon-783 380, Assam, India.

**Corresponding Author:** Manashi Hazarika, Department of Agricultural Statistics, Sarat Chandra Sinha College of Agriculture, Assam Agricultural University, Dhubri-783 376, Assam, India.
Email: manshihazarika@gmail.com

**How to cite this article:** Hazarika, M. and Phukon, K.K. (2024). Development of ARIMA Model for Forecasting Sugarcane Production in Assam. Indian Journal of Agricultural Research. DOI: 10.18805/IJARe.A-6169.

**Submitted:** 11-10-2023 **Accepted:** 08-04-2024 **Online:** 18-06-2024

The maximum area and production is recorded in Uttar Pradesh (45% area and 43% production) followed by Maharashtra (19% area and 21% production) and Karnataka (9% area and 10% production). These three states contributed about 73% area and 74% production of the country. There are two distinct agro-climatic regions of sugarcane cultivation in India, *viz*., tropical and subtropical. Tropical region has about 45% area and contributes around 55% of the total sugarcane production in the country. Thus, sub-tropical region accounts for 55% area and shares 45% of total production of sugarcane (GOI, 2020).

Assam falls under the Sub tropical regions of the country. The varieties of sugarcane grown in the state are Adhagathiya, Adhagathiya, Kolong, Seni Joba, Barak, Co 313, Co 997, Dhanshiri, Co JOR 2, CoBln 02173, CoBln

94063, CoBln 9006 and Kopilipar (GOI, 2020) depending on the suitability of soil and climatic conditions of the state. The sowing and harvesting time of sugarcane in the state are January-March and March-November respectively. In Assam, sugarcane is consumed in the form of fresh juice and as post harvest products. The nutritional contents present in sugarcane juice have multiple health benefits as it contains a good amount of energy, calcium, Potassium, Sodium and iron. In Assam, sugarcane has some socio cultural importance as its post harvest product Gur is used in preparing many traditional food items especially during the time of Bihu festivals. Although sugarcane is cultivated in all the districts of Assam, the major portion of states' sugarcane production comes from Karbi Anglong, Nagaon, Sonitpur,Dima Hasao and Golaghat district of the state. The cultivated area of sugarcane is 29,768 hectare in the state during 2021-22 with a production of 11,60,025 tones. The cultivated area of sugarcane in Assam accounts for 14 per cent of its total cultivated area of fruits and vegetables in the state (GoA, 2020).

## MATERIALS AND METHODS

Secondary data on production of sugarcane in Assam have been collected from various publications of Directorate of Economics and Statistics, Govt. of Assam for this study. This study is based on year wise time series data on production of sugarcane for a period of 60 years from 1962-63 to 2021-22. The objective of this paper is to develop a suitable time series model to predict the future production of sugarcane in Assam.

Statistical modeling based on time series data has always been considered as an important aid of analyzing and estimating the future values of the study variable under the assumption that the past pattern will continue to remain in the future. The process of estimating the future values of the study variable based on its past values is generally known as forecasting. Again forecasting is an important aspect of policy making process in different fields like business, education, health, government *etc*.

A time series is a set of observations $y_t$, each one being recorded at a specific time *t* (Peter Davis, 2001). Mathematically , a time series is defined by the functional relationship.

$$y_t = f(t)$$

Where,

$y_t$ ($t$ =0,1,2,3,….) = Value of the variable under consideration at time $t$.

In time series analysis, the recorded observations of the underlying variable and ordering of the observations possess equal importance.

One of the most widely used stochastic time series models for forecasting is Autoregressive Integrated Moving Average (ARIMA) model which is popularly known as Box-Jenkins methodology (Box and Jenkins, 1978). ARIMA models have been extensively used time to time to forecast various agricultural productions across the globe. In India ARIMA models were used for forecasting India's sugarcane productivity (Kumar *et al*., 2017), eggs production (Chaudhari and Tingre, 2015), milk production (Mishra *et al*., 2020), rice production (Mahajan *et al*., 2020) and fish production (Paul and Das, 2010).

The basic assumption of time series forecasting is that the series under study need to be stationary. If the mean and variance of a time series is time invariant, the series is considered as stationary. An ARIMA model is a combination of Autoregressive (AR) and Moving Average (MA) terms representing its own past values and the past errors respectively. The general notation for denoting an ARIMA model is ARIMA (p,d,q) where p represent the number of autoregressive terms , d the number of times the series has to be differenced before it becomes stationary and q the number of moving average terms. ARIMA processes incorporate a broad range of non stationary series that reduce to ARMA (Autoregressive Moving Average) processes when differenced ûnitely many times. Mathematically, a process $\{y_t\}$ is said to follow an ARIMA (p,d,q) if it satisfies a difference equation of the form:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t,\ \{\varepsilon_t\} \text{ follows WN } (0, \sigma^2),$$

WN indicating White Noise.

Where,

$\varphi(B)$ and $\theta(B)$ = Polynomials of degrees p and q respectively.

d = Non negative integration parameter.

Box-Jenkins Methodology *i.e.* ARIMA modeling of time series data consists of four steps *viz*. identification, parameter estimation, diagnostic checking and forecasting. A time series need to be tested for stationarity at this identification stage. There are different tests available for testing the stationarity of time series. This can be done by plotting the values of the variable under study against time points on graph or by plotting the values of Autocorrelation and Partial autocorrelation function for a specific lags. The ACF of $\{y_t\}$ at lag k is defined as:

$$ACF(y_k) = \frac{Cov(y_t, y_{t-k})}{Var(y_t)},\ k = 1,2,3....$$

Where,

$y_t$ = Original series.

$y_{t-k}$ = Lagged series at lag k.

$\mu$ = Mean of the data set.

Partial autocorrelation function between $y_t$ and $y_{t-k}$ is the autocorrelation between $y_t$ and $y_{t-k}$ after adjusting for $y_{t-1}, y_{t-2},...y_{t-k+1}$ (Douglas *et al*., 2008).

If ACF plot is positive and shows a very slow linear decay pattern, then the data are considered as non stationary (Nau, 2018). Statistical tests like Dickey-Fuller test, Augmented Dickey-Fuller test, Philips-Perron test are also available for testing the presence of staionarity in the time series. In our study we have applied ADF test for testing the stationarity of our data s*et al*ong with ACF and PACF of the series (Dickey and Fuller, 1979). If the series is found

not stationary, it must be differenced d times in order to make it stationary. Once the time series becomes stationary we go for determining the order of AR and MA terms *i.e* the values of p and q in ARIMA (p,d,q). This is done by examining the ACF and PACF values of the stationary series. After identifying the values of p and q in the ARIMA (p,d,q) model, the parameters are estimated by maximum likelihood estimation (MLE) at the stage of parameter estimation. Having chosen a particular ARIMA model and having estimated its parameters, the next step is to check whether the chosen model fits the data reasonably well or not. The process of validating the chosen model again consists of two measures of goodness of fit. One is based on ACF of residuals estimated from the model and the other is Akaike's Information Criterion (AIC) (Akaike, 1974). The ARIMA model with the smaller value of AIC is considered as a better fit of the data. The AIC is computed as:

$$AIC = \text{Log } \hat{\sigma}_k^2 + \frac{n + 2k}{n}$$

Where,

$\hat{\sigma}_k^2$ = Maximum Likelihood estimate of the error variance.

k = Number of parameters in the model.

n = Sample size. In this study n is equal to 60.

As an alternative to AIC, Bayesian Information Criterion (BIC) is used to identify the best fitted model which is also called as Schwarz Information Criterion (Schwarz, 1978). BIC is computed as:

$$BIC = \text{Log } \hat{\sigma}_k^2 + \frac{k + \log n}{n}$$

BIC does well at getting the correct order in large samples, whereas AIC tends to be superior in smaller samples where the relative number of parameters is large (McQuarrie and Tsai, 1998).

Another method of testing the randomness of the residuals is plotting the ACF of residuals against lags. If we find that about 95% of the sample autocorrelation are within the limits of ±1.96/√N where N is the number of observation that forms the model, then the model is considered as is a good fit. In addition to plotting the individual sample autocorrelation of the residuals (ρk) we can test the joint hypothesis that all the sample autocorrelations coefficients up to certain lags are simultaneously equal to zero. This can be done by using the Q statistic developed by Box and Pierce (1970) which is defined as:

$$Q = n \sum_{k=1}^{m} \hat{\rho}_k^2$$

Where

n = Sample size.

m = Lag length.

This Q statistic will be calculated for lag length m=10 in this research work. In large samples, Q is approximately distributed as $\chi^2$ (m) *i.e.* Chi-square distribution with m degrees of freedom. We reject the hypothesis that all the $P_k$ are zero if the computed Q exceeds the critical Q value from the Chi square distribution at the chosen level of significance.

A variant of the Box-Pierce Q statistic is the Ljung-Box (LB) Statistics (1978) which is defined as:

$$LB = n (n + 2) \sum_{k=1}^{m} \left( \frac{\hat{P}_k^2}{n - k} \right) \sim \chi^2 \text{ (m) for large samples.}$$

Once the chosen ARIMA model is found reasonably good fit to the data, forecasting is executed using that model.

The best fitted ARIMA model from the data collected has been developed by executing the above stated procedure which is presented in the results and discussion section of this paper. For the entire analysis of the collected data, R software has been used.

## RESULTS AND DISCUSSION

### Development of ARIMA Models

Year wise production of sugarcane for a period of sixty years from 1961-62 to 2021-22 has been collected and analyzed for making prediction of sugarcane production in Assam in coming years. Fig 1 depicts the behavior of sugarcane production in Assam during the study period. It reveals a vacillating behavior of the production of sugarcane in the state with respect to time. The graph also reveals the fact that the series is non stationary. The following two Fig 2 and 3 which depict the autocorrelation and partial autocorrelation function of the production of sugarcane also witness the presence of non-stationarity in the series. ADF (Augmented Dickey-Fuller) test has also been executed to the time series and found a non significant ( p-value >0.05) results which leads to the acceptance of the null hypothesis that the series is non stationary. The series became stationary after second order differencing of the original data. The stationarity has been tested again by applying ADF test to the differenced series and found p-value =0.012 <0.05.

Based on the ACF and PACF of second order differenced series (Fig 3) four different ARIMA models *viz*., ARIMA(1,2,1), ARIMA(2,2,1), ARIMA(4,2,1) and ARIMA(4,2,2) have been chosen for further analysis . For these ARIMA models AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion) values are calculated as presented in the Table 1. The ARIMA model with the least AIC and BIC values has been propped as the most suitable

**Table 1:** AIC and BIC values of different ARIMA models.

| Model | AIC | BIC |
|---|---|---|
| ARIMA (1,2,1) | 1608.49 | 1614.67 |
| ARIMA (2,2,1) | 1610.38 | 1618.62 |
| ARIMA (4,2,1) | 1613.90 | 1626.26 |
| ARIMA (4,2,2) | 1615.76 | 1630.19 |

model for sugarcane production in Assam. Thus the proposed model is ARIMA (1,2,1) which can be written as:

$$y_t = \beta_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

Where,

$y_t$ = Dependent variable, *i.e.* the production of sugarcane at time t.

$y_{t-1}$ = Production of sugarcane at lag 1, *i.e.* the production of sugarcane one year before.

$\varepsilon_{t-1}$ = Lagged error term of one year.

$\varepsilon_t$ = Current error term.

$\beta_1$ = Autoregreesive (AR) coefficient.

$\theta_1$ = Moving average (MA) coefficient.

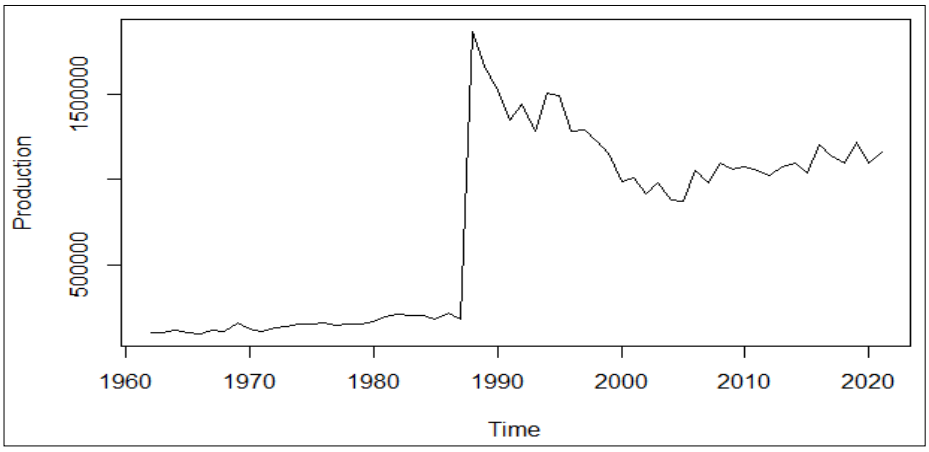


**Fig 1:** Sugarcane production (In tones) in Assam from 1962 to 2021.

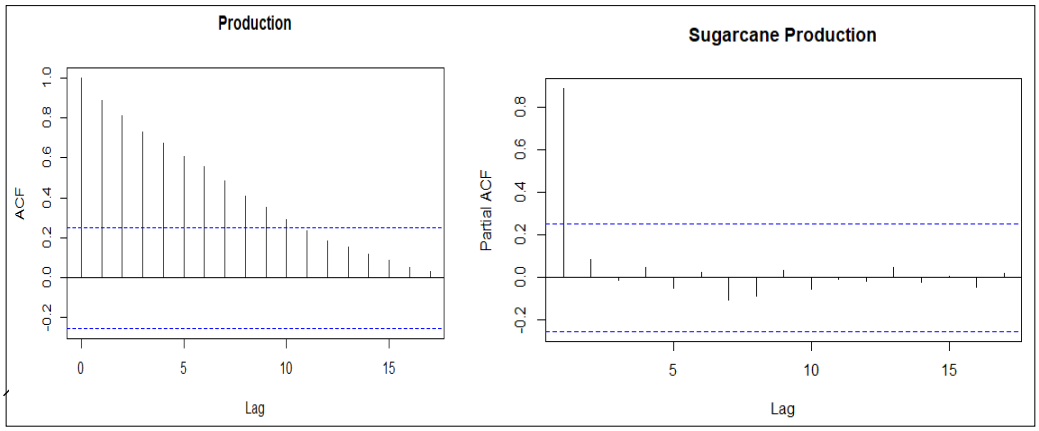


**Fig 2:** Plot of ACF and PACF of the sugarcane productions.

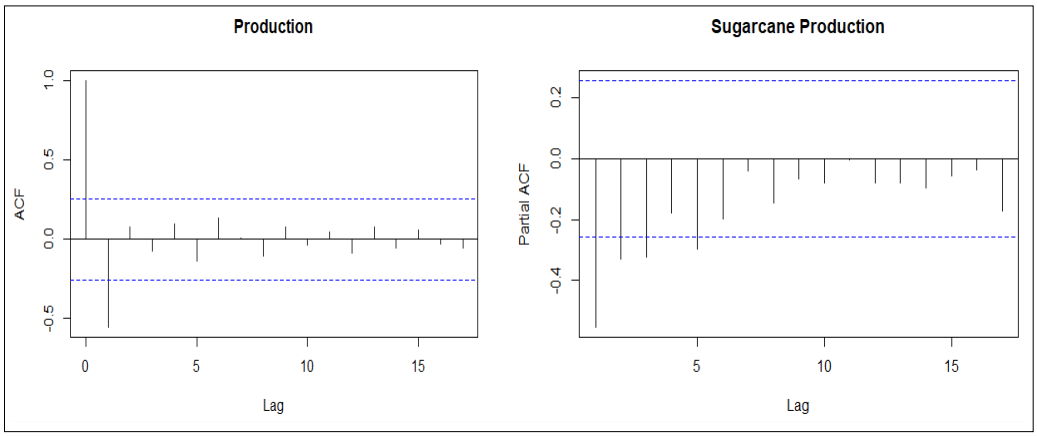


**Fig 3:** Plot of ACF and PACF of 2nd order differenced series.

**Table 2:** Estimates of the parameters of ARIMA (1,2,1).

| Parameters | Estimate | S.E |
|---|---|---|
| AR (1) | -0.143 | 0.1294 |
| MA (1) | -1.003 | 0.0521 |

**Table 3:** Forecast of sugarcane production in Assam from the fitted model ARIMA (1,2,1).

| Year | Forecasted values (In tones) | 95% confidence interval | |
|---|---|---|---|
| | | Lower | Upper |
| 2022-23 | 1170880 | 710690.165 | 1631069 |
| 2023-24 | 1189730 | 578577.571 | 1800883 |
| 2024-25 | 1207440 | 467035.816 | 1947844 |
| 2025-26 | 1225313 | 372543.178 | 2078082 |
| 2026-27 | 1243162 | 288268.416 | 2198056 |
| 2027-28 | 1261015 | 211332.042 | 2310697 |
| 2028-29 | 1278867 | 139885.808 | 2417848 |
| 2029-30 | 1296719 | 72729.502 | 2520709 |
| 2030-31 | 1314571 | 9024.581 | 2620118 |

From the Table (1 and 2), the estimated models for sugarcane production in Assam can be written as:

$$y_t = -0.1427\ y_{t-1} - 1.003\ \varepsilon_{t-1} + \varepsilon_t$$

The correlogram of ACF and PACF (Fig 4) of the residuals of the fitted ARIMA model also indicates a 'good fit' of the model as the estimated values of both ACF and PACF fall within the 95% confidence limit. We may say that ACF and PACF of residuals of the fitted model are non significant at 5% level of significance.

Further, Box-Pierce Q statistic is found non- significant with a p-value 0.9125 (>0.05) which also validates the good fit of the model.

### Forecasting with fitted ARIMA model

Once the diagnostic checks validate (correlogram of ACF and PACF (Fig 5), Box-Pierce Q statistic) that ARIMA (1,2,1) is the best fitted model, it has been used for forecasting sugarcane production in the state for the period 2022-23 to 2030-31. To check the predictive performance of the model, the actual observations were plotted with forecasted

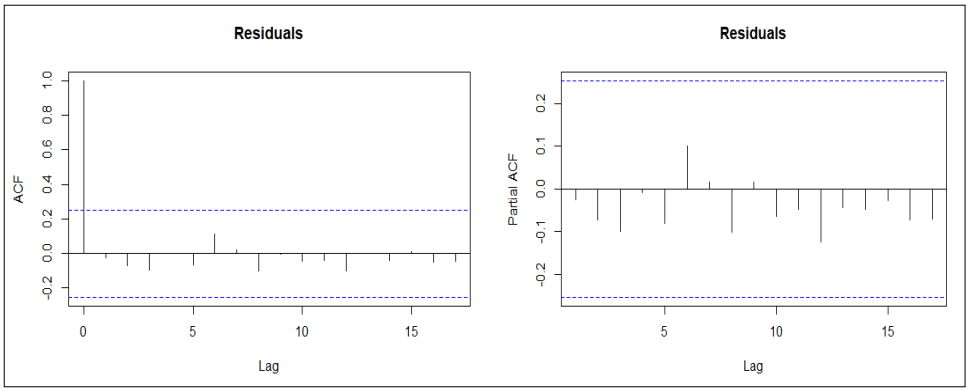


**Fig 4:** ACF and PACF of residuals of ARIMA (1,2,1).

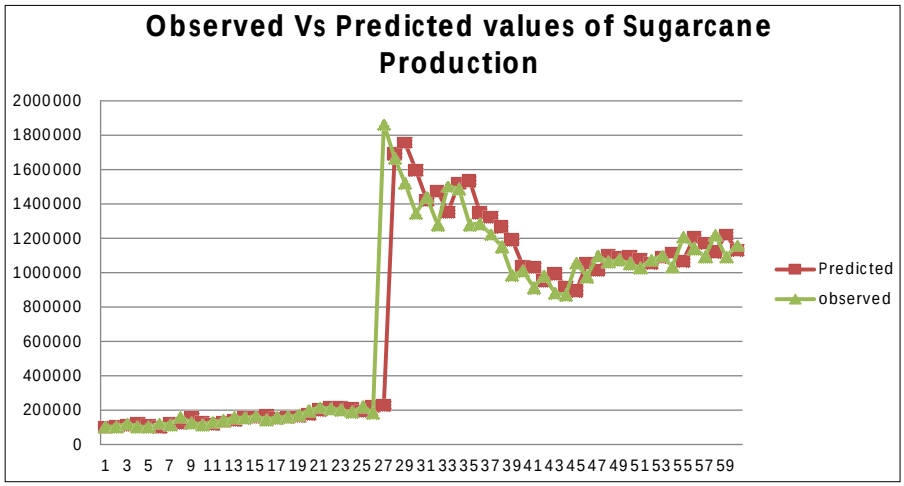


**Fig 5:** Observed and predicted values of sugarcane production in Assam.

values obtained from the fitted ARIMA (1,2,1) model for the study period *i.e.* 1962 to 2021 as shown in Fig 5. It has been observed that the forecasted values of productions of sugarcane in the state are on the average close to the actual productions data. The forecasted values of sugarcane production with 95% confidence interval for the period 2022-23 to 2030-2031 are shown in the Table 3. The forecasted values have revealed an increasing trend of production of sugarcane in the state (Table 3).

## CONCLUSION

In this study, ARIMA (1,2,1) has been found as the best fitted ARIMA model for forecasting sugarcane production in Assam which can be mathematically expressed as:

$$y_t = -0.1427\ y_{t-1} - 1.003\ \varepsilon_{t-1} + \varepsilon_t$$

This model explains that the production of sugarcane in the year *t* depends on its previous year *i.e.* (t-1 year) production as well as previous year error.

The forecast values obtained from this proposed forecasting model indicates that the production of sugarcane will increase in the state in the coming years. The production of sugarcane is expected to be 13,14,571 tones in the year 2030-31 on the basis of this fitted model. The demand for agricultural products is continuously increasing with a rapid growth of population in our country. Forecasting of agricultural production of different crops has always been considered as an important aspect for policy makers to formulate different schemes and programmes to feed the nation. This model can be used by researchers, policy makers for such purposes both at state and national levels. However, the model will need an updation from time to time with the inclusion of current data.

### Conflict of interest

This research work has been conducted solely by the researchers by collecting secondary data from different sources at their own interest without any support from any funding or sponsoring agency.

## REFERENCES


Akaike, H. (1974). A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, AC-19. 716-723.

Box, G.E.P. and Pierce, D.A. (1970) Distribution of residual autocorrelations in autoregressive integrated moving average time series models. Journal of the American Statistical Association. 65: 1509-1526.

Box, G.E.P., Jenkins, G.M, Reinsel, G.C., Ljung, G.M. (2016). Time Series Analysis: Forecasting and Control. John Wiley and Sons, Fifth Edition.

Chaudhari, D.J. and Tingre, A.S. (2015). Forecasting eggs production in India. Indian Journal of Animal Research. 49(3): 367-372. doi: 10.5958/0976-0555.2015.00143.0.

Dickey, A. and Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit-root. J. American Statistical Association. 74: 427-431.

Douglas, C.M. Cheryl, L., Jennings, M.K. (2008). Introduction to time series analysis and forecasting. Wiley Series in Probability and Statistics.

FAO, (2019). The State of Food and Agriculture 2019. Moving forward on food loss and waste reduction. Food and Agriculture Organization of United Nations.

FAO, (2021). World Food and Agriculture-Statistical Yearbook 2021. Rome.

GoA, (2020). Statistical Handbook of Assam. Directorate of Economics and Statistics.Govternment of Assam.

GoI, (2020). Agricultural Statistics at a Glance. Directorate of Economics and Statistics, Department of Agriculture, Cooperation and Farmers Welfare, Government of India.

GoI, (2020). Glimpses of Sugarcane Cultivation. Directorate of Sugarcane Development. Department of Agriculture, Cooperation and Farmers Welfare, Government of India.

Kumar, M., Raman, R.K. and Kumar, S. (2017). Sugarcane productivity in Bihar: A forecast through ARIMA model. Int. J. Pure App. Biosci. 5(6): 1042-1051.

Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. Biometrika. 65: 297-303.

Mcquarrie, A.D.R. and Tsai, C. (1998). Regression and Time Series Model Selection. World Scientific Publishing Company, Singapore.

Mahajan, S., Sharma, M. and Gupta, A. (2020). ARIMA modelling for forecasting of rice production: A case study of India. Agricultural Science Digest. 40(4): 404-407**.** doi: 10.18805/ag.D-5029.

Mishra, P., Fatih, C., Niranjan, H. K., Tiwari, S., Devi, M. and Dubey, A. (2020). Modelling and forecasting of milk production in Chhattisgarh and India. Indian Journal of Animal Research. 54(7): 912-917. doi: 10.18805/ijar.B-3918.

Nau, R. (2018). ARIMA models for time series forecasting. Retrieved from https://people.duke.edu/~rnau/411arim.htm.

Paul, R.K. and Das, M.K. (2010). Statistical modeling of Inland fish production in India. Journal Inland Fisheries Society of India. 42(2): 1-7.

Peter, J.B. and Davis, R A. (2001). Introduction to Time Series and Forecasting. Springer Texts in Statistics.

Purseglove, J.W. (1979). Tropical Crops: Monocotyledons. Longman Group Ltd., London, 607.

Schwartz, G. (1978). Estimating the dimension of a model. Annals of Statistics. 6: 461-464.