Innovations

Statistical Modelling for Prophesying Sugar Production in India

N Kausalya[1*] and S R Krishna Priya[2]

Research Scholar[1*], Assistant Professor[2]
Department of Statistics, PSG College of Arts & Science, Coimbatore
Corresponding author: **N Kausalya**

Abstract: *India relies extensively on agriculture. Seasonality, economy, and organic variables all affect agricultural yields. Examining India's sugar output trends was the aim of this study. Time-series data on sugar production (91 years) from 1930–1931 to 2019–20 were used. Two models, such as the Artificial Neural Network and the Auto Regressive Integrated Moving Average, were used to anticipate the sugar production in India. The outcomes for the two models were contrasted. The error measure results indicated that the ANN was the most adequate.*

Keywords: ***S****ugar Production, ARIMA, ANN, forecasting*

I. Introduction

With a 17 percent worldwide share, India is currently the world's second-largest producer of sugar. By 2022, it is predicted to produce a record 38 million tonnes of sugar, making it a surplus country. The second-biggest agro-based industry in India, the sugar industry provides 6.5 million families with industrial and agricultural workers with a living. The production of sugarcane has changed the economics of the nation in a number of ways. It has historically supplied raw materials for the gur and sugar industries, which have a positive impact on the rural economy. India's sugar statistics during 193–18 demonstrated a noteworthy 9.58-fold rise in sugarcane production compared to an almost four-fold increase in cane land. The key to this success has been the use of improved sugarcane cultivars. After the nation's historic achievement of creating its first commercial variety in 1918, more than a century has gone by, and over 190 kinds have found commercial success during that time. India's recent bioethanol policy, which aims to make the country a major player in the sugar and bioenergy sectors, benefited greatly from the integration of the best varieties, including the wonder varieties Co0238 and Co86032, with ongoing improvements in crop production and protection technologies. India's cane output in 2021–2022 was 81.98 t/ha, above the world average, and the country achieved 10% bioethanol blending in gasoline. Sugarcane has become a multiproduct crop from related sectors that produce bioethanol and electricity as fuel, fodder, fibre, feed, fertilizers, paper, biomolecules, etc., in addition to a larger sugar export. These all create prospects for a variety of vocations, including independent contractor work. The production of sugarcane is simultaneously complicated by a number of biotic and abiotic stresses, issues linked to climate change, dwindling natural resources, rising cultivation costs, a labour shortage, and a significant discrepancy between realized yield and theoretical yield potential. However, in addition to more reliable and productive cultivars, advancements in machinery, nanotechnology, decision support systems, and UAV-based crop management and protection have emerged recently to make sugarcane production more methodical, sustainable, and focused on the end result.

Local politics and economic policies have an impact on the amount of sugar produced in more than 110 nations. Cane grown in tropical and subtropical areas produces around 80% of the sugar produced

worldwide. Sugar beets, which are primarily farmed in the northern hemisphere's temperate zones, account for the remaining 20% of the crop. 1.87 billion tonnes of sugarcane were produced worldwide in 2020, with 40% coming from Brazil, 20% from India, and 6% from China. In 2020, 20 million hectares were planted with sugarcane worldwide. Global sugar production is predicted to reach 182.9 million metric tons in the 2022–2023 crop year, an increase of 1.7 million tons over the previous year.The top five sugar-producing countries in the world are the US, China, Thailand, India, and Brazil.

India ranked first in sugar output in 2020–21 and is not far behind Brazil. The amount used internally is rising gradually to 29 million tons (equal to raw sugar).

In order to anticipate sugar production in India, the Auto Regressive Integrated Moving Average (ARIMA) model and Artificial Neural Network (ANN) model are the main topics of this article. The article's remaining sections are arranged as follows: Section 2 outlines the techniques and methodology. Results and discussion are covered in Section 3, and a conclusion is provided in Section 4.

II. Materials and Methodology

2.1 Data Source& Area of Stuy dy

The external source from which the data was gathered is Co-operative Sugar Vol. 53, August 2022, a publically accessible record. This covered 91 years, from 1930–1931 to 2019–2020, when India produced sugar. In general, sugar production in India has been used as an input. In the market year 2022/2023, India is expected to manufacture roughly 36 million metric tons of sugar.

2.2 Box-Jenkins Modeling Procedure

An autoregressive integrated moving average model is a generalization of an autoregressive moving average model in statistics, specifically in time series analysis. In order to forecast future points in the series or to gain a better understanding of the data, these models are fitted to time series data. They are used when data exhibit non-stationary behaviour; in these situations, the non-stationary can be eliminated by applying an initial differencing step, which corresponds to the "integrated" portion of the model. The autoregressive, integrated, and moving average components of the model are denoted by the non-negative integer parameters p, d, and q, respectively. The model is commonly known as an ARIMA model. A key component of the Box-Jenkins method for time-series modelling is the ARIMA model.

Construction of ARIMA model

Selecting a specific mathematical model based on research and study results, as well as statistical metrics that describe the model for another. Use one of the estimation techniques to estimate the parameters of the model after nominating one or more suitable models to represent the time series of data that we are watching. This comprises testing the validity of the model's assumptions and doing a residuals analysis to determine how closely the computed values from the selected model match observations. The model is adopted as the final model and used to estimate future predictions if it passes these tests; if not, we go back to the initial phase for the appointment of a new model. Using the completed model, future predictions are made, and the prediction errors that transpired are subsequently computed.

2.3 Ann: A Succinct Summary

Artificial neural networks (ANNs) are a class of statistical learning algorithms that are used to estimate or approximate functions that may depend on a multitude of inputs. They are inspired by biological neural networks. ANNs are typically shown as networks of interconnected "neurons" that can compute values from inputs. Because of its adaptive nature, ANNs are also capable of machine learning and pattern recognition.

The human brain is made up of ANN, also known as neural networks, which are biologically inspired by artificial intelligence AI models. Artificial neural networks connect nerve cells in numerous layers of networks, analogous to how neurons in the human brain connect to each other.

Artificial Neural Networks can be applied in three different ways:

(a) As biological nerve system and intelligence models.

(b) In hardware as real-time adaptive signal processors or controllers for applications such as robots.

(c) In terms of data analytic approaches Nodes are the names given to the neurons. Without a lot of data or knowledge, ANN is a self-adaptive nonlinear nonparametric data-driven statistical method for regulating nonlinearity and approximating complex relationships.

The ways for processing the data through neural networks are given below,

- The information was processed by neurons, which are basic processing units.
- Neurons communicate by cross-linking lines.
- Incoming signals to neurons multiply the weight that each link has linked to it.
- To ascertain the output signal that results from it, each neuron applies an activation function to the entire input, which is the sum of the weighted input signals.

2.4 Akaike Information Criteria

One mathematical framework that can be applied to model selection and parsimony assessment in model construction is the Akaike information criterion.

When testing a data set is difficult, one can use an AIC score to identify which machine learning model performs best for the given data set. When working with a small data set or time series analysis, an AIC test is most helpful. The better, the lower the AIC score.

2.5 Auto Plot

Plotting the inverse characteristic roots is occasionally helpful when using ARIMA models to model data. Any fitted ARIMA model, including seasonal models, can have its inverse roots computed and plotted The moving average roots from the MA characteristic polynomial and the autoregressive roots from the AR characteristic polynomial are returned by the plot.

The complex unit circle will display the inverse of the roots plotted. All of the roots of a causal invertible model ought to be located outside of the unit circle. The inverse roots should, on the other hand, like inside the unit circle. The calculated ARMA is stable (stationary) and invertible, and will provide accurate estimates, if all roots have modulus less than one and are inside the unit circle.

2.6 Normal QQ-Plot

A normal probability plot, or more specifically a Quantile-Quantile (Q-Q) plot, shows the distribution of the data against the expected normal distribution.

The data sets univariate normality is indicated as points on the Normal QQ plot. A 45degree reference line will be where the points fall if the data is regularly distributed. In the event that the data is

not regularly distributed, the points will depart from the line of reference. Points at either end of the line that are far from the majority of the observations could be considered outliers.

III. Results and Discussion

Plotting the raw data is the initial step has to be made. Cleaning the data, fill the values if missing through the proper missing value techniques according to the time series dataset. Checking whether the data is stationary or non-stationary is the foremost step in the modeling of ARIMA model. Converting the non-stationary data into stationary data plays an important role in the ARIMA modeling. The collected raw data has been plotted for the checking process and the results indicates that India's sugar production dataare at non-stationary nature. Hence, it is concluded that the raw data is not fit modelling. Another test was performed for re-checking the data for stationary. For a second time, also the results from the Augumented Dickey Fuller test assures that the data is non-stationary. It is needed to make the time series data to stationary level and therefore, it was made by taking first differencing.

First Differencing

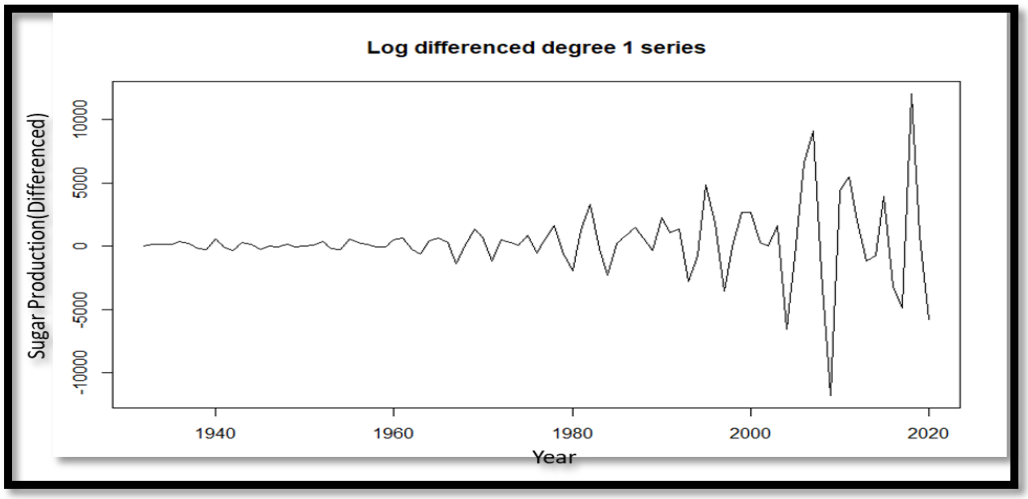

Figure 1. Time series plot of the first difference.

Augmented Dickey-Fuller Test for first differenced data

Table 1. Augumented Dicke Fuller Test

Augumented Dickey Fuller Test	**Lag Order**	***p*-value**
	4	0.01

Null Hypothesis: Non-Stationary
Alternative Hypothesis: Stationary

The *p*-value is 0.01 which is obviously less than 0.05, and considered to be a Stationary data. Therefore, taken as there is a significant difference and reject the null hypothesis

Normal QQ-Plot

Most of the points in the QQ-plot are lies on the line which authorises the data is stationary and can use for further forecasting modelling.

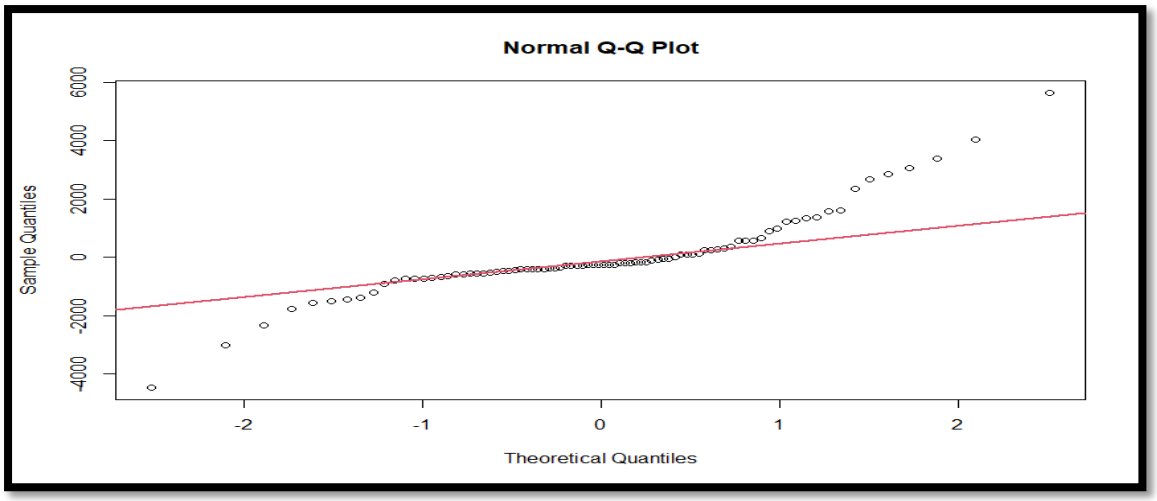

Figure 2. The computed QQ-plot for first differenced time series dataset

Auto-Plot for ARIMA

The red dot was present inside the circle of AR and MA root part which confirms the data is stationary and also presence of AR and MA part, it is considered to be AR value as 2 and MA as 1 from the plots.

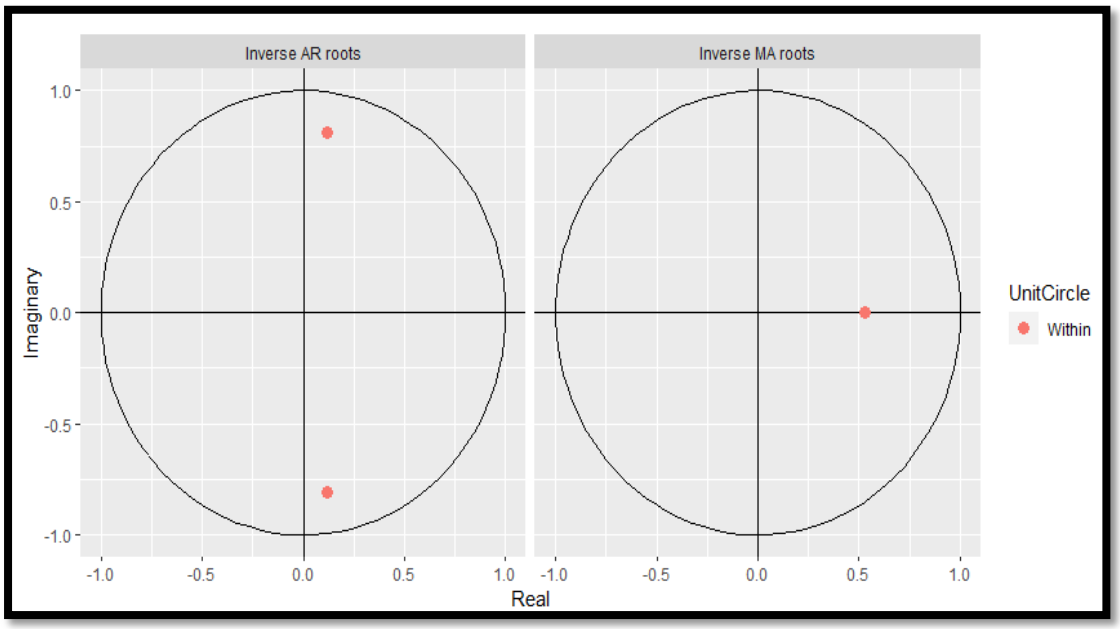

Figure 3. The Auto plot for ARIMA

From the information gathered from all the above processes, the AR is 2, MA is 1, and from the first differencing the integrated part is taken as 1. Finally, the (*p,d,q*) is (2,1,1).

Forecasted Plot for ARIMA (2,1,1) model

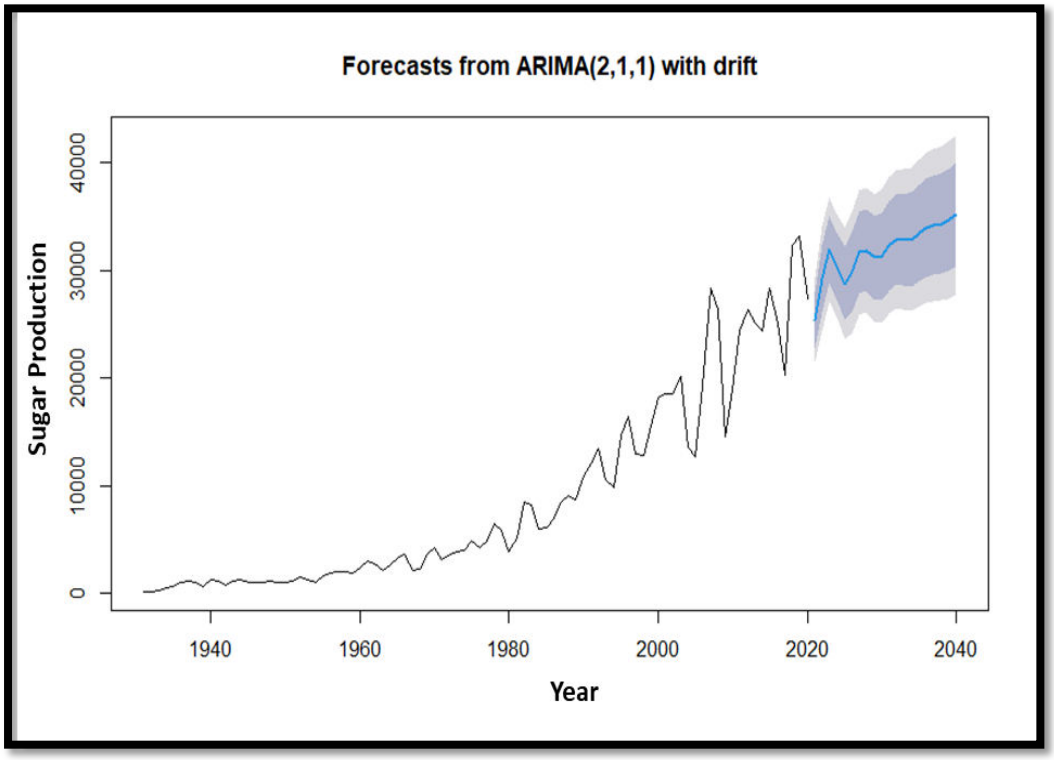

Figure 4. Forecasting for ARIMA (2,1,1) Model for next 20 years

Estimates offorecasted ARIMA (2,1,1)model

ARIMA(2,1,1) with drift

Table 2. Estimates of Sugar Production

Type	Coefficients	Standard Error
AR(1)	0.2371	0.0939
AR(2)	-0.6693	0.0819
MA(1)	-0.5283	0.0939

Summary Statistics of forecasted ARIMA (2,1,1) model

Table 3. Summary Statistics for Sugar Production

Sigma^2	Log likelihood	AIC	AICc	BIC
3995902	-801.5	1613	1613.73	1625.45

Error measures of forecasted ARIMA (2,1,1) model

Table 4. Error Measures

ME	-7.742992
RMSE	1942.655
MAE	1302.564
MPE	-26.04875
MAPE	33.10666
MASE	0.836996

Validation with Ljung-Box test

Table 5. Ljung Box Test

Lag order	5	6	7	8	9
X-Squared	12.613	13.966	17.895	23.127	28.258
df	4	4	4	4	4
***p*-value**	0.01333	0.007403	0.0012394	0.0001194	0.00000106

Artificial Neural Network

Model: **NNAR(10,6)**

Average of 20 networks, each of which is

Input layer is :10(The input neuron is the one third of the raw-data column)

Hidden Layer is : 6(Complex data, hence the better option is to choose the hidden neuron to 5 or 6)

Output layer is : 1 (Predicted ANN output layer should be one)

Forecast method: NNAR(10,6)

Error measures of forecasted NNAR (10,6) model

Table 6. Error Measures

ME	-0.8473831
RMSE	281.44
MAE	186.1505
MPE	-2.065659
MAPE	7.591274
MASE	0.1196158

Comparison between ARIMA & ANN

Error Measures

Table 7.ARIMA VS ANN

Models	**ME**	***RMSE***	***MAE***	***MPE***	***MAPE***	***MASE***
ARIMA	*-7.742992*	*1942.655*	*1302.564*	*-26.04875*	*33.1066*	*0.836996*
ANN	-0.8473831	*281.44*	*186.1505*	*-2.065659*	*7.591274*	*0.11961*

By paralleling the error measures between ARIMA & ANN, obviously the ANN values are comparatively lower than the ARIMA. Finally, it can be concluded that Artificial Neural Networks performs well for forecasting.

IV. Conclusion

Using data on sugar production in India, the Auto Regressive Integrated Moving Average and Artificial Neural Network models were applied. Error metrics were computed for each model and compared to determine which model performed better. After comparing the two models, it is found that the Artificial Neural Network model produced better prediction results, making it the superior model.These estimates will aid in the formulation of sound policies in India's sugar production industries.

References

[1] Paswan, S., Paul, A., Paul, A., & Noel, A. S. (2022). Time series prediction for sugarcane production in Bihar using ARIMA & ANN model. The Pharma Innovation Journal, 11

[2] Mrinmoy, R., Tomar, R. S., Ramasubramanian, V., & Singh, K. N. (2018). Forecasting sugarcane yield of India using ARIMA-ANN hybrid model. Bhartiya Krishi Anusandhan Patrika, 33(1/2), 120-127. (4), 1947-1956.

[3] Pannakkong, W., Huynh, V. N., & Sriboonchitta, S. (2016). ARIMA versus artificial neural network for Thailand's cassava starch export forecasting. Causal Inference in Econometrics, 255-277.

[4] Ravichandran, S., Yashavanth, B. S., & Kareemulla, K. (2018). Oilseeds production and yield forecasting using ARIMA-ANN modelling. The Indian Society Of Oilseeds Research, 57.

[5] Tyagi, S., Chandra, S., & Tyagi, G. (2023). Statistical modelling and forecasting annual sugarcane production in India: Using various time series models. Annals of Applied Biology, 182(3), 371-380.

[6] Fauziah, F. N., Gunaryati, A., & Sari, R. T. K. (2017). Comparison forecasting with double exponential smoothing and artificial neural network to predict the price of sugar. Int. J. Simul. Syst. Sci. Technol, 18(4), 1-8.

[7] Devi, M., Kumar, J., Malik, D. P., & Mishra, P. (2021). Forecasting of wheat production in Haryana using hybrid time series model. Journal of Agriculture and Food Research, 5, 100175.

[8] Paidipati, K., & Banik, A. (2019). Forecasting of rice cultivation in India–a comparative analysis with ARIMA and LSTM-NN models. EAI Endorsed Transactions on Scalable Information Systems, 7(24).

[9] Rathod, S. A. N. T. O. S. H. A., Singh, K. N., Patil, S. G., Naik, R. H., Ray, M. R. I. N. M. O. Y., & Meena, V. S. (2018). Modeling and forecasting of oilseed production of India through artificial intelligence techniques. Indian J. Agric. Sci, 88(1), 22-27.

[10] Rathod, S., Mishra, G. C., & Singh, K. N. (2017). Hybrid time series models for forecasting banana production in Karnataka State, India. Journal of the Indian Society of Agricultural Statistics, 71(3), 193-200.

[11] Rochman, E. M. S., Agustiono, W., Suryani, N., & Rachmad, A. (2021). Comparison between the backpropagation and single exponential smoothing method in sugar production forecasting case. Commun. Math. Biol. Neurosci., 2021, Article-ID.