Forecasting Soyabean Crop Production using Arima Model

Dr. B. Saidulu[1]; Dr. M. Raghavender Sharma[2]

[1,2] Department of Statistics, Osmania University

Publication Date: 2025/04/16

Abstract: In the realm of research, statistical modeling of non-stationary, non-linear statistics has grown to be a significant challenge. ANN and ARIMA are two of the most widely utilized models. This paper compares the Box-Jenkin's and Artificial Neural Network (ANN) approaches for estimating the actual value of the soybean harvest in India. The primary goal of this investigation is to create a forecasting model that can accurately anticipate India's agricultural production. In order to predict the annual production of the soybean crop in India, a statistical forecasting model utilizing Box-Jenkin's approach and artificial neural networks was created throughout this research. The model's ability to forecast was assessed using Mean Absolute Percent Error (MAPE) and Root Mean Squared Error (RMSE). The annual predictions recommend that, over a ten-year period, soybean crop production should be measured with an accuracy of 90% and a regular deviation of 13%.

Keywords: *ARIMA, Box-Jenkin's Methodology, ANN and MAPE.*

How to Cite: Dr. B. Saidulu[1]; Dr. M. Raghavender Sharma[2] (2025) Forecasting Soyabean Crop Production using Arima Model. *International Journal of Innovative Science and Research Technology*, 10(3), 3016-3023. https://doi.org/10.38124/ijisrt/25mar1605

I. INTRODUCTION

Statistics play a fundamental role in various sectors of analysis, serving as the basic object of study. Traditional statistical modeling often assumes a linear relationship between past and future values within a series. Agriculture has been the backbone of the Indian economy for many decades, driving growth and productivity across a wide range of crops. Soybean, one of the major crops in India, is cultivated over 29.94 million acres, making it a significant agricultural product in the country. It is primarily grown in the states of Madhya Pradesh, Maharashtra, Rajasthan, Karnataka, Telangana, and Chhattisgarh.

Crop rotation is a critical practice in soybean farming, promoting healthier oil production. Soybean oil is widely used in various applications, including baking, frying, and as a sustainable fuel. It also contributes to cleaner oceans, lakes, and rivers, and is a key ingredient in safer household products and better animal feed. Additionally, soybean has nutritional benefits for humans and is an essential component in the production of numerous food products.

Madhya Pradesh is the leading producer of soybean in India, followed by Maharashtra, Rajasthan, Karnataka, Telangana, and Chhattisgarh. Soybean's economic value extends beyond food; it is used in the production of tofu, soy milk, soy nuts, edamame, tempeh, and soy flour. The production of soybean in India is closely linked to the harvested area, and increasing production can significantly contribute to the development of rural areas by boosting the incomes of local producers.

Given the importance of soybean in the food industry, export revenues, and food security, it is crucial for India to adopt policies aimed at increasing soybean production for future sustainability. This paper aims to analyze the yearly production of soybean in India and forecast its future trends using statistical techniques. The following sections present the results based on the Box-Jenkins methodology and artificial neural networks, which were applied in this study.

II. MATERIALS AND METHODS

➢ *Box-Jenkin's Methodology:*

This section discusses modeling soybean production in India using the Box-Jenkins methodology, which involves fitting an ARIMA (Autoregressive Integrated Moving Average) model. The ARIMA model is defined by three parameters: p (autoregressive order), d (degree of differencing for stationarity), and q (moving average order). The model, ARIMA(p, d, q), helps capture patterns in historical production data to forecast future trends. The Box-Jenkins procedure includes selecting the appropriate values for p**,** d**,** and q, fitting the model, and using it for prediction.

$$\Phi(B)\nabla^{d}Zt = \theta(\mathrm{B})at$$

$$\text{Where } \Phi(B) = 1 - \Phi_{1}B - \Phi_{2}B^{2} - - - - - - \Phi_{p}B^{p}$$
$$\theta(B) = 1 - \theta 1B - \theta 2B^{2} - - - - \theta \mathrm{q}B^{\mathrm{q}}$$

$$\text{And } \nabla^{d} = (1 - B)^{d}$$

We have $B^{K}Zt = Zt–kandat$ is a white noise process with zero mean and variance σ^2.The Box-Jenkin's procedure consists of the subsequent four stages. The Box-Jenkins methodology involves four key steps:

- *Model Identification:*

The orders p**,** d**,** and q are determined by analyzing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

- *Estimation:*

Model parameters are estimated using the maximum likelihood method.

- *Diagnostic Checking:*

The adequacy of the model is tested using the Ljung-Box test, applied to the residuals.

- *Forecasting:*

Forecasts are made using the minimum mean square error method. (Box et al., 1994).

➢ *ANN Model:*

This study developed a statistical forecasting model using the Box-Jenkins method and artificial neural networks (ANN) to predict the yearly production of soybean in India. The model's performance was evaluated using Mean Absolute Percent Error (MAPE) and Root Mean Squared Error (RMSE). The projections indicated that soybean production would have a standard deviation of 13% and an accuracy of 90% over ten years.

ANNs are well-suited for predictions due to their non-linear structure, flexibility, and ability to estimate complex relationships. They can accurately approximate a wide range of practical relationships, making them a powerful tool for capturing the relationship between the target variable and relevant factors. Given these features, ANNs are highly adaptable for prediction, as most global systems are inherently non-linear (Naveen Kumar et al., 2011).

III. RESULTS

➢ *Forecasting Soyabean Crop Production Using ARIMA Model:*

The development of prediction models victimization is discussed in this paper. The annual production of the soybean crop using Box-Jenkin's methodology is discussed.

The Directorate of Economics and Statistics provided the information on the annual production of soybeans from 1961 to 2020, or 60 years. The model was built using the annual crop production from 1961 to 2010, and the model was validated using the annual soybean crop output from 2011 to 2020.

Prediction models for forecasting the yearly production of soybean crops were developed using Box-Jenkins methodology and Artificial Neural Networks (ANNs). The average annual production of soybeans was 96,000, though it fluctuated over time. The time trend of soybean production from 1961 to 2020 is shown in the following chart. The data reveals a non-stationary time series (Fig 1), with production varying due to factors like rainfall. In particular, production was relatively low in 2007 and high in 2012, reflecting the impact of varying rainfall levels during these years (Fig 1).

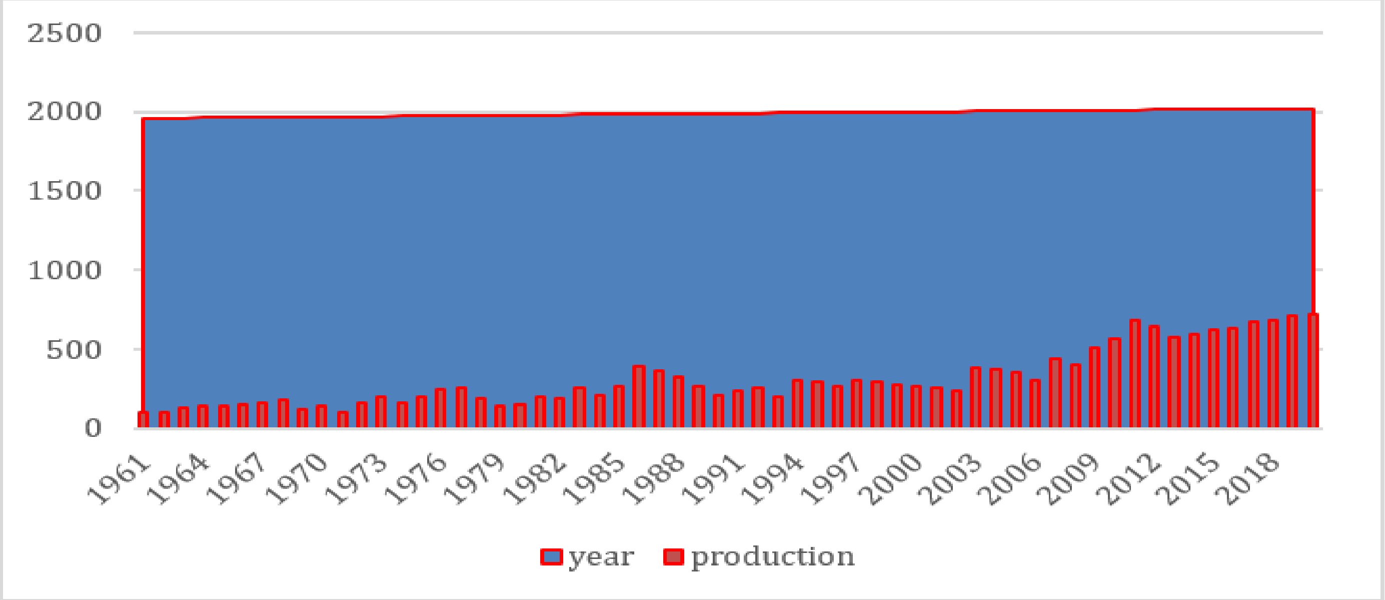


Fig 1 The Yearly Average for the Production of the Soyabean Crop (in Tonnes*)*

A proficient Python developer was employed to determine the optimal ARIMA model for predicting soybean production, as this approach automatically identifies and estimates the best-fitting ARIMA model for one or more variable series, eliminating the need for trial-and-error. The ARIMA (0, 1, 2) model, tested using ACF and PACF order-wise differencing and a validation set with the original series, was found to fit the data well. The model parameters are provided in Tables 1 and 2 below.

- *Original Series, Differencing, and ACF (Auto Correlation Function)*

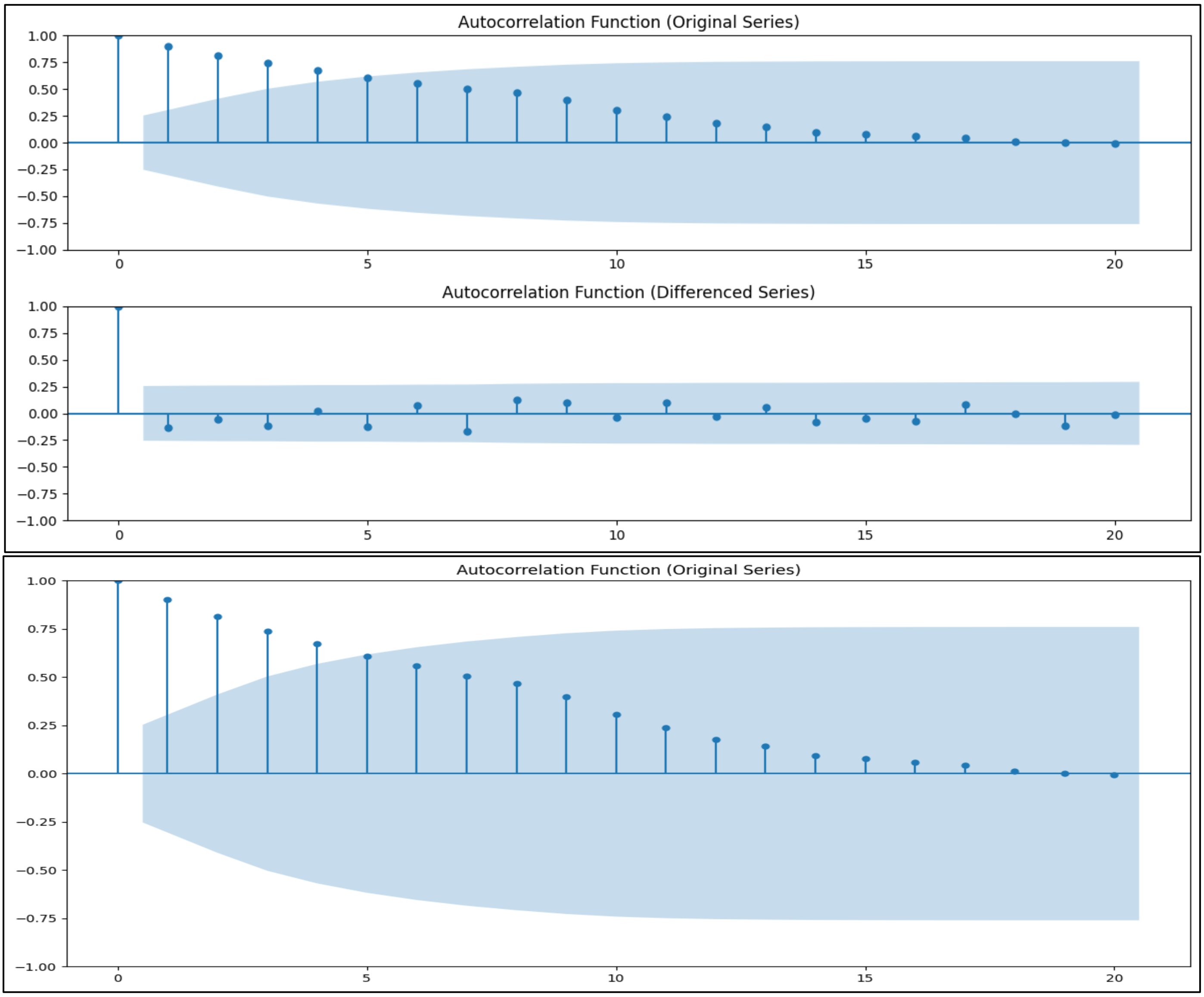


Fig 2 Time Series of Soyabean Production (Tonnes) in India Original Series with ACF

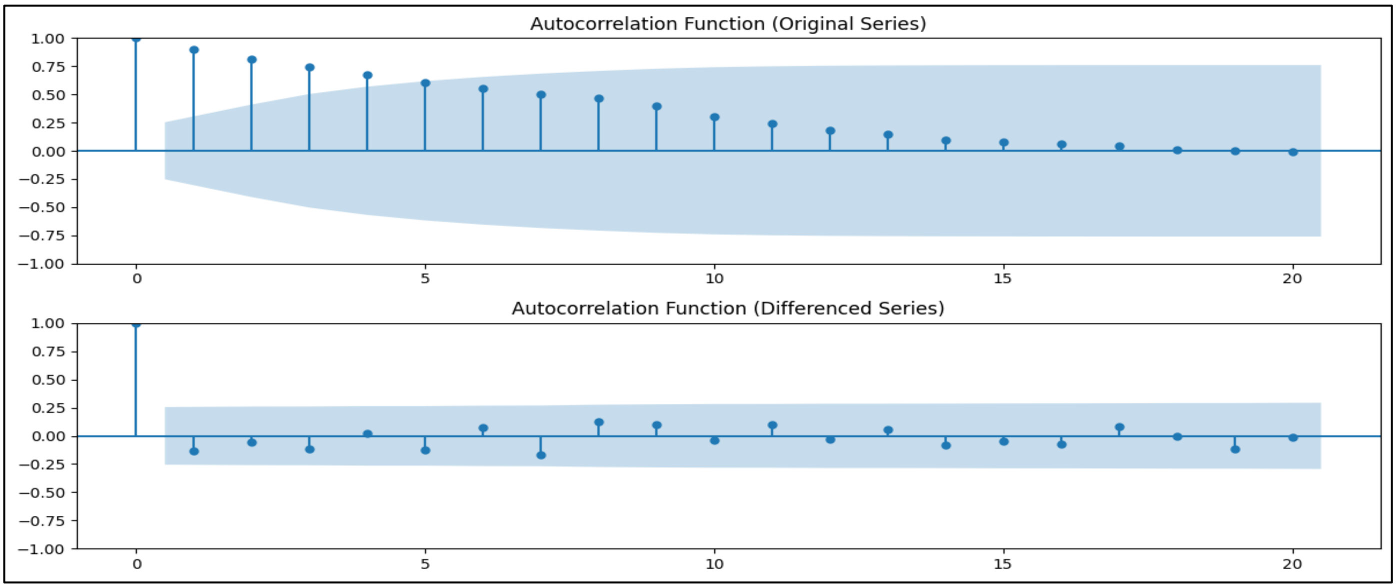


Fig 3 Time series of Soyabean production (tones) in India 1 st Order differencing with ACF

- *Original Series, Differencing, and PACF (Partial Auto Correlation Function)*

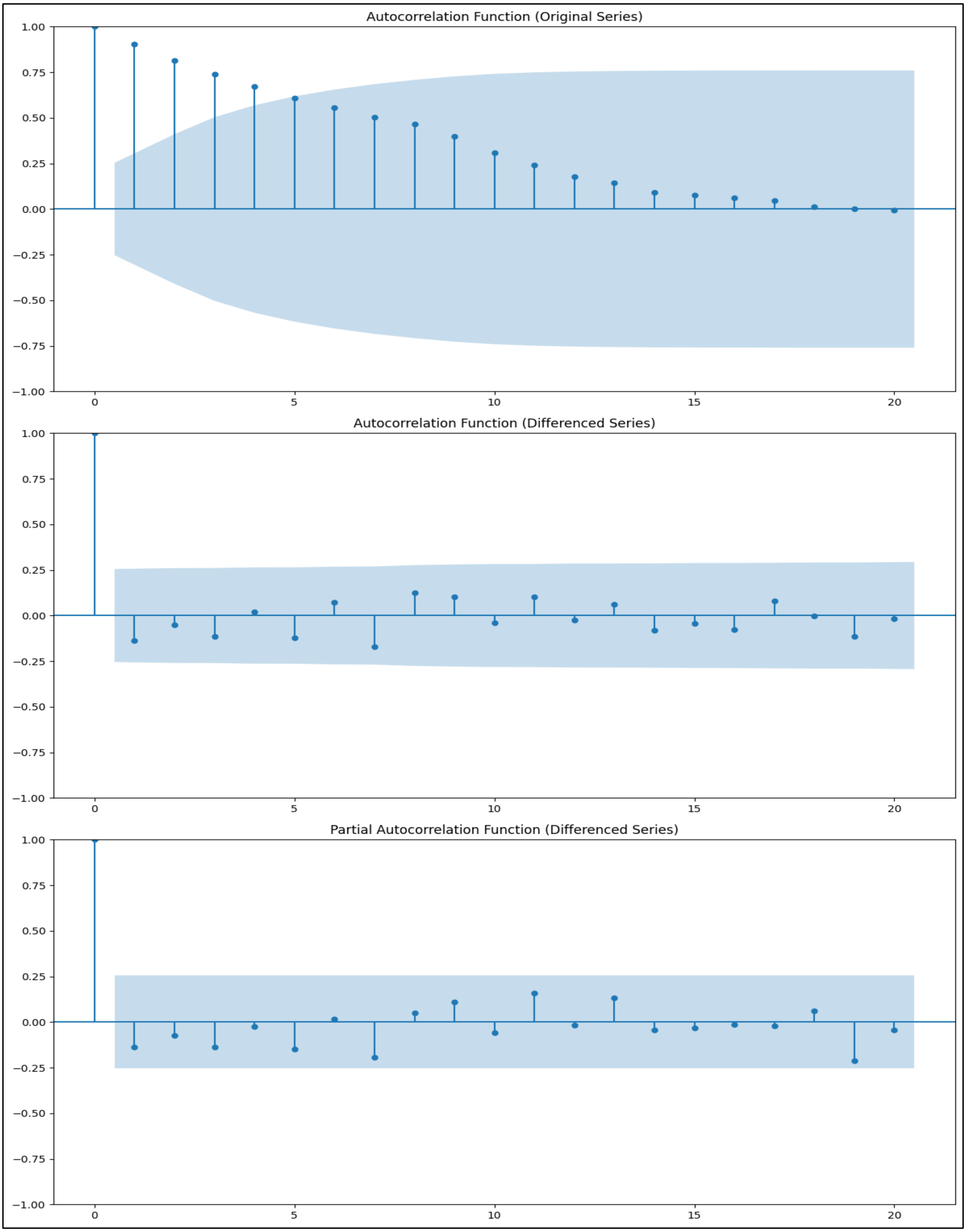


Fig 4 Time Series of Soyabean Production (Tones) in India, 1 st Order Differencing with PACF

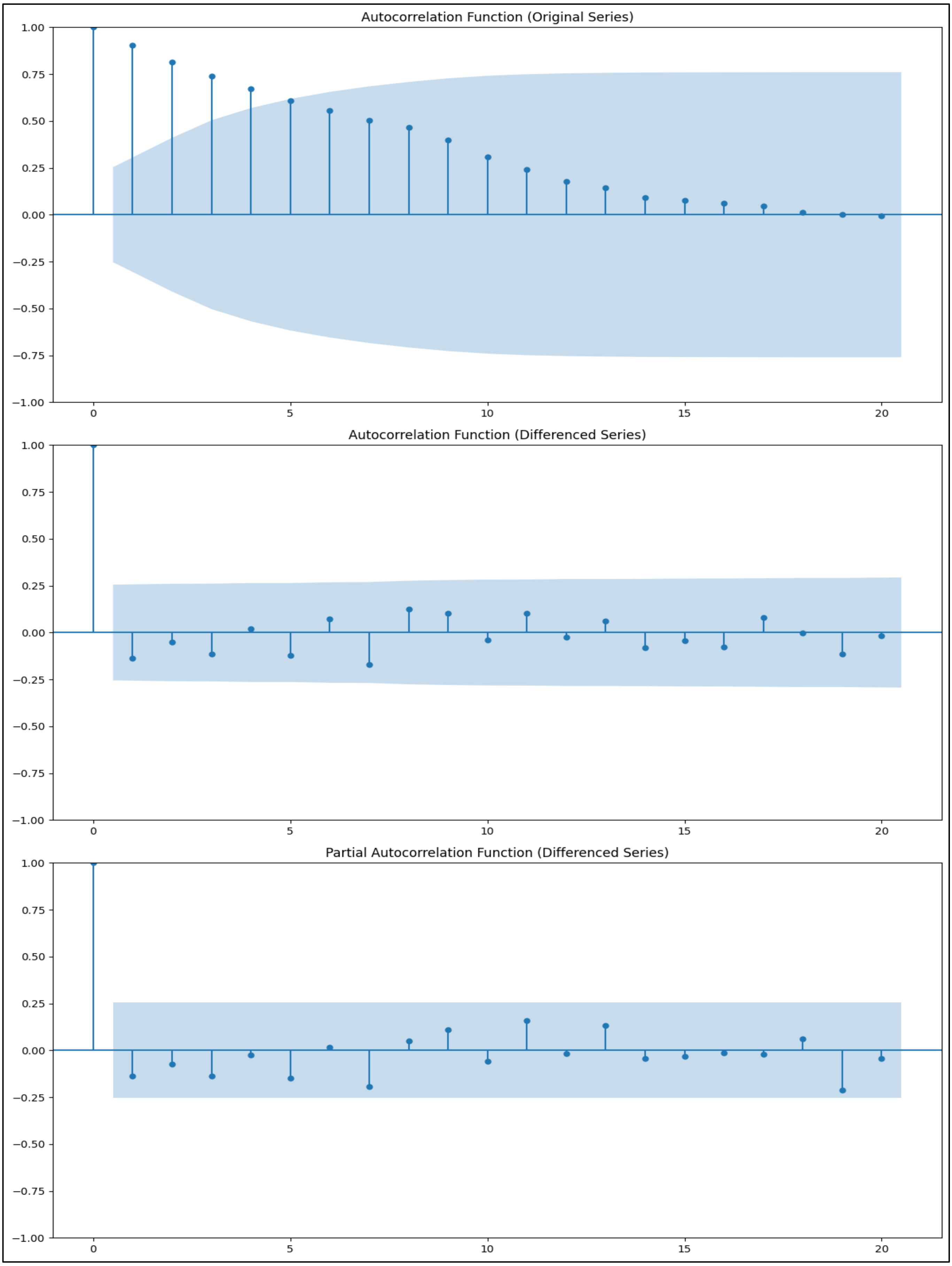
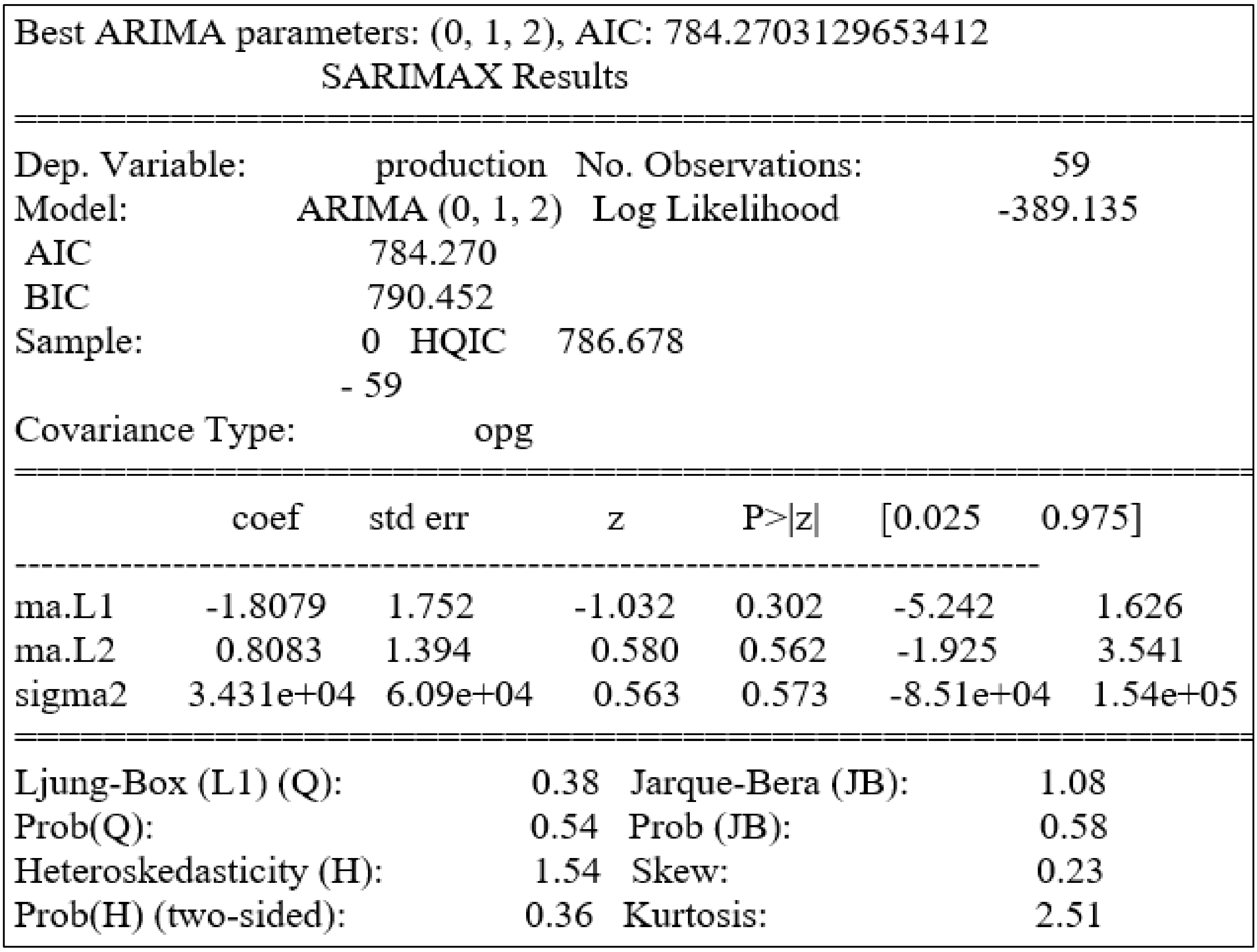


Fig 5 Time Series of Soyabean Production (Tonnes) in India 1st Order Differencing with ACF

Table 1 ARIMA Model Parameters



```
Best ARIMA parameters: (0, 1, 2), AIC: 784.2703129653412
                         SARIMAX Results
==============================================================================
Dep. Variable:          production  No. Observations:              59
Model:               ARIMA (0, 1, 2)  Log Likelihood             -389.135
 AIC                      784.270
 BIC                      790.452
Sample:                  0   HQIC     786.678
                       - 59
Covariance Type:               opg
==============================================================================
                 coef     std err          z        P>|z|     [0.025     0.975]
------------------------------------------------------------------------
ma.L1        -1.8079     1.752       -1.032      0.302     -5.242       1.626
ma.L2         0.8083     1.394        0.580      0.562     -1.925       3.541
sigma2     3.431e+04   6.09e+04       0.563      0.573   -8.51e+04   1.54e+05
==============================================================================
Ljung-Box (L1) (Q):                 0.38   Jarque-Bera (JB):            1.08
Prob(Q):                            0.54   Prob (JB):                   0.58
Heteroskedasticity (H):             1.54   Skew:                        0.23
Prob(H) (two-sided):               0.36   Kurtosis:                    2.51
```

Hence, the fitted model for the forecasting of Soyabean production in India is ARIMA (0,1,2) $\nabla^1 Zt = (1 + 0.043B^8 - 0.112B^{13})\, at$.

The Ljung-Box Q test statistic was used to evaluate the adequacy of the model, with the result showing Q = 11.376 at 20 degrees of freedom. The null hypothesis of a sufficient model was accepted, as the corresponding p-value of 0.85 was significantly greater than 0.05. Therefore, the ARIMA (0, 1, 2) model is considered a suitable model for predicting soybean production. Additionally, a Python-based model using artificial neural networks was developed to forecast soybean yield.

- *Artificial Neural Networks Model:*

The ANN model used for forecasting soybean production consists of an input layer, a hidden layer with two neurons to capture non-linearity, and an output layer. The one-step-ahead prediction is based on standardized prior observations (Lag-1). The output layer uses an identity function, and the hidden layer uses the hyperbolic tangent function. The model was trained using backpropagation until the testing sample's error was smaller than the training samples.

Table 2 Feed forward Neural Networks Model Parameters

Predictor		**Predicted**		
		Hidden Layer 1		**Output Layer**
		H(1:1)	H(1:2)	S_t
Input Layer	(Bias)	.249	.128	
	lag1	.196	.009	
Hidden Layer 1	(Bias)			.153
	H(1:1)			.163
	H(1:2)			.098

- *Comparison of Arima and Ann Models*

The training sample contrasted the predictions made by the two models, and the testing samples confirmed the mean absolute error, mean absolute percentage error, and root means square error. The error measures from the ARIMA and ANN prediction models are shown in the following Table 3.

Table 3 Comparison of the Forecasting Performance of ARIMA and ANN Models

Measure	**Training Sample**		**Testing Sample**	
	ARIMA	ANN	ARIMA	ANN
MAE	292.01	261.21	251.05	271.12
RMSE	264.13	214.01	261.03	209.12
MAPE	0.03	0.09	0.01	0.03

The ANN model shows lower error measures than the ARIMA model in the testing sample. While it fits the data better in both the training and testing samples, it doesn't outperform ARIMA in the testing sample. Figure 7 compares the out-of-sample predictions, with ARIMA showing a linear trend and ANN closely following the original values. The table below presents the predictions from both models.

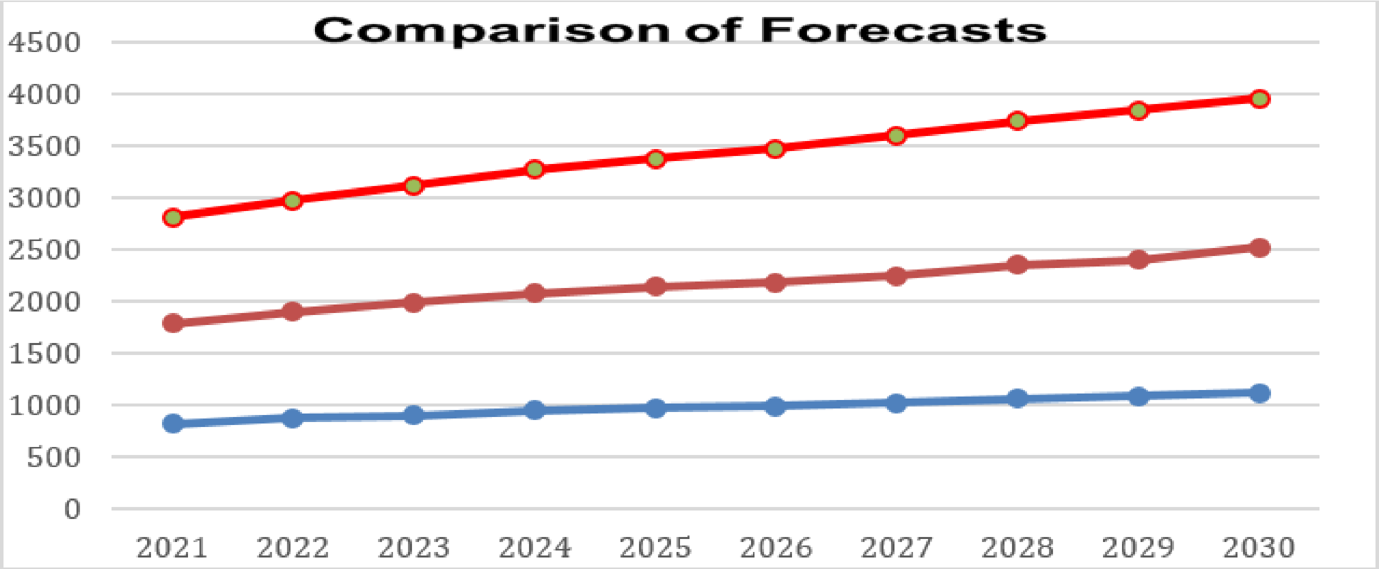


Fig 7 Comparison of Forecasts for Soyabean Production in India (in Tonnes)

Table 4 Out-of-Sample Forecasts of Soyabean Production in India using ARIMA and ANN Models

Production			
Year	**Soyabean (in tonnes)**	**ANN**	**ARIMA**
2021	821.25	968.28	1023.54
2022	881.02	1025.95	1068.12
2023	901.29	1094.21	1125.19
2024	956.12	1128.01	1185.91
2025	977.81	1168.01	1235.96
2026	997.12	1186.21	1289.01
2027	1028.16	1224.01	1356.01
2028	1068.25	1289.21	1382.16
2029	1086.56	1321.01	1436.2
2030	1125.65	1398.01	1431.98

IV. DISCUSSION AND CONCLUSION

According to the forecasts, the ANN model outperforms the ARIMA model in predicting India's soybean production. The ANN model effectively captures the nonlinear fluctuations and complex patterns within the data, providing a more accurate representation of the production trends over time. In contrast, the ARIMA model, while useful for capturing linear trends, fails to account for the nonlinear dynamics present in the soybean production data. As a result, the ANN model produces forecasts that are closer to the actual variations observed, reflecting the underlying complexity in production patterns, whereas the ARIMA model is limited to projecting simple linear trends that do not fully capture the seasonal and irregular fluctuations in the data. Therefore, the ANN model is more suited for forecasting soybean production, especially when dealing with non-linear and dynamic time series data.

REFERENCES

[1]. M. Raghavender Sharma et al (2016) Paddy Production in Telangana State: Current and Future Trends, Volume: 6 | Issue: 3 | March 2016 | ISSN - 2249-555X | IF: 3.919 | IC Value: 74.50

[2]. Ramu Yerukala (2008), ―Identification of Linear Time Series Models‖, unpublished M.Phil. Dissertation submitted to Osmania University, Hyderabad.

[3]. Satish, G. (2004), ― Application of time series and NN based short term load forecasting‖, Unpublished M.Tech. Project submitted to JNTU, Hyderabad.

[4]. Haykin, S.S., (1999), "Neural Networks: A Comprehensive Foundation", Upper Saddle River, N.J., PrenticeHall.

[5]. Hornik, K, (1993), Some new results on neural network approximation, Neural Networks, 6, 1069-1072.

[6]. K. Murali Krishna, Dr. M. Raghavender Sharma and Dr. N. Konda Reddy, forecasting of silver prices using Artificial Neural Networks. JARDCS, Volume 10, 06 issue 2018.R. J. Vidmar. (1992, Aug).

[7]. Makridakis S., Wheel Wright. S.C., Hyndman R.J., 2003, Forecasting Methods and Applications, John Wiley &Sons.

[8]. Manish kumar. and thenmozhi. M. (2012). Stock Index Return Forecasting and Trading Strategy using Hybrid ARIMA – Neural Network Model, Vol. 1(1).

[9]. Ramakrishna. R., Naveen Kumar.B and Krishna Reddy. M. (2011), Forecasting daily electricity load using neural networks, International Journal of Mathematical Archive, Vol.2, 1-11.

[10]. Tang, Z., Almeida, C.D. and Fishwick, P.A., 1991, "Time Series Forecasting using Neural Networks Vs. Box-Jenkins Methodology", Simulation, Vol.57, No.5,303-310.

[11]. Peter Zhang, G. (2004), ―Business Forecasting with Artificial Neural Networks: An Overview‖, Georgia State University, US, Idea Group Inc.