

EFFICIENCY AND PRECISION OF STOCHASTIC MODELS IN AGRICULTURAL PRODUCTIVITY

Vyshnavi Muraleedharan^{1,*} and Muthukumar Madaswamy²

Received: June 08, 2024; Revised: January 15, 2025; Accepted: March 03, 2025

Abstract

In this paper, we compare the Markov chain and ARMA models for predicting oilseed productivity, focusing on the area under cultivation, production, and yield. This study aims to evaluate the accuracy and efficiency of these two forecasting techniques and their capacity to predict future oilseed output patterns. Historical information on area, production, and yield was acquired and prepared for examination. The Markov chain model was created to depict the probability of changes between different productivity levels. On the other hand, the ARMA model was used to analyze time series data on the productivity of oilseeds, and suitable parameters for capturing temporal trends were discovered. Metrics including MAE, RMSE, MAPE, MSE, and SMAPE were used to evaluate the performance of both models. The research findings from the comparative analysis illuminate the pros and cons of each model, as well as their potential applications in forecasting oilseed output. Remarkably, the Markov chain model performed better in this experiment than the ARMA method.

Keywords: Markov Process; ARMA Modeling; Oilseed; Yield

Introduction

India is the world's fourth-largest producer of oilseeds. 10% of the world's production and 20.8% of the land area under cultivation are accounted for by it. The country produces a variety of oilseeds, including sesame, mustard, safflower, and niger seed. India is a significant supplier of oilseeds to several nations worldwide. Apart from the United States, the United Arab Emirates, the Netherlands, Belgium, China, Australia, Singapore, Germany, Hong Kong, Indonesia, Nepal, Saudi Arabia, and Italy, India also ships oilseeds to these nations. The cultivation and production of plants whose seeds, or

oilseed crops, are grown primarily for their oil content is known as oilseed agriculture. These oilseeds are necessary for producing vegetable oils used in food preparation, industrial uses, and as feedstock for biodiesel. A mathematical model that captures the unpredictable and erratic nature of oilseed productivity over time is known as a stochastic process. A series of random variables indicating the oilseed productivity at various time points define a stochastic process. $X(t)$, where 't' stands for the time index, is commonly used to refer to the variables. Each variable $X(t)$ is a random

Department of Statistics, PSG College of Arts and Science, Avinashi Road, Civil Aerodrome Post, Coimbatore – 641014, Tamil Nadu, India. E-mail: vyshnavimp@gmail.com; muthukumar@psgcas.ac.in

* Corresponding author

DOI: <https://doi.org/10.55766/sujst5229>

Suranaree J. Sci. Technol. 32(3):030310(1-12)

result that is affected by various hard-to-predict internal and external influences. Oilseed productivity can be modelled using the Markov chain and the Arma model, two different stochastic process types. Stochastic processes offer helpful tools for examining patterns in historical oilseed productivity data and forecasting future productivity. These models consider the inherent uncertainty in agricultural systems, which helps explain the unpredictable nature of oilseed productivity and the numerous variables that might affect yields.

Markov chain models' ability to simulate the stochastic nature of yield changes makes them useful in oilseed productivity. In contrast to standard time series models, Markov chains explicitly account for the probabilities of migrating from one productivity level to another, depending on previous observations (Waldemar *et al.*, 2018). Understanding yield transitions allows agricultural stakeholders to estimate the risk of potential yield changes better, optimize resource allocation, and design more informed strategies to increase productivity and profitability in the oilseed sector.

Time series analysis is a helpful technique for predicting the output of oilseeds because it enables the evaluation of historical trends and patterns to forecast future productivity. A popular time series forecasting method that works well in many fields, including meteorology, finance, and economics, is the Autoregressive Moving Average (ARMA) model. The ARMA model is a valuable tool when analyzing historical data, spotting trends, and forecasting future crop yields for the oilseed farming industry.

Oilseeds are crucial in supplying critical vegetable oils for various activities, including cooking, food processing, industrial uses, and feedstock for biodiesel. Applying sophisticated mathematical models like Markov chains and ARMA models is necessary due to the stochastic nature of crop productivity over time. Informed decision-making, resource optimisation, and strategic planning are made possible by these stochastic processes, which also allow for the analysis of historical production data. This helps maintain India's position as a significant participant in the global oilseed market while enhancing productivity. Given the importance of oilseeds to India's agriculture and economy, this study looks at their yields. The production of edible oils, which are necessary for nutrition and food security, depends on oilseeds. As a major contributor to agricultural exports and a source of income for millions of farmers, the industry makes a substantial economic contribution. To achieve its local needs, India still depends mostly on imported edible oil. Enhanced

agricultural planning, decreased reliance on imports, and increased production can all be achieved with accurate yield forecasts. This study evaluates advanced oilseed production prediction models to overcome these issues.

Aim and Objectives

This paper aims to forecast oilseed productivity by using an appropriate time series modelling approach. To achieve this aim, the following objectives have been identified:

- To propose oilseed forecasting methods using Markov and ARMA models.
- To inspect the accuracy of Markov and ARMA models in forecasting ability.

Review of Literature

Sabu and Kumar (2020) conducted a forecasting areca nut pricing in Kerala, India, which can assist the agricultural supply chain in controlling price swings and lowering output as a result of climate change and variable weather. Because of price swings and climate change, farmers in Kerala are switching from growing areca nuts to other crops. This study forecasts Kerala's monthly average prices using time series and machine learning algorithms. Using the arecanut dataset from 2007 to 2017, the models SARIMA, Holt-Winter's Seasonal approach, and LSTM neural network were assessed using the RMSE value. It was discovered that the LSTM neural network model fit the data the best. Mohammad Rahim Rahnama (2021) conducted a study on forecasts for 2020-2030 and models changes in land use and land cover in the Mashhad metropolitan region from 2016-2020. The study used the Markov chain model in ArcGIS and TerrSet software, Sentinel-2A satellite images, and Cellular Automata to find that land use decreased greatly with mass and light vegetation and grew dramatically with built-up regions and bare fields. According to the CA Markov chain model, LULC would alter from 2020 to 2030 in a positive way for bare areas and negatively for thick and light vegetation. According to the study, the north and east of the city will see the greatest spatial-temporal change from greenery and farmland to built-up areas and bare ground. These results lend assistance to planning and decision-making processes for sustainable land-use development.

Zulyadi and Sembiring (2015) developed the purpose of making decisions and implementing policies in the future, the government need accurate information on the farmer's terms of trade of the smallholder's estate crops subsector. Delays in FTTS provision, however, can be problematic. The ARIMA method was compared with a forecasting

model created utilizing Markov chains and the predictive, probabilistic architecture modeling framework (P2AMF). With reduced MSE and MAPE in both experiments, the Markov chain approach is superior to ARIMA.

Alani and Alhiyali (2021) forecasted Markov Chain, the study sought to forecast Iraq's wheat crop productivity from 2019 to 2022 and investigate methods for enhancing it. Self-sufficiency and food security are impacted by the crop's low production, which is concerning. The study discovered that changes in the region's overproduction were causing a continuous decline in output. Because it has been shown to have a positive impact on productivity, the research suggests vertical intensification in agriculture. Additionally, it implies that vertical intensification ought to work well with other elements such as better seeds, effective herbicides, and fertilizers. The study also highlights how crucial efficiency and quality are to management. Because it requires fewer statistical tests and less rigorous assumptions, the Markov chains method is suggested for forecasting. Papageorgiou *et al.* (2013) presented the application of fuzzy cognitive maps (FCMs) to apple yield modeling and prediction is investigated in this work. FCMs offer inference skills and help model dynamic systems. Expert knowledge was utilized to classify apple yield utilizing a data-driven non-linear FCM learning approach. The model's nodes, which represent soil variables influencing yield, are connected by directed edges. The study demonstrated the superiority of the FCM technique in yield prediction by contrasting the FCM learning algorithm with benchmark machine learning algorithms and traditional instruments.

Kumar *et al.* (2014) argued that for development to be sustainable, land use and land cover must be altered. Images from Indian Remote Sensing Satellites were used in research in Tiruchirappalli city between 1998 and 2006 to examine LU/LC dynamics. Future landscape distribution was calculated between 2014 and 2022 using a Markov model to analyze spatial and temporal change. Zhang *et al.* (2011) presented the Wetland distribution maps produced using CBERS-02B photos and the Land-sat 5 Thematic Mapper. These maps and the GIS were used to create a transition probability matrix. To forecast wetland kinds and distribution area trends, a Markov model was used. With a 98.5% accuracy rate, the model suggests that local governments can create wetland management policies. Zhao *et al.* (2001) presented the efficient determination criteria (EDC), a novel and more versatile criterion for shaping a Markov chain's order. The estimations' consistency and strength have been established in benign

circumstances. Zheng and Liu (2009) suggested that separate approaches cannot create accurate prediction models and that combined forecasting strategies should be the main focus of future perspectives. Combining non-linear models, such as the grey model and neural networks, offers promise for accident prediction.

Materials and Methods

Markov Chain

A mathematical framework called a Markov chain is used to depict the probability transitions of a system as it undergoes state changes over time. (Halilčević and Gubina, 2011). This stochastic process follows the Markov property, which states that the system's future state is solely influenced by its current state and is independent of previous circumstances (Zakaria *et al.*, 2019). The events that lead to the current state are not stored in the Markov chain, in other words. A Markov chain can be created to represent the various productivity levels of the crops (for example, low, medium, and high) as stated in the forecasting of oilseed output. Based on records of oilseed productivity, transition probabilities between these stages are determined. One can predict potential future yield distributions and evaluate the likelihood of switching from one level to another by examining the Markov chain model. The term 'Markov dependence' refers to a simple type of reliance between values of X_n in subsequent transitions that Russian mathematician A.A. Markov developed. Markov dependence is a type of dependence in which x_{n+1} is independent of $x_{n-1}, x_{n-2} \dots x_0$ and only depends on x_n when it is known. This suggests that regardless of the past, the process's present actions determine its future. Markov property is the name of this property (Zulyadi and Sembiring 2015; Łukasz *et al.*, 2020).

$$\begin{aligned} P\{X(t_n)=x_n \mid X(t_{n-1})=x_{n-1}, \dots, X(t_0)=x_0\} \\ = P\{X(t_n) = x_n \mid X(t_{n-1}) = x_{n-1}\} \end{aligned}$$

for all $X(t_0), X(t_1), \dots, X(t_n)$ $t_0 < t_1 < \dots < t_n$ indicates the time intervals, followed by the set of random variables $\{X(t_n)\}$ whose state space S .

Transition Probability Matrix (TPM)

The transition probability is the probability of changing from one condition to another or staying put for a specific amount of time (Du *et al.*, 2020).

$$P_{x(n-1),x(n)} = P\{X(t_n) = x_n \mid X(t_{n-1}) = x_{n-1}\}$$

is referred to as the first-order Markov chain transition probability. This exhibits the system's current state's conditional probability x_n at time t_n on condition that it was formerly in the state x_{n-1} at time t_{n-1} . The transition probabilities can be set up in a matrix of size $m \times n$, known as a one-step Transition Probability Matrix (TPM), and shown as follows (Al-sharif *et al.*, 2013):

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & & \ddots & \vdots \\ \vdots & & \cdots & P_{mn} \end{bmatrix}, \text{ where } m \text{ and } n \text{ represent}$$

the number of states. The members of the square matrix P are all non-negative, and the sum of the elements in each row is unity i.e. $\sum P_{ij} = 1$, $0 \leq P_{ij} \leq 1$ (Arumugam and Rajathi (2020)).

The Markov Chain forecasting process follows a series of steps, starting with the classification of historical data into different states. A transition probability matrix is then developed to capture the likelihood of moving between states over time. Based on this matrix and the initial state, the model estimates future states, which are ultimately converted into numerical predictions. The complete procedure is illustrated in Figure 1.

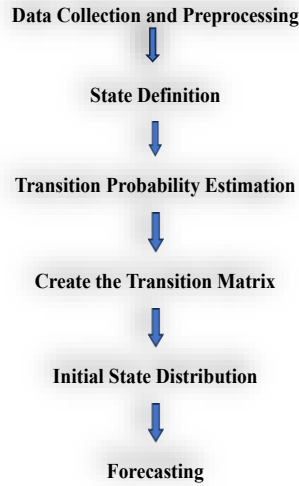


Figure 1. The Markov Chain flowchart for the forecasting procedure

Algorithm for Markov Chain Forecasting

- Categorize numerical data into discrete states based on values, dividing data points into low and high.
- Calculate the transition probabilities between discrete states based on the data.
- Evaluate the model's performance on the testing data using a performance metric.

Choose the model order that minimizes either the AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). Lower AIC/BIC values indicate a better-fitting state.

- Based on the data's first state, calculate each state's initial probability distribution.
- Forecast the next state using calculated TPM and initial probabilities, updating probabilities based on the forecasted state.

Autoregressive Moving Average (ARMA)

An autoregressive model is comparable to a linear regression model. Yule initially presented them in 1926. Let X_t stand for the level of the series at time t , which is the value of a time series with no trend but fluctuations around a fixed value. Current X_t values depend on previous values from the same series at any time " t " when an autoregressive model (Obasohan, 2020) is used.

$$Y_t = C + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots \phi_p X_{t-p} + e_t$$

Where level, X_{t-1} , X_{t-2} , ..., X_{t-p} are past series values (lags) and e_t denote the error term or residual, which captures the random fluctuations or unmodeled influences on the data. It reflects the difference between the actual values and the predicted values, typically assumed to have a zero mean and constant variance. The C term represents the constant or intercept, serving as the baseline value of the series when all lagged terms are zero. This allows the model to account for any inherent offset or non-zero mean in the data. Lastly the $\phi_1, \phi_2, \dots, \phi_p$ are the autoregressive coefficients, indicating the extent to which past values of the series influence the current value. These coefficients are key in autoregressive models, determining the strength and nature of the relationships between past and present data points. If X_{t-p} is the furthest lag with a nonzero coefficient, the AR model is said to order p , denoted $AR(p)$ (Padhan, 2012). There are specific restrictions on the allowable values of the autoregressive parameters.

- $p = 1$: For a simple AR (1) model, the autoregressive parameter (ϕ_1) must satisfy $-1 < \phi_1 < 1$. This ensures that the process remains stationary and doesn't diverge or blow up over time.
- $p = 2$: In this case, two autoregressive parameters (ϕ_1 and ϕ_2). The following three conditions need to be satisfied simultaneously, $-1 < \phi_2 < 1$, $\phi_1 + \phi_2 < 1$, $\phi_1 - \phi_2 < 1$.
- $p = 3$: The conditions for p greater than or equal to 3 can become more complicated and

involve more autoregressive parameters. The goal is still to maintain stationarity and stability in the model. These conditions might vary based on the specific order of the AR(p) model and the interplay between the autoregressive coefficients.

ARIMA is an extension of the ARMA model that incorporates differencing to handle non-stationarity in the data. It stands for Autoregressive Integrated Moving Average. The ARMA model analyzes and forecasts time-dependent data by combining autoregressive (AR) and moving average (MA) components. It is especially beneficial for stationary time series data when statistical features such as mean and variance do not fluctuate over time. The ARMA forecasting process follows a series of steps from data preparation to model fitting and prediction, as shown in Figure 2.

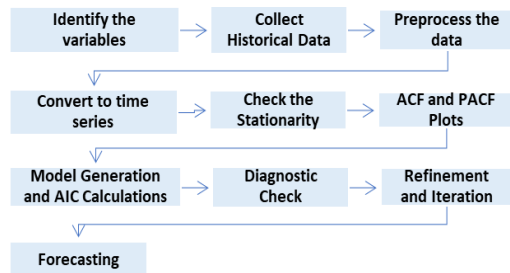


Figure 2. ARMA flow chart for the forecasting procedure

The construction of an autoregressive moving average (ARMA) model involves a series of steps:

- Assessing data stationarity through the KPSS Test.
- Identifying potential ARMA models by determining order specifications through lag correlation analysis performed on the ACF (Auto Correlation Function) and PACF (Partial Auto Correlation Function) plots.
- Estimating the significance of parameters in the ARMA model at a predetermined significance level, typically $\alpha = 5\%$.
- Select the optimal ARMA model by choosing the one with the lowest Akaike Information Criterion (AIC) value, which is determined by the following formula (Darekar and Amarender Reddy 2017):

$$AIC = n \ln(\sigma_{\epsilon}^2) + 2k$$

For n is the number of observations, σ_{ϵ}^2 is the maximum likelihood estimator, and k is the number of parameters estimated.

When the model has been validated, use it to forecast future values.

Stationary Time Series Autocorrelation Function (ACF)

An analytical tool for measuring the correlation between a time series sequence and a lagged version of itself is the autocorrelation function. In other words, it measures how closely a data point is related to a data point from a preceding time step (lag). The ACF is frequently used in signal processing and time series analysis to comprehend the patterns and connections within a dataset. According to mathematics, the autocorrelation function of a time series y_t at lag k is given by,

$$ACF(k) = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}$$

The ACF is between -1 and 1. The time series at the current time step and the time series at the lagged time step are said to be positively correlated when the value is positive. A low number denotes a poor correlation. Zero means there is no correlation.

Partial Autocorrelation Function (PACF)

After reducing correlations at shorter delays, the PACF, a statistical tool used in time series analysis, measures the direct link between two time series data. For autoregressive (AR) models, which are frequently used to model time series data, it aids in determining the proper lag order. The correlation between the time series value at time t and the value at time $t-k$, while accounting for the intermediate values between t and $t-k$, is known as the PACF at lag k .

Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS)

The KPSS test is a statistical method used to assess the stationarity of a time series. It tests whether a series is stationary around a deterministic trend or at a constant level. The null hypothesis assumes that the time series is stationary, while the alternative hypothesis suggests that the series exhibits a trend or unit root, indicating non-stationarity. A p-value below 0.05 typically rejects the null hypothesis, implying non-stationarity.

Methods of Accuracy Measures

In statistics, the MSE, MAPE, MAE, RMSE, and SMAPE are used to measure how well each forecasting technique matches the data. These statistics don't offer much information on their own, but they can be applied to the fits discovered by various techniques (Jon and David, 1983).

Mean Square Error

The Mean Square Error (MSE) for an estimator evaluates the average of the squared errors or variances.

$$MSE = \sum_{i=1}^n \frac{(A_i - F_i)^2}{n}$$

Mean Absolute Percentage Error

MAPE, known as the Mean Absolute Percentage Error, assesses the percentage difference between the actual values and the predicted values, computed as an average across all observations (Shahwan and Odening, 2007).

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - F_i}{A_i} \right| \times 100\%$$

Where n is the number of observations.

A_i is the actual value of the observation i.

F_i is the forecasted or predicted value for observation i.

Mean Absolute Error

MAE, or Mean Absolute Error, quantifies the average absolute disparity between predicted and actual values.

$$MAE = \frac{1}{n} \sum_{i=1}^n |A_i - F_i|$$

Root Mean Square Error

RMSE, which stands for Root Mean Square Error, serves as a metric for gauging the average magnitude of the discrepancies between predicted values and actual values. It shares similarities with MAE but places more emphasis on penalizing larger errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (A_i - F_i)^2}{n}}$$

Symmetric Mean Absolute Percentage Error

SMAPE calculates the absolute percentage error for each observation and then divides it by the average of the actual and predicted values.

$$SMAPE = \frac{1}{n} \sum_{i=1}^n \frac{|A_i - F_i|}{(A_i + F_i)/2} \times 100\%$$

Results and Discussion

Source of Data

The data utilized in this research was obtained from the website of the Economics & Statistics Division within the Ministry of

Agriculture & Farmers Welfare, accessible at <http://desagri.gov.in>. This research utilizes historical data on oilseed productivity, focusing on three key aspects: area under cultivation, production, and yield. The data is collected annually, spanning a total of 17 years, from the 2005-2006 to 2021-2022 period. The study uses 17 data points for each variable (area, production, and yield). This dataset provides the foundation for analyzing long-term trends and accurately predicting oilseed productivity. The data is organized and stored in a structured format, facilitating its use in forecasting models such as the Markov Chain and ARMA methods. This research uses three data sets with the following units: Area is measured in Million Hectares (Million Ha), representing the land used for oilseed cultivation; Production is quantified in Million Tonnes (Million MT), indicating the total harvested amount of oilseeds; and Yield is expressed in Kilograms per Hectare (Kg/Ha), reflecting the oilseed output per unit of land. These units are crucial for analyzing and forecasting oilseed productivity trends.

Descriptive Statistics

The statistical analysis reveals moderate variability across the datasets, as summarized in Table 1. The means are close to the medians, indicating near-symmetric distributions. There is a slight positive skew in some variables, while others exhibit a negative skew, suggesting a concentration of higher values. The kurtosis values point to light-tailed distributions. The largest range and variance highlight greater fluctuation in certain variables, while others show more consistency. These findings suggest that the data is suitable for modeling with time series methods such as Markov Chains and ARMA models to explore trends and make predictions.

Table 1. Shows the summary statistics for the three datasets

Statistic	Area	Production	Yield
Mean	26.76	30.26	1130.41
Median	26.51	30.94	1168
SE	0.309	0.902	29.678
S.D	1.275	3.721	122.367
Skewness	0.175	0.147	-0.376
Kurtosis	-0.213	-0.274	-1.169
Sample Var	1.626	13.844	14973.76
Range	4.66	13.41	376

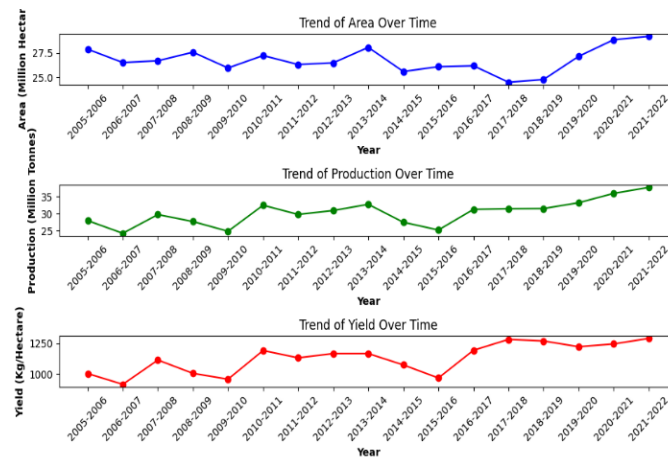


Figure 3. Time Series Trends of Oilseed Data Productivity

The long-term patterns in oilseed productivity offer important insights into historical changes and underlying trends. Figure 3 displays the time series plot of oilseed productivity, highlighting periods of fluctuation, phases of growth, and possible cyclical behavior evident in the dataset.

Markov chain- Model selection

As shown in Table 2, State 2 achieves the highest log-likelihood and lowest AIC and BIC scores for the 'Area', 'Yield' and 'Production', making it the best fit for the model. When the measures are examined further, state 2 has a higher log-likelihood and the lowest AIC and BIC scores among the studied states. This convergence of measures emphasizes state 2's consistency in terms of both goodness of fit and model complexity.

Table 2. Shows the model selection evaluation for oilseed forecasting: Log-Likelihood and Information Criteria Analysis

Field	States	Log-likelihood	AIC	BIC
Area	2	-5.534	17.068	18.763
	3	-7.755	29.050	28.965
	4	-8.677	47.354	55.828
Production	2	-5.227	17.045	18.740
	3	-9.463	34.926	39.445
	4	-10.003	50.007	58.482
Yield	2	-8.487	22.975	24.670
	3	-10.229	36.459	40.978
	4	-6.133	42.267	50.742

Markov Chain - State Boundaries

The state transition matrix categorizes variable values into discrete states for 'Area', 'Production', and 'Yield', with ranges of values for

'Low' and 'High' for 'Area' and 'Production' and numerical ranges for 'Yield' as shown in Table 3.

Table 3. Shows the boundaries of yield values

State Yield	Low	High
Area	(21.49, 24.50)	(24.50, 27.52)
Production	(14.84, 20.02)	(20.02, 25.19)
Yield	(691, 877.5)	(877.5, 1064)

Transition Probability Matrix

The probabilities of changing between various states within each variable are quantified by these matrices, providing helpful information on how these variables change over time and enabling predictive modeling and analysis. The initial step is to define the states to establish the TPM for a two-state. In the case of oilseed productivity, these could be 'Low Productivity' (L) and 'High Productivity' (H). Afterwards, historical data must be collected, tracking how frequently the system transitions from one state to another over time. This data should capture the frequency of transitions such as Low to High, High to Low, or remaining in the same state (Low to Low or High to High). With this information, calculate the probabilities of each transition by dividing the number of occurrences of each transition by the total occurrences from the starting state. These probabilities are then placed into a 2×2 matrix. Each row of the matrix represents the probabilities of moving from one state to the others (Low to Low, Low to High, High to Low, and High to High). The sum of each row must equal 1, ensuring that the system will transition from one state to one of the two available states. The TPM is essential for predicting future states in a Markov

Chain model, as it helps estimate future transitions based on observed past data.

$$\text{Area, TPM} = \begin{matrix} L & H \\ \text{Area, TPM} = \begin{matrix} L & H \\ 0.8 & 0.2 \\ 0.14 & 0.86 \end{matrix} \end{matrix}$$

$$\text{Production, TPM} = \begin{matrix} L & H \\ \text{Production, TPM} = \begin{matrix} L & H \\ 0 & 1 \\ 0.2 & 0.8 \end{matrix} \end{matrix}$$

$$\text{Yield, TPM} = \begin{matrix} L & H \\ \text{Yield, TPM} = \begin{matrix} L & H \\ 0.5 & 0.5 \\ 0.75 & 0.25 \end{matrix} \end{matrix}$$

Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS Test)

Time series data stationarity is evaluated using the KPSS test. However, it's crucial to remember that the KPSS test focuses primarily on whether there are any trends in the data. R software validates the data.

In all three fields-‘Area’, ‘Production’, and ‘Yield’-the KPSS test outcomes indicate that the data is stationary, as shown in Table 4. This stationary behaviour suggests that the data exhibits a consistent pattern over time, which is vital for accurate time series analysis and forecasting.

Table 4. Shows the boundaries of yield values

Field	KPSS Level	Lag order	P-Value
Area	0.234	2	0.100
Production	0.0911	2	0.100
Yield	0.243	2	0.100

The generation of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

In time series analysis, ACF and PACF are fundamental methods for comprehending and modelling the autocorrelation and partial autocorrelation patterns within a specific time series dataset.

Figures 4 - 6, present the correlograms of the ACF and PACF for oilseed productivity. These plots help identify the presence of autocorrelations within the production data, providing valuable insights into potential seasonal patterns and lag dependencies.

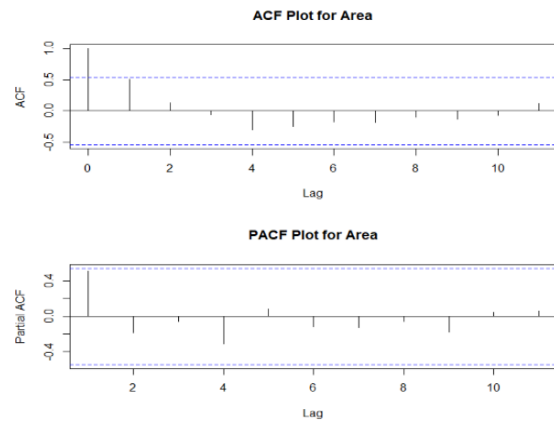


Figure 4. ACF and PACF correlogram – Area

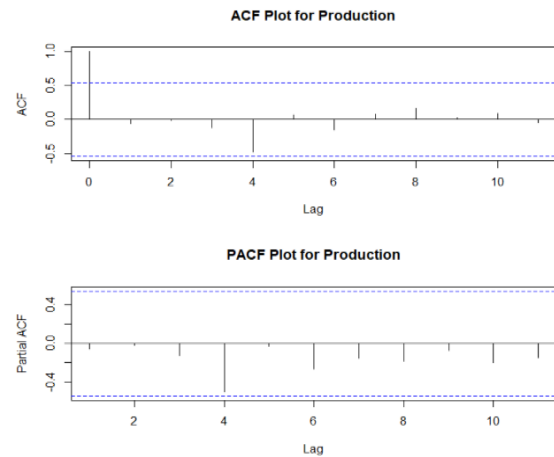


Figure 5. ACF and PACF correlogram – Production

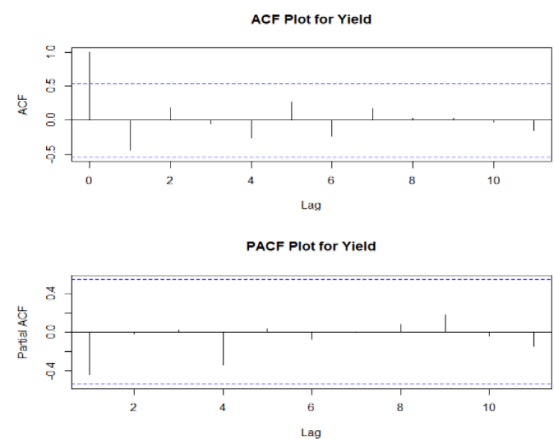


Figure 6. ACF and PACF correlogram – Yield

Model Selection Criteria

As shown in Table 5, ARMA (1,1) was selected for Area based on its lowest AIC (51.91), indicating it provides the best fit for predicting oilseed area. For Production, ARMA (3,2) was chosen with an AIC of 69.12, while for Yield, ARMA (1,1) was selected with an AIC of 155.86. These models were preferred because they offered the best balance of predictive performance and model simplicity, demonstrating the importance of AIC in model selection to ensure accurate and interpretable forecasts.

Forecasted Value

Table 6 presents a comparative analysis of the actual values alongside the forecasted values obtained from the Markov Chain and ARMA models for the three key parameters across the study period from 2005-2006 to 2021-2022. The Markov Chain forecasts demonstrate better alignment with observed data across all parameters, highlighting its superior predictive performance. In contrast, the ARMA model tends to underestimate the actual

values, particularly in several years, suggesting its limitations in capturing the complex dynamics influencing oilseed productivity. This comparison further supports the selection of the Markov Chain model as a more reliable approach for forecasting within this study.

Accuracy Metrics Comparison for Oilseed Productivity Forecasting Model

As shown in Table 7, analysis of the accuracy metrics reveals that the choice of forecasting model has a substantial influence on the precision of oilseed productivity projections in India. The Markov Chain model consistently outperforms the ARMA models, showing better performance across all evaluation metrics, such as MSE, RMSE, MAE, and MAPE. This underscores the importance of model selection in agricultural forecasting, as it significantly affects the accuracy and reliability of predictions. In conclusion, the Markov Chain model offers a more dependable and precise approach for forecasting oilseed productivity, making it the more suitable choice for this study.

Table 5. Shows the AIC values for different models

Model	AIC(Area)	AIC(Production)	AIC(Yield)
ARMA (1,1)	51.91	69.24	155.86
ARMA (2,1)	53.81	70.85	157.72
ARMA (1,2)	53.06	71.69	157.11
ARMA (3,1)	55.74	70.06	159.21
ARMA (1,3)	53.12	70.76	159.05
ARMA (2,3)	52.07	71.41	160.44
ARMA (3,2)	54.09	69.12	158.68
ARMA (2,2)	53.38	72.21	158.49
ARMA (3,3)	53.70	72.17	162.15
ARMA (2,0)	52.45	71.10	157.90

Table 6. Results of Forecasting the data using the Markov Chain and ARMA Model (Area in Million Hectare, Production in Million Tonnes, and Yield in Kg/ Hectare)

Year	Actual Value			Forecasted Markov Chain			Forecasted ARMA model		
	Area	Production	Yield	Area	Production	Yield	Area	Production	Yield
2005-2006	27.86	27.98	1004	25.161	21.432	864.239	21.360	24.924	785.61
2006-2007	26.51	24.29	916	25.013	21.442	863.152	21.408	23.354	859.86
2007-2008	26.69	29.76	1115	24.910	21.439	863.424	21.436	21.267	862.39
2008-2009	27.56	27.72	1006	24.838	21.440	863.356	21.453	19.290	862.47
2009-2010	25.96	24.88	958	24.786	21.439	863.373	21.463	18.573	862.480
2010-2011	27.22	32.48	1193	25.161	21.440	863.369	21.470	19.276	862.481
2011-2012	26.31	29.80	1133	25.013	21.443	863.370	21.473	21.045	862.484
2012-2013	26.48	30.94	1168	24.910	21.442	863.369	21.475	22.943	862.481
2013-2014	28.05	32.75	1168	24.838	21.441	863.370	21.477	24.07	862.480
2014-2015	25.60	27.51	1075	24.786	21.440	863.374	21.478	23.92	862.485
2015-2016	26.09	25.25	968	24.746	21.444	863.38	21.478	22.65	862.482
2016-2017	26.18	31.28	1195	24.715	21.447	863.36	21.478	20.917	862.480
2017-2018	24.51	31.46	1284	24.690	21.444	863.37	21.479	19.57	862.482
2018-2019	24.79	31.52	1271	24.669	21.450	863.37	21.479	19.25	862.481
2019-2020	27.14	33.22	1224	24.650	21.443	863.38	21.477	20.04	862.489
2020-2021	28.83	35.95	1247	24.632	21.442	863.37	21.479	21.503	862.480
2021-2022	29.17	37.70	1292	24.616	21.45	863.33	21.480	22.904	862.482

As shown in Table 7, analysis of the accuracy metrics reveals that the choice of forecasting model has a substantial influence on the precision of oilseed productivity projections in India. The Markov Chain model consistently outperforms the ARMA models, showing better performance across all evaluation metrics, such as MSE, RMSE, MAE, and MAPE. This underscores the importance of model selection in agricultural forecasting, as it significantly affects the accuracy and reliability of predictions. In conclusion, the Markov Chain model offers a more dependable and precise approach for

forecasting oilseed productivity, making it the more suitable choice for this study.

Data Visualization: Actual and Forecasted Values

Figures 7-9 show encouraging agreement between actual and anticipated data points in the area, production, and yield dimensions, demonstrating the model's predictive capability in considering complex factors affecting oilseed crop production, efficiency, and culture.

Table 7. Accuracy Measurements for Markov Chain and ARMA Models

Field	Model	MSE	MAPE	RMSE	MAE	SMAPE
Area	Markov Chain	5.258	7.103%	2.293	1.951	7.459%
	ARMA (1,1)	29.634	19.632%	5.443	5.299	21.863%
Production	Markov Chain	90.901	28.132%	9.534	8.824	33.324%
	ARMA (3,2)	94.312	27.862%	9.711	8.763	33.37%
Yield	Markov Chain	85,389.88	22.728%	292.21	267	26.15%
	ARMA (1,1)	87,526.17	23.279%	295.8	272.6	26.81%

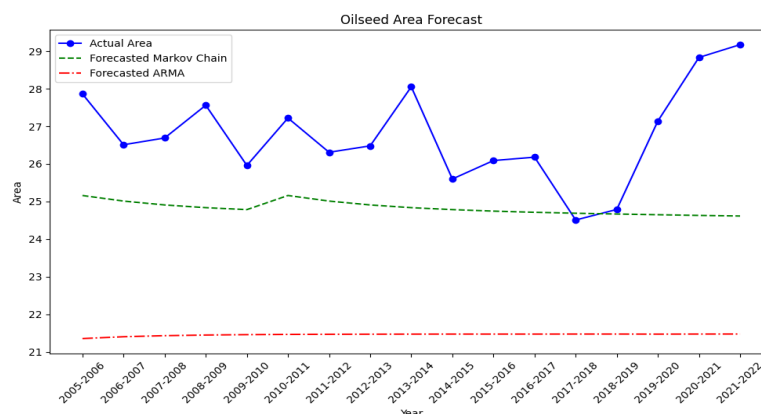


Figure 7. Assessment of actual and forecasted oilseed area trends using Markov Chain and ARMA approaches

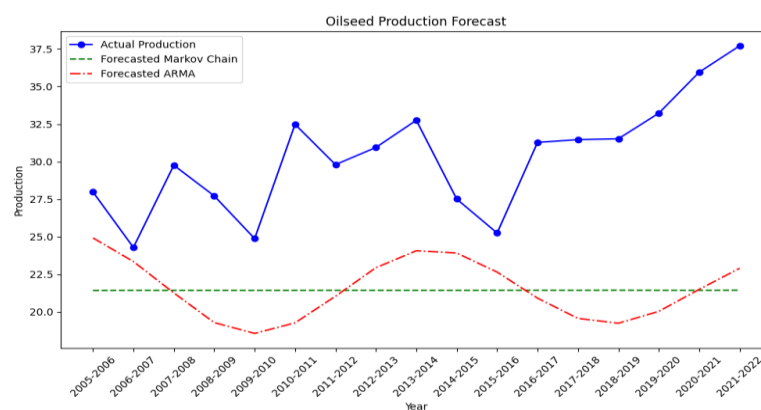


Figure 8. Actual and forecast oilseed production productivity over the years is represented

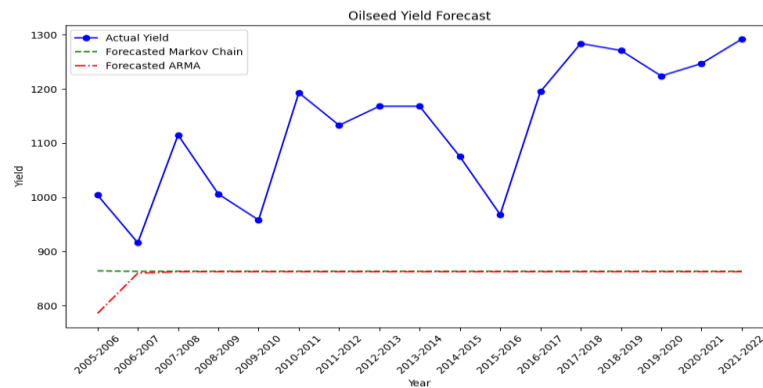


Figure 9. Shows actual and forecast oilseed yield productivity over the years

Conclusions

ARMA models may not be suitable for long-term forecasting due to their reliance on historical data trends. Long-term trends and variations in factors like climate change may not be well-expressed in terms of oilseed productivity. ARMA assumes stationary data, which may not apply to agricultural productivity data due to exogenous influences like climate change and governmental interventions. The Markov Chain model consistently exceeds the ARMA model regarding forecast accuracy in all three categories (Area, Production, and Yield), as shown by the lower values of the MSE, RMSE, MAPE, MAE, and SMAPE metrics. This implies that the Markov Chain model performs better across these aspects when forecasting oilseed productivity in India. It's essential to keep in consideration, nevertheless, that the model selection process should also take domain expertise, the integrity of the data, and the unique features of the Indian oilseed market into account. Even if the Markov Chain model has potential, it is advisable to view these findings completely and consider any possible upgrades or alternative modelling techniques, if any are available. The findings of this study are consistent with prior research, which has shown that Markov chain models tend to outperform ARMA models in agricultural forecasting. Previous studies have highlighted the Markov chain's ability to accurately capture changes in productivity states, whereas ARMA models often face challenges with the complexities of agricultural time series data. This study supports the effectiveness of Markov chains for forecasting oilseed productivity.

Acknowledgement

I thank the Economics & Statistics Division within the Ministry of Agriculture & Farmers Welfare for providing the data utilized in this research.

References

- Alani, L.A.F. and Alhiyali, A.D.K. (2021). Forecasting wheat productivity in Iraq for the period 2019-2025 using Markov chains. *Iraqi Journal of Agricultural Sciences*, 52(2):411-421. <https://doi.org/10.36103/ijas.v52i2.1302>
- Al-sharif, Abubakr A.A. and Pradhan, Biswajeet. (2014). Monitoring and predicting land use change in Tripoli Metropolitan City using an integrated Markov chain and cellular automata models in GIS. *Arabian Journal of Geosciences*, 7(10):4291-4301. <https://doi.org/10.1007/s12517-013-1060-7>
- Arumugam, R. and Rajathi, M. (2020). A Markov model for prediction of coronavirus COVID-19 in India: A statistical study. *Journal of Xidian University*, 14(4):1,422-1,426. <https://doi.org/10.37896/jxu14.4/164>
- Brandt, Jon A., and Bessler, David A. (1983). Price forecasting and evaluation: An application in agriculture. *Journal of Forecasting*, 2(3), 237-248. <https://doi.org/10.1002/for.3980020306>
- Darekar, A. and Reddy, A Amarender (2017). Forecasting of Common Paddy Prices in India. *Journal of Rice Research*, 10(1):71-75. <https://doi.org/10.2139/ssrn.3064080>
- Du, H., Zhao, Z., and Xue, H. (2020). ARIMA-M: A new model for daily water consumption prediction based on the autoregressive integrated moving average model and the Markov chain error correction. *Water*, 12(3):760.
- Halilčević, S.S., and Gubina, A.F. (2011). The short-term electricity prices forecasting using Markov chains. In 2011 8th International Conference on the European Energy Market (EEM), 1-6. <https://doi.org/10.1109/EEM.2011.5953008>

- Kumar, S., Radhakrishnan, N., and Mathew, S. (2014). Land use change modelling using a Markov model and remote sensing. *Geomatics, Natural Hazards and Risk*, 5(2):145-156. <https://doi.org/10.1080/19475705.2013.795502>
- Lukasz S., Gola, A., and Kozłowski, E. (2020). Predictive scheduling with Markov chains and ARIMA models. *Applied Sciences*, 10(17):6,121. <https://doi.org/10.3390/app10176121>
- Obasohan, P.E. (2020). Comparing weighted Markov chain and auto-regressive integrated moving average in the prediction of under-5 mortality annual closing rates in Nigeria. *International Journal of Statistics and Probability*, 9(3):13-22. <https://doi.org/10.5539/ijsp.v9n3p13>
- Padhan, P.C. (2012). Application of ARIMA model for forecasting agricultural productivity in India. *Journal of Agriculture and Social Sciences*, 8(2), 50-56.
- Papageorgiou, E.I., Aggelopoulou, K.D., Gemtos, T.A., and Nanos, G.D. (2013). Yield prediction in apples using Fuzzy Cognitive Map learning approach. *Computers and Electronics in Agriculture*, 91(5):19-29. <https://doi.org/10.1016/j.compag.2012.11.008>
- Rahnama, M.R. (2021). Forecasting land-use changes in Mashhad Metropolitan Area using Cellular Automata and Markov chain model for 2016-2030. *Sustainable Cities and Society*, 64:102548. <https://doi.org/10.1016/j.scs.2020.102548>
- Sabu, Kiran M. and Kumar, T.K. Manoj. (2020). Predictive analytics in agriculture: Forecasting prices of Areca nuts in Kerala. *Procedia Computer Science*, 171:699-708. <https://doi.org/10.1016/j.procs.2020.04.076>
- Shahwan, T., and Odening, M. (2007). Forecasting agricultural commodity prices using hybrid neural networks. In Chen, S.H., Wang, P.P., and Kuo, T.W. (Eds.), *Computational Intelligence in Economics and Finance*, p 109-128. Springer.
- Waldemar, B., Knopik, L., Jacek, Ż., Kuśmierk-Tomaszewska, R., and Żarski, W. (2018). Markov chain as a tool for forecasting daily precipitation in the vicinity of the city of Bydgoszcz, Poland. *ITM Web of Conferences*, 23(2018):00003. <https://doi.org/10.1051/itmconf/20182300003>
- Zakaria, N.N., Othman, M., Sokkalingam, R., Daud, H., Abdullah, L., and Abdul Kadir, E. (2019). Markov chain model development for forecasting air pollution index of Miri, Sarawak. *Sustainability*, 11(19):5,190. <https://doi.org/10.3390/su11195190>
- Zhang, R., Tang, C., Ma, S., Yuan, H., Gao, L., and Fan, W. (2011). Using Markov chains to analyze changes in wetland trends in arid Yinchuan Plain, China. *Mathematical and Computer Modelling*, 54(3-4):924-930. <https://doi.org/10.1016/j.mcm.2010.11.017>
- Zhao, L.C., Dorea, C.C. Y., and Gonçalves, C. R. (2001). On determination of the order of a Markov chain. *Statistical Inference for Stochastic Processes*, 4(3):273-282. <https://doi.org/10.1023/A:1012245821183>
- Zheng, X., and Liu, M. (2009). An overview of accident forecasting methodologies. *Journal of Loss Prevention in the Process Industries*, 22(4):484-491. <https://doi.org/10.1016/j.jlp.2009.03.005>
- Zulyadi, and Sembiring, J. (2015). Design of FTTS forecasting model using Markov chain and P2AMF framework: Case study - Farmer's Terms of Trade of Smallholders Estate Crops Subsector in Riau. In 2015 International Conference on Information Technology Systems and Innovation (ICITSI), 1-6. IEEE. <https://doi.org/10.1109/ICITSI.2015.7437725>