STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 119–135, https://doi.org/10.59139/stattrans-2026-007
Received – 08.12.2024; accepted – 22.09.2025

The comparison of the hidden Markov model with machine learning techniques in agricultural prediction

Muraleedharan Vyshnavi[1], **Madaswamy Muthukumar**[2]

Abstract

This study compares hidden Markov models (HMMs) with various machine learning approaches to assess their effectiveness in forecasting agricultural data based on Python. Accurate forecasts are essential to promote sustainability and increase agricultural productivity. Through the use of an extensive dataset of agricultural parameters, specifically the cultivated area of oilseeds, the study explores historical trends and correlations. Model performance is evaluated using the Mean Absolute Error (MAE) along with the R-squared, and residual analysis is used to analyse how well the models represent the underlying trends. The findings demonstrate that HMMs are able to predict agricultural trends with higher accuracy than their other counterparts, thereby providing useful information for improved agricultural planning and decision-making. Future studies should concentrate on improving forecast accuracy and resolving any issues associated with agricultural data prediction.

Key words: machine learning, mean absolute error, R-squared, hidden Markov model, residual analysis.

1. Introduction

Technological developments in data science and machine learning (ML) have revolutionized the agriculture industry in the last few decades. With these advanced tools, farmers, agronomists, and policymakers can now maximize productivity, control risks, and improve decision-making. Due to their capacity to manage the intricate temporal and spatial interactions present in agricultural systems, HMM and other ML techniques have become essential tools for assessing agricultural data among the variety of methodologies available. The agricultural sector faces distinct difficulties that are defined by unpredictability and are impacted by market forces, environmental variables, and technological advancements. While traditional statistical techniques have yielded valuable

[1] Department of Statistics, PSG College of Arts & Science, Coimbatore-641014, Tamil Nadu, India. E-mail: vyshnavimp@gmail.com. ORCID: https://orcid.org/0009-0000-1098-8161.
[2] Department of Statistics, PSG College of Arts & Science, Coimbatore-641014, Tamil Nadu, India. E-mail: muthukumar@psgcas.ac.in.

© M. Vyshnavi, M. Muthukumar. Article available under the CC BY-SA 4.0 licence

insights, they are still limited in their ability to capture complex patterns and interconnections seen in dynamic agricultural systems. This constraint has prompted the use of new computational techniques that can better describe uncertainties, nonlinear relationships, and stochastic processes.

A strong foundation for simulating temporal sequences in which underlying states are deduced from observable data is provided by HMMs. HMMs, which were first created in the fields of speech recognition and computational linguistics, have found extensive use in agriculture for tasks like predicting crop yield, identifying disease outbreaks, and evaluating the effects of climate change. Because HMMs are probabilistic, state transitions and emissions can be flexibly modeled to account for the inherent complexities and uncertainties seen in agricultural datasets. Techniques for ML include a broad range of algorithms that can recognize patterns in data and forecast outcomes without the need for explicit programming. In agriculture, ML models such as SVMs, NNs, LR, and KNN, have been extensively applied across various domains including crop classification, soil fertility prediction, yield estimation, and pest management. These techniques leverage large datasets to identify patterns and relationships, offering valuable insights for optimizing agricultural practices and resource allocation.

Poonam Somani, Shreyas Talele, and Suraj Sawant (2014) investigated the application of SVMs, NNs, and HMM to stock market forecasting. It emphasizes how crucial these methods are for reporting on changes in stock prices and suggests an HMM to increase precision. Choukri Djellali and Mehdi Adda (2020) presented the RHMM deep learning model, which leverages ANN and the HMM for recommender systems. Through experimentation, the model outperforms testing models by optimizing the bias-variance conflicts and enhancing training stability and accuracy. Pasak Senawongse, Andrew R. Dalby, and Zheng Rong Yang (2005) predicted phosphorylation sites in amino acid sequences, the study used HMMs which outperform decision trees and neural network methods and produce predictions with greater certainty than HMMs. Lefteris Benos, et al. (2008) examined the potential of machine learning in agriculture, emphasizing managing crops, water, soil, and livestock. It emphasizes the effectiveness of algorithms using Artificial Neural Networks, with the most researched crops being sheep, cattle, wheat, and maize. To summarize, this study aims to extend our understanding of how machine learning approaches, such as HMMs, might improve the analysis and prediction of long-term oilseed area patterns. This research aims to give practical insights to stakeholders in the agricultural industry by assessing the strengths and limitations of various models, ultimately contributing to more informed decision-making.

2. Methodology

2.1. Hidden Markov Model

A Hidden Markov Model is a statistical model describing systems with unobserved (hidden) states. It works particularly well with time series data, where the system is considered to follow a Markov process with unknown parameters. It is utilized in this study to model and forecast the actual annual oilseed area values. The model begins by defining hidden states S = $\{S_1, S_2, \ldots, S_N\}$ that represent various underlying conditions impacting oilseed areas, such as different weather patterns or economic factors. These hidden states correspond to observable data O = $\{O_1, O_2, \ldots, O_T\}$, which in this case are the measurements of the oilseed area. The HMM (Leite et al. 2008) parameters are the following,

- Transition Probabilities(A): A matrix A = $[a_{ij}]$ where a_{ij}= P $(q_{t+1} = S_j / q_t = S_i)$ indicates the probability of transitioning from state $S_i to\ S_j$.
- Emission Probabilities (B): A matrix B = $[b_j(k)]$ where b_j(k) = P $(o_t = O_k / q_t = S_j)$ specifies the probability of observing an oilseed area value O_k given that the system is in state S_j.
- Initial State Distribution (π): A vector π = $[\pi_i]$ where $\pi_i = P(q_1 = S_i)$ denotes the probability of starting in each hidden state.

The model is trained using the Baum-Welch algorithm, which is an expectation-maximization method that iteratively estimates these parameters based on the observed oilseed area data. This algorithm enhances the model's comprehension of state transitions and emissions to maximize the likelihood P(O/ʎ) of the observed data, where ʎ represents the model parameters A, B, and π. After, training HMM can predict the actual oilseed area value for the years 1992 to 2022 by identifying the sequence of hidden states $q_{1:T}$ = arg $max_{q1:T}$ P($q_{1:T}$ / $O_{1:T}$, ʎ) that best explains the observed data. This method utilizes the most probable sequence of hidden states and their corresponding emissions, providing insights into historical trends and fluctuations in oilseed areas.

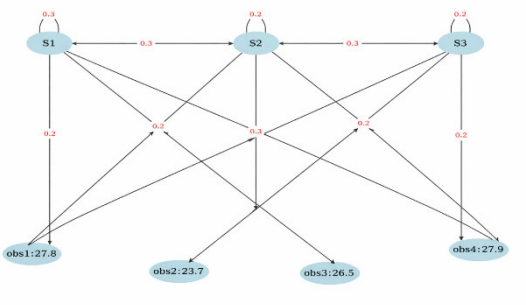


Figure 1. Structural diagram of HMM

2.2. Linear regression

A modeling approach called linear regression (Tranmer and Elliot 2008) is used to analyze data and generate predictions. To predict a response variable (Y) from an explanatory variable (X), a bivariate model is constructed using simple linear regression. The scatterplot's points are fitted with a straight line using linear regression, and predictions are made using the line's equation. The equation for the line in regression modeling is as follows:

$$Y = \beta_0 + \beta_1 X + e_i \tag{1}$$

where the slope of the regression line is the regression parameter β_1 and the intercept is the regression parameter β_0. With a mean of zero and a constant variance, it is assumed that the random error term e_i is uncorrelated. The assumption that the mistakes are distributed regularly is frequently added to analyses to make inference easier and enhance estimation efficiency. The data may be transformed to approach normality (Zou et al. 2003).

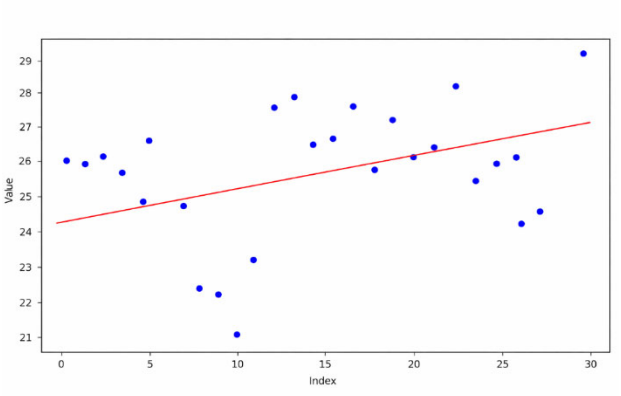


Figure 2. Visualization of linear regression model

2.3. Support vector machines

SVMs (Meyer and Wien 2001) for binary classification were created by Cortes and Vapnik (1995). In machine learning, support vector machines are an effective technique that may be applied to both regression and classification problems. The first step in using SVMs for prediction is gathering and preparing the dataset to ensure it is appropriately scaled for analysis. The data type is then considered while choosing an appropriate kernel function, such as radial basis function (RBF), polynomial, or linear. Additionally, the regularization parameter (C) is initialized. It regulates the trade-off between maximizing the margin and decreasing classification errors. The data is then transformed into a higher-dimensional space where it is simpler to separate using a hyperplane by computing the kernel matrix using the chosen kernel function. When

input data points x_i and x_j are used, the kernel matrix element K_{ij} is calculated as follows (Jakkula 2006):

$$K_{ij} = K(x_i, x_j) \tag{2}$$

The next step is to formulate a quadratic optimization problem, whose goal is to identify the best hyperplane that maximizes the margin between classes while adhering to the restrictions. For the quadratic optimization problem, the objective function (Jakkula 2006) is:

$$\min_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{3}$$

Based on the limitations;

- $\sum_{i=1}^{n} \alpha_i y_i = 0$
- $0 \leq \alpha_i \leq C$, for all i.

The solution function for regression problems is:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \, K(x_i, x_j) + b \tag{4}$$

SVMs are certain to generate reliable and accurate predictions on fresh data through this systematic technique.

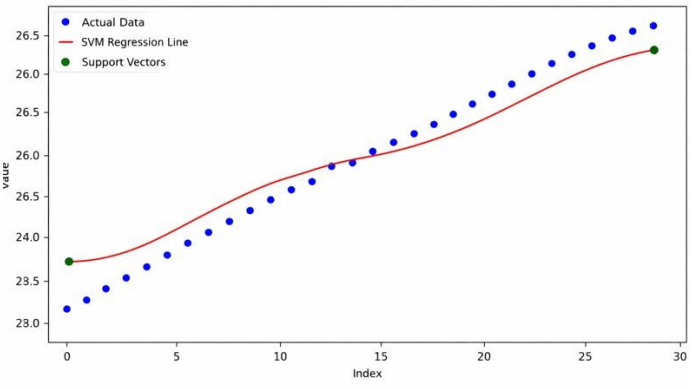


Figure 3. Support vector machine structure

2.4. Neural networks

The neural network (Mark et al. 1992) was designed with three layers: input, hidden, and output, as illustrated in the following figure. Layers consist of many nodes and are connected by correlation weights. Nodes accept input from either outside the model or interconnections. Nodes process the input to produce an analogue output, determining the firing rate. The weight function multiplies the incoming firing rate

before it arrives at the next layer. Each node's transformation is a sigmoid function as specified below:

$$f(x) = \frac{1}{1+ exp(-a(x-b))} \tag{5}$$

When the node's input is denoted by x, its output, or firing rate, is represented by f(x), the gain is 'a', and the bias is 'b'. The values of a and b are calculated using in the below equation:

$$(H_{input}) = f\left(\sum_{j-1}^{n} w_{ij}x_j + b_i\right) \tag{6}$$

where w_{ij} are the weights connecting the input x_j.

$$(O_{output}) = \sum_{k=1}^{m} w_k h_k + b_o \tag{7}$$

In this neural network model, raw data is processed by the input layer, non-linear changes are applied by the hidden layer using the sigmoid function, and predictions are produced by the output layer. The network gains strong predictive capabilities through training, which teaches it to modify its weights and biases to minimize the error between real and anticipated values.

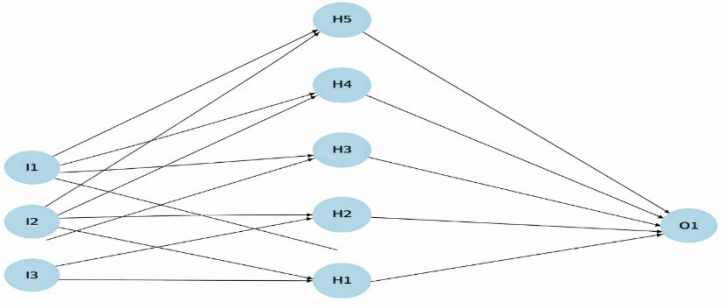


Figure 4. Structural overview of neural networks: input, hidden, and output layers

2.5. K-Nearest Neighbors

The K-Nearest Neighbors (Sadegh and Bolandraftar 2013) technique predicts the outcome for a query point X based on the average of its K nearest neighbors. To improve this, one can weigh the contributions of these neighbors so that closer neighbors have a bigger impact on the forecast than those further away. Regression problems involve predicting the value of a dependent variable (y) using independent variables (x). Consider a scheme where a series of points reflects the relationship between the independent variable and the dependent variable. KNN uses a set of known examples to predict the outcome of a query point (X). The k-nearest neighbor approach predicts the value of the nearest neighbor. If x_4 is the nearest neighbor to X, then:

$$\bar{y}_X = y_4 \tag{8}$$

In the 2-nearest neighbor approach, we average the results of the two nearest neighbors x_3 *and* x_4.

$$\bar{y}_X = \frac{y_3 + y_4}{2} \tag{9}$$

This strategy can be applied to K nearest neighbors. The predicted value is obtained as the average of the outcomes of the K nearest neighbors.

$$\bar{y}_X = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{10}$$

where y_i denotes the outcome of the i^{th} nearest neighbor. The distance between the query points can be measured using the Euclidean distance, Squared Euclidean distance, Manhattan distance, and Chebyshev distance.

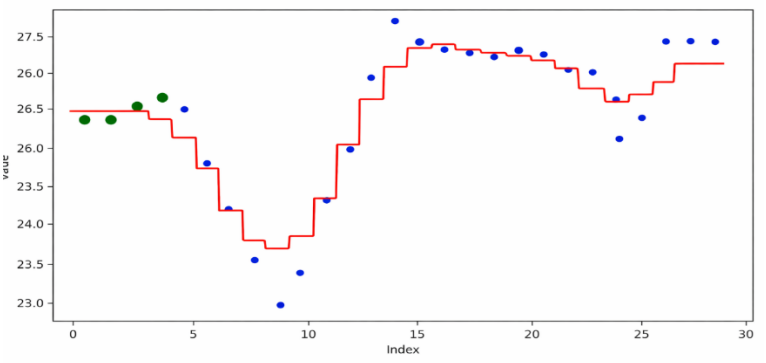


Figure 5. Schematic diagram of K-Nearest Neighbors Model

2.6. Residual analysis

Residual analysis (Rebekka and Gomez 2004) is a valuable set of tools for evaluating the goodness of a fitted model. The purpose of residual analysis is to verify if statistical models applied to data satisfy the underlying assumptions of the models. These presumptions include independent and identically distributed model errors, as well as a regression function that is appropriately defined. Additionally, since most regression estimators only maintain consistency under the assumption that the error distribution is true, residual analysis is particularly crucial. The usage of residual plots, or plots of residuals against explanatory variables or the corresponding fitted values, has become commonplace in the detection of regression model deficiencies evaluation. Residuals are the variations between observed values (y_i) and the predicted values ($\hat{y}_i$) from the models. The mathematical representation is as follows:

$$\text{Residual} = y_i - \hat{y}_i \tag{11}$$

To evaluate how well the model's predictions match the actual data, residuals are utilized. The residuals, when they are randomly distributed around zero, show that there are no systematic errors in the model's representation of the underlying structure of the data. Several methods are used in residual analysis to validate a model's key assumptions, which include independence, normality and homoscedasticity.

A statistical test called the Durbin-Watson (DW) test is used to determine whether the residuals from a regression analysis contain autocorrelation, which is a link between values that are separated from one another by a certain amount of time. The DW (Walter 2011) is calculated as:

$$\mathrm{DW} = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{12}$$

The Shapiro-Wilk (W) test is a statistical tool for evaluating the normality of residuals in regression analysis. Residuals must be normally distributed because this validates the use of statistical tests and confidence intervals in the model. By running this test, we can find potential problems with the assumptions in the model and fix them to increase the model's accuracy and reliability. The test statistic (Zofia et al. 2016) is calculated as:

$$\mathrm{W} = \frac{(\sum_{i=1}^{n} a_i e_i)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2} \tag{13}$$

The Breusch-Pagan (BP) test is a statistical method for detecting homoscedasticity in regression model residuals. Homoscedasticity happens when the variance of the residuals fluctuates between observations, which contradicts the regression analysis assumption that there is constant variance. The Breusch-Pagan test statistic is given by:

$$\mathrm{BP} = \mathrm{n}\ \mathrm{R}^2 \tag{14}$$

where n is the number of values and R^2 is the coefficient of determination from the auxiliary regression of e_i^2 on the independent variables. The computed test statistic BP follows a chi-squared distribution with k degrees of freedom, where k represents the number of independent variables in the supplementary regression.

3. Results and discussions

The dataset used in this study comprises annual historical records of oilseed cultivation area in India, covering the period from 1992 to 2022. Each data point represents the total cultivated area measured in million hectares for the respective agricultural year. This data was sourced from the Economics & Statistics Division of the

Ministry of Agriculture & Farmers Welfare and is publicly accessible via their official website: http://desagri.gov.in.

Table 1. Effectiveness of various models in predicting values in given- years

Year	Area-Actual Value	Predicted Value				
		HMM	KNN	LR	NN	SVMs
1992-93	24.26	25.85	25.764	24.431	24.836	25.940
1993-94	26.20	25.85	25.764	24.538	24.915	25.965
1994-95	26.09	25.85	25.764	24.648	24.993	25.990
1995-96	26.31	25.85	25.972	24.762	25.072	26.014
1996-97	25.96	25.85	26.112	24.877	25.151	26.039
1997-98	25.30	25.85	25.942	24.986	25.229	26.064
1998-99	26.90	25.85	25.234	25.087	25.308	26.088
1999-00	25.24	25.85	24.570	25.189	25.386	26.113
2000-01	22.77	22.93	23.808	25.290	25.465	26.138
2001-02	22.64	23.70	23.160	25.391	25.544	26.162
2002-03	21.49	22.93	23.616	25.493	25.622	26.187
2003-04	23.66	23.70	24.634	25.594	25.701	26.212
2004-05	27.52	27.30	25.408	25.696	25.779	26.237
2005-06	27.86	27.30	26.448	25.798	25.858	26.261
2006-07	26.51	27.30	27.228	25.896	25.937	26.286
2007-08	26.69	27.30	26.916	25.994	26.015	26.311
2008-09	27.56	27.30	26.788	26.092	26.094	26.335
2009-10	25.96	27.30	26.748	26.190	26.172	26.360
2010-11	27.22	27.30	26.706	26.288	26.251	26.385
2011-12	26.31	27.30	26.804	26.386	26.329	26.410
2012-13	26.48	27.30	26.732	26.483	26.408	26.434
2013-14	28.05	27.30	26.506	26.581	26.487	26.459
2014-15	25.60	25.85	26.480	26.679	26.565	26.484
2015-16	26.09	25.85	26.086	26.777	26.644	26.508
2016-17	26.18	25.85	25.434	26.875	26.722	26.533
2017-18	24.51	22.93	25.742	26.973	26.801	26.558
2018-19	24.79	23.70	26.290	27.084	26.880	26.582
2019-20	27.14	27.30	26.888	27.196	26.958	26.607
2020-21	28.83	27.30	26.888	27.307	27.037	26.632
2021-22	29.17	27.30	26.888	27.419	27.115	26.657

Model evaluation

R-squared measures how well a model explains the variability in the observed data, making it sensitive to how accurately the model captures overall trends and fluctuations. Even small changes in the pattern of prediction errors can significantly impact its value. A higher R-squared indicates a stronger fit, meaning the model accounts for more of the

data's variability, while a low or negative R-squared suggests a poor fit. In contrast, Mean Absolute Error (MAE) represents the average size of prediction errors regardless of their direction, offering a clear measure of accuracy without reflecting how well the model explains the variance in the data. Together, a high R-squared and low MAE provide complementary insights into a model's explanatory power and predictive accuracy.

Table 2. Performance comparison of predictive models

Models	MAE	R^2- Valued
Hidden Markov Model	0.7041	0.7449
K- Nearest Neighbors	0.9353	0.5760
Linear Regression	1.307	0.14
Neural Networks	1.2783	0.1529
Support Vector Machines	1.246	0.0466

With the greatest R-squared of 0.7449 and the lowest MAE of 0.7041 when compared to other models, Hidden Markov Models proved to have greater prediction accuracy. These results highlight how well they predict agricultural oilseed areas.

Table 3. The Durbin-Watson test results for the independence of residuals

Model	Durbin-Watson Statistic	Autocorrelation Interpretation
Residual-HMM	1.324	Mild positive autocorrelation
Residual-KNN	1.536	Mild positive autocorrelation
Residual-NN	0.817	Significant positive autocorrelation
Residual-LR	0.832	Significant positive autocorrelation
Residual-SVMs	0.744	Significant positive autocorrelation

These results show variable degrees of positive autocorrelation in the residuals among models. A Durbin-Watson statistic closer to 2 suggests less autocorrelation, whereas values that deviate significantly from 2 imply more autocorrelation.

Table 4. The normality test results for residual errors

Model	Statistic	p-value	Conclusion
Residual-HMM	0.978	0.767	Fail to reject H_0
Residual-KNN	0.973	0.621	Fail to reject H_0
Residual-NN	0.908	0.013	Reject H_0
Residual-LR	0.921	0.028	Reject H_0
Residual-SVMs	0.953	0.199	Fail to reject H_0

Shapiro-Wilk tests show that residuals from HMM, KNN, and SVMs follow a normal distribution ($p > 0.05$), however, residuals from NN and LR do not ($p < 0.05$).

Table 5. The results of the Breusch-Pagan test for homoscedasticity

Model	Breusch-Pagan Test	Homoscedasticity (Interpretation)
Residual-HMM	0.128	Likely homoscedastic
Residual-KNN	0.231	Likely homoscedastic
Residual-NN	0.615	Likely homoscedastic
Residual-LR	0.801	Likely homoscedastic
Residual-SVMs	0.790	Likely homoscedastic

This assumption indicates that the residuals from each model have a constant variance. The p-values (all greater than 0.05) indicate that there is no meaningful evidence to challenge this assumption for any of the models. As a result, each model's residuals are likely to be homoscedastic, which supports their dependability in producing consistent predictions.

3.1. Model fitting and sequence estimation

This study uses the hidden state space S = {Low, Medium, High} and observable state space O = {Decrease, Neutral, Increase}. The hidden states reflect categorized levels of the target agricultural variable (area). Specifically, the Low state corresponds to area values ranging from 21 to 23.83, indicating the lower end of the distribution. The Medium state includes values between 23.83 and 26.66, reflecting the mid-range or average levels. The High state covers values from 26.66 to 29.5, representing the upper portion of the dataset. The observable states describe the direction of year-to-year changes: Decrease indicates a drop in value, Neutral reflects minimal or no change, and Increase denotes a rise in value. These classifications support the model in linking hidden state transitions with observable patterns, enhancing both estimation and interpretability.

The model was trained after determining that HMM was the most suitable based on performance metrics. During training, the following parameters were estimated:

1. Transition Probability Matrix (A): It defines the likelihood of moving from one hidden state to another over time. In this study, the hidden states are categorized as Low, Medium, and High, representing different levels of agricultural area. Each element in the matrix indicates the probability of transitioning from a current state to a future state. It is calculated by analyzing how often the model transitions from one hidden state to another in a given sequence. Each value in the matrix represents the probability of moving from one state (e.g., Low) to another (e.g., Medium or High). To compute it, the number of transitions between states is counted and then normalized by dividing each row by the total transitions from that state. This provides a probability-based understanding of how the system evolves over time.

$$A = \begin{array}{c|ccc} & \text{Low} & \text{Medium} & \text{High} \\ \text{Low} & 0.75 & 0 & 0.25 \\ \text{Medium} & 0.062 & 0.625 & 0.312 \\ \text{High} & 0 & 0.55 & 0.44 \end{array}$$

2. Emission Probability Matrix (B): It shows the likelihood of observing a particular output given a specific hidden state. It connects the hidden states (e.g., Low, Medium, High) to the observable outcomes (e.g., Decrease, Neutral, Increase). To calculate it, the number of times each observation occurs while the system is in a specific hidden state is counted. These counts are then normalized by dividing each by the total number of observations for that hidden state, resulting in probabilities that reflect how likely each observation is under each state.

$$B = \begin{array}{r} \\ \text{Low} \\ \text{Medium} \\ \text{High} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Decrease} & \text{Neutral} & \text{Increase} \end{array} \\ \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.30 & 0.70 \end{bmatrix} \end{array}$$

3. Initial Probability Matrix (π): It represents the probability distribution over hidden states at the starting point (time t = 0). It indicates how likely the system is to begin in each state (e.g., Low, Medium, or High). To calculate it, the number of times each state appears as the first state in multiple sequences is counted and then divided by the total number of sequences.

$$\pi = [0.133, 0.533, 0.333]$$

Using the initial state distribution, transition matrix, and emission matrix, a 10-year forecast can be made by repeatedly updating the state probabilities through the transition matrix. Expected values are then calculated by combining these state probabilities with the emission matrix. This method captures the system's time-dependent behavior and can be efficiently performed using MATLAB's built-in HMM tools.

3.2. Graphical representation

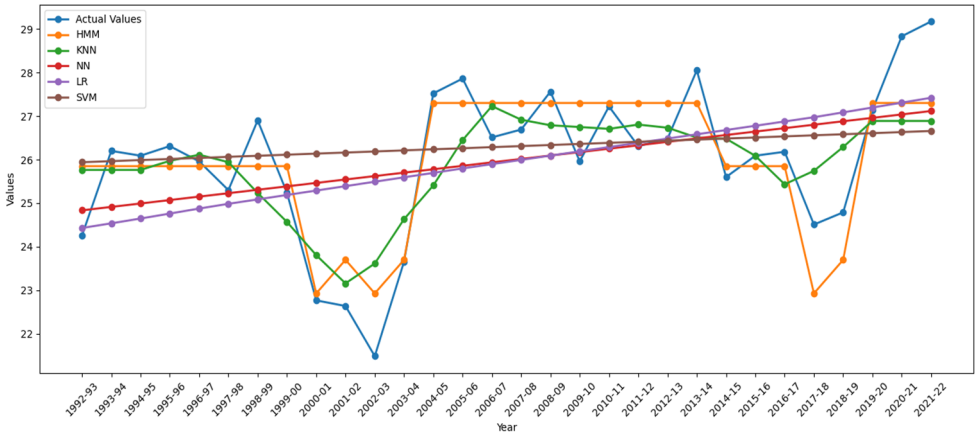


Figure 6. The actual and predicted values of different models

The graph presents a comparison between actual oilseed area data and the predictions made by five models over the period from 1992 to 2022. The actual data shows significant variability, with a noticeable decline during the early 2000s. Among all models, the HMM closely tracks the real data, effectively reflecting both declines and recoveries. KNN and NN provide reasonably accurate predictions but exhibit some inconsistencies. In contrast, SVM and Linear Regression tend to produce smoother or more linear trends, overlooking key changes in the actual data. Overall, HMM provides the most reliable fit to the values.

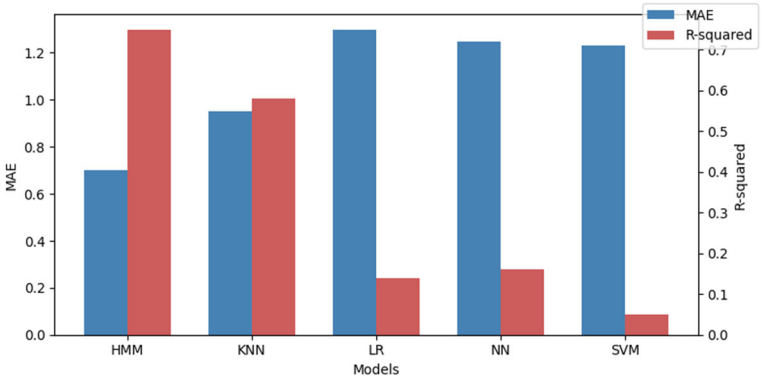


Figure 7. Comparison of MAE and R-squared Value

The bar chart presents a comparison of predictive performance based on MAE and R-squared values. One model achieves the lowest error and highest fit, indicating strong accuracy and reliability. Others show moderate to poor performance, either with higher errors or weaker data alignment. Overall, the best-performing model stands out in both accuracy and consistency.

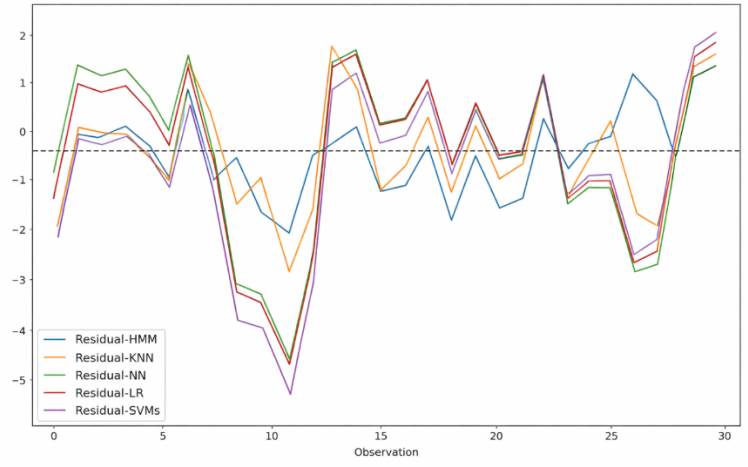


Figure 8. The residuals for different models across the observations

The line graph illustrates residual variations across different prediction models. Residuals close to zero indicate higher prediction accuracy. The HMM model shows relatively smaller and more stable residuals, especially after the 10th observation, indicating better consistency. In contrast, other models exhibit larger fluctuations, particularly around the 10th and 30th observations, suggesting less reliable predictions. Overall, the HMM model demonstrates more controlled and balanced residual behavior.

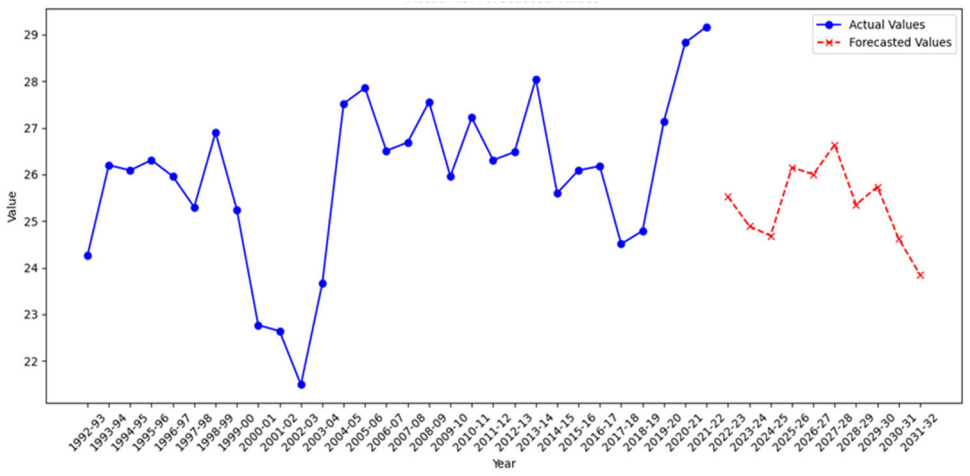


Figure 9. Comparison of actual and forecasted values over time using HMM parameters

The line chart presents both the actual and forecasted oilseed area values over time. The solid blue line represents the observed data, showing noticeable fluctuations with periods of sharp declines followed by recoveries. The forecasted values, indicated by a red dashed line, generally follow the overall trend of the historical data while displaying some variability. This suggests that the model effectively captures the main movement of the data and adapts to changing patterns, providing a useful projection for future values.

4. Conclusions

The findings of this study demonstrate that Hidden Markov Models (HMMs) significantly outperform other predictive methods in forecasting agricultural data. HMMs exhibit the lowest Mean Absolute Error (MAE) and the highest R-squared values, signifying their superior accuracy and reliability. The residual analysis further confirms the model's robustness, as the residuals are randomly distributed with no identifiable patterns, indicating a strong model fit. HMMs have proven to be highly effective in capturing the temporal and stochastic characteristics of agricultural data,

surpassing traditional machine learning techniques in performance. These results underline the potential of HMMs as a reliable tool for agricultural forecasting, enabling stakeholders to make data-driven decisions and develop strategic plans based on accurate predictions.

Additionally, this study utilizes HMM parameters to forecast agricultural trends for the next 10 years, providing valuable insights for long-term planning. Future research can expand on these findings by applying HMMs to different agricultural domains and exploring further enhancements to improve their predictive capabilities.

References

Agarwal, S., Tarar, S., (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. In Journal of *Physics*. Conference Series, Vol. 1714, page 012012.IOP Publishing.

Bafandeh, I. S., Bolandraftar, M., (2013). Application of K-Nearest Neighbor (KNN) approach for predicting economic events: Theoretical Background. *International Journal of Engineering Research and Applications*, Vol. 3, issue 5, pp. 605–610.

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D. and Bochtis, D., (2021). Machine Learning in Agriculture: A Comprehensive updated review. *Sensors*, 21(11), 3758.

Chai, T., Ren, H., (2023). Risk prediction of financial management in agricultural companies based on RBF neural network and Markov. *Pakistan Journal of Agricultural Sciences*, 60(4), pp. 739–749.

Choukri, D., Adda, M., (2020). A new hybrid deep learning model-based recommender system using Artificial Neural Network and Hidden Markov model. *Procedia Computer Science*, 175, pp. 214–220.

Elavarasan, D., Vincent, P. M. D., (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. In *IEEE* Access, Vol. 8, pp. 86886–86901.

Erivwo, O., Makis, V. and Kwon, R., (2024). Bayesian change point prediction for downhole drilling pressures with hidden Markov models. *Applied Stochastic Models in Business and Industry*, 40(3), pp. 772–790.

French, M. N., Krajewski, W. F. and Cuykendall, R. R., (1992). Rainfall forecasting in space and time using a neural network. *Journal of Hydrology*, 137, pp. 1–31.

Hanusz, Z., Tarasinska, J. and Zielinski, W., (2016). Shapiro-Wilk test with known mean. *Revstat-Statistical Journal*, Vol. 14, No. 1, pp. 89–100.

Jakkula, V., (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5), 3.

Khosla, E., Dharavath, R. and Priya, R., (2020). Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environment, Development and Sustainability*, 22, pp. 5687–5708.

Kramer, W., (2011). Durbin-Watson test. In International Encyclopedia of Statistical Science. *Springer*, pp. 408–409.

Leite, P. B. C., Feitosa, R. Q., Formaggio, A. R., Da Costa, G. A. O. P., Pakzad, K. and Sanches, I. D. A., (2008). Hidden Markov models applied in agricultural crops classification, proceeding of GEOBIA (GEOgraphic Object-Based Image Analysis for the 21St Century).

Meyer, D. and Wien, F. T., (2001). Support vector machines. *R News*, 1(3), pp. 23–26.

Mithra, C. and Suhasini, (2023). Machine Learning based oilseed crop yield forecasting with recommendation system using organic manures in Tamil Nadu. *Journal of Survey in Fisheries Sciences*, 10(2S), pp. 1601–1620.

Senawongse, P., Dalby, A. R. and Yang, Z. R., (2005). Predicting the phosphorylation sites using Hidden Markov Models and machine learning methods. *Journal of chemical information and modelling*, 45(4), pp. 1147–1152.

Sharma, P., Dadheech, P., Aneja, N. and Aneja, S., (2023). Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning, in *IEEE Access*, Vol. 11, pp. 111255–111264.

Somani, P., Talele, S. and Sawant, S., (2014). Stock market prediction using Hidden Markov Model. *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, pp. 89-92, Chongqing, China.

Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T. and Rahimi, J., (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Sci Rep*, 12, 3215.

Topp, R., Gomez, G., (2004). Residual analysis in linear regression models with an interval-censored covariate. *Statistics in medicine*, 23, pp. 3377–3391.

Tranmer, M., Elliot, M., (2008). Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research*, 5(5), pp. 1–5.

Yuan, H., Yang, G., Li, C., Wang, Y., Liu, J., Yu, H., Feng, H., Xu, B., Zhao, X. and Yang, X., (2017). Retrieving soybean leaf area index from unmanned aerial vehicle hyperspectral remote sensing: Analysis of RF, ANN, and SVM regression models. *Remote Sensing*, 9(4), 309.

Zhu, D., Ding, Z. and Huang, X., (2024). Probabilistic model for fatigue damage estimation of wind turbines with hidden Markov model and neural network. *Ocean Engineering*, 310, 118663.

Zou, K. H., Tuncali, K. and Silverman, S. G., (2003). Correlation and simple linear regression. *Radiology*, 227(3), pp. 617–628.