Journal of Mathematical Problems, Equations and Statistics

E-ISSN: 2709-9407
P-ISSN: 2709-9393
JMPES 2025; 6(2): 08-15

www.mathematicaljournal.com
Received: 03-05-2025
Accepted: 07-06-2025

Vyshnavi M
Research Scholar & Assistant Professor, Department of Statistics, PSG College of Arts & Science, Coimbatore, Tamil Nadu, India

Muthukumar M
Research Scholar & Assistant Professor, Department of Statistics, PSG College of Arts & Science, Coimbatore, Tamil Nadu, India

Corresponding Author:
Vyshnavi M
Research Scholar & Assistant Professor, Department of Statistics, PSG College of Arts & Science, Coimbatore, Tamil Nadu, India

Investigating temporal patterns using hidden Markov models with lag analysis

Vyshnavi M and Muthukumar M

DOI: https://www.doi.org/10.22271/math.2025.v6.i2a.215

Abstract
Analyzing temporal patterns in time series data is crucial for improving forecasting accuracy and understanding underlying trends. This study explores the use of Hidden Markov Models (HMMs) with lag integration to better capture sequential dependencies. By incorporating past observations, the proposed method aims to enhance predictive accuracy and state identification. The performance of the lag-integrated HMM is evaluated using historical agricultural data and compared against conventional HMMs. This research contributes to time series modeling by demonstrating the benefits of lag incorporation in capturing complex temporal dynamics. A trained HMM, implemented in MATLAB, generates future projections based on learned patterns and lag-based adjustments. The study further assesses the impact of increasing lag on predictive performance using divergence measures such as Kullback-Leibler (KL) divergence, Jensen-Shannon (JS) divergence, and Bhattacharyya distance. Findings indicate that as the lag increases, prediction accuracy declines. The results enhance agricultural forecasting methodologies, supporting efficient resource allocation and sustainable farm management.

Keywords: Hidden markov models, lag analysis, kullback-leibler divergence, jensen-shannon divergence, bhattacharyya distance

Introduction

Agriculture is a vital pillar of global economies, providing food security, raw materials, and employment to millions worldwide. Accurate forecasting of agricultural variables, such as crop area and yield, is essential for effective decision-making, resource allocation, and policy formulation. However, agricultural systems are inherently complex due to their dependence on various dynamic factors, including climate variability, soil conditions, market demand, and government policies. These interdependencies introduce challenges in predicting future trends, necessitating the use of advanced computational techniques for reliable forecasts. Traditional statistical methods often struggle to capture the intricate temporal patterns of agricultural data, leading to the adoption of sophisticated stochastic models such as Markov Chains and Hidden Markov Models.

The Markov Chain, introduced by Russian mathematician Andrey Markov in the early 20th century, is a stochastic process that models state transitions based on a probabilistic framework. A key characteristic of Markov Chains is the Markov Property, which states that the probability of transitioning to a future state depends solely on the present state, independent of past states. This property allows Markov Chains to efficiently model dynamic systems across various domains, including finance, physics, economics, and engineering. Over the years, researchers have refined Markovian techniques to analyze transition probabilities and forecast long-term system behavior, making them a foundational tool for sequential data analysis.

Building upon this concept, Leonard E. Baum and colleagues extended the Markov framework in the 1970s by introducing the HMM. Unlike traditional Markov Chains, HMMs incorporate hidden states, which are not directly observable but generate measurable outputs through probabilistic mechanisms. These models are particularly advantageous in scenarios where underlying processes are not explicitly visible but influence observed data, such as speech recognition, bioinformatics, financial modeling, and weather prediction. In agriculture, where multiple latent factors contribute to variations in crop yield and cultivated area, HMMs offer a powerful method to uncover hidden patterns and dependencies within the data.

One critical aspect of agricultural time series analysis is the influence of historical trends on future values. To address this, Lag Analysis is integrated with HMMs to enhance predictive accuracy by considering past fluctuations in crop area. Lag Analysis helps quantify the extent to which previous values influence current states, capturing seasonal variations, cyclical patterns, and external market dynamics. Integrating Lag Analysis into HMMs improves the understanding of long-term agricultural trends, resulting in more precise and dependable forecasting models. This method strengthens the ability to identify temporal patterns, seasonal fluctuations, and hidden dependencies.

This study aims to develop a six-state Hidden Markov Model enhanced with Lag Analysis to analyze historical oilseed area data. It emphasizes estimating essential model parameters, analyzing transition probabilities, and evaluating the model's forecasting capability for agricultural trends. By leveraging the strengths of HMMs and Lag Analysis, this research provides valuable insights into the evolving patterns of oilseed cultivation. The proposed methodology not only enhances predictive accuracy but also provides a framework for policymakers and stakeholders to make informed, data-driven decisions for sustainable agricultural planning.

1.1 Objectives of the Study

- Develop a six-state HMM incorporating lag analysis to capture temporal dependencies in agricultural time series data.
- Estimate key parameters, including the Transition Probability Matrix (TPM), Emission Probability Matrix (EPM), and Initial State Probability (π), to establish a structured probabilistic framework for state transitions and observations.
- Extend the model to a ten-state HMM and generate synthetic sequences to analyze state transitions and compare predictive performance with the six-state model.
- Evaluate model accuracy and effectiveness using statistical divergence measures such as Kullback-Leibler (KL) Divergence, Jensen-Shannon (JS) Divergence and Bhattacharyya Distance to quantify prediction reliability.
- Examine the influence of varying lag structures on forecasting performance, identifying the most effective lag configuration for enhancing long-term prediction accuracy.

1.2 Review of Literature

Hidden Markov Models (HMMs) have become an important tool for analyzing sequential data and forecasting temporal trends in various fields, including agriculture. Rabiner (1989) [9] provided foundational insights into HMMs, highlighting their ability to model sequential dependencies through hidden states. This makes HMMs well-suited for forecasting agricultural data, where patterns are often complex and evolve over time. In contrast, traditional time series models like ARMA (Box & Jenkins, 1976) [10] focus primarily on observed data relationships and lack the flexibility to model hidden states and intricate temporal patterns.

In the context of agriculture, HMMs have shown their effectiveness in both forecasting and risk management. Gupta and Dhingra (2012) [8] explored the potential of HMMs for predicting stock market trends, illustrating their versatility in forecasting dynamic, sequential data. Zhang, Yang, and Liang (2021) [7] applied HMMs to assess the risks associated with agricultural droughts, demonstrating the model's usefulness in risk evaluation. Additionally, Kavitha, Udhayakumar, and Nagarajan (2013) [3] used HMMs to analyze stock market trends, a method that can also be applied to agricultural trends. More recent work by Vyshnavi and Muthukumar (2024) [6] incorporated lag analysis into HMMs, improving the accuracy of agricultural forecasts by accounting for delayed temporal effects. Shumway and Stoffer (2017) [11] further contributed by offering statistical methods for time series analysis that can complement HMMs. Together, these studies emphasize how HMMs, particularly when integrated with techniques like lag analysis, can significantly improve the forecasting of temporal patterns in agricultural data.

While HMMs have been extensively applied in fields like finance and climate, there remains a significant gap in their use for agricultural time series forecasting. This research seeks to fill that gap by investigating the application of lag-integrated HMMs to agricultural data, with a particular focus on oilseed production. The study introduces lag-integrated HMMs into agricultural forecasting, where capturing temporal dependencies is crucial. It reveals that while incorporating lag improves prediction accuracy, excessive lag can degrade the model's performance. These findings are consistent with previous research, highlighting the need to select an optimal lag length to avoid overfitting. This work contributes to the field by enhancing the understanding of how lag can be effectively utilized in agricultural time series forecasting, offering new insights for refining prediction models in this domain.

2. Materials and Methods

2.1 Lag Analysis: Lag analysis is an important technique in time series studies that examines the connection between observations recorded at different times. It involves introducing a lag, which is the time interval between two observations, to understand how earlier values influence later ones. By creating lag variables, where the original data is shifted by a certain number of time points, analysts can discover patterns, dependencies, or trends within the data. Lag analysis not only helps in developing effective forecasting models but also assists in identifying unusual observations or patterns that may need special attention during modeling. Mathematically, lagging means finding the difference between consecutive values in a time series. For a series Z observed at different times t, the lag value LZ_t at time t is given by:

$$LZ_t = Z_t - Z_{t-1}$$

where: LZ_t represents the lag value at time t, Z_t is the value of the series at the current time t and Z_{t-1} is the value of the series at the previous time t−1. This calculation measures how the series changes over time and provides important insights into the dynamic behavior of the data.

Lag analysis plays a crucial role in HMMs, particularly when trying to capture the temporal dependencies between observations at different time points. In the context of HMMs, the hidden states evolve over time, and these states are used to generate the observations. Each observation at time t is dependent on the hidden state at that time, and the hidden states themselves are interdependent. Lag analysis allows us to investigate how past states or observations influence current and future observations. By introducing lagged variables into the model, we can account for longer-term dependencies and improve the model's predictive capability. Mathematically, this can be represented as follows: let S_t denote the hidden state at time t, and O_t represent the observation at time t. The standard HMM setup is defined by the following joint probability:

$P(O_1, O_2, \dots, O_T, S_1, S_2, \dots, S_T) = P(S_1)\prod_{t=2}^{T} P(S_t/S_{t-1})P(O_t/S_t)$

Now, introducing lag analysis means incorporating lagged observations or hidden states into the transition or emission matrices. For instance, a lag of p steps would introduce the lagged state S_{t-P} or the lagged observation O_{t-P} into the state transition or emission probability. The lagged transition probabilities and the lagged emission probabilities can be modified as follows:

$P(S_t/S_{t-P})$ and $P(O_t/S_{t-P})$

By adjusting the transition and emission matrices with such lagged terms, the HMM can capture the temporal dependencies over different time scales. This helps improve the model's ability to predict future states and observations based on a more comprehensive view of past behaviors. Additionally, in HMM-based time series forecasting, lag analysis can be employed to determine the appropriate number of previous states or observations to consider when predicting future states. The inclusion of multiple lags, for example $S_{t-1}, S_{t-2}, \dots, S_{t-p}$ may improve the predictive performance of the model by enabling it to account for longer-term dependencies in the data. The extended transition probability can be written as:

$P(S_t/S_{t-1}, S_{t-2}, \dots, S_{t-p})$ and $P(O_t/S_{t-1,}S_{t-2}, \dots, S_{t-p})$

This allows the model to incorporate more comprehensive historical data when predicting future states, improving both its interpretability and forecasting accuracy. By understanding the relationships between past and present states, lag analysis enhances the predictive power of HMMs, particularly in applications where long-term dependencies are crucial.

2.2 Kullback-Leibler (KL) Divergence: Kullback-Leibler Divergence is a measure of how one probability distribution P (actual steady-state probability) differs from another probability distribution Q (generated steady-state probability). It quantifies the information loss when Q is used to approximate P. Mathematically, it is given by:

$D_{KL}(P\|Q) = \sum_i P_i \log \frac{P_i}{Q_i}$

Interpretation
- $D_{KL} = 0$ means P and Q are identical.
- A small value indicates high similarity.
- A larger value suggests greater divergence or inefficiency in using Q instead of P.

KL Divergence in HMM
In the HMMs, KL divergence plays a crucial role in:
- Comparing the steady-state probabilities of different models to assess accuracy.
- Evaluating the effectiveness of training algorithms in learning transition and emission probabilities.
- Measuring the difference between observed and predicted state sequences.

By computing KL divergence for different lags, one can analyze how well the model retains information over time, which is essential for time series forecasting and sequence modeling.

2.3 Jensen-Shannon (JS) Divergence
Jensen-Shannon divergence is a symmetric measure used to quantify the similarity between two probability distributions. Unlike KL divergence, JS divergence ensures symmetry and produces a finite value even when the distributions have zero probabilities in certain regions. It is widely applied in machine learning, statistical modeling, and information theory to assess how different two probability distributions are. Given two probability distributions P and Q, the JS divergence is defined as:

$D_{JS}(P\|Q) = \frac{1}{2} D_{KL}(P\|M) + \frac{1}{2} D_{KL}(Q\|M)$

Where, $M = \frac{P+M}{2}$ is the average distribution of P and Q, $D_{KL}(P\|M)$ is the KL divergence between P and M and $D_{KL}(Q\|M)$ is the KL divergence between Q and M.

Interpretation
- $D_{JS}(P\|Q) = 0$, the actual and generated distributions are identical, meaning the model is perfectly replicating the real data.
- Small $D_{JS}(P\|Q)$ value, the two distributions are very similar, indicating that the model closely approximates the actual steady-state probabilities.
- Large $D_{JS}(P\|Q)$ value, there is a noticeable difference between the two distributions, suggesting the model's generated probabilities do not accurately reflect the actual probabilities.

JS divergence is a refined and symmetric alternative to KL divergence, making it a more stable and interpretable measure of distribution similarity.

2.4 Bhattacharya Distance
The Bhattacharyya distance is a statistical measure used to evaluate the similarity between two probability distributions. It determines the degree of overlap between the distributions, making it valuable in classification, pattern recognition, and probabilistic modeling. For two probability distributions, P and Q, the Bhattacharyya distance is given by:

$D_B(P, Q) = -\ln(\sum_i \sqrt{P_i\, Q_i}$

The term $\sum_i \sqrt{P_i\, Q_i}$ is called the Bhattacharyya coefficient, which quantifies the overlap between the two distributions. The logarithmic transformation ensures that larger values correspond to greater dissimilarity.

Interpretation
- $D_B(P, Q) = 0$, the two distributions are identical, meaning the generated probabilities perfectly match the actual ones
- Small $D_B(P, Q)$ value, indicates high similarity, suggesting an effective model.
- Large $D_B(P, Q)$ value, suggests that the generated and actual distributions differ significantly, which may indicate model inaccuracies.

Results & Discussions
Understanding past patterns helps improve forecasting accuracy in time series data. This section examines the performance of the lag-integrated HMM, focusing on its predictive accuracy, state transitions, and the effect of different lag values.

Table 1: Values of the area at different lags

Year	Area	Lag-1	Lag-2	Lag-3	Lag-4	Lag-5	Lag-6
2000-01	22.77	-	-	-	-	-	-
2001-02	22.64	-0.13	-	-	-	-	-
2002-03	21.49	-1.15	-1.28	-	-	-	-
2003-04	23.66	2.17	1.02	0.89	-	-	-
2004-05	27.52	3.86	6.03	4.88	4.75	-	-
2005-06	27.86	0.34	4.2	6.37	5.22	5.09	-
2006-07	26.51	-1.35	-1.01	2.85	5.02	3.87	3.74
2007-08	26.69	0.18	-1.17	-0.83	3.03	5.2	4.05
2008-09	27.56	0.87	1.05	-0.3	0.04	3.9	6.07
2009-10	25.96	-1.6	-0.73	-0.55	-1.9	-1.56	2.3
2010-11	27.22	1.26	-0.34	0.53	0.71	-0.64	-0.3
2011-12	26.31	-0.91	0.35	-1.25	-0.38	-0.2	-1.55
2012-13	26.48	0.17	-0.74	0.52	-1.08	-0.21	-0.03
2013-14	28.05	1.57	1.74	0.83	2.09	0.49	1.36
2014-15	25.60	-2.45	-0.88	-0.71	-1.62	-0.36	-1.96
2015-16	26.09	0.49	-1.96	-0.39	-0.22	-1.13	0.13
2016-17	26.18	0.09	0.58	-1.87	-0.3	-0.13	-1.04
2017-18	24.51	-1.67	-1.58	-1.09	-3.54	-1.97	-1.8
2018-19	24.79	0.28	-1.39	-1.3	-0.81	-3.26	-1.69
2019-20	27.14	2.35	2.63	0.96	1.05	1.54	-0.91
2020-21	28.83	1.69	4.04	4.32	2.65	2.74	3.23
2021-22	29.17	0.34	2.03	4.38	4.66	2.99	3.08

This table shows the Year, Area, and several Lag values (from Lag-1 to Lag-6) for each year. These lag values represent the change in the area of oilseeds from one year to another, calculated as the difference between the area in the current year and the area in previous years. Here's how the lag values are determined: Lag-1: Difference between the current year's area and the previous year's area. Lag-2: Difference between the current year's area and the area from two years ago. Similarly, Lag-3 to Lag-6 are calculated as the difference between the current year's area and the area from the respective earlier years. The dashes (-) in the table indicate missing data for those lags due to the lack of prior years' data.

Table 2: States (S) and Observing Symbol (OS)

Area	Lag-1		Lag-2		Lag-3		Lag-4		Lag-5		Lag-6	
	OS	S	OS	S	OS	S	OS	S	OS	S	OS	S
22.77	-	-	-	-	-	-	-	-	-	-	-	-
22.64	D	S_3	-	-	-	-	-	-	-	-	-	-
21.49	D	S_2	D	S_1	-	-	-	-	-	-	-	-
23.66	I	S_5	I	S_3	I	S_3	-	-	-	-	-	-
27.52	I	S_6	I	S_6	I	S_5	I	S_6	-	-	-	-
27.86	I	S_3	I	S_5	I	S_6	I	S_6	I	S_6	-	-
26.51	D	S_2	D	S_1	I	S_4	I	S_6	I	S_6	I	S_5
26.69	I	S_3	D	S_1	D	S_1	I	S_5	I	S_6	I	S_5
27.56	I	S_4	I	S_3	D	S_2	I	S_3	I	S_6	I	S_6
25.96	D	S_1	D	S_1	D	S_1	D	S_2	D	S_2	I	S_4
27.22	I	S_4	D	S_2	I	S_2	I	S_4	D	S_2	D	S_2
26.31	D	S_2	I	S_2	D	S_1	D	S_3	D	S_3	D	S_1
26.48	I	S_3	D	S_1	I	S_2	D	S_2	D	S_3	D	S_2
28.05	I	S_4	I	S_3	I	S_2	I	S_4	I	S_3	I	S_3
25.60	D	S_1	D	S_1	D	S_1	D	S_2	D	S_3	D	S_1
26.09	I	S_3	D	S_1	D	S_2	D	S_3	D	S_2	I	S_3
26.18	I	S_3	I	S_2	D	S_1	D	S_3	D	S_3	D	S_1
24.51	D	S_1	D	S_1	D	S_1	D	S_1	D	S_1	D	S_1
24.79	I	S_3	D	S_1	D	S_1	D	S_2	D	S_1	D	S_1
27.14	I	S_5	I	S_4	I	S_3	I	S_4	I	S_4	D	S_1
28.83	I	S_4	I	S_5	I	S_5	I	S_5	I	S_5	I	S_4
29.17	I	S_3	I	S_4	I	S_5	I	S_6	I	S_5	I	S_4

The table displays Area values for each year along with Lag-1 to Lag-6, which represent the changes in area compared to previous years. Each lag value is associated with a state label $(S_1, S_2, \dots S_6)$ and a trend indicator (D for decrease, I for increase). These indicators show the direction of change in the area, while the states categorise different patterns or conditions. For example, in Lag-1 for a particular year, the value might indicate an increase (I) with a corresponding state like S_3, suggesting a specific behavior observed in that year's data. The states and trends help identify patterns in the changes of the area over time, which can be used for modeling and forecasting future trends.

3.1 Computation of HMM parameters

The probability values of TPM, EPM and π for lag 1 to lag 6 values on the area were calculated given below,

Probability values of TPM, EPM and π for lag 1

$$TPM = \begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0 & 0 & 0.66 & 0.33 & 0 & 0 \\ 0 & 0 & 0.66 & 0 & 0.33 & 0 \\ 0.14 & 0.28 & 0.14 & 0.28 & 0.14 & 0 \\ 0.5 & 0.25 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$EPM = \begin{array}{cc} & \begin{array}{cc} I & D \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0.875 & 0.125 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array}$$

Steady-state probability distribution of lag 1 values:
$\pi = [\, 0.14 \quad 0.142 \quad 0.38 \quad 0.19 \quad 0.09 \quad 0.047 \,]$

1. **Probability values of TPM, EPM and π for lag 2**

$$TPM = \begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0.33 & 0.22 & 0.33 & 0.11 & 0 & 0 \\ 0.66 & 0.33 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

$$EPM = \begin{array}{cc} & \begin{array}{cc} I & D \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0 & 1 \\ 0.66 & 0.33 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array}$$

Steady-state probability distribution of lag 2 values:
$\pi = [\, 0.45 \quad 0.15 \quad 0.15 \quad 0.1 \quad 0.1 \quad 0.05 \,]$

2. **Probability values of TPM, EPM and π for lag 3**

$$TPM = \begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0.28 & 0.57 & 0.14 & 0 & 0 & 0 \\ 0.8 & 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

$$EPM = \begin{array}{cc} & \begin{array}{cc} I & D \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0 & 1 \\ 0.6 & 0.4 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array}$$

Steady-state probability distribution of lag 3 values:
$\pi = [\, 0.36 \quad 0.26 \quad 0.10 \quad 0.05 \quad 0.157 \quad 0.05 \,]$

3. **Probability values of TPM, EPM and π for lag 4**

$$TPM = \begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0.25 & 0.5 & 0.25 & 0 & 0 & 0 \\ 0 & 0.33 & 0.33 & 0 & 0.33 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0.33 & 0.66 \end{bmatrix} \end{array}$$

$$EPM = \begin{array}{cc} & \begin{array}{cc} I & D \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0.25 & 0.75 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array}$$

Steady-state probability distribution of lag 4 values:

$\pi = [\, 0.05 \quad 0.22 \quad 0.22 \quad 0.166 \quad 0.11 \quad 0.22 \,]$

3. **Probability values of TPM, EPM and π for lag 5**

$$TPM = \begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0.5 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0.33 & 0.66 & 0 & 0 & 0 \\ 0.2 & 0.2 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0.25 & 0 & 0 & 0 & 0.75 \end{bmatrix} \end{array}$$

$$EPM = \begin{array}{cc} & \begin{array}{cc} I & D \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0.2 & 0.8 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array}$$

Steady-state probability distribution of lag 5 values:
$\pi = [\, 0.117 \quad 0.176 \quad 0.29 \quad 0.05 \quad 0.117 \quad 0.235 \,]$

4. **Probability values of TPM, EPM and π for lag 6**

$$TPM = \begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0.5 & 0.16 & 0.16 & 0.16 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

$$EPM = \begin{array}{cc} & \begin{array}{cc} I & D \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array}$$

Steady-state probability distribution of lag 6 values:
$\pi = [\, 0.37 \quad 0.125 \quad 0.125 \quad 0.187 \quad 0.125 \quad 0.062 \,]$

3.2 Generate a sequence of HMM

A random sequence of emission symbols and states can be generated using the **"hmmgenerate"** function in MATLAB. The syntax for this function is:

[Sequence, States] = hmmgenerate (L, TPM, EPM)
where L represents the desired length of both the sequence and the states. The function utilizes the Transition Probability Matrix and Emission Probability Matrix to produce the sequence.

Lag-1										
Generated Sequence	I	D	I	I	D	I	I	I	D	I
States Sequence	S_4	S_1	S_3	S_4	S_3	S_5	S_6	S_3	S_2	S_3
Lag-2										
Generated Sequence	I	I	D	D	D	D	I	I	I	I
States Sequence	S_2	S_2	S_1	S_2	S_2	S_1	S_3	S_6	S_5	S_4
Lag-3										
Generated Sequence	I	D	D	I	D	D	I	I	I	I
States Sequence	S_2	S_1	S_1	S_2	S_1	S_1	S_3	S_5	S_6	S_4
Lag-4										
Generated Sequence	D	I	I	D	D	I	D	D	I	D
States Sequence	S_2	S_4	S_5	S_3	S_2	S_4	S_3	S_2	S_4	S_2
Lag-5										
Generated Sequence	I	I	I	I	I	I	I	I	I	I
States Sequence	S_4	S_5	S_5	S_5	S_5	S_5	S_5	S_5	S_5	S_5
Lag-6										
Generated Sequence	D	I	D	D	I	I	D	D	D	D
States Sequence	S_1	S_4	S_2	S_1	S_4	S_4	S_2	S_1	S_1	S_1

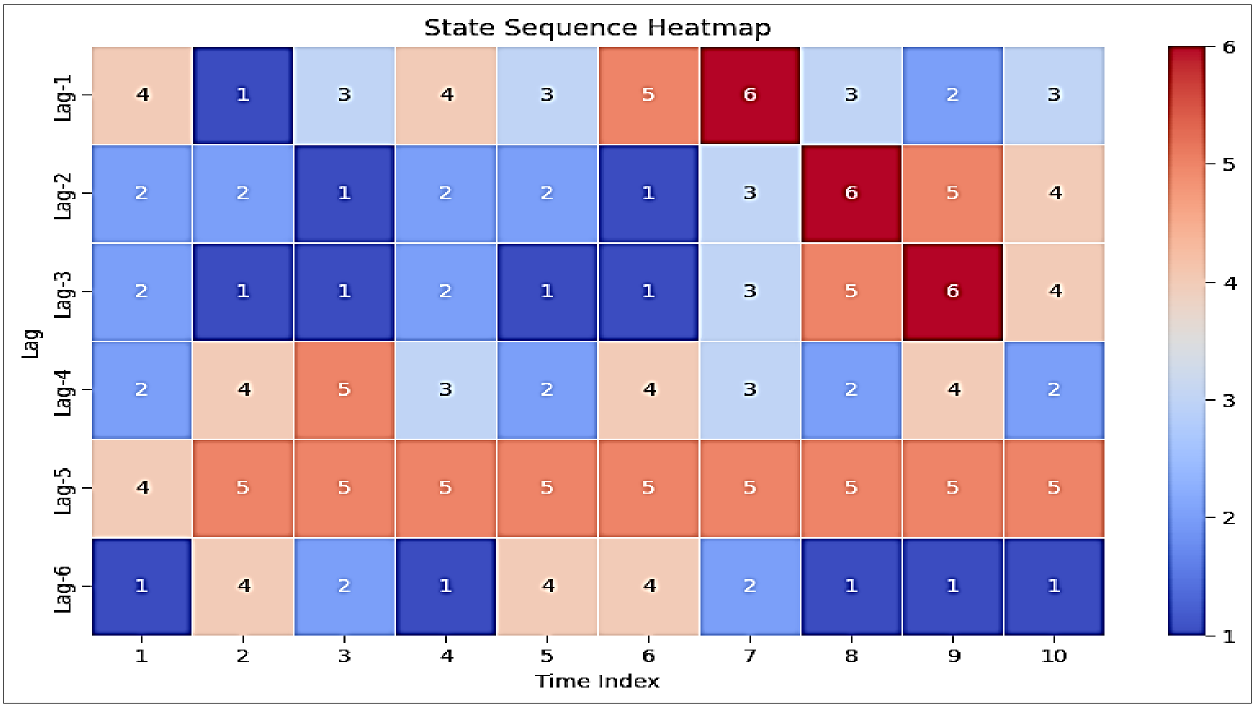

Fig 1: Variation of State Sequences Over Time and Lag Intervals

The heatmap visualizes the state sequences corresponding to different lags (Lag-1 to Lag-6) across ten-time indices. Each row represents a different lag, while each column corresponds to a specific time index. The colour variations indicate different states (S_1 to S_6), helping to identify patterns and transitions in the state sequences over time. The heatmap serves as an essential analytical tool in this study, providing a visual representation of state transitions across different lags. It effectively simplifies complex patterns, facilitating the analysis of state stability, dependencies, and variations over time. By offering a clear depiction of these transitions, the heatmap enhances the understanding of sequence behavior and contributes to more accuracy.

Table 3: Probability Distribution of States across Different Lags

States	Lag-1		Lag-2		Lag-3		Lag-4		Lag-5		Lag-6	
	A_i	G_i	A_i	G_i	A_i	G_i	A_i	G_i	A_i	G_i	A_i	G_i
S_1	0.14	0.1	0.45	0.2	0.36	0.4	0.05	0	0.117	0	0.37	0.5
S_2	0.142	0.1	0.15	0.4	0.26	0.2	0.22	0.4	0.176	0	0.125	0.2
S_3	0.38	0.4	0.15	0.1	0.10	0.1	0.22	0.2	0.29	0	0.125	0
S_4	0.19	0.2	0.1	0.1	0.05	0.1	0.166	0.3	0.05	0.1	0.187	0.3
S_5	0.09	0.1	0.1	0.1	0.157	0.1	0.11	0.1	0.117	0.9	0.125	0
S_6	0.047	0.1	0.05	0.1	0.05	0.1	0.22	0	0.235	0	0.062	0

The table displays the steady-state probabilities of actual (A_i) and generated (G_i) states over six lag periods. While most states show stable transitions, some fluctuations indicate variations in state behavior. The differences between A_i and G_i reflect distinct transition patterns, emphasizing the role of past states in improving predictive modeling and capturing temporal dependencies more effectively.

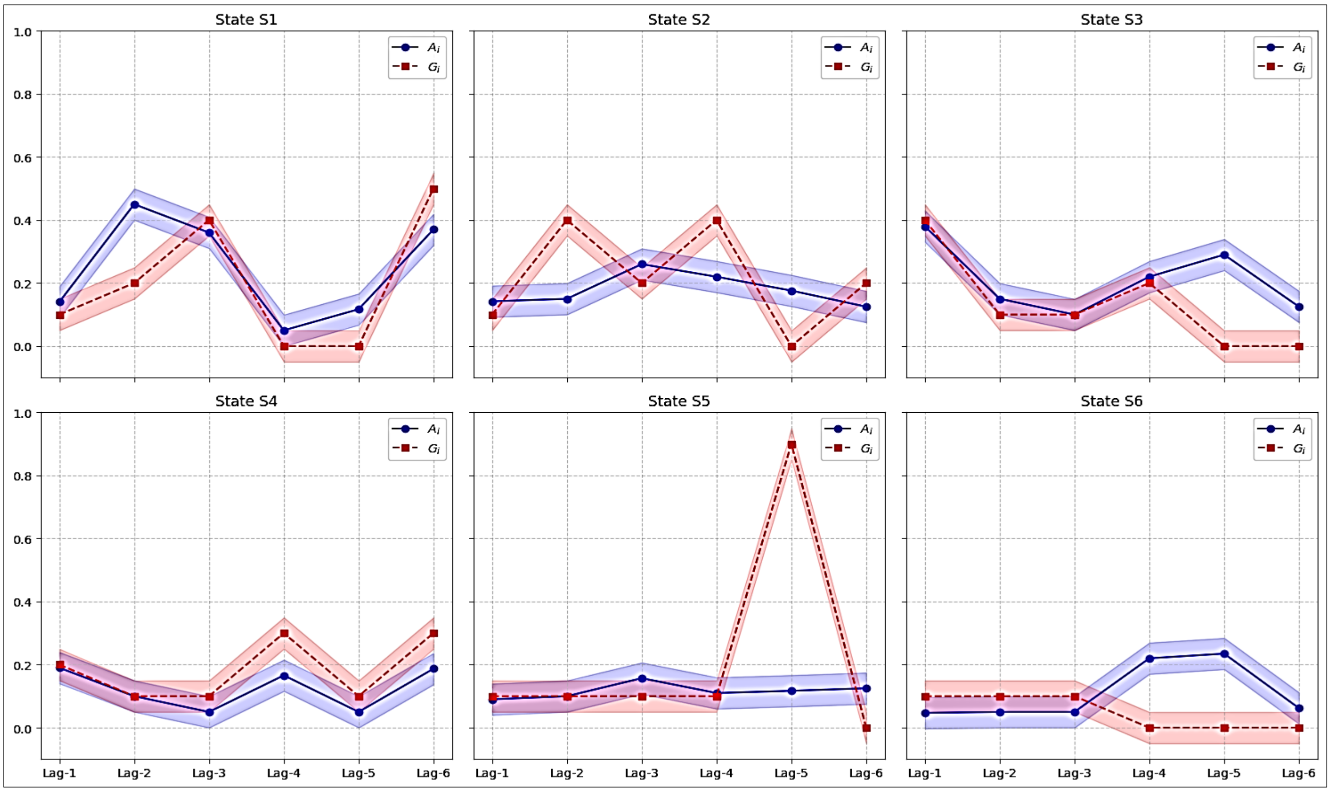

Fig 2: Probability Trends Across States with Lag-Integrated Analysis

The graph presents probability trends for six states (S_1 to S_6) over six lag periods, incorporating confidence intervals to depict variability and uncertainty. The blue line represents the steady-state probability of the actual value, while the red dashed line represents the steady-state probability of the generated sequence. Shaded confidence intervals indicate uncertainty, with wider bands reflecting greater variability. These trends help analyze state dependencies and improve predictive modeling. Each state follows a unique probability trajectory, reflecting differing influences from previous states.

The confidence bands illustrate uncertainty levels, where broader intervals indicate higher variability, and narrower ones suggest more stable probability estimates. A key observation is the sharp increase in A_i and G_i probabilities at Lag-5 for S_5 , possibly signalling a significant transition or anomaly. While most states exhibit similar trends in A_i and G_i , some deviations suggest differences in state-transition dynamics. Overall, the graph effectively captures probability changes over time, offering insights into state dependencies and potential forecasting implications.

Table 4: Evaluation of Divergence Measures for Prediction Accuracy Across Lags

Lag	KL Divergence	JS Divergence	Bhattacharyya Distance	Interpretation
1	0.0227	0.0087	0.0144	Very low divergence, Highly accurate predictions.
2	0.2439	0.0609	0.0638	Slight divergence, predictions remain reliable.
3	0.0317	0.0142	0.0261	Low divergence, stable and effective forecasting.
4	5.53	0.1169	0.1809	Significant divergence, reduced prediction accuracy.
5	17.3056	0.4587	0.9282	Very high divergence, predictions become unreliable.
6	6.233	0.123	0.1922	High divergence, weak prediction performance.

The probability distribution of states across different lags quantifies the likelihood of transitioning from one state to another over time. Higher probabilities at shorter lags indicate stronger immediate state persistence, while lower probabilities or zero values at longer lags suggest diminishing influence. The distribution varies across states, with some states maintaining stable probabilities and others fluctuating. As lag increases, the probabilities become more dispersed, reflecting increased uncertainty in transitions. This probabilistic measure provides insights into the temporal behavior of states, essential for analyzing sequential dependencies in modeling frameworks.

Conclusion

This study evaluated the effectiveness of HMMs with lag integration in capturing temporal dependencies for improved forecasting. By incorporating past observations, the model demonstrated enhanced state identification and predictive accuracy. The steady-state probabilities of actual and generated states highlighted distinct transition patterns, emphasizing the influence of lag structures on state behavior. The results indicate high accuracy for short-term predictions (lags 1-3) with minimal divergence, while higher lag values led to significant deviations. KL Divergence, JS Divergence, and Bhattacharyya Distance confirmed strong short-term reliability but weaker long-term performance, suggesting potential instability with excessive lag incorporation.

These findings highlight the critical role of optimal lag selection in HMM-based forecasting. While incorporating lag enhances prediction accuracy, excessive lag values can reduce model stability. This research contributes to time series modeling by demonstrating the advantages and limitations of lag integration. Future studies should focus on refining long-term forecasting methodologies to improve stability and reliability, ensuring more effective applications in agriculture, resource planning, and decision-making.

References

1. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics. 1966;37(6):1554-1563.
2. Hou W, Fan W, Amayri M, Bouguila N. A Novel Continuous Hidden Markov Model for Modeling Positive Sequential Data. In: Bouguila N, Fan W, Amayri M, editors. Hidden Markov Models and Applications. Unsupervised and Semi-Supervised Learning. 2022,199-210. https://doi.org/10.1007/978-3-030-99142-5_8
3. Kavitha G, Udhayakumar A, Nagarajan D. Stock market trend analysis using Hidden Markov Models. arXiv preprint arXiv:1311.4771. 2013.
4. Pastell M, Frondelius L. A Hidden Markov Model to estimate the time dairy cows spend in feeder based on indoor positioning data. Computers and Electronics in Agriculture. 2018;152:182-185.
5. Sosiawan AY, Nooraeni R, Sari LK. Implementation of using HMM-GA in time series data. Procedia Computer Science. 2021;179:713-720.
6. Vyshnavi M, Muthukumar M. Dynamic Forecasting through Hidden Markov Models and Lag Analysis: Explicitly Focusing on Agriculture. Indian Journal of Natural Science. 2024;15(87):86560-86568.
7. Zhang L, Yang W, Liang C. Application of the Hidden Markov Model to Agricultural Drought Disaster Risk Assessment. Available at SSRN 3975980. 2021.
8. Gupta A, Dhingra B. Stock market prediction using Hidden Markov Models. In: 2012 Students Conference on Engineering and Systems. IEEE. 2012, 1-4.
9. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77(2):257-286.
10. Box GE, Jenkins GM. Time Series Analysis: Forecasting and Control. Holden-Day. 1976.
11. Shumway RH, Stoffer DS. Time Series Analysis and Its Applications: With R Examples. 2017.