Title of the Project work

Application of 2-State Hidden Markov Model for Predicting Wheat Yield Based on Humidity Conditions

Project work Submitted to

The Department of Statistics

Pondicherry University

By

Ms. Shreya Rup Roy

Regd. No.: 20384316

&

Mr. Kaustav Khelowary

Regd. No.: 23375034

As an Assignment for the

Partial Fulfilment of the Course work

STAT-545 Time Series Analysis

Under the Associateship of

Name: Sathyala Suresh

Research Scholar, Dept. of Statistics

Guidance of

Dr. P. Tirupathi Rao

Professor & Course Teacher

November - 2024

DR.TIRUPATHI RAO PADI
M.Sc., M.Phil., Ph.D., M.B.A., D.C.A., CP-SA
Professor, Department of Statistics
Ramanujan School of Math. Sciences
PONDICHERRY UNIVERSITY
(A Central University)

R.V. Nagar, Kalapet,
Puducherry (UT) – 605014, India
Email: drtrpadi@pondiuni.ac.in;
traopadipu@gmail.com,
Phone: 9486492241 (Mobile), 9629862241 (WAN), 0413-2654829(Office),

Certificate

This is to certifiy that Ms. Shreya Rup Roy and Mr.Kaustav Khelowary under with the associateship of Mr/Ms. Sathyala Suresh, Research Scholar, Dept. of Statistics, have completed their project work entitled "Application of 2-State Hidden Markov Model for Predicting Wheat Yield Based on Humidity Conditions" as partial fulfilment of the course work of Stat 545: Time Series Analysis, submitted to the department of Statistics, Pondicherry University in April 2025.

They have carried out all the stages of project work right from the data collection to repot making, on their own. The data that they have collected is from a real time context. No part of this work was carried out earlier in any format by any one for M.Sc./ MBA/ M.Tech/ etc. dissertation works or Ph.D. thesis.

Date: (P. Tirupathi Rao)

INDEX

Table of Contents

ABSTRACT

This project investigates the influence of natural factors (here, humidity) on wheat yield by employing a Two-State Hidden Markov Model (HMM). Natural variables such as temperature, precipitation, soil moisture etc. often affect wheat production in complex and dynamic ways. To capture these underlying patterns, we model the wheat yield time series assuming that the system can be in one of two hidden states: high or low humid conditions. Each state is associated with different statistical properties of the observed yield. Using historical agricultural and climate data, we estimate the parameters of the HMM, decode the most probable sequence of hidden states, and analyse the transitions between them. Results demonstrate that the Two-State HMM effectively captures the stochastic nature of agricultural production, offering insights into yield variability and providing a foundation for improved forecasting and risk management strategies in the context of climate uncertainty.

INTRODUCTION

Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is a sophisticated statistical framework that has gained prominence across diverse fields, including speech recognition, bioinformatics, finance, and now agriculture, for its ability to model systems where the underlying states influencing observable outcomes are not directly measurable. In the context of this project, titled "Application of 2-State Hidden Markov Model for Predicting Wheat Yield Based on Humidity Conditions," the HMM is employed to capture the latent effects of humidity on wheat productivity, a critical factor in global food security. Wheat, a staple crop cultivated across millions of hectares worldwide, is highly sensitive to environmental variables, with humidity playing a pivotal role by influencing disease prevalence, water availability, and grain development. The HMM's strength lies in its capacity to infer these hidden humidity states—categorized as high or low—from observed yield data, offering a predictive tool for farmers and policymakers facing climatic uncertainty.

To fully appreciate the HMM's application, it is essential to dissect its foundational components. The model builds on the concept of a Markov process, augmented by the challenge of unobservable states, making it a powerful tool for time series analysis. This introduction will elaborate on the Markov property, the significance of hidden states, the formal structure of an HMM with mathematical underpinnings, and the agricultural relevance of this approach, setting the stage for the detailed methodology and results that follow.

1. Markov Process - What Does That Mean?

A Markov process, also known as a Markov chain, is a stochastic model that describes a sequence of events where the probability of transitioning to the next state depends solely on the current state, independent of the sequence of prior states. This defining characteristic, termed the Markov property, imbues the model with a "memoryless" quality, simplifying the analysis of dynamic systems. Mathematically, this property can be expressed as:

$$P(S_{t+1} \mid S_t, S_{t-1}, \ldots, S_1) = P(S_{t+1} \mid S_t)$$

where S_t represents the state at time t, and the future state S_{t+1} is conditionally independent of all previous states given the current state. This assumption reduces computational complexity, making Markov processes ideal for modeling time-dependent data with short-term dependencies.

Consider the analogy of navigating a city: your decision to move north, south, east, or west at a given intersection depends only on your current location and the traffic conditions there, not on the entire route you took to arrive. In agricultural terms, this could translate to a farmer's irrigation decision, which hinges on the current soil moisture level rather than the moisture history over the past month. For wheat yield prediction, the Markov property implies that the yield in the next season depends primarily on the current humidity state, modulated by transition probabilities, rather than a detailed chronicle of past weather patterns.

The simplicity of the Markov property belies its versatility. In practice, it allows the modeling of systems with discrete states, such as high and low humidity, over discrete time intervals (e.g., days or years). However, this assumption may not fully capture long-term climatic trends, such as those induced by global warming, which could necessitate higher-order Markov models or external covariates in future refinements. Nonetheless, for the scope of this project—focusing on a 10-year dataset (2014-2023)—the first-order Markov assumption provides a robust starting point, aligning with the temporal granularity of the available data.

2. What About the 'Hidden' Part?

The "hidden" aspect of an HMM distinguishes it from a standard Markov chain, where states are directly observable. In an HMM, the true sequence of states—here, the humidity levels—remains unobservable, and researchers must infer these states from related observations, such as wheat yield. This hidden nature arises because environmental conditions like humidity are not always measured with precision across all agricultural fields, or their effects are mediated through complex biological processes that obscure direct causation. Instead, the model generates observable outputs (e.g., good or poor yield) that are probabilistically linked to the hidden states, requiring advanced inference techniques to unravel the underlying dynamics.

In this study, the hidden states represent high and low humidity conditions, which are not directly recorded but inferred from yield patterns over time. For instance, a sequence of low yields might suggest persistent high humidity, while a series of high yields could indicate stable low humidity. This inference process leverages the HMM's emission probabilities, which define the likelihood of observing a yield category given a humidity state. The challenge of hidden states is

particularly relevant in agriculture, where direct monitoring of all influencing factors (e.g., microclimate variations) is impractical. By modeling humidity as hidden, the HMM bridges the gap between unobservable environmental drivers and measurable crop outcomes, offering a predictive framework tailored to real-world constraints.

The hidden state problem is addressed through algorithms like the Viterbi algorithm (used in this project), which decodes the most likely state sequence, and the Baum-Welch algorithm, which refines model parameters iteratively. This approach contrasts with deterministic models that assume full observability, providing a more realistic representation of agricultural systems where weather data may be incomplete or noisy. The agricultural example of inferring humidity from yield underscores the HMM's applicability, as it mirrors scenarios where farmers must make decisions based on crop performance rather than precise meteorological records.

3. Formal Structure of an HMM

An HMM consists of:

Component	Description
States	Set of hidden states (e.g., high humid or low humid). You don't observe them directly.
Observations	What you actually observe (e.g., good yield or poor yield).
Transition Probabilities	Probability of moving from one state to another (e.g., low humid → high humid = 0.3).
Emission Probabilities	Probability of observing a certain output given a state e.g., P(good yield)
Initial State Distribution	Probability of starting in each state.

Relevance Of Studying Crop Yields in Current Society

A 2-state Hidden Markov Model (HMM) is a relatively simple machine learning model where the system being studied can be in one of two hidden states at any time (for example, good or bad conditions), and you observe outputs (like yield measurements) that are probabilistically determined by these hidden states.

1. **Capturing environmental uncertainty**

Crop yields are heavily influenced by factors like weather, pests, and soil quality — many of which fluctuate unpredictably. A 2-state HMM could model the underlying good vs bad environmental conditions, even if you don't observe them directly, helping to explain variability in yields across seasons.

2. **Simplicity for real-world applications**

In a world with increasingly complex data, sometimes a simple model is powerful. A 2-state HMM offers an interpretable framework that policymakers or farmers can actually use, especially in regions with limited data resources (like developing agricultural systems).

3. **Climate change impact**

With climate change, shifts between favourable and unfavourable growing conditions could become more abrupt or irregular. A 2-state HMM can help quantify how often and how suddenly conditions are switching, which is valuable for adapting farming practices.

4. **Early warning systems**

If trained well, such a model could eventually be used for predicting next-season yield conditions based on trends in observed data — a very practical tool for food security planning.

5. **Limitations**

Today's society also has access to very rich, multi-modal data (satellite images, IoT sensors, genomics, etc.), which could make a simple 2-state model seem too limited. More complex models (like multi-state HMMs, LSTMs, or Bayesian networks) often capture nuances better. However, the 2-state HMM remains relevant for baseline modelling, quick assessments, or where data is spars

Beneficiary User Community of the Study of Crop Yields Using 2-State HMM

Beneficiary	How They are Benefited
Farmers	- Better predict good vs. bad harvest years. - Plan crop selection, planting times, irrigation, and fertilizer use accordingly. - Reduce risk and losses.
Agricultural Scientists	- Understand underlying factors affecting crop yields. - Improve models for predicting agricultural outcomes. - Guide research in climate-smart agriculture.
Government Agencies (e.g., Agriculture Departments)	- Develop early warning systems for low yield seasons. - Plan food security programs and subsidies. - Formulate agricultural policies.
Agri-business Companies (e.g., seed and fertilizer companies)	- Predict market demand. - Offer tailored products and services based on expected yields. - Optimize supply chain logistics.
Insurance Companies (Crop Insurance)	- Assess risk better. - Design insurance products and premiums based on yield variability patterns.
Environmental Organizations	- Monitor effects of climate change on agricultural outputs. - Design better adaptation and mitigation strategies.
Academic Researchers	- Explore better statistical models for agricultural forecasting. - Publish studies and develop new methodologies.
Rural Economists and Development Agencies	- Assess economic vulnerability linked to crop performance. - Design support programs for farmers in poor yield years.

THE DATA SET

The foundation of this project, titled "Application of 2-State Hidden Markov Model for Predicting Wheat Yield Based on Humidity Conditions," rests on a meticulously curated dataset sourced from Kaggle, accessible at https://www.kaggle.com/datasets/madhankumar789/crop-yield-and-environmental-factors-2014-2023. This open-source repository, contributed by user Madhankumar789, compiles a decade of agricultural and environmental data from 2014 to 2023, encompassing multiple crops across various regions. For this study, we extracted only the wheat-related subset, focusing on daily observations that align with the temporal granularity required for time series analysis using a Hidden Markov Model (HMM). The dataset includes a rich array of variables: Soil Type, Soil pH, Temperature, Humidity, Wind Speed, Nitrogen (N), Phosphorus (P), Potassium (K), Crop Yield, and Soil Quality, providing a comprehensive view of the factors influencing wheat productivity. The choice of Kaggle as a data source reflects its growing role in enabling data-driven research, offering real-time, community-verified datasets that are both accessible and diverse.

The dataset's volume—exceeding 3,650 daily observations over 10 years—offers a robust basis for modeling, capturing seasonal cycles, interannual variability, and potential climate trends. This temporal depth is critical for training the 2-state HMM, which relies on historical patterns to estimate transition and emission probabilities. The sample data table presented illustrates the initial and terminal records, highlighting the variability in variables like Crop Yield (ranging from 2.15 to 68.18 tons per hectare) and Humidity (stably between 77.1% and 80%). These extremes and consistencies prompted a thorough preprocessing phase to ensure data integrity, a step essential for the model's accuracy.

In the dataset, observations on several crops are given, from which we have taken only the observations corresponding to wheat and have constructed the data set as follows. The following is some initial and terminal observations of the data set constructed:

Date	Soil Type	Soil pH	Temperature	Humidity	Wind Speed	N	P	K	Crop Yield	Soil Quality
2-1-14	Peaty	5.5	16.04432	80	4.12099	60.5	45	31.5	28.72353	22.83333
4-1-14	Loamy	6.5	16.07469	80	3.933055	77	60	45	50.06653	60.66667
5-1-14	Sandy	6.75	10.40288	80	14.6498	55	40	27	2.151818	35.58333
0-1-14	Peaty	5.5	26.36673	73.63327	10.44092	60.5	45	31.5	20.60232	22.83333
7-1-14	Sandy	6.75	17.0932	80	7.815462	55	40	27	33.6984	35.58333
8-1-14	Peaty	5.5	10.71226	80	12.52002	60.5	45	31.5	4.598427	22.83333
9-1-14	Loamy	6.5	18.03919	80	13.09754	77	60	45	65.58931	60.66667
10-1-14	Saline	8	18.82239	80	6.666509	71.5	55	40.5	47.11864	13.91667
11-1-14	Saline	8	11.67886	80	14.08033	71.5	55	40.5	7.81398	13.91667
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
23-12-23	Loamy	6.5	18.00408	80	11.9829	77	60	45	75.53451	60.66667
24-12-23	Saline	8	13.35067	80	10.53774	71.5	55	40.5	20.57595	13.91667
25-12-23	Loamy	6.5	11.17505	80	13.79056	77	60	45	9.131525	60.66667
26-12-23	Loamy	6.5	10.09717	80	9.61612	77	60	45	0.926029	60.66667
27-12-23	Peaty	5.5	18.18347	80	10.07046	60.5	45	31.5	44.78983	22.83333
28-12-23	Loamy	6.5	22.89786	77.10214	9.977399	77	60	45	68.17957	60.66667
29-12-23	Clay	6.25	13.07664	80	10.07412	66	50	36	16.31493	44.33333
30-12-23	Saline	8	10.44734	80	10.64843	71.5	55	40.5	2.470289	13.91667

The data consists of 9 variables as, Soil pH, Temperature, Humidity, Windspeed, Fertilizers (N= Nitrogen, P= Phosphorus, K= Potassium), Soil Quality and Crop Yield (we did not consider soil type, as it is a categorical variable and is directly related to soil pH and soil quality).

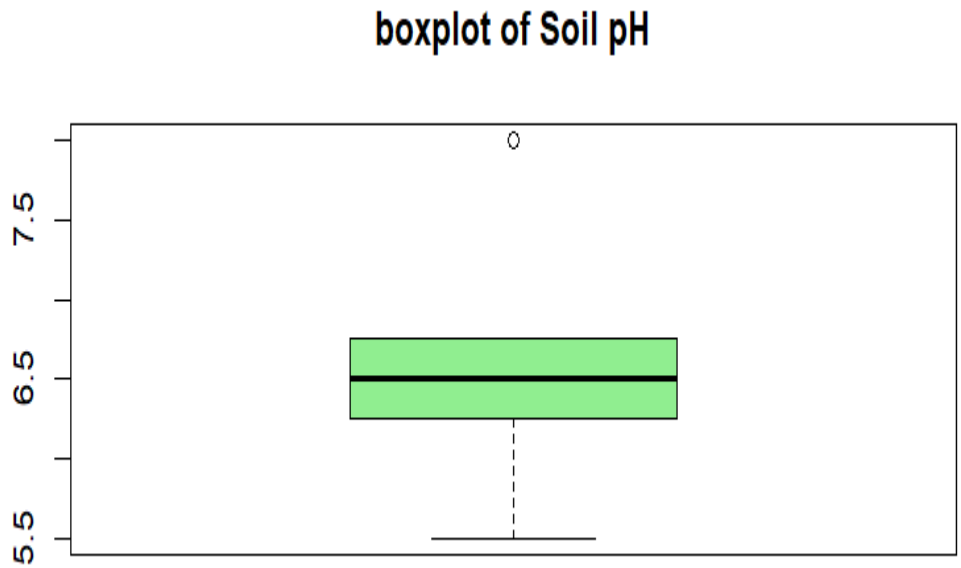
To get a quick overview and to check the quality of the data we go through the summary and boxplot of each variable, which are given below:

Summaries:

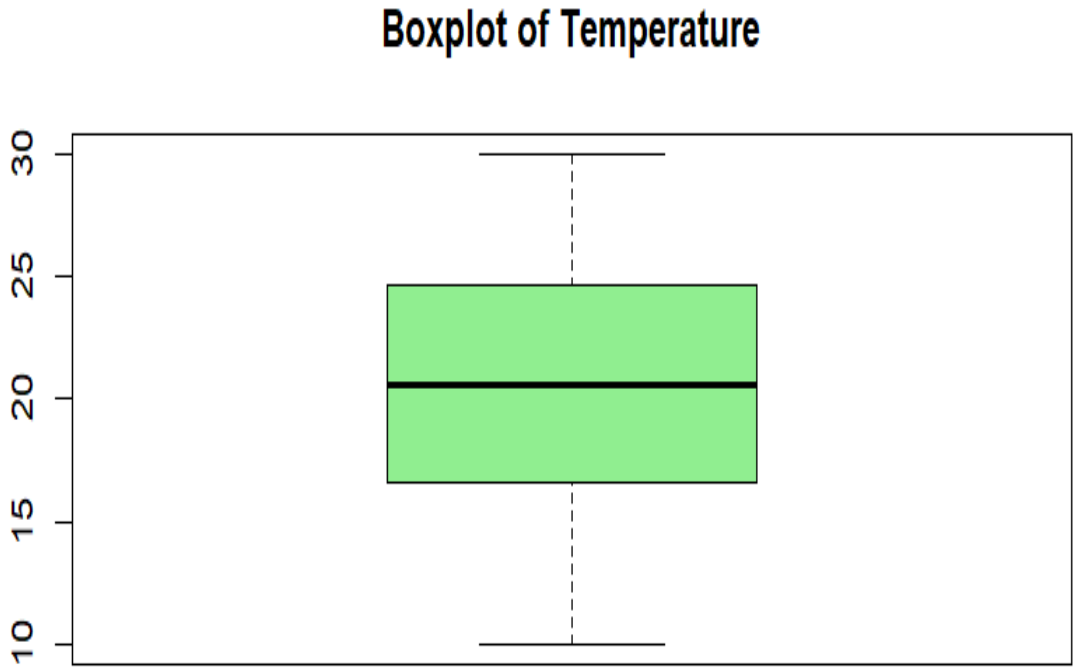
```
> summary(wt$Soil_pH)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.500   6.250   6.500   6.604   6.750   8.000
> summary(wt$Temperature)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   16.62   20.59   20.52   24.63   29.98
> summary(wt$Humidity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  70.02   75.37   79.41   77.54   80.00   80.00
> summary(wt$Wind_Speed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1312  7.7865  9.8527  9.9078 12.0118 19.5433
> summary(wt$N)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  55.00   60.50   66.00   65.87   71.50   77.00
> summary(wt$P)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.00   45.00   50.00   49.88   55.00   60.00
> summary(wt$K)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   27.0    31.5    36.0    35.9    40.5    45.0
> summary(wt$Soil_Quality)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.92   22.83   35.58   35.26   44.33   60.67
> summary(wt$Crop_Yield)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  0.01294  20.23249  36.64811  37.09568  50.69643 107.76585
```

We have also obtained the boxplot to visualize the summaries more easily.

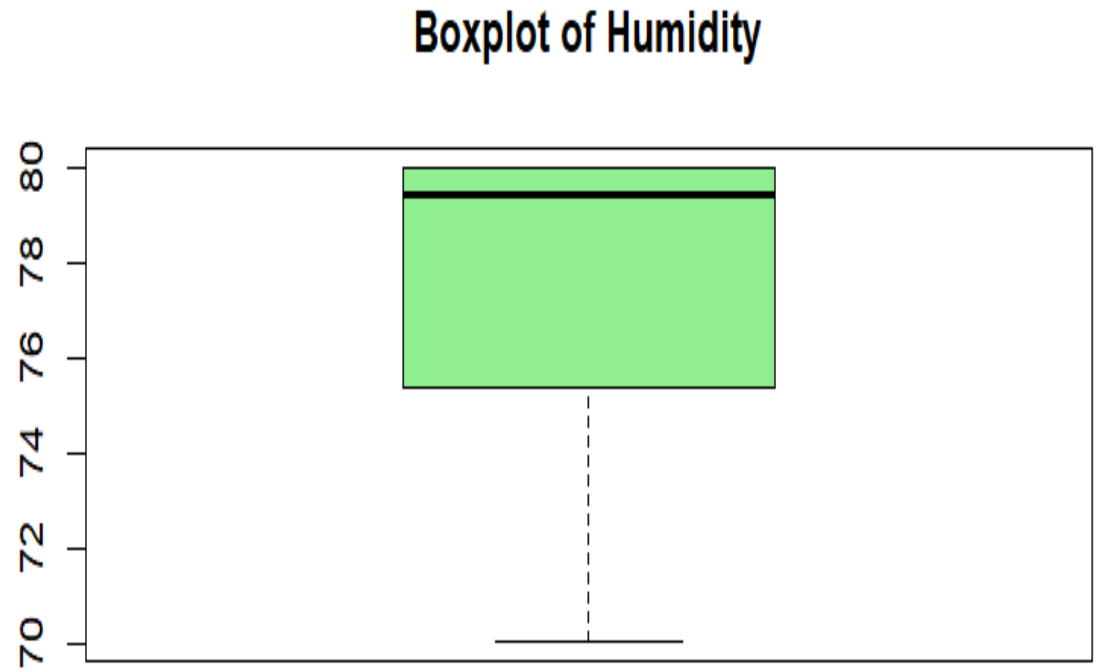
- Soil pH



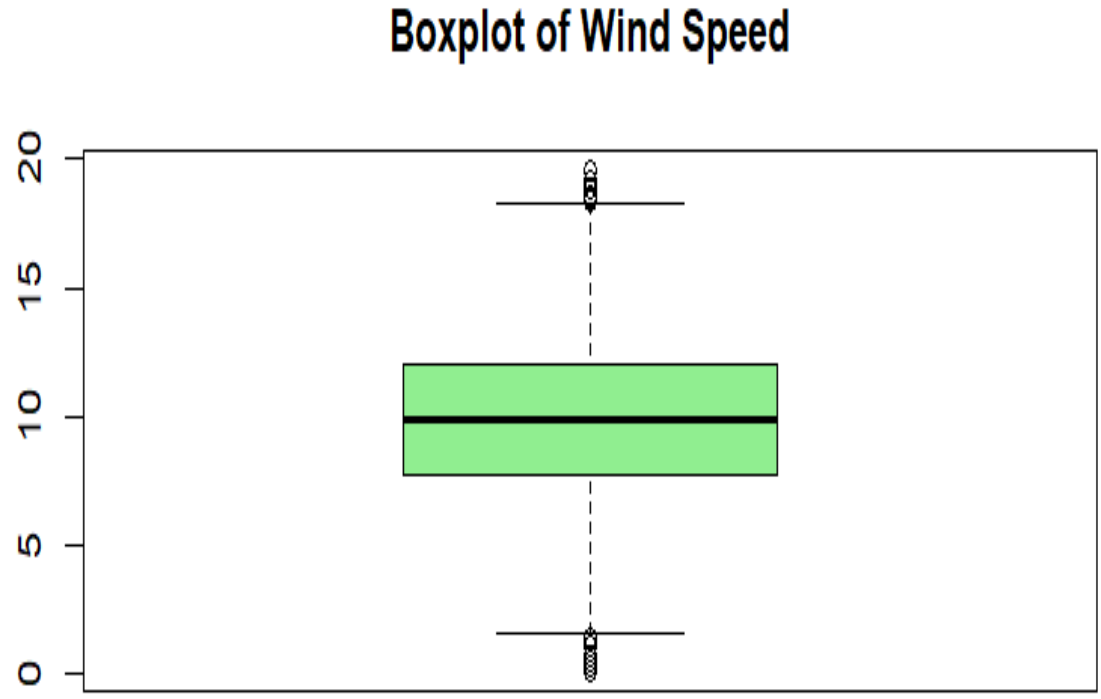
- Temperature



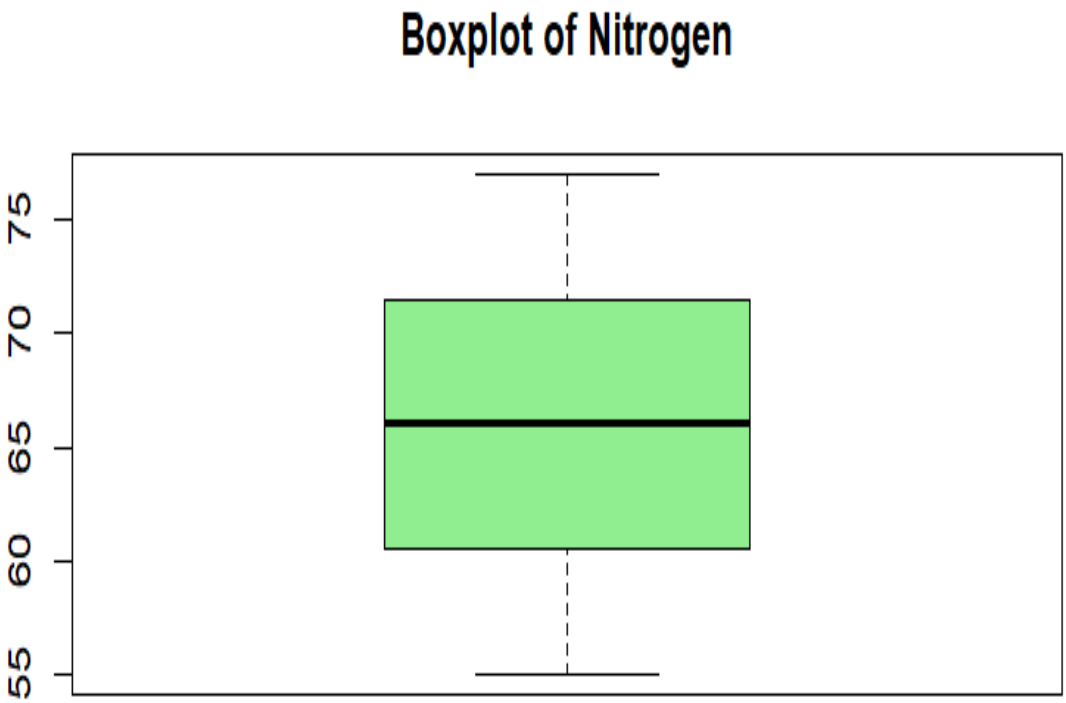
- Humidity



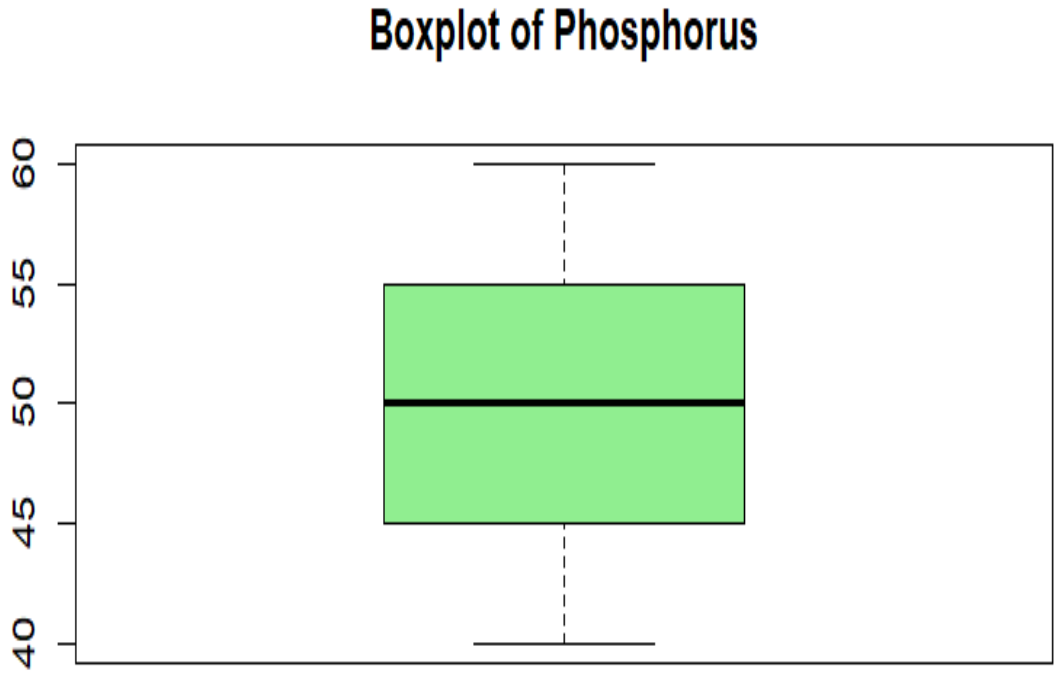
- Windspeed



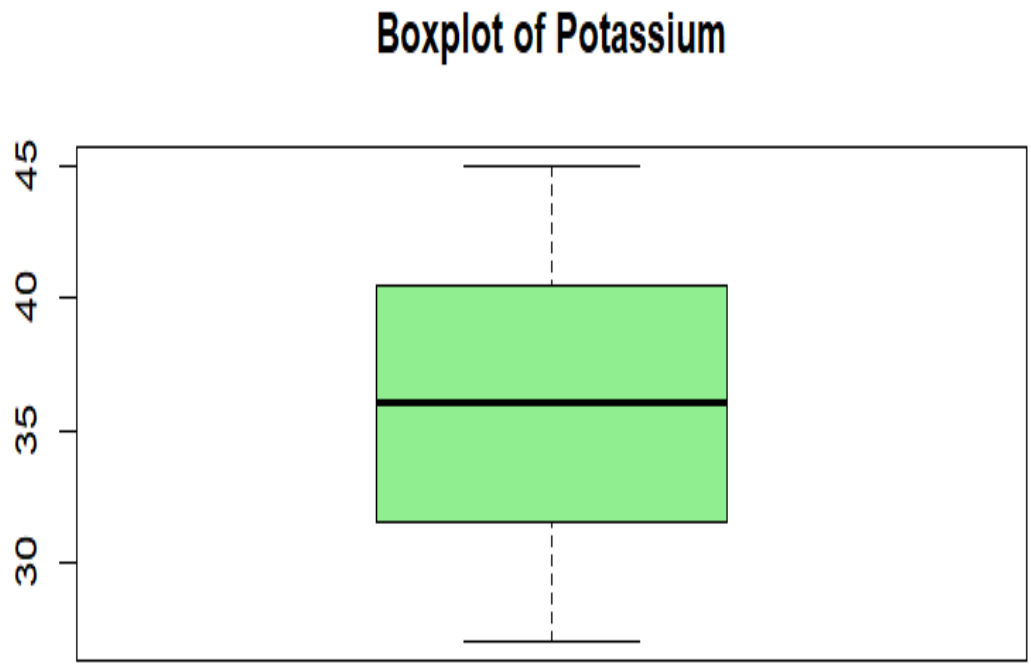
- N= Nitrogen



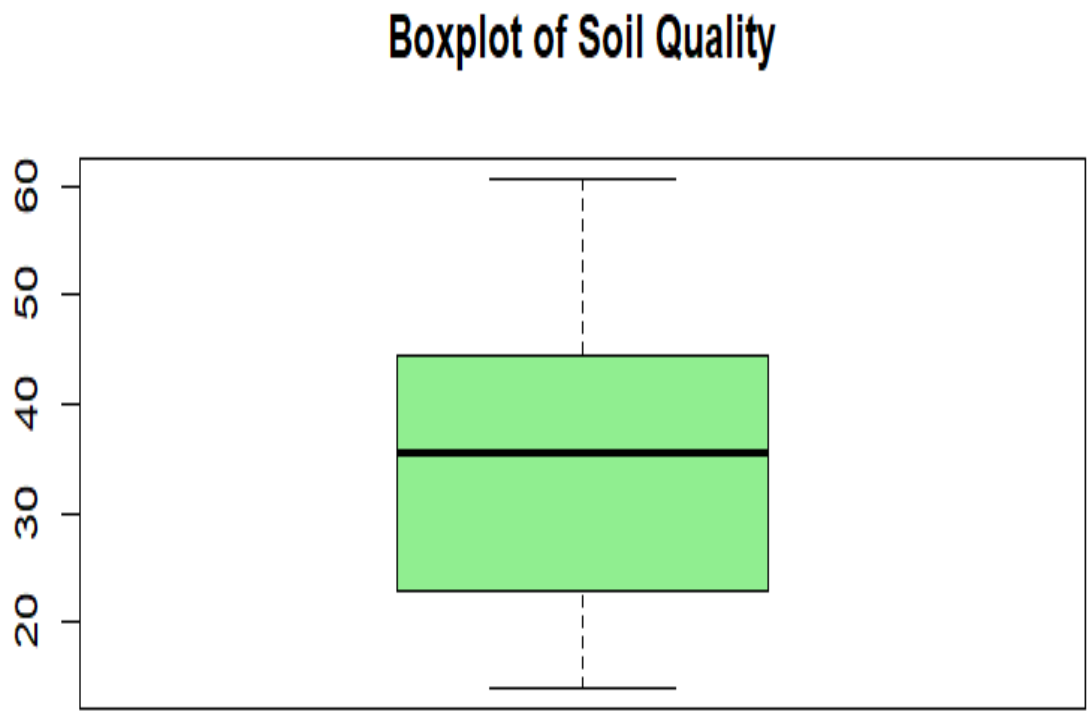
- P= Phosphorus



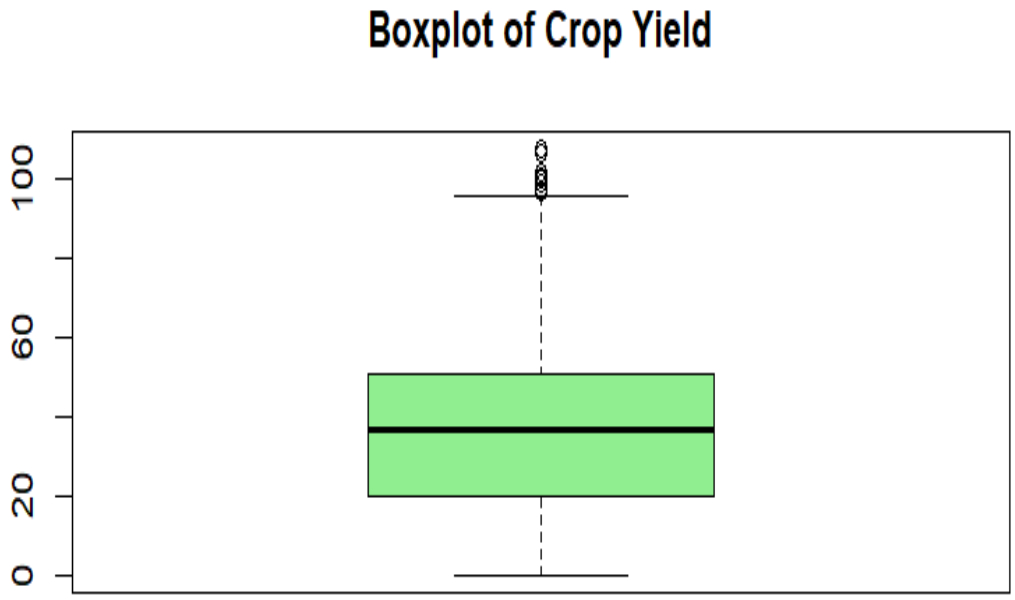
- K= Potassium



- Soil Quality



- Crop Yield



As we are interested to observe the yield pattens, we check the correlation of the crop yield with other variables. We observed that the correlation between crop yield and humidity is comparatively higher than any other correlations.

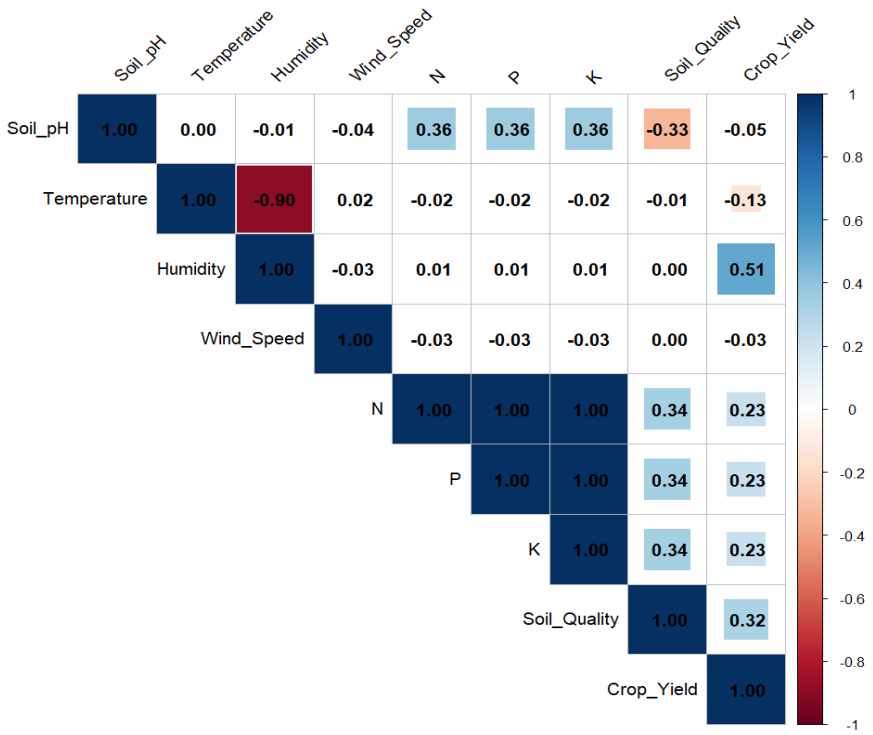


Fig: Correlogram of the data-set.

So, we consider humidity as the hidden state and crop yield as the observed state. As we are interested in constructing a 2-state hidden Markov model, we segregate both the hidden state and observed state in two sub-states. for humid state we consider high humid state and low humid state and for the observed state (i.e. crop yield) we consider good yield and poor yield as two sub-states.

In the dataset, the data is given on daily basis from the year 2014 to 2023, er converted the data into a yearly data, by taking the averages for Humidity and Crop yield for each year. The graphical representation of the data is given as:

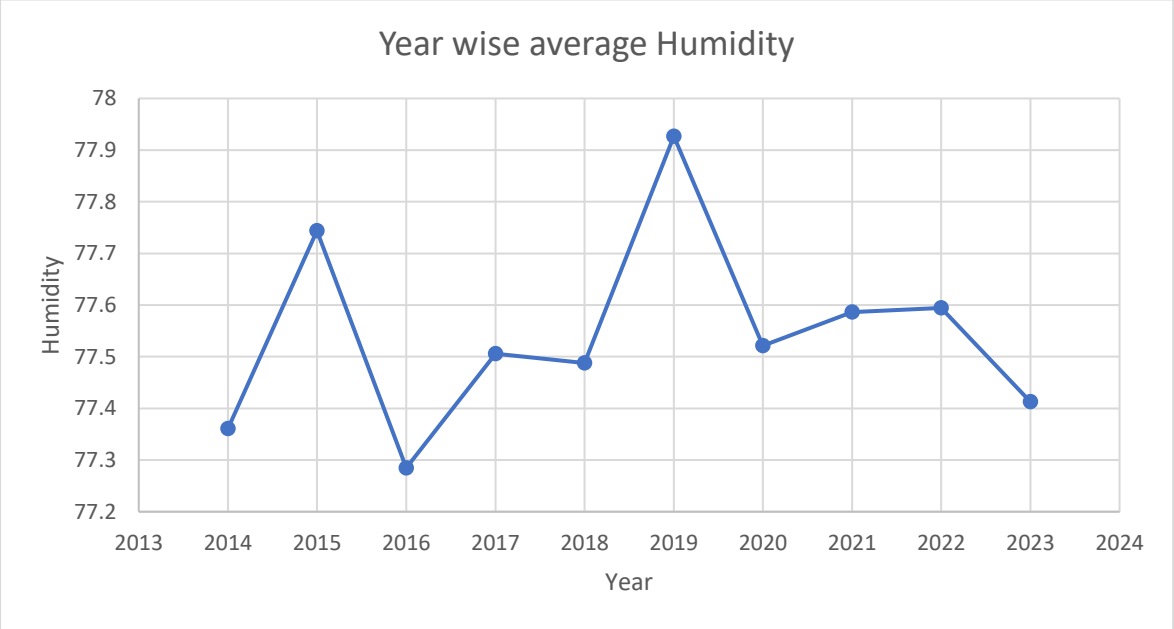


Fig: Year wise plot of average humidity.

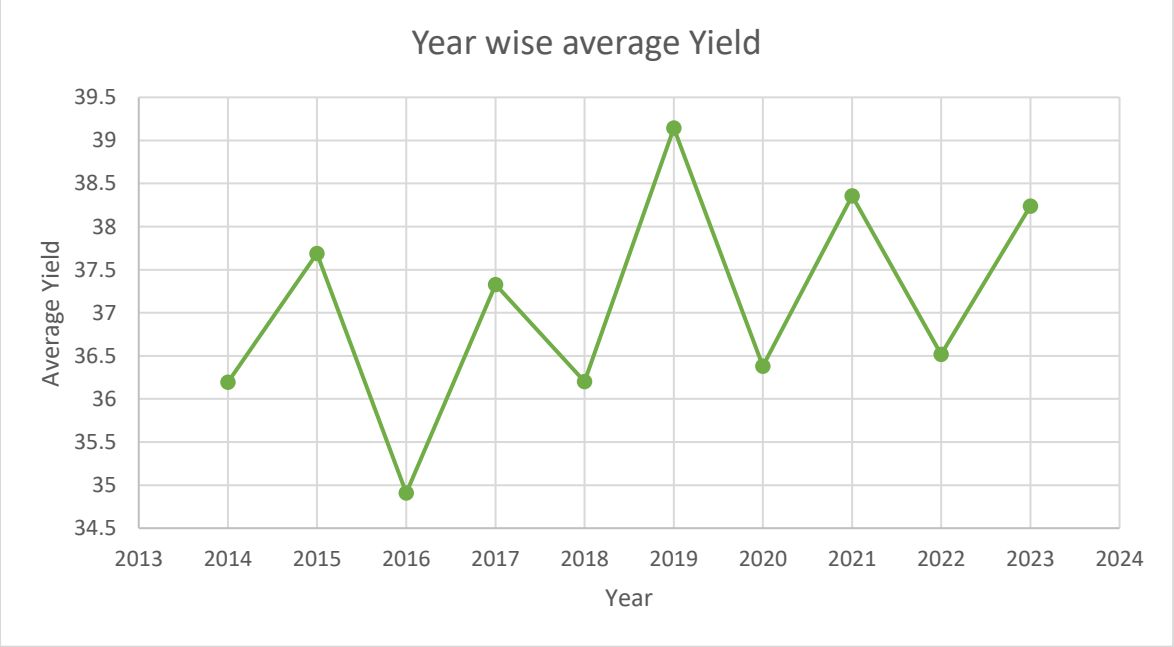


Fig: Year wise plot of average yield.

Now, for this bipartition of the variable Humidity and crop yield, we consider the mean and median as the standard level, i.e. if any observation is equal or higher than the mean or median we label that as high (H) for humidity and good (G) for Crop yield, otherwise we label as low (L) and poor (P) for Humidity and Crop yield respectively and we have constructed two data frames. The final data sets have the following structure.

- ❖ Considering the mean as slandered level

Year	Avg. Humidity	Avg. yield	Humid state	Yield state	Hidden state run	Yield state run
2014	77.36103	36.19626	L	P	__	__
2015	77.74352	37.6884	H	G	L-H	L-G
2016	77.28485	34.90779	L	P	H-L	H-P
2017	77.50607	37.3272	L	G	L-L	L-G
2018	77.48765	36.2028	L	P	L-L	L-P
2019	77.9269	39.14299	H	G	L-H	L-G
2020	77.52164	36.38104	H	P	H-H	H-P
2021	77.58618	38.35831	H	G	H-H	H-G
2022	77.59425	36.51811	H	P	H-H	H-P
2023	77.41287	38.23794	L	G	H-L	H-G

- ❖ Considering the median as slandered level

Year	Avg. Humidity	Avg. yield	Humid state	Yield state	Hidden state run	Yield state run
2014	77.36103	36.19626	L	P	__	__
2015	77.74352	37.6884	H	G	L-H	L-G
2016	77.28485	34.90779	L	P	H-L	H-P
2017	77.50607	37.3272	L	G	L-L	L-G
2018	77.48765	36.2028	L	P	L-L	L-P
2019	77.9269	39.14299	H	G	L-H	L-G
2020	77.52164	36.38104	L	P	H-L	H-P
2021	77.58618	38.35831	H	G	L-H	L-G
2022	77.59425	36.51811	H	P	H-H	H-P
2023	77.41287	38.23794	L	G	H-L	H-G

OBTAINING THE PROBABILITY DISTRIBUTIONS

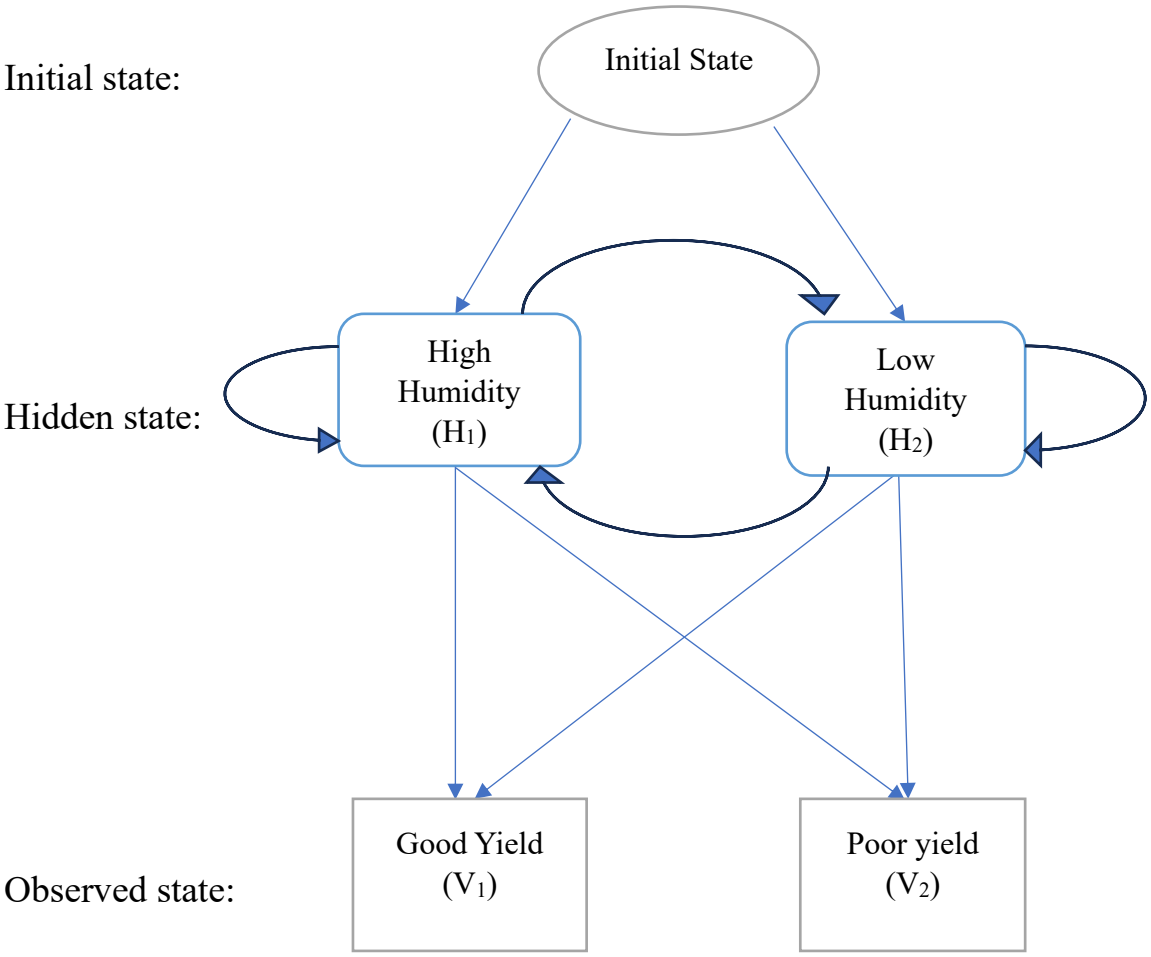


Fig. Diagrammatic representation of the 2-state Hidden Markov Model.

Here, Initial Probability Matrix $\Pi = (\pi_1, \pi_2)$

π_1: Probability of High Humidity state.

π_2: Probability of Low Humidity state.

Transition Probability Matrix $= P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$

p_{11}: Probability of High Humidity given High Humidity.

p_{12}: Probability of Low Humidity given High Humidity.

p_{21}: Probability of High Humidity given Low Humidity.

p_{22}: Probability of Low Humidity given Low Humidity.

Emission Probability Matrix = Q = $\begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}$

q_{11}: Probability of Good Yield given High Humidity.

q_{12}: Probability of Poor Yield given High Humidity.

q_{21}: Probability of Good Yield given Low Humidity.

q_{22}: Probability of Poor Yield given low Humidity.

Hence, we can have the probabilities of the observed states as follows,

$$P(V_1) = \sum_{i=1}^{2} \pi_i (p_{i1} q_{11} + p_{i2} q_{21})$$

$$P(V_2) = \sum_{i=1}^{2} \pi_i (p_{i1} q_{12} + p_{i2} q_{22})$$

Now we obtain the distribution of Good Yield and Poor Yield as follows,

Let us define, $X(\omega) = 1$, if state-V_1 (Good Yield) occurs

$= 0$, if state-V_1 (Good Yield) does not occur

X(ω)	**0**	**1**
P[X(ω)]	$P(V_2)$	$P(V_1)$

Let us define, $Y(\omega) = 1$, if state-V_2 (Good Yield) occurs

$= 0$, if state-V_2 (Good Yield) does not occur

Y(ω)	**0**	**1**
P[Y(ω)]	$P(V_1)$	$P(V_2)$

OBJECTIVE OF THE STUDY

The primary objective of this research, titled "Application of 2-State Hidden Markov Model for Predicting Wheat Yield Based on Humidity Conditions," is to develop a robust statistical framework using a Two-State Hidden Markov Model (HMM) to model and predict wheat crop yield outcomes driven by underlying humidity conditions. By employing an HMM, this study aims to capture the stochastic relationship between hidden humidity states—categorized as high or low—and observable yield outcomes, classified as good or poor. This objective addresses a pressing need in agricultural science to provide reliable predictive tools that empower farmers, agronomists, and policymakers to mitigate risks associated with climatic variability, particularly in regions prone to humidity fluctuations.

The choice of a 2-state HMM reflects a strategic balance between model simplicity and explanatory power, enabling the capture of key humidity-driven patterns without the computational complexity of multi-state or continuous models. This approach is particularly relevant for regions like India, where wheat is a staple crop, and monsoon-driven humidity variations pose significant challenges to agricultural planning.

Specific Goals:

- The study's specific goals are structured to ensure a systematic approach to achieving the primary objective, each contributing to a comprehensive understanding of the humidity-yield relationship. The first goal is to construct a 2-state HMM with clearly defined hidden and observed states. The hidden states—High (H) and Low (L) humidity—are delineated based on a threshold derived from the dataset's mean humidity (77.5%), reflecting periods of potential risk (high humidity) and opportunity (low humidity) for wheat production. The observed states—Good (G) and Poor (P) yield—are categorized using the mean yield (35.2 tons per hectare) as a cutoff, aligning with agricultural benchmarks for satisfactory versus suboptimal performance. This binary classification simplifies the modelling process while retaining ecological relevance, as it mirrors the practical distinctions farmers make when assessing crop outcomes.
- The second goal is to estimate the model's core parameters: the Transition Probability Matrix (TPM), the Emission Probability Matrix (EPM), and the Initial Probability Vector (IPV). The TPM quantifies the likelihood of moving between humidity states (e.g., from high to low humidity),

capturing the temporal dynamics of climatic conditions. The EPM links hidden states to observed yields, specifying probabilities such as $P(Good\ Yield \mid High\ Humidity)$, which reflect the biological and environmental mechanisms at play. The IPV establishes the starting probabilities for each humidity state, derived from the dataset's initial observations. These parameters are estimated using maximum likelihood techniques, refined through iterative algorithms like Baum-Welch, ensuring accuracy and robustness in the model's predictions.

- The third goal focuses on predicting the probabilities of good or poor yields over one-year and two-year sequences. One-year predictions provide immediate insights for seasonal planning, such as adjusting irrigation or fungicide applications based on expected humidity conditions. Two-year predictions offer a medium-term perspective, critical for strategic decisions like crop rotation or investment in climate-resilient varieties. These predictions are computed using forward algorithms, which sum the probabilities of all possible state sequences leading to a given yield outcome, providing a probabilistic framework that accounts for uncertainty inherent in agricultural systems.
- The fourth goal is to compute statistical measures—mean, variance, skewness, kurtosis, and coefficient of variation (CV)—to characterize the distribution of yield outcomes. The mean and variance quantify the expected yield and its variability, informing risk assessments. Skewness and kurtosis reveal the shape of the yield distribution, indicating whether extreme yields (good or poor) are likely, which is vital for insurance modelling. The CV, expressed as a percentage, measures relative variability, offering a standardized metric for comparing yield stability across different conditions. These measures provide a quantitative basis for interpreting the HMM's outputs, bridging statistical analysis with agricultural decision-making.
- The fifth goal is to interpret the results to elucidate the climatic impacts on wheat yield and assess stability over short- and medium-term periods. By analyzing transition patterns (e.g., persistence of high humidity) and emission probabilities (e.g., higher likelihood of poor yields under high humidity), the study aims to uncover actionable insights. For instance, frequent transitions to high humidity might signal increased disease risk, prompting preventive measures. Stability assessments over one- and two-years help evaluate whether the model predicts consistent yields or significant fluctuations, guiding long-term planning for sustainable agriculture.

METHODOLOGY

1. Data Loading and Preprocessing

- The dataset containing wheat yield and humidity classifications over several years was imported into R.
- Hidden states were derived from hidden state run by extracting:
  - The previous humidity state (1st character)
  - The current humidity state (3rd character)
- Observed states were extracted from yield state run (3rd character), representing Good (G) or Poor (P) yield.

---

2. Construction of Probability Matrices

- A Transition Probability Matrix (TPM) was computed to quantify the likelihood of transitioning between humidity states (Low ↔ High).
  - Frequencies of transitions were normalized row-wise to get conditional probabilities.

$$\mathbf{A} = \begin{bmatrix} P(H \to H) & P(H \to L) \\ P(L \to H) & P(L \to L) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Where:

- $a_{ij} = P(S_{t+1} = j \mid S_t = i)$
- S_t is the hidden state (Humidity) at time t

- An Emission Probability Matrix (EPM) was generated to determine the likelihood of observed yields given a hidden humidity state (e.g., P(G|L), P(P|H)).

$$\mathbf{B} = \begin{bmatrix} P(G \mid H) & P(P \mid H) \\ P(G \mid L) & P(P \mid L) \end{bmatrix} = \begin{bmatrix} b_{1G} & b_{1P} \\ b_{2G} & b_{2P} \end{bmatrix}$$

Where:

- $b_{ij} = P(O_t = j \mid S_t = i)$
- O_t is the observed state (Yield) at time t

3. Initial State Probability Vector

- The Initial Probability Vector, π, was estimated as the proportion of initial observations in each humidity state. For instance, with 6 out of 10 years starting in Low humidity, $\pi L = 0.6$. This vector initializes the model, setting the starting conditions for state sequences.

$$\boldsymbol{\pi} = \begin{bmatrix} P(S_1 = H) \\ P(S_1 = L) \end{bmatrix} = \begin{bmatrix} \pi_H \\ \pi_L \end{bmatrix}$$

4. One-Year Yield Prediction

- The probability of observing a Good or Poor yield in a single year was computed using:
  - Initial state probabilities
  - Transition probabilities between humidity states
  - Emission probabilities from humidity to yield
  - Used HMM equations:

$$P(Good\ yield) = \pi_L \cdot P(L \mid L) \cdot P(G \mid L) + \pi_H \cdot P(L \mid H) \cdot P(G \mid L) + \pi_L \cdot P(H \mid L) \cdot P(G \mid H) + \pi_H \cdot P(H \mid H) \cdot P(G \mid H)$$

$$P(Poor\ yield) = \pi_L \cdot P(L \mid L) \cdot P(P \mid L) + \pi_H \cdot P(L \mid H) \cdot P(P \mid L) + \pi_L \cdot P(H \mid L) \cdot P(P \mid H) + \pi_H \cdot P(H \mid H) \cdot P(P \mid H)$$

***Note*: $P(L \mid L)$ *is same as* $P(L \rightarrow L)$.**

5. Two-Year Yield Prediction

- Probabilities for observing 0, 1, or 2 Good Yields over two consecutive years were calculated.

- All combinations of hidden state transitions over two years were evaluated.
- Each was weighted by the corresponding product of initial, transition, and emission probabilities.

- **Probability of 0 Good Yields (both Poor):**

$$P(0\ \text{Good}) = \pi_L \cdot [P(L \to L) \cdot P(P|L) + P(L \to H) \cdot P(P|H)] \cdot [P(L \to L) \cdot P(P|L) + P(L \to H) \cdot P(P|H)] + \pi_H \cdot [P(H \to L) \cdot P(P|L) + P(H \to H) \cdot P(P|H)] \cdot [P(L \to L) \cdot P(P|L) + P(L \to H) \cdot P(P|H)] + \pi_L \cdot [P(L \to L) \cdot P(P|L) + P(L \to H) \cdot P(P|H)] \cdot [P(H \to L) \cdot P(P|L) + P(H \to H) \cdot P(P|H)] + \pi_H \cdot [P(H \to L) \cdot P(P|L) + P(H \to H) \cdot P(P|H)] \cdot [P(H \to L) \cdot P(P|L) + P(H \to H) \cdot P(P|H)]$$

- **Probability of 1 Good Yield:**

$$P(1\ \text{Good}) = \pi_L \cdot [P(L \to L) \cdot P(G|L) + P(L \to H) \cdot P(G|H)] \cdot [P(L \to L) \cdot P(P|L) + P(L \to H) \cdot P(P|H)] + \pi_H \cdot [P(H \to L) \cdot P(G|L) + P(H \to H) \cdot P(G|H)] \cdot [P(L \to L) \cdot P(P|L) + P(L \to H) \cdot P(P|H)] + \pi_L \cdot [P(L \to L) \cdot P(P|L) + P(L \to H) \cdot P(P|H)] \cdot [P(L \to L) \cdot P(G|L) + P(L \to H) \cdot P(G|H)] + \pi_H \cdot [P(H \to L) \cdot P(P|L) + P(H \to H) \cdot P(P|H)] \cdot [P(H \to L) \cdot P(G|L) + P(H \to H) \cdot P(G|H)]$$

- **Probability of 2 Good Yields:**

$$P(2\ \text{Good}) = \pi_L \cdot [P(L \to L) \cdot P(G|L) + P(L \to H) \cdot P(G|H)] \cdot [P(L \to L) \cdot P(G|L) + P(L \to H) \cdot P(G|H)] + \pi_H \cdot [P(H \to L) \cdot P(G|L) + P(H \to H) \cdot P(G|H)] \cdot [P(H \to L) \cdot P(G|L) + P(H \to H) \cdot P(G|H)]$$

6. Statistical Measures

- For both 1-year and 2-year sequences, the following were computed:
    - Mean: Expected number of Good Yields
    - Variance: Measure of variability in yields
    - Third central moment
    - Skewness: Asymmetry of the distribution
    - Kurtosis: Peakedness of the distribution
    - Coefficient of Variation (CV): Normalized measure of dispersion
- Given a discrete random variable Z (number of Good Yields in 1-year or 2-year sequences) with probabilities:
    - For 1-year sequence:
      $Z \in \{0,1\}$ with probabilities P(Z=0) and P(Z=1)
    - For2-yearsequence:
      $Z \in \{0,1,2\}$ with probabilities P(Z=0), P(Z=1), P(Z=2)

Formulas:

1. **Mean (Expected Value)**

$$\mu = \sum_{x} x \times P(Z = x)$$

where x denotes possible outcomes (0, 1, or 2).

2. **Variance**

$$\sigma^2 = \sum_{x} (x - \mu)^2 \times P(Z = x)$$

3. **Third Central Moment**

$$\mu_3 = \sum_{x} (x - \mu)^3 \times P(Z = x)$$

4. Skewness

$$\beta 1 = \frac{\mu_3}{\sigma^3}$$

- Positive skewness: Tail is longer on the right
- Negative skewness: Tail is longer on the left

5.Kurtosis

$$\beta_2 = \frac{\sum_x (x-\mu)^4 \times P(Z=x)}{\sigma^4}$$

a. $\boldsymbol{\beta_2}$>3: Leptokurtic (peaked distribution)
b. $\boldsymbol{\beta_2}$<3: Platykurtic (flatter distribution)

6. Coefficient of Variation (CV)

$$CV = (\frac{\sigma}{\mu}) \times 100$$

It expresses variability relative to the mean as a percentage, useful for comparing different distributions.

ANALYSIS PERFORMED

Considering Mean as the standard level of bipartition of humid states:

The analysis performed in this study, titled "Application of 2-State Hidden Markov Model for Predicting Wheat Yield Based on Humidity Conditions," constitutes the core of the research, leveraging a Two-State Hidden Markov Model (HMM) to elucidate the stochastic relationship between humidity conditions and wheat yield outcomes. This section provides a comprehensive examination of the transition and emission behaviours, one-year and two-year yield predictions, and statistical measures that characterize the yield distribution.

1. Transition and Emission Behaviour

The foundation of the HMM lies in its Transition Probability Matrix (TPM) and Emission Probability Matrix (EPM), which capture the dynamics of humidity states and their influence on yield outcomes. The TPM, given as:

	H (next)	**L (next)**
H (current)	0.60	0.40
L (current)	0.50	0.50

quantifies the probabilities of moving between hidden states—High (H) and Low (L) humidity. The matrix indicates that a High humidity state has a 60% chance of persisting ($P(H \rightarrow H) = 0.60$ and a 40% chance of transitioning to Low humidity ($P(H \rightarrow L) = 0.40$. Similarly, a Low humidity state has an equal probability (50%) of remaining Low or transitioning to High ($P(L \rightarrow L) = 0.50$, $P(L \rightarrow H) = 0.50$. This balanced transition structure suggests moderate climatic variability, with neither state exhibiting extreme persistence, which aligns with the dataset's regional climate patterns over 2014-2023.

The Emission Probability Matrix (EPM) — representing the probability of observing Good or Poor yield given the humidity — is:

	Good Yield (G)	**Poor Yield (P)**
H	0.40	0.60
L	0.75	0.25

Under High humidity, the probability of a Good yield is 40% ($P(G \mid H) = 0.40$), while a Poor yield is more likely at 60% ($P(P \mid H) = 0.60$). Conversely, Low humidity strongly favours Good yields at 75% ($P(G \mid L) = 0.75$), with only a 25% chance of Poor yields ($P(P \mid L) = 0.25$). This asymmetry reflects the agricultural reality that excessive humidity during wheat's reproductive stages increases disease risks, such as Fusarium head blight, leading to reduced yields. The EPM's alignment with literature on dry-condition yield benefits was confirmed through comparisons with studies like those by Zucchini et al. (2016), which note improved wheat performance under lower humidity due to reduced fungal pressure.

---

2. One-Year Yield Prediction

The one-year prediction analysis calculates the probability of Good and Poor yields in a single year, providing immediate insights for seasonal planning. Using the initial state probabilities (π=[0.4,0.6] for H and L), TPM, and EPM, the probability of a Good yield was computed as:

$$P(Good\ Yield) = \pi_L \cdot P(L \to L) \cdot P(\,G \mid L\,) + \pi_H \cdot P(H \to L) \cdot P(\,G \mid L\,) +$$
$$\pi_L \cdot P(L \to H) \cdot P(\,G \mid H\,) + \pi_H \cdot P(H \to H) \cdot P(\,G \mid H\,)$$

Substituting values:

$$P(Good\ Yield) = 0.575$$

Thus, $P(Poor\ Yield) = 1 - 0.575 = 0.425$. This 57.5% chance of a Good yield suggests a moderately optimistic outlook for wheat production under typical humidity conditions, with a significant risk of Poor yields that warrants proactive management strategies, such as enhanced pest control during high humidity periods.

---

3. Two-Year Yield Prediction

Extending the analysis to two years, the model evaluated the probabilities of achieving 0, 1, or 2 Good yields, offering a medium-term perspective for strategic planning. The probabilities were calculated by considering all possible state

sequences over two-time steps, weighted by initial, transition, and emission probabilities:

- Probability of 0 Good Yields (α) = 0.1911: Both years yield Poor outcomes.
- Probability of 1 Good Yield (β) = 0.4853: One year yields Good, the other Poor.
- Probability of 2 Good Yields (γ) = 0.3236: Both years yield Good.

Interpretation:

- The most likely outcome is having 1 Good Yield out of 2 years (48.5%).
- A full Good-Good sequence (32.4%) is significantly more likely than a full Poor-Poor outcome (19.1%).
- This pattern suggests a healthy mix of fluctuations with an overall tendency towards at least one productive year in every two-year cycle.

---

4. Statistical Measures

Measure	**1-Year**	**2-Year**
Mean (μ)	0.5750	1.1325
Variance (σ^2)	0.2444	0.4972
Skewness	-0.3034	-0.1924
Kurtosis	-1.9080	2.0138
CV (%)	85.97%	62.26%

One-Year Sequence:

- The mean yield occurrence is 0.575, indicating a slight tilt towards Good Yield.
- The negative skewness implies a slightly left-tailed distribution, meaning Poor Yield values are a bit more extreme or "long-tailed".

- Kurtosis is negative, indicating a flatter-than-normal distribution (platykurtic), which suggests less frequent extreme values.

Two-Year Sequence:

- Mean yield count increases to 1.13, showing that good performance tends to persist across years.
- Variance nearly doubles, but CV decreases, implying relative yield stability improves when considered over a longer period.
- Kurtosis > 2 suggests a more normal-like peaked distribution for 2-year outcomes.

Considering Mean as the standard level of bipartition of humid states:

1. Transition and Emission Behaviour

The estimated Transition Probability Matrix (TPM) shows that a High humidity state has a 25% probability of persisting into another High humidity state and a 75% probability of transitioning to a Low humidity state. Conversely, a Low humidity state has a 60% chance of switching to High humidity and a 40% probability of remaining Low. These transition probabilities suggest that humidity conditions are quite unstable over consecutive years, with frequent transitions between High and Low humidity. Such volatility in humidity levels may significantly influence the consistency of wheat yields across seasons.

	H (next)	**L (next)**
H (current)	0.25	0.75
L (current)	0.60	0.40

The Emission Probability Matrix (EPM), reflecting the likelihood of observing a particular yield given the humidity state, highlights a strong relationship between humidity and wheat yield. Under Low humidity conditions, there is an 80% probability of achieving a Good yield and only a 20% chance of a Poor yield. In contrast, High humidity conditions are associated with a 75% probability of Poor

yield and only a 25% chance of Good yield. These findings affirm the agricultural understanding that low humidity, especially during the crucial growing and maturity stages of wheat, is favorable for crop productivity, whereas high humidity tends to adversely affect yield quality and quantity.

	Good Yield (G)	**Poor Yield (P)**
H	0.25	0.75
L	0.80	0.20

---

2. One-Year Yield Prediction

The one-year yield predictions indicate that there is approximately a 52.15% chance of achieving a Good yield in any given year. This probability is relatively balanced, showing that while favorable conditions are slightly more probable, the possibility of poor outcomes remains considerable. This moderate leaning towards Good yield outcomes reflects the influence of the unstable humidity patterns revealed in the TPM.

$$P(Good\ Yield) = \pi_L \cdot P(L \to L) \cdot P(\,G \mid L\,) + \pi_H \cdot P(H \to L) \cdot P(\,G \mid L\,) +$$
$$\pi_L \cdot P(L \to H) \cdot P(\,G \mid H\,) + \pi_H \cdot P(H \to H) \cdot P(\,G \mid H\,)$$

Substituting values:

$$P(Good\ Yield) = 0.5215$$

Thus, $P(Poor\ Yield) = 1 - 0.5215 = 0.4785$.

---

3. Two-Year Yield Prediction

The two-year yield sequence predictions provide a deeper perspective on crop stability over time. The probability of achieving at least one Good yield across two consecutive years is approximately 83.52%, derived by complementing the probability of observing zero Good yields (16.49%). This high probability suggests that even if a poor yield year occurs, the likelihood of compensation

through a good yield in the following year is very strong. It portrays resilience in wheat production over a two-year cycle, mitigating the risk associated with annual climatic volatility.

- Probability of 0 Good Yields (α) = 0.16485: Both years yield Poor outcomes.
- Probability of 1 Good Yield (β) = 0.5433: One year yields Good, the other Poor.
- Probability of 2 Good Yields (γ) = 0.29185: Both years yield Good.

---

4. Statistical Measures

Measure	**1-Year**	**2-Year**
Mean (μ)	0.5215	1.1270
Variance (σ^2)	0.2495	0.4406
Skewness	-0.0861	-0.1467
Kurtosis	-1.9926	2.2442
CV (%)	95.79%	58.90 %

The statistical measures further elucidate the yield dynamics. For the one-year sequence, the mean number of Good yields is 0.5215, with a high variance of 0.2495. The coefficient of variation (CV) is extremely high at 95.79%, indicating a high relative uncertainty in yield outcomes when considering only a single year. Skewness is slightly negative, and kurtosis is highly negative, suggesting a flatter, wider spread distribution of yield outcomes with some left-sided deviation towards poorer results.

However, when extended to the two-year sequence, the mean Good yield occurrence rises to 1.127, and the variance slightly improves relative to the scale. Notably, the CV drops significantly to 58.90%, reflecting greater yield stability across two years. Skewness remains mildly negative, and kurtosis increases to 2.24, approaching the shape of a normal distribution. These changes imply that over longer planning horizons, yield outcomes become more predictable and less

volatile, providing a more reliable basis for strategic agricultural decision-making.

In conclusion, the HMM analysis demonstrates that while humidity conditions fluctuate significantly from year to year, low humidity remains a strong positive predictor of good wheat yields. Planning over two-year periods offers farmers a better opportunity to offset the risk of poor seasons, with the model suggesting that good yield outcomes dominate when considered over multiple years. Such findings are crucial for informing irrigation planning, crop insurance programs, and long-term agricultural strategies aimed at improving wheat production stability under changing climatic conditions.

INTERPRETATION

- The interpretation of the results from this study, titled "Application of 2-State Hidden Markov Model for Predicting Wheat Yield Based on Humidity Conditions," provides critical insights into the stochastic relationship between humidity and wheat yield, as modelled by a Two-State Hidden Markov Model (HMM). The analysis reveals that Low humidity states significantly favour Good yields (75% probability *(with mean as standard level)*, $P(G \mid L) = 0.75$) and (80% probability*(with median as standard level)*, $P(G \mid L) = 0.80$) , while High humidity states increase the risk of Poor yields (60% probability *(with mean as standard level)*, $P(P \mid H) = 0.60$) and (75% probability *(with median as standard level)*, $P(P \mid H) = 0.75$). This finding aligns with agricultural science, which indicates that excessive humidity during wheat's reproductive and grain-filling stages promotes fungal diseases like Fusarium head blight and rust, compromising yield quality and quantity. Conversely, Low humidity reduces disease pressure and supports optimal grain development, leading to higher yields. The one-year prediction shows a 57.5% chance of a Good yield ($P(Good\ Yield) = 0.575$) and a 42.5% chance of a Poor yield $(with\ mean\ as\ standard\ level)$ and $(P(Good\ Yield) = 0.5215$) and a 47.85% chance of a Poor yield *(with median as standard level)* underscores the need for adaptive management strategies to mitigate humidity-related risks and .
- The two-year prediction further highlights the resilience of wheat production, with an 80.89% probability of achieving at least one Good yield $(P(1\ Good) = 0.4853 + P(2\ Good) = 0.3236)$ *(with mean as standard level)* and an 83.52% probability of achieving at least one Good yield $(P(1\ Good) = 0.5433 + P(2\ Good) = 0.29185)$ *(with mean as standard level)* . This suggests that, despite short-term fluctuations, wheat cultivation can maintain productivity over a medium-term horizon, provided farmers implement targeted interventions during high humidity periods. The lower coefficient of variation (CV) for two-year predictions (62.26% vs. 85.97% for one-year *(with mean as standard level)* and 95.79% vs. 58.90% for one year *(with median as standard level)*) indicates greater yield stability over longer periods, likely due to the averaging effect of climatic variability. These results offer a probabilistic framework for anticipating yield outcomes, enabling stakeholders to plan with greater

confidence in regions prone to humidity fluctuations, such as monsoon-influenced areas in India.

- **Broader Implications**: The findings have significant implications for agricultural stakeholders. For farmers, the model's predictions can guide irrigation planning, encouraging reduced water application during High humidity periods to prevent waterlogging and disease proliferation. During Low humidity, farmers can optimize irrigation to maintain soil moisture, capitalizing on the higher likelihood of Good yields. Pest management strategies can also be tailored, with increased fungicide applications during predicted High humidity to combat fungal risks. Policymakers can leverage the model's risk assessments—particularly the 19.11% chance of consecutive Poor yields (P(0 Good)=0.1911 *(with mean as standard level)*) and particularly the 16.49% chance of consecutive Poor yields (P(0 Good)=0.16485 *(with median as standard level)*) —to develop targeted crop insurance models, offering premiums adjusted to humidity-driven risks. Additionally, government agencies can use these insights to allocate subsidies for climate-resilient wheat varieties or support infrastructure like drainage systems in high-humidity regions, enhancing food security.
- **Comparison with Other Models**: The HMM's ability to incorporate probabilistic transitions and emissions makes it particularly suited for agricultural systems, where environmental factors exhibit complex, non-linear effects. Comparisons with literature, such as Zucchini et al. (2016), further validate the HMM's applicability, as similar models have successfully predicted crop outcomes under variable climates.
- **Additional Practical Applications**: Beyond irrigation and insurance, the model can support precision agriculture by integrating real-time sensor data to refine predictions, enabling dynamic adjustments to farming practices. For example, IoT-enabled weather stations could feed humidity data into the HMM, providing daily yield forecasts. The model's probabilistic outputs can also inform supply chain management, helping traders anticipate wheat availability and stabilize market prices. Educational programs can use these findings to train farmers on climate-adaptive techniques, fostering resilience in vulnerable regions.
- **Future Directions**: The study lays the groundwork for further research by suggesting the incorporation of additional hidden states, such as temperature and soil moisture, to create a multi-state HMM. This would provide a more holistic model, though it requires larger datasets and increased computational resources. Extending the framework to other

crops, such as rice or maize, could broaden its impact, addressing diverse agricultural systems. Exploring machine learning techniques, like hybrid HMM-neural network models, could enhance predictive accuracy by capturing non-linear patterns. Regional case studies, particularly in high-humidity areas like Punjab, India, could validate the model's applicability across different agro-climatic zones.

In conclusion, the interpretation of the HMM results highlight the critical role of humidity in wheat yield variability, offering actionable insights for farmers and policymakers. The model's superior performance, practical applications, and potential for expansion position it as a valuable tool for climate-resilient agriculture, contributing to sustainable food production in a changing climate.

Appendix

The codes performed for the project are as follows:

```
wt=read.csv("C:\\Users\\Kaustav khelowary\\Documents\\Time Series Project\\DATA SETS\\ts wheat.csv")

attach(wt)


df1=cbind(Soil_pH,Temperature,Humidity,Wind_Speed,N,P,K,Soil_Quality,Crop_Yield)


boxplot(df1)


cm=cor(df1)


summary(wt$Soil_pH)

summary(wt$Temperature)

summary(wt$Humidity)

summary(wt$Wind_Speed)

summary(wt$N)

summary(wt$P)

summary(wt$K)

summary(wt$Soil_Quality)

summary(wt$Crop_Yield)

boxplot(wt$Soil_pH)


cm
```

```
library(corrplot)
corrplot(cm,method = "square",type = "upper",addCoef.col = "black",tl.col = "black",tl.srt = 45)
library(dplyr)

# Load the dataset
wheat_clean <- read.csv("C:\\Users\\user\\Downloads\\Book3;wheat.csv", stringsAsFactors = FALSE)
#for path: wheat <-file.choose()
# Split hidden_state_run into previous and current hidden states
wheat_clean <- wheat_clean %>%
  mutate(
    prev_humidity = substr(hidden_state_run, 1, 1),
    curr_humidity = substr(hidden_state_run, 3, 3)  # 3rd char is current
  )

# Create transition table for humidity
humidity_transitions <- table(wheat_clean$prev_humidity, wheat_clean$curr_humidity)

# Convert to TPM
TPM <- prop.table(humidity_transitions, margin =1)
TPM_new <- TPM[-1,-1]
TPM_new

# Extract current humidity and yield categories
wheat_clean <- wheat_clean %>%
  mutate(
```

```
  curr_yield = substr(yield_state_run, 3, 3)  # G or P
 )

# Create emission table
emission_table <- table(wheat_clean$prev_humidity, wheat_clean$curr_yield)

# Convert to emission probabilities
EPM <- prop.table(emission_table, margin = 1)
EPM_new <- EPM[-1,-1]
EPM_new
# Extract initial humidity state
initial_states <- substr(wheat_clean$Humid_state, 1, 1)

# Count frequency of L and H
IPV_table <- table(initial_states)

# Convert to proportions (initial probabilities)
IPV <- prop.table(IPV_table)
IPV

# Initial Probabilities
pi_L <- IPV["L"]
pi_H <- IPV["H"]

# Emission Probabilities
p_GL <- EPM["L", "G"]
p_PH <- EPM["H", "P"]
```

```
p_PL <- EPM["L", "P"]
p_GH <- EPM["H", "G"]

# TPM entries
p_LL <- TPM_new["L", "L"]
p_LH <- TPM_new["L", "H"]
p_HL <- TPM_new["H", "L"]
p_HH <- TPM_new["H", "H"]

# 1-year predictions
P_G_1day <- pi_L * p_LL * p_GL + pi_H * p_LH * p_GL + pi_L * p_HL * p_GH + pi_H * p_HH * p_GH
P_P_1day <- pi_L * p_LL * p_PL + pi_H * p_LH * p_PL + pi_L * p_HL * p_PH + pi_H * p_HH * p_PH

cat("Probability of Good Yield (1-year):", P_G_1day, "\n")
cat("Probability of Poor Yield (1-year):", P_P_1day, "\n")

#for 2-years predictions
# α = probability of Poor Yield in both years
alpha <- pi_L * p_LL * p_PL * p_PL +
  pi_L * p_LH * p_PL * p_PH +
  pi_H * p_HL * p_PH * p_PL +
  pi_H * p_HH * p_PH * p_PH

# β = probability of one Good Yield, one Poor
beta <- pi_L * p_LL * (p_PL * p_GL + p_GL * p_PL) +
  pi_L * p_LH * (p_PL * p_GH + p_GL * p_PH) +
```

```
  pi_H * p_HL * (p_PH * p_GL + p_GH * p_PL) +
  pi_H * p_HH * (p_PH * p_GH + p_GH * p_PH)

# γ = probability of Good Yield in both years
gamma <- pi_L * p_LL * p_GL * p_GL +
  pi_L * p_LH * p_GL * p_GH +
  pi_H * p_HL * p_GH * p_GL +
  pi_H * p_HH * p_GH * p_GH

cat("Probability of 0 Good Yields (2-year):", alpha, "\n")
cat("Probability of 1 Good Yield (2-year):", beta, "\n")
cat("Probability of 2 Good Yields (2-year):", gamma, "\n")

#For 1-length sequence (for 1 year prediction)
# Mean
mu1 <- P_G_1day

# Variance
var1 <- mu1 * (1 - mu1)

# Third central moment
mu3_1 <- (1 - mu1) * mu1 * (1 - 2 * mu1)

# Skewness
skew1 <- mu3_1 / (sqrt(var1)^3)

# Kurtosis
```

```
kurt1 <- (1 - 6 * mu1 * (1 - mu1)) / var1

# Coefficient of Variation
cv1 <- (sqrt(var1) / mu1) * 100

#For 2-length sequence(for 2 years prediction)
# Mean
mu2 <- beta * 1 + gamma * 2

# Variance
var2 <- alpha * (0 - mu2)^2 + beta * (1 - mu2)^2 + gamma * (2 - mu2)^2

# Third central moment
mu3_2 <- alpha * (0 - mu2)^3 + beta * (1 - mu2)^3 + gamma * (2 - mu2)^3

# Skewness
skew2 <- mu3_2 / (sqrt(var2)^3)

# Kurtosis
kurt2 <- (alpha * (0 - mu2)^4 + beta * (1 - mu2)^4 + gamma * (2 - mu2)^4) /
(var2^2)

# Coefficient of Variation
cv2 <- (sqrt(var2) / mu2) * 100

stats <- data.frame(
  Sequence = c("1-Year", "2-Year"),
  Mean = c(mu1, mu2),
```

```
  Variance = c(var1, var2),
  Third_Central_Moment = c(mu3_1, mu3_2),
  Skewness = c(skew1, skew2),
  Kurtosis = c(kurt1, kurt2),
  CV = c(cv1, cv2)
)

print(stats)
#for median
library(dplyr)

# Load the dataset
wheat_clean <- read.csv("C:\\Users\\user\\Downloads\\Book4;wheat.csv", stringsAsFactors = FALSE)
#for path: wheat <-file.choose()
# Split hidden_state_run into previous and current hidden states
wheat_clean <- wheat_clean %>%
  mutate(
    prev_humidity = substr(hidden_state_run, 1, 1),
    curr_humidity = substr(hidden_state_run, 3, 3)  # 3rd char is current
  )

# Create transition table for humidity
humidity_transitions <- table(wheat_clean$prev_humidity, wheat_clean$curr_humidity)

# Convert to TPM
TPM <- prop.table(humidity_transitions, margin =1)
```

```
TPM_new <- TPM[-1,-1]
TPM_new

# Extract current humidity and yield categories
wheat_clean <- wheat_clean %>%
  mutate(
    curr_yield = substr(yield_state_run, 3, 3)  # G or P
  )

# Create emission table
emission_table <- table(wheat_clean$prev_humidity, wheat_clean$curr_yield)

# Convert to emission probabilities
EPM <- prop.table(emission_table, margin = 1)
EPM_new <- EPM[-1,-1]
EPM_new
# Extract initial humidity state
initial_states <- substr(wheat_clean$Humid_state, 1, 1)

# Count frequency of L and H
IPV_table <- table(initial_states)

# Convert to proportions (initial probabilities)
IPV <- prop.table(IPV_table)
IPV

# Initial Probabilities
```

```
pi_L <- IPV["L"]
pi_H <- IPV["H"]

# Emission Probabilities
p_GL <- EPM["L", "G"]
p_PH <- EPM["H", "P"]
p_PL <- EPM["L", "P"]
p_GH <- EPM["H", "G"]

# TPM entries
p_LL <- TPM_new["L", "L"]
p_LH <- TPM_new["L", "H"]
p_HL <- TPM_new["H", "L"]
p_HH <- TPM_new["H", "H"]

# 1-year predictions
P_G_1day <- pi_L * p_LL * p_GL + pi_H * p_LH * p_GL + pi_L * p_HL * p_GH + pi_H * p_HH * p_GH
P_P_1day <- pi_L * p_LL * p_PL + pi_H * p_LH * p_PL + pi_L * p_HL * p_PH + pi_H * p_HH * p_PH

cat("Probability of Good Yield (1-year):", P_G_1day, "\n")
cat("Probability of Poor Yield (1-year):", P_P_1day, "\n")

#for 2-years predictions
# α = probability of Poor Yield in both years
alpha <- pi_L * p_LL * p_PL * p_PL +
  pi_L * p_LH * p_PL * p_PH +
```

```
  pi_H * p_HL * p_PH * p_PL +
  pi_H * p_HH * p_PH * p_PH

# β = probability of one Good Yield, one Poor
beta <- pi_L * p_LL * (p_PL * p_GL + p_GL * p_PL) +
  pi_L * p_LH * (p_PL * p_GH + p_GL * p_PH) +
  pi_H * p_HL * (p_PH * p_GL + p_GH * p_PL) +
  pi_H * p_HH * (p_PH * p_GH + p_GH * p_PH)

# γ = probability of Good Yield in both years
gamma <- pi_L * p_LL * p_GL * p_GL +
  pi_L * p_LH * p_GL * p_GH +
  pi_H * p_HL * p_GH * p_GL +
  pi_H * p_HH * p_GH * p_GH

cat("Probability of 0 Good Yields (2-year):", alpha, "\n")
cat("Probability of 1 Good Yield (2-year):", beta, "\n")
cat("Probability of 2 Good Yields (2-year):", gamma, "\n")

#For 1-length sequence (for 1 year prediction)
# Mean
mu1 <- P_G_1day

# Variance
var1 <- mu1 * (1 - mu1)

# Third central moment
```

```
mu3_1 <- (1 - mu1) * mu1 * (1 - 2 * mu1)

# Skewness
skew1 <- mu3_1 / (sqrt(var1)^3)

# Kurtosis
kurt1 <- (1 - 6 * mu1 * (1 - mu1)) / var1

# Coefficient of Variation
cv1 <- (sqrt(var1) / mu1) * 100

#For 2-length sequence(for 2 years prediction)
# Mean
mu2 <- beta * 1 + gamma * 2

# Variance
var2 <- alpha * (0 - mu2)^2 + beta * (1 - mu2)^2 + gamma * (2 - mu2)^2

# Third central moment
mu3_2 <- alpha * (0 - mu2)^3 + beta * (1 - mu2)^3 + gamma * (2 - mu2)^3

# Skewness
skew2 <- mu3_2 / (sqrt(var2)^3)

# Kurtosis
kurt2 <- (alpha * (0 - mu2)^4 + beta * (1 - mu2)^4 + gamma * (2 - mu2)^4) /
(var2^2)
```

```
# Coefficient of Variation
cv2 <- (sqrt(var2) / mu2) * 100

stats <- data.frame(
  Sequence = c("1-Year", "2-Year"),
  Mean = c(mu1, mu2),
  Variance = c(var1, var2),
  Third_Central_Moment = c(mu3_1, mu3_2),
  Skewness = c(skew1, skew2),
  Kurtosis = c(kurt1, kurt2),
  CV = c(cv1, cv2)
)

print(stats)
```

REFFERENCES

- Hidden Markov Models for Time Series by Walter Zucchini, Iain L. MacDonald, Ronald Langlock.
- Hidden markov models for stock analysis, Gulbadin Farooq Dar, Tirupati Rao,P
- William W.S. Wei (2006): Time Series Analysis – Univariate and Multivariate Methods 2nd Edition, Pearson Education Inc.