SeasonWatch Project Report

Public Interest Technology-New England (PIT-NE) Fellowhip 2024

Zach Meurer, Shivansh Soni, Jacob Epstein

June 28, 2024

Abstract

Citizen science efforts like SeasonWatch provide researchers valuable citizen observations, which can be used to study the effect of climate change on trees. Previous work has been done to create a reference database to verify SeasonWatch's citizen observations, however, the reference data were collected by hand and were from years ranging between 1950 and 2019. In this project, we created an automated pipeline to select representative citizen observations as reference data using isolation trees for anomaly detection. Furthermore, we developed a novel probabilistic method to identify the start and end weeks of tree phenophases from SeasonWatch's citizen observations, and generated a dataset of phenophase transition times. Our reference data and phenophase transition time datasets can serve as baselines for verifying new citizen observations collected by SeasonWatch. We hope that our work will improve researchers' understanding of when phenophases occur, and pave the way for a year-by-year analysis of how tree phenophases shift, allowing the effects of climate change to be studied.

Contents

1 Overview of the project

1.1 Introduction

Over the course of the past six weeks, our team has worked with SeasonWatch's citizen and reference tree phenology data to complete a wide array of tasks, such as data cleaning, data visualization, machine learning, and statistical inference. These processes have led to the discovery of trends, flaws, and intriguing behaviors within the citizen data.

This project relied entirely on Python, utilizing industry standard tools and libraries to extrapolate meaningful conclusions from data. To do so, our team first had to clean SeasonWatch's citizen and reference data. Following this cleaning process, machine learning methods were developed to detect and filter out outlying and anomalous citizen observations.

Following this citizen data validation process, our team produced a model to select representative citizen observations to use as up-to-date reference data. Additionally, our team explored mathematical models to infer annual phenophase start and end times. Throughout development, numerous plots have been produced offering unique insights into both SeasonWatch's citizen and reference data. Our team's applications have set up a foundation for understanding how climate change can shift phenophase timeframes.

1.2 Main Project Tasks

This is a brief overview of the main project tasks:

- Clean the citizen and reference data
  - Reformatted the reference data to have the same structure as the citizen data for consistency and ease of comparison.
  - Organized citizen data by week and year to be consistent with the reference data (48 weeks per year).
  - Identified and fixed incorrectly labeled -2 values in the citizen dataset.
  - Identified the state associated with citizen observations missing the state name phenophase (Null/NA) by using an observation's longitude and latitude coordinates.
  - Removed any observations with missing phenophase values (Null/NA) in the citizen dataset.
  - Made the species names consistent between citizen and reference data following the format of 'common name-scientific name'.
- Visualized the citizen and reference data (Time-scale: Weekly)
  - Bar plots showing discrepancies in phenophase transitions between citizen and reference data over time.
  - Line plots depicting percentage presence of phenophases in citizen data compared to transitions in the reference data over time.
  - Line plots visualizing the difference between the average citizen data and the reference data of phenophases over time.
  - Line plots depicting percentage presence of phenophases in citizen data compared to bar plots depicting the number of observations over time.

  - Line plots depicting percentage presence of phenophases in citizen data compared to up-to-date selected reference data.

- Develop a data validation system for citizen observations
  - Used Isolation Forests as an anomaly detection algorithm to filter out outlier observations
  - Used K-Means clustering to select representative observations from citizen data to create up-to-date reference data.
  - Used Principal Component Analysis for visualizing the results of each of these methods.

- Generate a dataset of mean and standard deviation of transitions of phenophases
  - Created a probability distribution to approximate the most likely week where a phenophase transition occurrs.
  - Used this probability distribution to generate a dataset of mean and standard deviations of phenophase transition times.

1.3 Project Pipeline

Shown below is the complete project pipeline, detailing precisely the steps we have taken to obtain our results starting from the citizen and reference data.

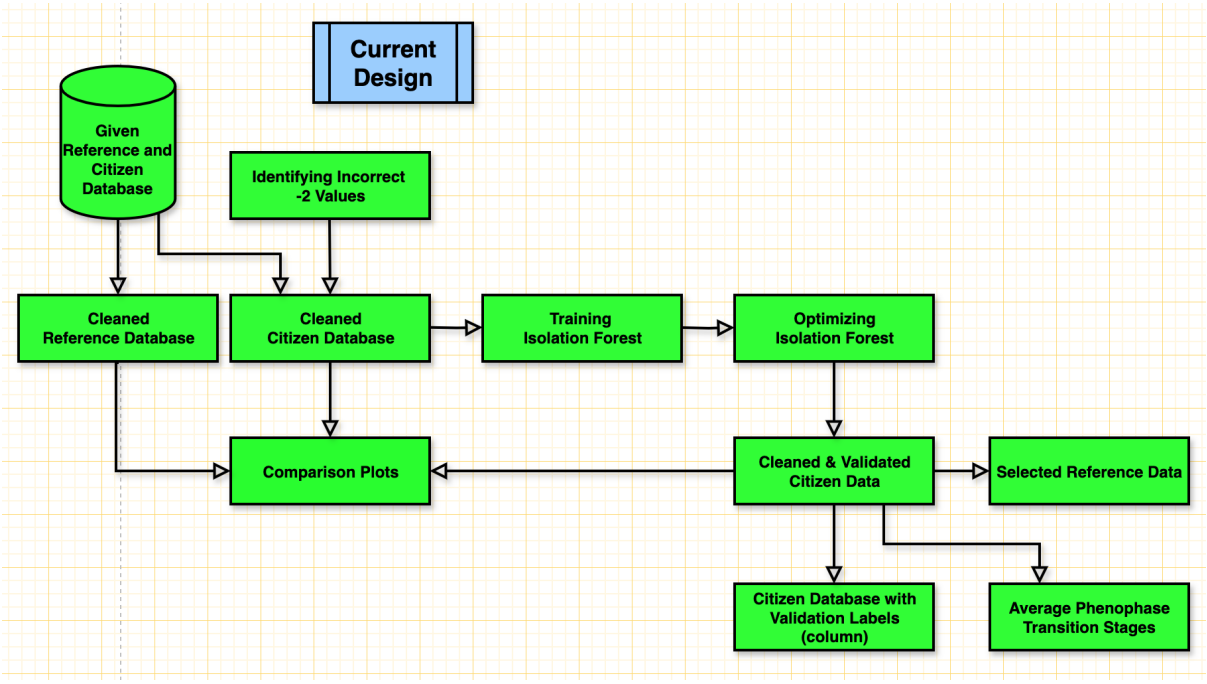

Figure 1: The project pipeline

We started off by cleaning the reference and citizen database. A major component of cleaning the citizen database was identifying and correcting false positive and false negative -2 values which took a lot of manual effort. Then on the cleaned data, we trained and optimized the Isolation forest algorithm for outlier detection to remove all unwanted datapoints. We used this

cleaned and validated data to generate plots, and then used clustering on this data to select representative citizen observations as reference data. We also used the cleaned citizen observations to generate an average phenophase transition stages dataset.

1.4 Dataset of Citizen Observations

Column	Meaning	Example
Species_name	Common & scientific name of tree species	African Tulip-Spathodea campanulata
State_name	Indian state in which observation was recorded	Jammu and Kashmir
Coordinates (Lat, Long)	2 columns for latitude and longitude of observation	(9.38520, 76.58480)
Phenophases (Flower buds, ripe fruits, etc.)	10 columns for different tree phenophases: Fresh leaves, mature leaves, old leaves, flower buds, open flowers, male flowers, female flowers, unripe fruits, ripe fruits, open fruits. Consists of values -2, -1, 0, 1, & 2 to indicate does not appear in species, user doesn't know, none, few, and many respectively	2.0
Year	Year observation was recorded (range 2014-2023)	2014
Week	Week observation was recorded (range 0-47)	23

Figure 2: Table describing the dataset

The main data that we were dealing with was SeasonWatch's dataset of citizen observations. It consists of species names, their date of observation, latitude and longitude coordinates of location, and each of their 10 phenophases. In total, the dataset had around 592,000 observations. The data on the phenophases was given in a categorical form. Each entry was either:

- 2 - Many observed
- 1 - Few observed
- 0 - None observed
- -1 - Don't know
- -2 - Phenophase not seen in this species

However, the data was not immediately ready for visualization and machine learning. It needed to be cleaned, as it contained problems like:

- A large number observations with null entries
- Mislabeled -2 entries
- Latitude and longitude coordinates that lie outside of India
- Inconsistent species names across data

Therefore, we designed a data cleaning pipeline where we dealt with all of these issues to clean the data.

2 Methods

In this section we will go into more detail about the machine learning algorithms we used, why we used them, and how they work.

2.1 Isolation Forests

Isolation Forests are classical machine learning algorithms used primarily for anomaly detection tasks. They filter out anomalous data points and preserve similar data.

The principle these work on is based on picking a random dimension and splitting the data along that dimension. Each isolation tree does this process randomly, and an isolation forest is made of several such trees. Take the following plot for example:

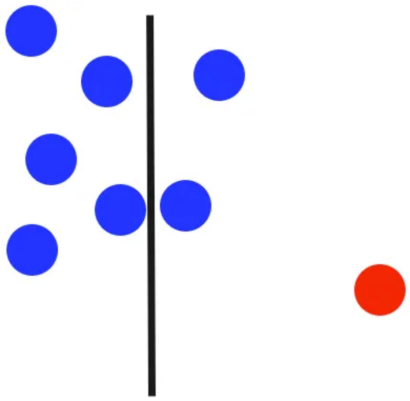

Figure 3: Red depicts the anomalous data point. There is a split done along a dimension separating the two sets of data points. Other trees may split across a different dimension. The two resulting subspaces create their own tree.

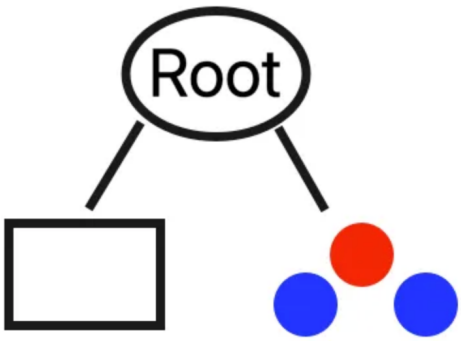

Figure 4: The tree after splitting

This process of splitting continues until every data point of the graph is a leaf node of the tree. The logic with filtering out anomalous data points is that they will be filtered out sooner than non-anomalies, since anomalous data points are generally farther away. So the leaf nodes closer to the root are treated as anomalies.

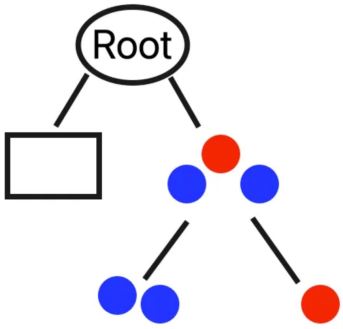

Figure 5: The tree after another splitting step.

In this example, we have successfully isolated an anomalous data point. We use this anomaly detection algorithm in our project to filter out outlier data points.

2.2 K-Means Clustering

K-Means Clustering is an algorithm that groups similar data points into clusters. It divides the data points into K distinct, non-overlapping subgroups by following these steps:

1. **Initialization**: Randomly select K data points as initial *centroids* or the centers of the clusters.
2. **Assignment**: Assign each data point to the nearest centroid, forming K clusters.
3. **Update**: Calculate the mean of the data points in each cluster to update the centroids.
4. **Iteration**: Repeat the assignment and update steps until the centroids no longer change positions, indicating convergence.

The algorithm stops when the centroids stabilize and no longer move, meaning the clusters have been defined. In this project, we used K-Means clustering to extract reference data points from a cluster of filtered, validated data. Once our isolation forest has given us the validated data, we can then use K-Means clustering with $K = 1$ to find a centroid in the data. We do that since the centroid is the 'average' data point, and then we find the 3 closest data points to the centroid and use their median as the reference data point.

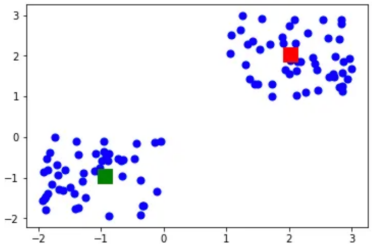

Figure 6: Plot showing clusters and centroids

2.3 Detecting Anomalous Observations

To detect anomalies in citizen data, we used the following process:

1. Preprocess the citizen observations by removing all fields except categorical phenophases (Leaves_fresh, Leaves_mature, Leaves_old, Flowers_bud, Flowers_open, Flowers_male, Flowers_Female, Fruits_unripe, Fruits_ripe, Fruits_open), and geographical information (Latitude, Longitude)

2. Filter citizen observations to only those of the given species during the given week

3. Using an ensemble of isolation trees, identify anomalies in the data. Through experimentation, we found that using an ensemble with 500 isolation trees, and a contamination index. of 0.8 worked best. The contamination index controls the sensitivity of the Isolation Tree model to outliers.

2.4 Selecting Reference Data

For a given species, phenophase, and year, we select a representative citizen observation of the species and phenophase for each week of the given year to serve as reference data.

To select representative citizen observations for a given week, we follow a 5-step process:

1. Preprocess the citizen observations by removing all fields except categorical phenophases (Leaves_fresh, Leaves_mature, Leaves_old, Flowers_bud, Flowers_open, Flowers_male, Flowers_Female, Fruits_unripe, Fruits_ripe, Fruits_open), and geographical information (Latitude, Longitude)

2. Filter citizen observations to only those of the given species during the given week

3. Compute a centroid of the observations using K-means clustering, with $K = 1$.

4. Find the k closest citizen observations to the centroid with respect to the euclidean distance metric.

5. Compute the values of each of these observations for the phenophase in question (which are all either zero, one, or two) and take a median over all these values to serve as the reference data for that week, species, and phenophase

2.5 Transition Distributions

In order to extract phenophase transition times from graphs of citizen data, we tried multiple approaches, such as using derivatives to mark the start and end of fruiting/flowering seasons. However, due to the noise and irregularity of the data, none of these approaches worked consistently. Subsequently, we designed a score function that returns the probability that a transition occurred in a particular week. Since the start of a season is demarcated by a sudden spike in growth after a period of relative stagnation, the way this function works by design is to measure a stagnation and spike score for each week, and the week with the highest score is most likely to be the transition week. A plot depicting this distribution is provided below:

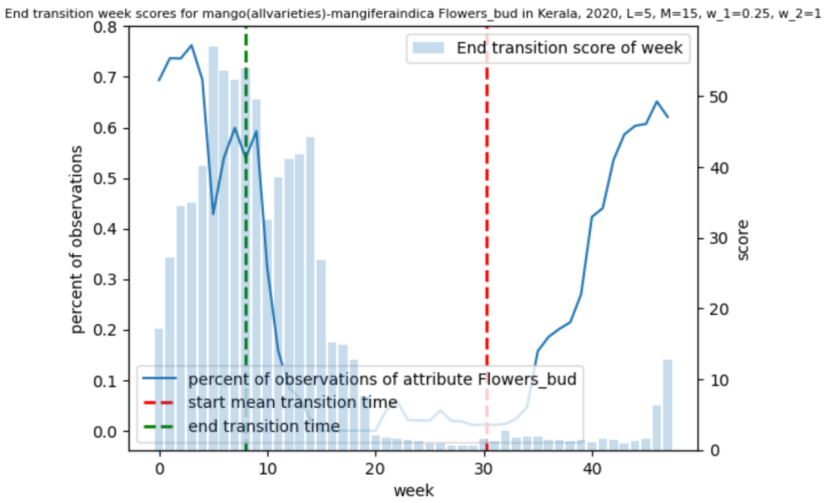

Figure 7: The line represents percent of observations, and the bars represent the score function - the likelihood of a week to be an end transition week.

The equation for the start score function is the following:

$$s_{start}(w) = w_1 \left(\frac{1}{0.1 + (\max_{w-L \le k \le w} y(k) - \min_{w-L \le k \le w} y(k))^2} \right) + w_2 \max(y(w+M) - y(w), 0)$$

and the equation for the end score function is the following:

$$s_{end}(w) = w_1 \left(\frac{1}{0.1 + (\max_{w \le k \le w+L} y(k) - \min_{w \le k \le w+L} y(k))^2} \right) + w_2 \max(y(w-M) - y(w), 0)$$

The left term of these functions measures the stagnation - how similar the observations have been in the recent weeks. The right measures the spike - how quickly the observations have grown in the recent weeks. Based on these, we can calculate the mean start and end times by normalizing these functions to turn them into a probability distribution:

$$p_{start}(w) = \frac{s_{start}(w)}{\sum_{w \in Y} s_{start}(w)}$$

$$p_{end}(w) = \frac{s_{end}(w)}{\sum_{w \in Y} s_{end}(w)}$$

From these probability distributions, we can calculate the mean and standard deviations through the following formulas:

$$\mu_{start/end} = \sum_{w=w_a}^{w_a+48} w \cdot p_{start/end}(w)$$

$$\sigma_{start/end} = \sqrt{\sum_{w=w_a}^{w_a+48} (w - \mu_{start/end})^2 \cdot p_{start/end}(w)}$$

where w_a is the start week of a 48 week window. In practice, we center our 48 week window over which means and standard deviations are computed around local maxima in the score function. So w_a ends up being 24 weeks before each local maximum.

2.5.1 Tuning Transition Distribution Parameters

The parameters to tune are L, the size of the interval of weeks over which the stagnation term is computed, M, a parameter controlling the number of weeks ahead we look when computing the spike term, w_1, which controls how much the stagnation is factored into s_{start} and s_{end}, and w_2, which controls how much the spike is factored into s_{start} and s_{end}.

Our best practice currently is to set $w_2 = 1$ and $w_1 = 0.25$, which reduces the amount stagnation factors into the score, ensuring that flat periods at the peaks of the percentage plot do not get high scores. We set $L = 5$, which we found to be a reasonable size of the interval over which stagnation is computed. Most significantly, we set $M = 15$, which we found to be a reasonable estimate of the number of weeks between the transition week and the peak week of the phenophase.

3 Results

3.1 Anomaly Detection

Using our method of anomaly detection, we found multiple anomalies in the citizen data. Many of these anomalies were geographic outliers, possibly caused by malfunctions of the Google Maps API, which we corrected for during data cleaning. The anomalies (or outliers) in the upper right and bottom left of Figure 8, a plot of Mango observations in the first week of 2023 projected into 2-D space using Principal Component Analysis, had significantly different latitudes and longitudes than the other points.

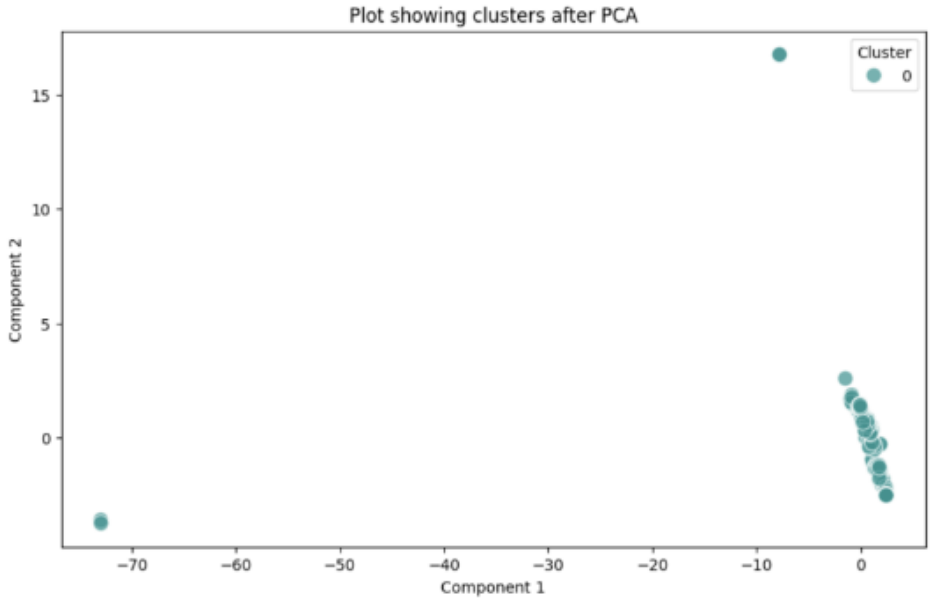

Figure 8: Example of anomalies in citizen data, Mango observations during first week of 2023

3.2 Selected Reference Data

We ran our automated method for selecting representative citizen observations to be reference data on the top 10 species in Kerala in terms of number of citizen observations available, which includes Mango, Jackfruit, and Coconut. We selected reference data for the year 2023 only.

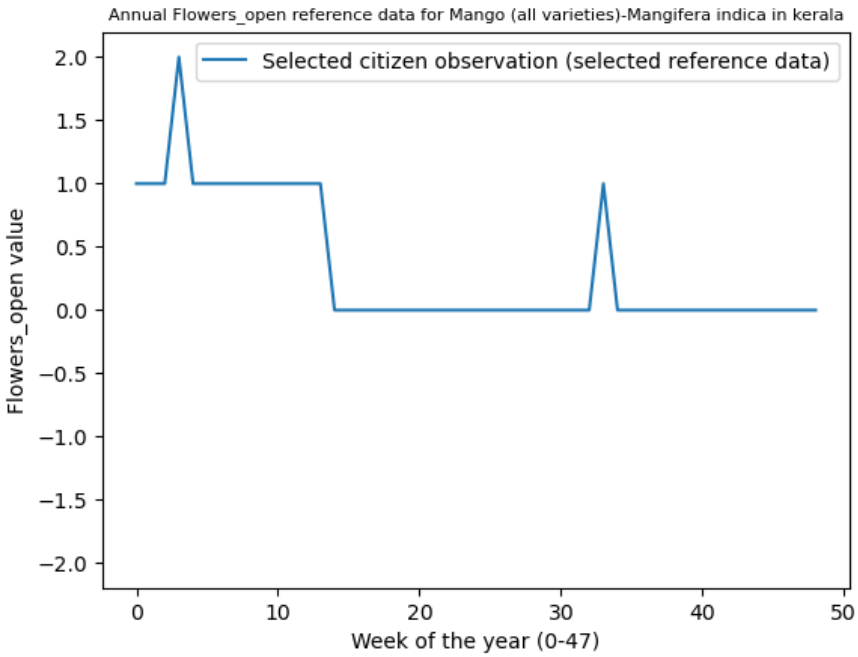

Figure 9: Selected Reference Data

Figure 7 depicts the reference data we selected for the Flowers open phenophase of Mango, in 2023. The x-axis represents weeks, and the y-axis represents the value of the reference data (either -2, 0, 1, or 2).

3.3 Comparing Selected Reference Data and Citizen Observations

Overall, the selected reference data and citizen data align well. Figure 8 compares the reference data for the budding phenophase of mangoes (in Kerala, 2023) with the percentage of observations of mangoes each week which observe the budding flowers phenophase (in Kerala, 2023). The reference data at values of 1 or 2 in weeks where this percentage exceeds 50 percent, and is zero in the majority of weeks of the year, save one week where this percentage is below 50 percent, save one week at the end of the year.

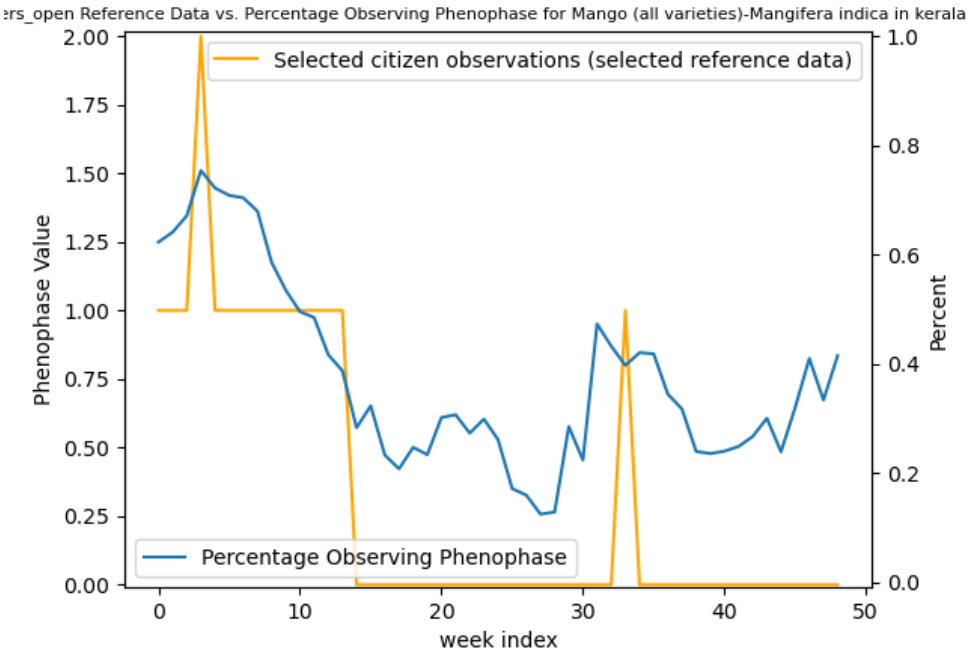

Figure 10: Selected Reference Data

3.4 Mean Phenophase Transition Times

Using our method of generating transition distributions from time-series data of the percentages of citizen observations of a species which observe a given phenophase, we created a dataset of mean phenophase transition times for the following species and phenophases in Kerala, for the years 2018 through 2022:

- **Mango**: Flowers_bud, Flowers_open, Fruits_unripe, Fruits_ripe
- **Jackfruit**: Flowers_bud, Flowers_male, Flowers_Female, Fruits_unripe, Fruits_ripe

Our method proved to be unreliable for 2023, as some species have phenophase windows that extend across the boundary of a year and may not have had a phenophase start or end time occuring in 2023.

Figure 11 shows s_{start} computed each week for Mangoes in 2022. Figure 12 shows p_{start}, which is created by normalizing the values of s_{start} in Figure 11 in a 48-week window which extends across year boundaries around the week of maximum s_{start}. Figure 13 shows p_{end}, the probability distribution of the end transition week. Figure 14 shows the mean transition times

and s_{start} for the phenophase Fruits_unripe for mangoes in Kerala for each of the approximately 300 weeks from 2018 through 2023.

Our pipeline for generating mean phenophase transition times in general produced reasonable results. The distributions p_{start} and p_{end} were often high variance, as in Figures 12 and 13, and the mean transition times, were occasionally innaccurate. For example, the end mean transition time in Figure 13 should be approximately 5 weeks further to the right, as it was around this time when citizens began to cease observing the phenophase Fruits_unripe in Mangoes in Kerala.

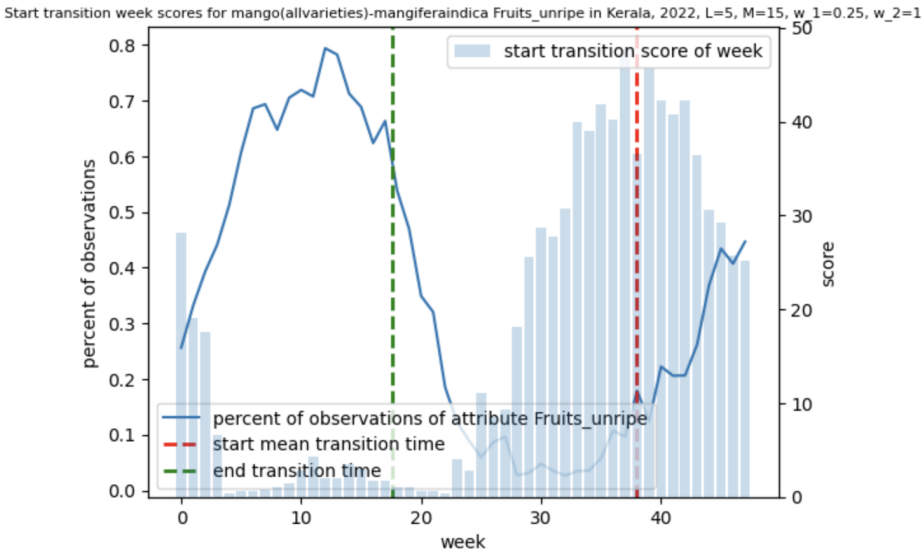

Figure 11: s_{start} for Fruits_unripe phenophase in mangoes, Kerala, 2022

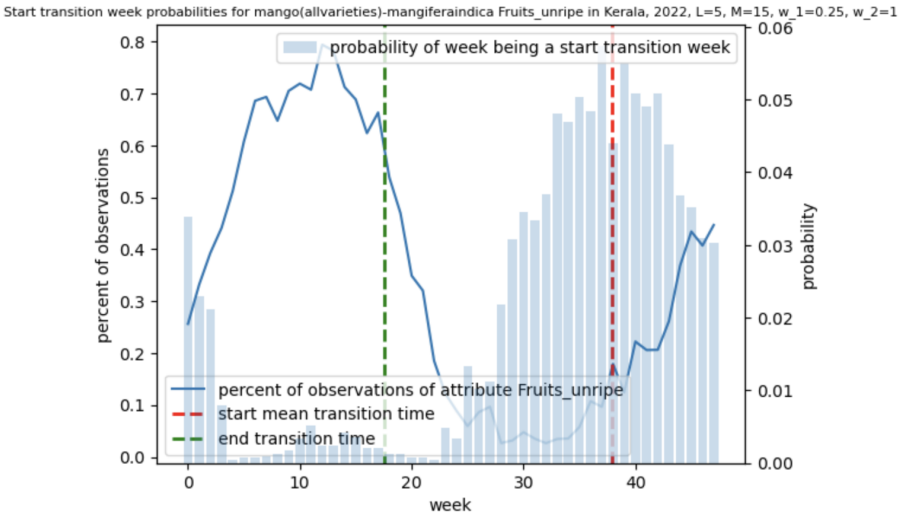

Figure 12: p_{start} for Fruits_unripe phenophase in mangoes, Kerala, 2022

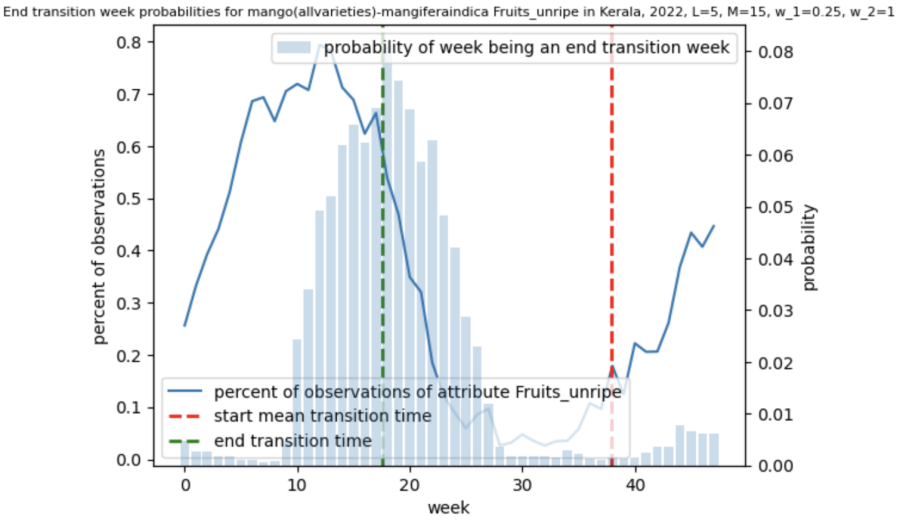

Figure 13: p_{end} for Fruits_unripe phenophase in mangoes, Kerala, 2022

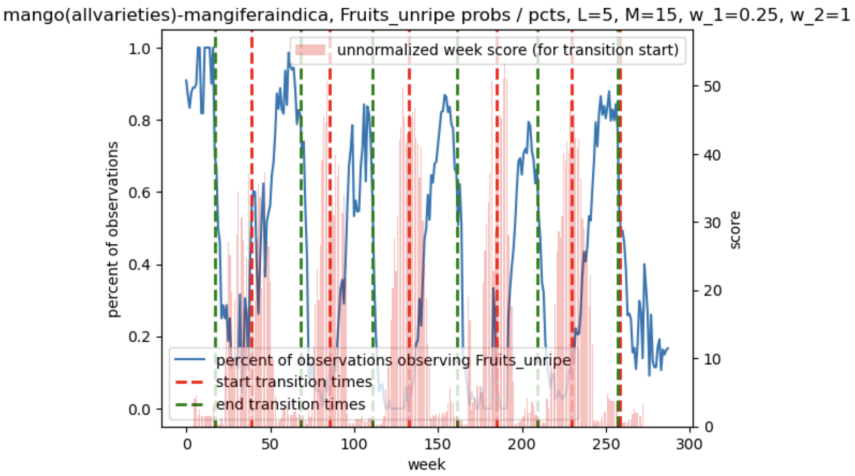

Figure 14: s_{start} for Fruits_unripe phenophase in mangoes, Kerala, years 2018-2023

4 Avenues for Further Work

After a thorough exploration of the data and understanding the project, we have identified important avenues where we would have been excited to do more work, had we more time.

4.1 Improving the Phenophase Transition Times Dataset

The score function works a lot of the time but its not perfect. For example, in the following graph:

The score function is improperly predicting the end transition to be a earlier than it actually is. This could be addressed by altering the score function.

Other avenues of work include:

- Altering the score function to directly encode a transition week as the first week we dip below a threshold percentage

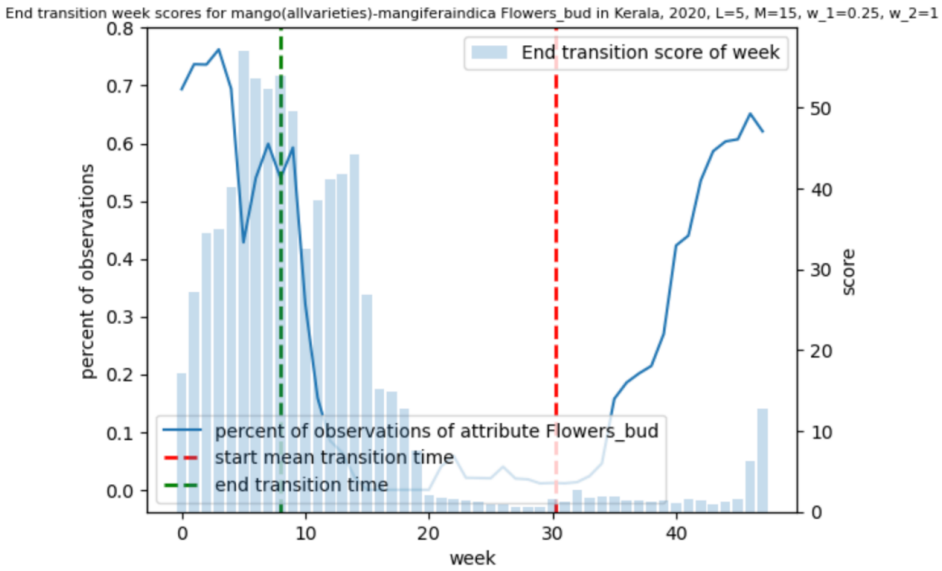

Figure 15: Plot showing imperfect end transition prediction

- Combining the distributions for starting and ending transition weeks to get a distribution over phenophases

4.2 More visualizations

Future visualizations could use climate data along with the citizen observation data to see how the change in climate is corresponding to the change in various phenophases of a species. Our angle was primarily to directly compare the reference and citizen data to notice shifts in phenophases over the years, so there is a lot more exploration that could be done here.

4.3 Smoothing the Synthetic Reference Data

The reference data matches closely with citizen observations for most species and phenophases. However, there are times where the synthetic reference data is noisy and can be smoothed out to give a better representation of the phenophase in that week. An example of this is shown below:

Note that there is a sudden dip in the reference data at week 17, which is most likely noise. Ideally, it should be flat around that week in order to most closely match citizen observations. This problem can be tackled with obtaining more citizen data, or by using more than 3 data points to calculate the median to select the reference data.

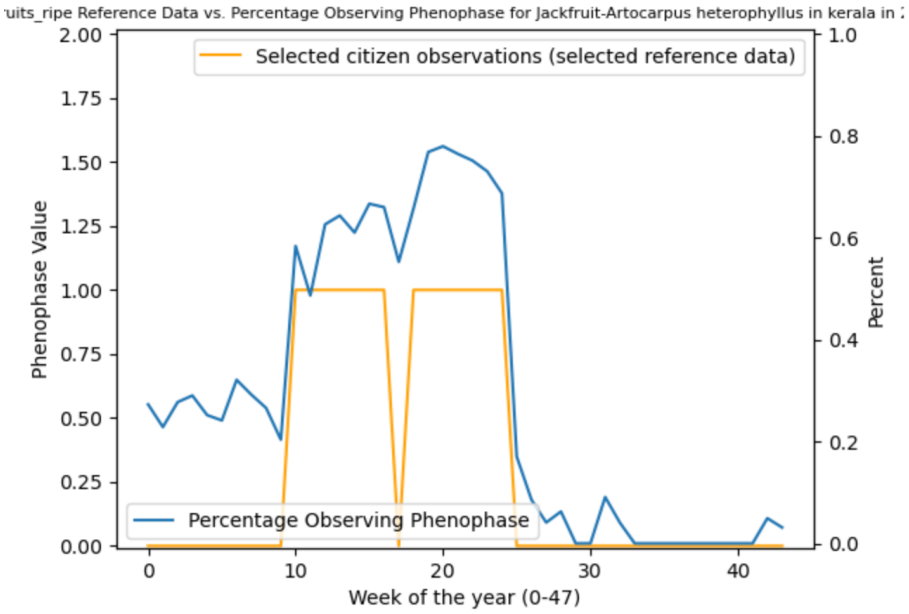

Figure 16: Plot showing citizen vs selected reference data

5 Bibliography

References

[1] Cory Maklin, *Introduction to Isolation Forests*, Medium, 2023. Available: `https://medium.com/@corymaklin/isolation-forest-799fceacdda4`.

[2] Education Ecosystem (LEDU), *Understanding K-means Clustering in Machine Learning*, Towards Data Science, 2018. Available: `https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1`.

[3] Yamini Bansal, *Introduction to K-Means Clustering*, Towards Data Science, 2021. Available: `https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawb`

Tools Used

Data Cleaning

- **Pandas** - Python library for dataset manipulation and cleaning
- **Shapely** - Python library for analysis of planar geometric objects
- **GeoPandas** - Python library for working with geospatial data

Data Visualization

- **Matplotlib** - Python data visualization library
- **Seaborn** - Python data visualization library

Machine Learning

- **Scikit-learn** - Python library for classical machine learning methods

Acknowledgements

A special thank you to Geetha Ramaswami and the SeasonWatch team for being so responsive and helpful throughout the process of this project, and to our incredible advisors for their valuable insight. Finally, a big thank you to the amazing Public Interest Technology-New England (PIT-NE) foundation for making this opportunity available to us!