**Phenology Report**

**Jean Lafitte National Historical Park and Preserve**

UNO ISQA8086 Project, Fall 2019, Team: Parts Per Million

By: Priyanka Choudhary, Pratheeba Nalligander Palanisami, and Mike Rouw

# Introduction

As a part of UNO’s data analytics program, the university offered a special topics course designed to take a data analysis project from start to finish using some specific technology tools. As a part of this project, Jean Lafitte National Historical Park and Preserve partnered with UNO to provide data they had collected over a two-year period relating to phenophase events observed in the park. This report is the summary of the completed project, and provides some analysis information with regards to the data provided.

# Background

Located in the state of Louisiana just southwest of New Orleans is the Jean Lafitte National Historical Park and Preserve. The area where the park is located is called the Barataria Preserve and consists of wetlands, swamps, forests and prairie land that cover over 26,000 acres of the southern Mississippi River delta. The park was first established as a state park in 1966 by the State of Louisiana. Later, in 1978, the United States Congress designated the park as the Jean Lafitte National Historical Park and Preserve. The area is named after Jean Lafitte, a rather infamous character who resided in and dominated the area in the early 1800s. The stated mission of the park is to preserve the natural and cultural resources in and around the Mississippi River delta region. The park has six distinct sites, including the Barataria Preserve, the Chalmette Battlefield, the Acadian Cultural Center, the Prairie Acadian Cultural Center, the Wetlands Acadian Cultural Center, as well as the visitor center in the French Quarter of New Orleans. For purposes of this study, the focus will be on the Barataria Preserve and the natural resources, habitat, and wildlife contained within the preserve. The Barataria Preserve invites visitors to explore the area via nine different hiking and walking trails. The trails include the Bayou Coquille Trail, the Marsh Overlook Trail, the Visitor Center Trail, the Palmetto Trail, the Ring Levee Trail, the Wood Duck Trail, the Plantation Trail, the Old Barataria Trail, and the Twin Canals Trail. For purposes of this study, data observations were taken over time along four of these different trails [1] [2].

# Phenology

Phenology is the study of periodic plant and animal life cycle events and how these are influenced by seasonal and interannual variations in climate, as well as habitat factors. An observable stage in the annual life cycle of a plant or animal that can be **defined** by a start and end point. **Examples** include the period over which newly emerging leaves are visible, or the period over which open flowers are present on a plant. Measurable elements for a phenophase event include start day, end day, and duration. Specific phenophase events can then be described over time and/or across different locations. Scientists are now beginning to study inter-relationships of phenophase events between species [3].

# Audience Profile

The primary contact at Jean Lafitte NHPP was Dr. Elizabeth (Liz) Marchio. She is currently working as Citizen Science Coordinator at the park. She has a PhD in the Department of Recreation, Park, and Tourism Science with a specialty in Environmental Sociology. In addition to Dr. Marchio, others who may have an interest in this report include the park employees, who work for the National Park Service. Many of these may be interacting with the public, so information about the timing and duration of specific phenophase events may be of interest. Also, Dr. Marchio and the park use citizen volunteers to collect the primary data used for this study, so they too may have an interest in viewing some results and possibly improving on their data collection methods.

# Research Questions

After a close review of the metadata provided, and an initial exploration of the provided dataset, the team proposed some specific research questions for this study. These were then reviewed and approved by Dr. Marchio to be some of the questions that would be of interest. While the metadata included descriptions of weather, temperature, and rainfall data, these were not included in the dataset, and so were not considered in this study due to time constraints. The following are the research questions which are the focus for this study:

* What is the time of year when the phenological event begins for the different species grouped by categories?
* What is the length of time observed for the phenological event for the species grouped by categories?
* Is there a time or duration change of the phenological event for any species during the three years of observations?
* Is there a difference in the time or duration for the species found in multiple regions of the phenological events?

# Data Overview

The data was collected by volunteers with some training while hiking the park’s trails and making observations. Data was collected from four separate trails and provided from late-summer 2017 through mid-summer 2019. It contained information about over two dozen different species of plants and birds with over two dozen different types of phenophase events measured and was collected specifically to study phenophase events (primary data). Metadata information was provided giving solid definitions for each column. Observation data was well organized with single rows showing specific observation information. The given dataset was NOT the original observations, but rather statistical values for single and multiple observations of the same event on the same species at the same location during the same year. The project used the provided mean values for the study, as most observations had a sample size of only one or two samples.

# Data Structure

The data contained category information for location, species, kingdom, and phenophase, with most of these categories having multiple descriptors. For each observation within these categories, the data contained the year of observation, the first “yes” day when the event was observed occurring (First Yes), and the last “yes” day when the event occurred (Last Yes). Also, within the data was the number of days since the last “no” observation before the First Yes, and the number of days until the first “no” observation after the Last No. As mentioned earlier, these were all mean values from the original observation data which was not provided. In addition to this provided data, the team identified other data that would be useful for our analysis. Namely, this included the day-of-year for the last “no” observation before the event (Last No), the day-of-year for the first “no” observation after the event (First No), the number of days between the First Yes and the Last Yes (Positive Duration), and the number of days between the Last No and the First No (Total Duration).

# Data Cleaning

As the data was reviewed, it was discovered that a large portion of the dataset was missing either a First Yes day or a Last Yes day. In order to continue, if either of these values were missing, the missing value was set equal to the other “yes” day and the lag time to the nearest “no” day was set to zero. This would give a positive duration of a single day on these observations. After this was corrected, and the additional data elements noted above were added, the data was checked for temporal integrity. Some observations were showing negative durations, so these were removed from the dataset. Finally, the dataset was narrowed to include only the columns that would be needed for the analysis.

# Data Enrichment

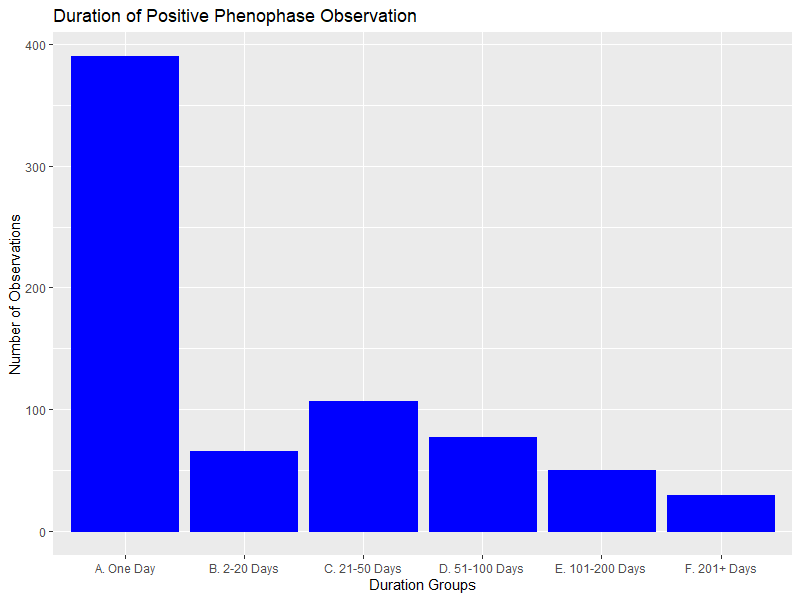
Based on the results from the data cleaning whereby many records were found to have either a missing Yes Day or No Day, an additional research question arose – What is the extent of this missing information within the dataset? To answer this question, the data was enriched to provide ranges of days for the Positive Duration. This would allow later data plotting to see the extent of the problem. In addition, a three species and four phenophases were identified for the focus of the study to narrow the scope. These were all selected from the group of deciduous plants in the dataset to provide commonality. Some basic statistics were computed for these focus areas and a table of means was compiled to provide plotting information on phenophase start times and durations. For this mean table, the years 2017-2018 were used for fall phenophase events while the years 2018-2019 were used for spring phenophase events.

# Information Plots

This section will show a sampling of key results found within the dataset in visual format, describe the plots for clarity, and provided a brief analysis of the plot results as it relates to the research questions noted earlier. During the project, plots were produced for many more species and more phenophase events than are included in this write up, and all are available for review.

## Positive Duration by Range

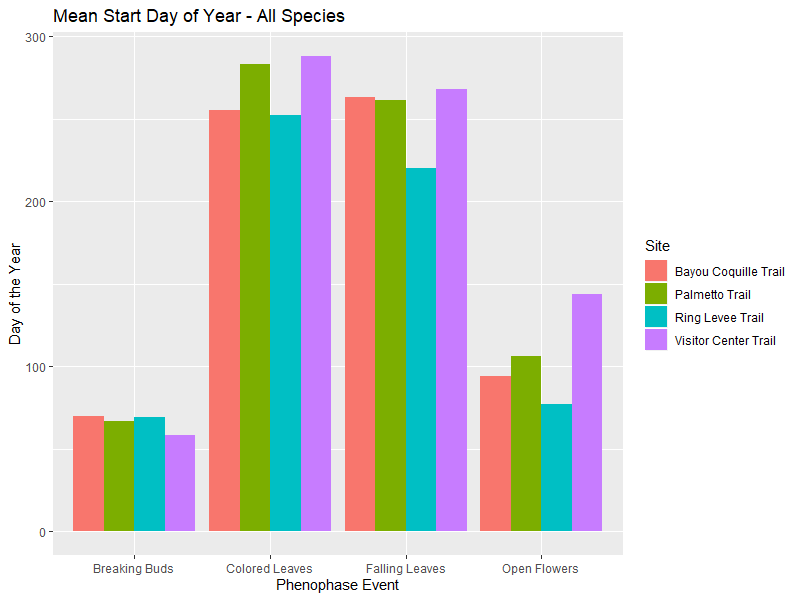
This plot was created to address the fifth research question – namely what is the extent of the missing “yes” values in the dataset? For this plot, the entire dataset was categorized into ranges of Positive Duration, and then the number of observations (frequency) is shown by range.



The Positive Duration in the first column of One Day will indicate all the observations missing either a First Yes day or a First No day. As can be seen above, over half of the observations in the dataset are missing one of these values, giving a Positive Duration of one day. This most likely indicates an issue with the method of data collection, or in the aggregation of the original data into the mean values found in the given dataset. If the problem is in collection, process changes could be made to improve training, increase observation frequencies, or otherwise improve the observations. If the problem is in the aggregation of the original data into the given dataset, these should be addressed. While some phenophase events may last only one (or just a few days), the volume of these short duration records would indicate there is likely an issue with the quality of the information in the dataset. Note that this will affect the plots and analysis throughout this section of the report.

## Mean Start Day

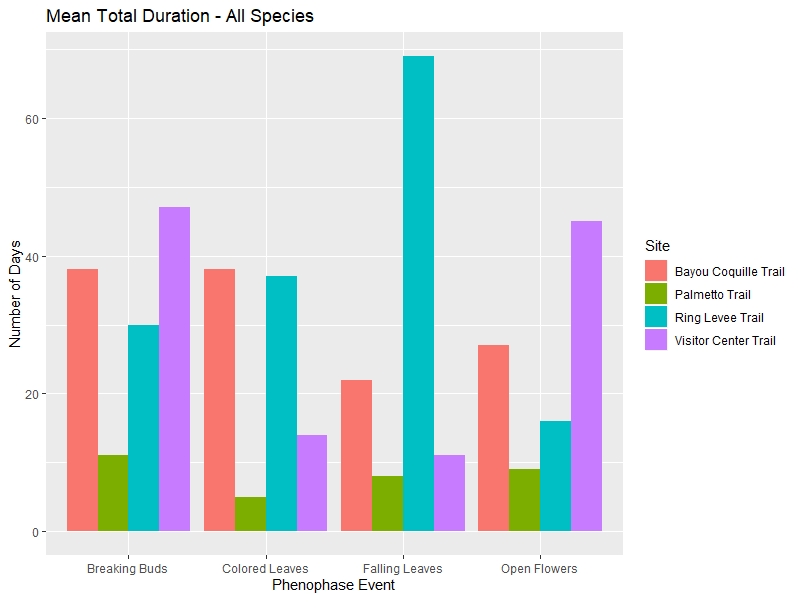
In this plot the First Yes day is averaged for multiple species with the same phenophase events. Again, only four events are selected, but all species having this type of phenophase event are included in the calculations. The mean values have been grouped into site locations to enable comparisons between sites within the park. The plot shows the **day-of-the-year** when the phenophase was starting (First Yes).



As can be seen in the plot above, we have four different phenophase events with their start day-of-year mapped by site location. For the Breaking Buds phenophase, this appears to be the earliest start in the year with an early-March start time across all the trail locations. Shortly afterward, the Open Flowers phenophase begins around early-April. While there is a bit more location variation in the data for Open Flowers, this is still pretty consistent for the start of this phenophase. For the fall phenophases of Colored Leaves and Falling Leaves, these seem to start at about the same time across all locations – namely around late-August or early-September. This plot addresses the research questions related to phenophase start times, and any variation that might occur among the observation sites.

## Total Duration

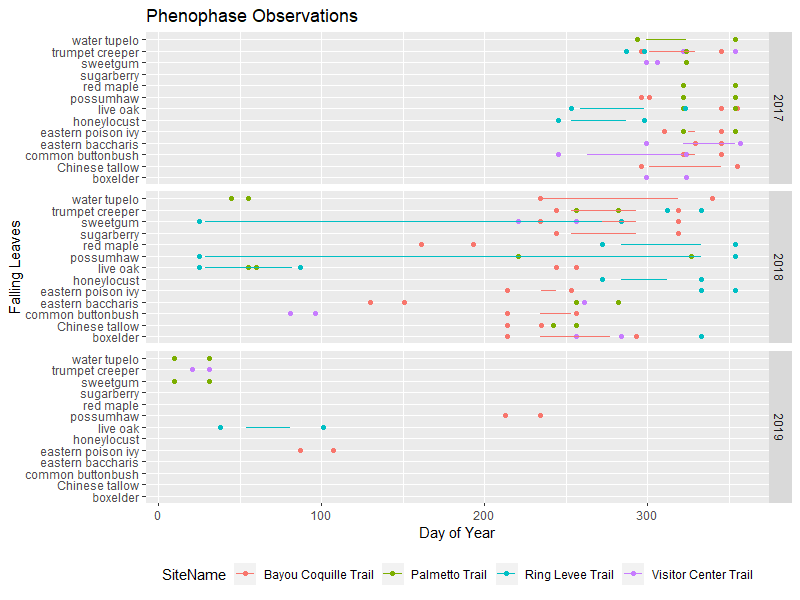
As with the prior plot, this chart will show the same four select phenophases with mean data from all species exhibiting that event over two years. However, where the last plot shows the day-of-year for the start of the phenophase, this plot below shows the **total number of days** (Total Duration) between the Last No and the First No observations. As before, the plot is grouped by site location.



This plot shows the Total Duration of the noted phenophase events. As can be seen above, the durations vary widely by location. This is more likely due to the data quality issues noted earlier in this report, rather than actual large differences in durations among the site locations. It is pretty difficult to derive any conclusions from this chart due to the questions relating to the underlying data. It may have been beneficial to ignore the observations with a single day of Positive Duration, however, those were over 50% of the records in the dataset, so again the validity of the results would be somewhat suspect.

## Time Chart by Phenophase

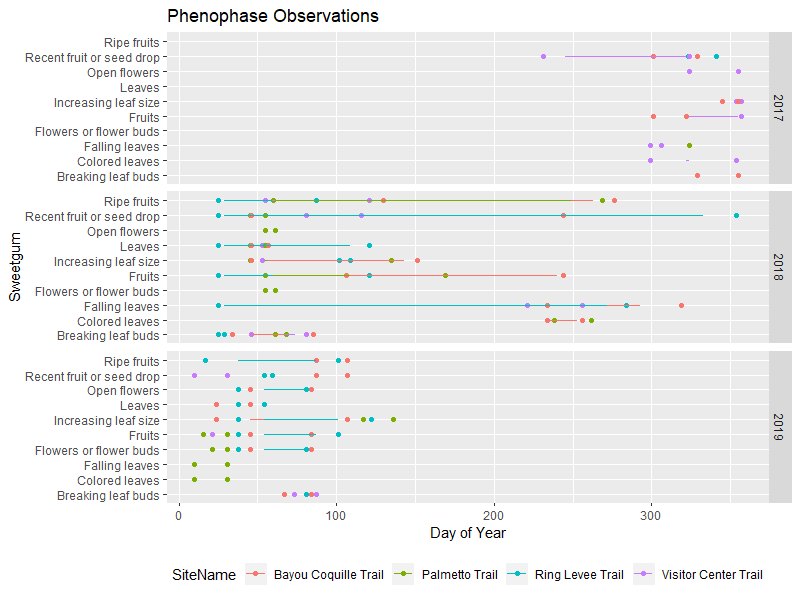
This plot attempts to show all the relevant data where the x-axis is time (day-of-year) and the y-axis shows a single phenophase across multiple species. The plot is segmented both by observation year and by site location. The “dots” on the plot represent the “no” values, while the “lines” on the plot represent the Positive Duration between “yes” values.



This is a fairly complex view of the data, but should be appropriate for the intended audience. Dr. Marchio sent a template view to work toward which was pretty similar to this view. The observations above without any “line” are again those with a Positive Duration of only one day (data missing a First Yes or a Last Yes). Also, some other likely data problems are visible in this view – namely spring observations for a phenophase event (Falling Leaves) that largely occurs in the fall. While there may be some species that hold leaves all winter and shed them in the spring, it may also be a problem with the observations. This plot is focused on the phenophase Falling Leaves, but three additional charts are available for Colored Leaves, Open Flowers, and Breaking Buds.

## Time Chart by Species

Like the prior chart, this plot shows a view of all the data over time in multiple dimensions. However, this plot is showing a single species with all its associated phenophase events. As before, the plot is segmented both by observation year and by site location. The “dots” on the plot represent the “no” values, while the “lines” on the plot represent the Positive Duration between “yes” values.



Now for the first time the data is shown with all the phenophase events associated with a single species. Again, the observations that do not include a “line” have a Positive Duration of just one day which is attributed to the missing First Yes or First No data. By studying this plot, along with the same plot for all the other species, all the research questions can be addressed in terms of phenophase timing, duration, variance by location, and changes over time.

# Summary

Based on the results of the analysis, the answers to our research questions were partially able to be addressed. The largest issue was the large number of observations with a missing First Yes day or First No day, which certainly skewed some of the results. It is recommended that the root cause for this be investigated and addressed to make future analysis more relevant. Conversely, if there is another way to make assumptions when this data is missing, the analysis could be rerun with different data cleaning processes to insert these new assumptions. Thank you to Dr. Marchio of Jean Lafitte NHPP and Dr. Wiggins of UNO for assisting in this project.

# References

[1] Wikipedia. Jean Lafitte National Historical Park and Preserve. (2019, August 19). Retrieved from <https://en.wikipedia.org/wiki/Jean_Lafitte_National_Historical_Park_and_Preserve>.

[2] Jean Lafitte National Historic Park & Preserve. (n.d.). Retrieved from <https://www.nationalparks.org/explore-parks/jean-lafitte-national-historic-park-preserve>.

[3] Wikipedia. Phenology. Retrieved from <https://en.wikipedia.org/wiki/Phenology>.

Word count before References = 2530 words.