Modified preregistration template for Data Mastery Challenge course

Source of original: [OSF](https://osf.io/jea94)

[Study Information](#_3o9ssh2swz09)

[Design Plan](#_4mzf79vx2q6j)

[Sampling Plan](#_hu8o0vkz41nk)

[Variables](#_pec3rgxfolor)

[Analysis Plan](#_3mtn7m44krsg)

[Other](#_6wujw18ggcuz)

### Study Information

1. Title (required) "A Comparative Analysis of Process-Based and Machine Learning Phenological Models for Red Maple, North United States"
2. Authors (required):

Fabian Moßner

Medha Iyer

Yvonne Ongera

1. Description

Region – North United States (states)

Plant species – Red Maple (Acer rubrum)

Phenophase – Spring leaf-out and autumn senescence

Phenological modelling is an important consideration for ecological and climate change studies. This paper aims at understanding the timing of phenological events and the drivers of phenological patterns for Red Maple trees in the North United States. Simultaneously, it also tests and compares both process based and machine learning based phenological models. Red maple trees have strong response to temperate cues and are widely studied species to understand climate change impacts.

1. Hypotheses (required)

Directional –

* The predictive accuracy of both model types will improve when incorporating multi-source data (e.g., satellite remote sensing, field observations, weather station data), but machine learning models will have a higher capacity for utilizing diverse data sources effectively.
* The spring leaf out time may be earlier due to warming temperatures and autumn senescence may be delayed.

Non directional

* Machine learning models may perform better in areas with greater climatic variability due to their ability to incorporate multiple non-linear variables.
* There may be interannual variations in the models due to rising temperatures.

### Design Plan

1. Study type

This study aims to replicate the original study to emphasize on the importance of phenology-based modelling to understand the influence of environmental changes on plants by adopting a mixed approach involving machine learning and process-based models explained later.

1. Study design
2. Data preprocessing:

Extract red maple phenology data from USA-NPN database with focus on spring leaf-out and autumn senescence.

Cleaning and normalizing the data.

1. Objectives:

Compare the predictive performance of machine learning models and process-based phenology models for Acer Rubrum trees

Analyze the spring leaf-out and autumn senescence phenophases from years 200-2021.

Predicting DOY with bilinear logistic regression and observing changes in trends over the year to observe pattens indicating influence of climate change.

1. Variables:

Possible Independent variables (Experimental):

Temperature, Cumulative growing degree days, precipitation

Elevation, location (latitude, longitude), land cover type

1. Analysis:

* Training and testing:

Split data into training(70%) and testing(30%) datasets

Using cross validation to ensure robustness.

* Prediction:

Applying the trained models to predict phenophase DOY transition.

1. Comparative analysis:

* As an extended outcome various machine learning models will be implemented to evaluate which approach better predicts DOY for red maple apart from comparing the performance of machine learning and process based models.
* Analyzing the regional and temporal variations in prediction accuracy and evaluate which approach performs better under varying conditions.

### Sampling Plan

In this section we’ll ask you to describe how you plan to collect samples, as well as the number of samples you plan to collect and your rationale for this decision. Please keep in mind that the data described in this section should be the actual data used for analysis, so if you are using a subset of a larger dataset, please describe the subset that will actually be used in your study.

1. Existing data (required)
   * Preregistration is designed to make clear the distinction between confirmatory tests, specified prior to seeing the data, and exploratory analyses conducted after observing the data. Therefore, creating a research plan in which existing data will be used presents unique challenges. **Please select the description that best describes your situation.** Please do not hesitate to contact us if you have questions about how to answer this question ~~(~~[~~prereg@cos.io~~](mailto:prereg@cos.io) -> discussion board of Canvas course).
     + Registration prior to creation of data: As of the date of submission of this research plan for preregistration, the data have not yet been collected, created, or realized.
     + Registration prior to any human observation of the data: As of the date of submission, the data exist but have not yet been quantified, constructed, observed, or reported by anyone - including individuals that are not associated with the proposed study. Examples include museum specimens that have not been measured and data that have been collected by non-human collectors and are inaccessible.
     + [Choose this option or the next] Registration prior to accessing the data: As of the date of submission, the data exist, but have not been accessed by you or your collaborators. Commonly, this includes data that has been collected by another researcher or institution.
     + [Choose this option or the former] Registration prior to analysis of the data**:** As of the date of submission, the data exist and you have accessed it, though no analysis has been conducted related to the research plan (including calculation of summary statistics). A common situation for this scenario when a large dataset exists that is used for many different studies over time, or when a data set is randomly split into a sample for exploratory analyses, and the other section of data is reserved for later confirmatory data analysis.
     + Registration following analysis of the data: As of the date of submission, you have accessed and analyzed some of the data relevant to the research plan. This includes preliminary analysis of variables, calculation of descriptive statistics, and observation of data distributions. Please see cos.io/prereg for more information.
2. Explanation of existing data (optional)
   * If you indicate that you will be using some data that already exist in this study, please describe the steps you have taken to assure that you are unaware of any patterns or summary statistics in the data. This may include an explanation of how access to the data has been limited, who has observed the data, or how you have avoided observing any analysis of the specific data you will use in your study.
   * **Example**: An appropriate instance of using existing data would be collecting a sample size much larger than is required for the study, using a small portion of it to conduct exploratory analysis, and then registering one particular analysis that showed promising results. After registration, conduct the specified analysis on that part of the dataset that had not been investigated by the researcher up to that point.
   * **More info**: An appropriate instance of using existing data would be collecting a sample size much larger than is required for the study, using a small portion of it to conduct exploratory analysis, and then registering one particular analysis that showed promising results. After registration, conduct the specified analysis on that part of the dataset that had not been investigated by the researcher up to that point.
3. Data collection procedures (required)

Data will be obtained from the USA National Phenology Network (USA-NPN), which maintains an extensive database of phenological observations for plant and animal species across the United States. Individual phenometrics will be filtered to meet the specific requirements of this study. The selected criteria include:

* **Species**: *Acer rubrum* (Red Maple)
* **States**: Illinois, Minnesota, and Wisconsin
* **Date Range**: January 2000 to December 2021
* **Phenophase**: Leaves

The output fields will be kept as default, with the addition of a few climate-specific variables relevant to spring and fall phenology, such as minimum temperature (Tmin), maximum temperature (Tmax), and precipitation.

As an alternative approach, the USA-NPN API documentation will be utilized to access data in instances where the dataset is too large to load and process efficiently through standard download methods.

### Variables

In this section you can describe all variables (both manipulated and measured variables) that will later be used in your confirmatory analysis plan. In your analysis plan, you will have the opportunity to describe how each variable will be used. If you have variables which you are measuring for exploratory analyses, you are not required to list them, though you are permitted to do so.

1. Manipulated variables (optional)
   * Describe all variables you plan to manipulate and the levels or treatment arms of each variable. This is not applicable to any observational study.
   * **Example:** We manipulated the percentage of sugar by mass added to brownies. The four levels of this categorical variable are: 15%, 20%, 25%, or 40% cane sugar by mass.
   * **More information**: For any experimental manipulation, you should give a precise definition of each manipulated variable. This must include a precise description of the levels at which each variable will be set, or a specific definition for each categorical treatment. For example, “loud or quiet,” should instead give either a precise decibel level or a means of recreating each level. 'Presence/absence' or 'positive/negative' is an acceptable description if the variable is precisely described.
2. Measured variables (required)
   * Describe each variable that you will measure. This will include outcome measures, as well as any predictors or covariates that you will measure. You do not need to include any variables that you plan on collecting if they are not going to be included in the confirmatory analyses of this study.
   * **Example**: The single outcome variable will be the perceived tastiness of the single brownie each participant will eat. We will measure this by asking participants ‘How much did you enjoy eating the brownie’ (on a scale of 1-7, 1 being ‘not at all’, 7 being ‘a great deal’) and ‘How good did the brownie taste’ (on a scale of 1-7, 1 being ‘very bad’, 7 being ‘very good’).
   * **More information**: Observational studies and meta-analyses will include only measured variables. As with the previous questions, the answers here must be precise. For example, 'intelligence,' 'accuracy,' 'aggression,' and 'color' are too vague. Acceptable alternatives could be 'IQ as measured by Wechsler Adult Intelligence Scale' 'percent correct,' 'number of threat displays,' and 'percent reflectance at 400 nm.'
3. Indices (optional)
   * If any measurements are going to be combined into an index (or even a mean), what measures will you use and how will they be combined? Include either a formula or a precise description of your method. If your are using a more complicated statistical method to combine measures (e.g. a factor analysis), you can note that here but describe the exact method in the analysis plan section.
   * **Example**: We will take the mean of the two questions above to create a single measure of ‘brownie enjoyment.’
   * **More information**: If you are using multiple pieces of data to construct a single variable, how will this occur? Both the data that are included and the formula or weights for each measure must be specified. Standard summary statistics, such as “means” do not require a formula, though more complicated indices require either the exact formula or, if it is an established index in the field, the index must be unambiguously defined. For example, “biodiversity index” is too broad, whereas “Shannon’s biodiversity index” is appropriate.

### Analysis Plan

You may describe one or more confirmatory analysis in this preregistration. Please remember that all analyses specified below must be reported in the final article, and any additional analyses must be noted as exploratory or hypothesis generating.

A confirmatory analysis plan must state up front which variables are predictors (independent) and which are the outcomes (dependent), otherwise it is an exploratory analysis. You are allowed to describe any exploratory work here, but a clear confirmatory analysis is required.

1. Statistical models (required)
   * What statistical model will you use to test each hypothesis? Please include the type of model (e.g. ANOVA, multiple regression, SEM, etc) and the specification of the model (this includes each variable that will be included as predictors, outcomes, or covariates). Please specify any interactions, subgroup analyses, pairwise or complex contrasts, or follow-up tests from omnibus tests. If you plan on using any positive controls, negative controls, or manipulation checks you may mention that here. Remember that any test not included here must be noted as an exploratory test in your final article.
   * **Example**: We will use a one-way between subjects ANOVA to analyze our results. The manipulated, categorical independent variable is 'sugar' whereas the dependent variable is our taste index.
   * **More information**: This is perhaps the most important and most complicated question within the preregistration. As with all of the other questions, the key is to provide a specific recipe for analyzing the collected data. Ask yourself: is enough detail provided to run the same analysis again with the information provided by the user? Be aware for instances where the statistical models appear specific, but actually leave openings for the precise test. See the following examples:
     + If someone specifies a 2x3 ANOVA with both factors within subjects, there is still flexibility with the various types of ANOVAs that could be run. Either a repeated measures ANOVA (RMANOVA) or a multivariate ANOVA (MANOVA) could be used for that design, which are two different tests.
     + If you are going to perform a sequential analysis and check after 50, 100, and 150 samples, you must also specify the p-values you’ll test against at those three points.
2. Transformations

* Data preprocessing for machine learning: Cleaning and normalizing data to address missing or anomalous values.
* Assigning labels for bilinear logistic regression: Binary label 0 or 1 to DOY values based on whether phenophase event has occurred by a given day. This is also used for prediction where the first change from 0 to 1 occurs in the DOY value.

1. Inference criteria

Evaluation metrics:

1. Model comparison through:

* R-squared
* Root Mean Square Error (RMSE)
* Mean Absolute Error (MAE)

1. Assess phenophase based performance metrics for spring and autumn transitions.
   1. What criteria will you use to make inferences? Please describe the information youÍll use (e.g. p-values, bayes factors, specific model fit indices), as well as cut-off criterion, where appropriate. Will you be using one or two tailed tests for each of your analyses? If you are comparing multiple conditions or testing multiple hypotheses, will you account for this?
   2. **Example**: We will use the standard p<.05 criteria for determining if the ANOVA and the post hoc test suggest that the results are significantly different from those expected if the null hypothesis were correct. The post-hoc Tukey-Kramer test adjusts for multiple comparisons.
   3. **More information:** P-values, confidence intervals, and effect sizes are standard means for making an inference, and any level is acceptable, though some criteria must be specified in this or previous fields. Bayesian analyses should specify a Bayes factor or a credible interval. If you are selecting models, then how will you determine the relative quality of each? In regards to multiple comparisons, this is a question with few “wrong” answers. In other words, transparency is more important than any specific method of controlling the false discovery rate or false error rate. One may state an intention to report all tests conducted or one may conduct a specific correction procedure; either strategy is acceptable.
2. Data exclusion (optional)
   1. How will you determine what data or samples, if any, to exclude from your analyses? How will outliers be handled? Will you use any awareness check?
   2. **Example**: No checks will be performed to determine eligibility for inclusion besides verification that each subject answered each of the three tastiness indices. Outliers will be included in the analysis.
   3. **More information**: Any rule for excluding a particular set of data is acceptable. One may describe rules for excluding a participant or for identifying outlier data.
3. Missing data (optional)
   1. How will you deal with incomplete or missing data?
   2. **Example**: If a subject does not complete any of the three indices of tastiness, that subject will not be included in the analysis.
   3. **More information**: Any relevant explanation is acceptable. As a final reminder, remember that the final analysis must follow the specified plan, and deviations must be either strongly justified or included as a separate, exploratory analysis.
4. Exploratory analysis (optional)
   1. If you plan to explore your data set to look for unexpected differences or relationships, you may describe those tests here. An exploratory test is any test where a prediction is not made up front, or there are multiple possible tests that you are going to use. A statistically significant finding in an exploratory test is a great way to form a new confirmatory hypothesis, which could be registered at a later time.
   2. **Example**: We expect that certain demographic traits may be related to taste preferences. Therefore, we will look for relationships between demographic variables (age, gender, income, and marital status) and the primary outcome measures of taste preferences.

### Other

1. Other (Optional)
   1. If there is any additional information that you feel needs to be included in your preregistration, please enter it here. Literature cited, disclosures of any related work such as replications or work that uses the same data, or other context that will be helpful for future readers would be appropriate here.

**Labels Explanation**:

* **Label 0**: Assigned if the **date** in the time series (end of the window) is **before DOY** (Day of Year). This indicates that the phenological event has **not yet occurred** at the observation site.
* **Label 1**: Assigned if the **date** in the time series (end of the window) is **on or after DOY**. This indicates that the phenological event **has occurred**.

**Prediction Phase**:

* The **predicted DOY** is determined as the first date where the binary classification changes from **0 (event not occurred)** to **1 (event occurred)**. This transition point is critical for identifying when the phenological event occurs.

**Additional Context**:

* The **DOY** (Day of Year) is the specific day when the event of interest is observed.
* The process uses time-series data such as daily mean temperature (from datasets like Daymet) integrated with phenological records.

**Example:**

**Context:**

* A phenological event being studied is the blooming of a specific plant species.
* **DOY (Day of Year)** is the day the event (blooming) occurs, e.g., DOY = 120 (the 120th day of the year, approximately April 30 in a non-leap year).
* The daily mean temperature is used as part of the time-series data.

**Binary Classification:**

1. **Time-Series Data**: Let's consider a sliding window of time-series data for days leading up to and after DOY = 120:
   * Day 115: No blooming observed → **Label = 0**
   * Day 116: No blooming observed → **Label = 0**
   * Day 117: No blooming observed → **Label = 0**
   * Day 118: Blooming observed → **Label = 1**
   * Day 119: Blooming observed → **Label = 1**
   * Day 120: Blooming observed → **Label = 1**
2. **Training Phase**:
   * The dataset includes observations from multiple years and locations, with labels assigned as **0** for days before DOY and **1** for days on or after DOY.
   * Features extracted from the sliding windows (e.g., temperature, prior conditions) are used to train the binary classifier.
3. **Prediction Phase**:
   * For a new time series (e.g., from another location or year), the binary classifier is applied to predict labels for each day.
   * The **predicted DOY** is identified as the first day when the label changes from **0** (event not occurred) to **1** (event occurred).

**Title:**

Comparative Analysis of Machine Learning and Process-Based Phenology Modeling for Predicting Red Maple Phenophases in the Northeastern United States

**Objectives:**

1. Compare the predictive performance of machine learning models and process-based phenology models for red maple (*Acer rubrum*) trees.
2. Analyze the spring leaf-out and autumn senescence phenophases from 2000 to 2021.
3. Determine the feasibility of predicting the day of year (DOY) for phenophase transitions using bilinear logistic regression.

**Study Area and Dataset:**

* **Study Area**: Northeastern United States
* **Phenophase**: Spring leaf-out and autumn senescence
* **Data Source**: USA National Phenology Network (USA-NPN) datasets from 2000 to 2021.
* **Dependent Variable**: Day of Year (DOY) of the phenophase transition.

**Study Framework:**

1. **Data Preprocessing**:
   * Extract red maple phenology data from the USA-NPN database.
   * Focus on phenophase-specific events: spring leaf-out and autumn senescence.
   * Clean and normalize data to address missing or anomalous values.
2. **Independent Variables for Machine Learning Models**:
   * **Climatic Variables**:
     + Temperature (mean, minimum, maximum)
     + Cumulative growing degree days (GDD)
     + Precipitation (monthly and seasonal totals)
     + Solar radiation
   * **Temporal Variables**:
     + Day of year (DOY) as a sequence number
     + Year
   * **Site-Specific Variables**:
     + Elevation
     + Latitude and longitude
     + Land cover type
   * **Derived Variables**:
     + Seasonal trends in temperature and precipitation
     + Winter chilling hours
     + Onset of spring warming
3. **Modeling Approaches**:
   * **Process-Based Model**:
     + Use existing phenology models that simulate red maple phenophases based on climatic thresholds.
   * **Machine Learning Model**:
     + Train a bilinear logistic regression model:
       - Assign a binary label (0 or 1) to DOY values based on whether the phenophase event has occurred by a given day.
       - During prediction, determine the first DOY where the label changes from 0 to 1.
4. **Evaluation Metrics**:
   * Compare models using:
     + Mean Absolute Error (MAE)
     + Root Mean Square Error (RMSE)
     + R² (Coefficient of Determination)
   * Assess phenophase-specific performance metrics for spring and autumn transitions.
5. **Analysis Framework**:
   * **Training Phase**:
     + Split data into training (70%) and testing (30%) datasets.
     + Perform hyperparameter tuning for the machine learning model.
   * **Validation Phase**:
     + Use cross-validation to ensure robustness.
   * **Prediction Phase**:
     + Apply the trained models to predict phenophase DOY transitions.
   * **Comparative Analysis**:
     + Evaluate which approach better predicts DOY for red maple phenophases.
     + Analyze regional and temporal variations in prediction accuracy.

**Expected Outcome:**

* Identification of the most effective method for predicting red maple phenophases.
* Insights into the role of climatic and site-specific factors in phenological events.
* Recommendations for integrating machine learning and process-based modeling approaches.

**Comparative Analysis**:

* Analyze performance across regions, years, and phenophases.
* Evaluate which approach performs better under varying climatic and geographic conditions.