Data Mastery Challenge Course
PRE-REGISTRATION

Title
“A Comparative Analysis of Process-Based and Machine Learning Phenological Models for Red Maple(Acer rubrum)”, North United States.

Authors
Medha Iyer
Fabian Moßner
Yvonne Ongera

Description
Region- North United States (Illinois, Minnesota, and Wisconsin)
Plant Species- Red Maple (Acer rubrum)
Phenophases- Spring leaf out and autumn senescence
Phenological modeling is an important consideration for ecological and climate change studies. This paper aims to understand 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 using the Springtime Python Package. Red maple trees have strong responses to temperate cues and are widely studied species to understand climate change impacts.

Hypotheses
1. 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.
- Spring leaf out time may be earlier due to warming temperatures and autumn senescence may be delayed.

2. 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 and extend the original study (Khodadadzadeh et al., 2024) to emphasize the importance of phenology-based modeling to understand the influence of environmental changes on plants by adopting a mixed approach involving machine learning and process-based models explained later.

2. Study Design:

I. Data preprocessing:

- Extract red maple phenology data from the USA-NPN database with a focus on spring leaf-out and autumn senescence (2000-2021)
- Cleaning and normalizing the data to remove outliers and missing data.

II. 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.

III. Variables:

- Possible Independent variables (Experimental):
- Temperature, Cumulative growing degree days, precipitation
- Elevation, location (latitude, longitude), land cover type

IV. 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.

V. 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

1. Existing Data

- Phenology observations from USA-NPN.
  The USA National Phenology Network provides extensive observational data on plant phenophases across multiple years. These datasets contain detailed metadata including the location and timestamps. The dataset will cover the Northern US region due to the widespread distribution (Illinois, Minnesota, and Wisconsin).

2. Data Collection

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
- Phenophases: 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.

3. Sample size

   The sample size ensures sufficient representation for the training, validation, and testing of the models. Machine learning models typically require a large number of samples to avoid overfitting, while process-based models may work effectively with smaller datasets if the data quality is high(Rajput et al., 2023).To ensure robust model training and evaluation, 1,500–2,000 observations are suggested. The original study used 2010 observations.

Variables

I. Independent Variables:

Average daily temperature, Cumulative growing degree days, Precipitation, Elevation, Location, Landcover type

II. Dependent Variables:

Day of Year (DOY) for spring leaf-out and autumn senescence.

III. Measured variables:

Spring Leaf-out DOY: The day of the year (DOY) when red maple trees first show signs of leaf emergence, measured using data from the USA-NPN database.

Autumn Senescence DOY: The day of the year (DOY) when red maple trees first show signs of leaf senescence, measured using data from the USA-NPN database.

IV. Covariate (Year):

Year of observation (2000-2021), included to account for inter-annual variation in phenology.

Analysis Plan

1. Statistical Models:

- 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 phenophases event has occurred by a given day. During prediction, determine the first DOY where the label changes from 0 to 1.

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 the phenophase event has occurred by a given day. This is also used for predicting where the first change from 0 to 1 occurs in the DOY value.

3. Inference criteria:

- Evaluation metrics:
  Model comparison through: R-squared, Root Mean Square Error (RMSE), Mean Absolute Error (MAE)
- Assess phenophase-based performance metrics for spring and autumn transitions.

Other

- Software: The study uses the Springtime Python Package to harmonize data preparation and modeling workflows.
- Reproducibility: All scripts, workflows, and datasets will be made publicly available upon publication.

References

1. Khodadadzadeh, M., Kalverla, P., & Zurita-Milla, R. (2024). Harmonizing Machine Learning BasedPhenological Modeling: A Unified Workflow for Comparative Analyses. IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, 5333–5336. https://doi.org/10.1109/IGARSS53475.2024.10641356
2. Rajput, D., Wang, W.-J., & Chen, C.-C. (2023). Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics, 24(1), 48. https://doi.org/10.1186/s12859-023-05156-9