**Measuring the Right Thing: Identifying Predictors of Latent Phenological**

**Variation in a Hawaiian Rainforest**

Caleb Charpentier, Colby Stakun-Pickering, Jianyang Gu

**Background:**

To predict how global forests might respond to a changing climate, we must first understand their phenology. The phenology of trees in temperate climates, particularly deciduous trees, is highly pronounced and is generally well understood. However, phenological patterns also occur in tropical forests, where seasonality is less obvious. In these forests, including the forest at the Pu'u Maka'ala Natural Area Reserve, the drivers of phenological variation are poorly documented. This lack of documentation is, in a large part, due to difficulties in measurement. Recording phenophase – the state of a tree during its phenological cycle, such as flowering or fruiting – has historically required researchers to walk through the forest and to manually identify the species-specific phenophases of individual trees. This process is time consuming and is difficult to scale in a way that is useful for understanding the phenological processes of the forest as whole. Additionally, the concept of “whole-forest” phenology is a rather abstract concept to begin with, as researchers have performed limited work to understand how large regions of Hawaiian forests respond to seasonal variation. Generally, these studies use only the average greenery of the whole canopy as a phenological proxy, which is likely an incomplete description of its total phenological variation.

In this study, we hope to accelerate the process of phenophase measurement and to develop a machine learning pipeline that leverages those measurements to create a latent indicator of canopy-level phenology. We will then regress this indicator against a variety of climatic factors that NEON has collected over the last half-decade, to understand how climate can influence the phenology of whole forests beyond the level of individual trees.

**Research Question:**

In this study, we hope to address two questions:

1. Can we use PhenoCam canopy images to measure patterns of forest phenology beyond a simple greenness index?
2. If these larger scale phenological patterns exist, what climatic features are important for predicting them?

**Project Goals:**

To address our questions, we intend to accomplish 3 goals:

1. We hope to automate and increase the temporal resolution of phenophase observations of understory plants via ML models and camera trap footage. This task will both provide data for our own analysis, and will provide future NEON researchers with an easier and more rapid method for recording plant phenophase.
2. Using these fine-scaled observations of individual phenology, we will apply techniques akin to those used for latent phenotyping in agricultural response-to-treatment studies to identify latent, canopy-level patterns of phenological variation. This process will result in a new model (or at least architecture) for identifying phenological patterns from PhenoCam observations above the forest canopy, which is much more efficient than manual understory observations.
3. After identifying this measure of phenology, we hope to regress it against a variety of climatic and environmental features to identify its potential drivers (such as carbon uptake, water use, surface-energy balance, etc…). If successful, it will provide researchers with a much more nuanced understanding of canopy level phenology beyond simple regressions of greenness against climatic predictors, which can contribute to more nuanced approaches to conservation and will eventually result in a publication.

**Fieldwork/Analytic Approach:**

Phenological data is only meaningful over the scale of long time periods. As such, most of the data we will use for analysis has already been recorded by NEON, in the form of PhenoCam images or tabular climatic observations. However, for our project, we intend to perform three separate tasks, described below. For the first task, we will need images of plants with phenophase labels from the forest understory. This sort of data is already being recorded by Dan Rubenstein’s Phenology and Use by Animals project, so, as suggested by our project lead, we will primarily be helping with camera trap setup and management during the 2-week field experience in Hawaii.

For Task 1, we will use cameras shared with Dan Rubenstein’s Phenology and Use by Animals project to generate images of plants at NEON’s PUUM site and automatically extract phenophase (e.g. leaves, flowers, fruit) with a neural network-based vision model. While we may have to explicitly train a vision model to predict phase from scratch, we will first try an existing zero shot method, such as Grounding-Dino + SAM, to label fruits, flowers, and leaves in images, with no training at all on our end. If this method is inadequate, we will train a simple classifier (likely logistic regression or a support vector machine) on the embeddings of a pretrained image model (possibly BioCLIP, but more likely a simple pretrained ResNet model from PyTorch). If neither of these methods are sufficient, we will fine-tune a pretrained image model outright.

For Task 2, we will adapt the Latent Phenotyping (LP) framework developed by Ubbens et al for small plant populations to whole canopy data. We have many images (1 every 15 minutes for the past 5 years) of a section of forest canopy from the tower camera at the PUUM site. Using these images, in combination with phenological data from Task 1 and existing NEON data, we will train an autoencoder to capture latent representations of the canopy, which distinguish canopy images where understory plants are predominantly in particular phenophases (leaves only, flowering, fruiting, and 2 transitional phases). If we are unable to collect this data in the allotted time period, we will simply use the rainy and dry season as responses in lieu of dominant phenophase indicators for the third section of the class. To avoid unwanted interactions between day-to-day weather conditions and the appearance of trees in the canopy image, we will use domain adaptation techniques to make the model invariant to weather conditions on individual days, and will use a time series model (probably an LSTM, as was used by Ubbens et al 2020) to capture variation across entire seasons in our embeddings.

For Task 3, we will regress the latent, canopy-level phenology embeddings taken in Task 2 against a variety of climatic variables that NEON has already measured (for example, the amount of carbon uptake happing at a given time, or the amount of solar radiation hitting the canopy that can contribute to photosynthesis). If the regression model we use is not scale invariant, we may use a specialized activation function in our original Latent Phenotyping network so that all embedded dimensions are scaled equally. We do not want to perform standard scaling on these embeddings after learning an initially non-scaled phenological representation, as the dimensions in the embedding vector of an autoencoder are generally dependent on each other in complex, nonlinear ways, and there is no guarantee that they are scale or location invariant unless this is explicitly enforced to begin with. The regression model we use will be a more traditional statistical or machine learning technique, such as a support vector regressor or GLM, so that we can more easily interpret the influence of our predictors on our derived metric of canopy-level phenology.

Additionally, we may also regress these same predictors against the raw phenological indicators themselves (raw phenophase measures or wet/dry season) as well as against embeddings of canopy images that do not rely on phenological indicators, to ensure that our learned features are sufficiently picking up on phenological variation. Since our dataset spans several years, we can split training and testing sets by year or use stratified cross-validation across distinct phenological seasons.

After analysis, we intend to report our results in a research paper, which provides a clear example of how Imageomics approaches can be used for understanding large-scale phenological patterns that would otherwise be difficult to measure.

**Relevance:**

The first three objectives of this course, as stated in the syllabus, are to:

1. Explain why AI can help solve ecology problems.
2. Apply and evaluate relevant AI models for different data types.
3. Develop ecology research questions that can be answered using both unstructured and structured data.

One of the primary goals of this project is to develop a new way to measure whole-canopy phenology using guided by pre-existing knowledge of phenological phases. Accomplishing this aim will allow us to find new indicators of canopy phenology that are more holistic than traditional greenness indices and more efficient and temporally precise than fully manual understory surveys. We are using both raw image data of the forest canopy and tabular data from pre-existing NEON records and phenophase measurements, and we are analyzing them with a variety of models from image-based autoencoders to “classic” support vector machines and logistic regression. Using these techniques, we hope to show that knowledge-guided Imageomics approaches allow researchers to understand phenomena across spatial and temporal scales that would otherwise be intractable.

**Citations**

1. Ubbens, J. *et al.* Latent Space Phenotyping: Automatic Image-Based Phenotyping for Treatment Studies. *Plant Phenomics* **2020**, 5801869 (2020).