# Predicting Invasive Understory Species in Boston's Deciduous Forest

## A Comparison of Random Forest Models with Sentinel-2 Time Series Data for 2023

Tufts University | Gerald J. and Dorothy R. Friedman School of Nutrition Science and Policy

# Background

Forest understories provide crucial services like habitat and forage for wildlife, carbon cycling, and biodiversity. Their effective management relies on being able to identify and track invasive species, but field surveys can be expensive and impractical.

Recently there has been increased interest in remote sensing techniques that rely on an 'observation window' when the tree canopy is still bare, but the understory has greened-up. This window is short, often only 7 – 16 days!

A new method from the University of Connecticut uses harmonic regression and random forest classifiers to differentiate native and invasive understory species in Connecticut deciduous forests.[1] Harmonic regression fits a complex periodic function using sines and cosines to extract phenology over the year. The regression is then used as inputs to the machine learning classifier. This work applies a simplified version of their method as a proof of concept in Boston and answers the question:

**Can supervised classification and harmonic regression effectively differentiate invasive and native understory species in Boston using Sentinel-2 time series data?**

# Methodology Overview

**Pre-process Sentinel-2 imagery from Google Earth Engine**
Filter, mask, and clip to study area; mosaic same-day tiles together

↓

**Collect iNaturalist training data for each species of interest**

↓

**Calculate indices**

↓

**Conduct pixelwise harmonic regression in SciPy to extract seasonality coefficients**

↓

**Train 4 different random forest (RF) models on 70% of data and test on remaining 30%**

↓

**Use the models to predict the remaining forest pixels that did not have iNaturalist observations**

# Training Data

iNaturalist is a global online database of citizen science observations of species. Using a python package (pyinaturalist) to access to the website's API, the locations of 5 important understory species in New England deciduous forests were collected from 2021 – 2024.

Glossy Buckthorn
*Frangula alnus*

Multiflora Rose
*Rosa multiflora*

Invasives

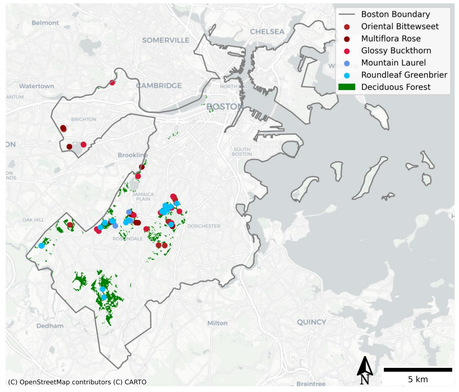
Asian Bittersweet
*Celastrus orbiculatus*

Mountain Laurel
*Kalmia latifolia*

Natives

Roundleaf Greenbrier
*Smilax rotundifolia*

These species were selected after consultation with local ecology and forestry colleagues and are of specific concern in our region. The location of deciduous forests was obtained from the National Land Cover Database (2019) and clipped to Boston using MassGIS towns data.




**Map 1: Study area and training sites**

# Remote Sensing Data and Vegetation Indices

Using 10 bands from Sentinel-2, 4 different vegetation indices were calculated, including:

1. Normalized difference vegetation index (**NDVI**)
2. Red edge normalized difference vegetation index (**RENDVI**)
3. Soil adjusted vegetation index (**SAVI**)
4. Normalized burn ratio (**NBR**).

These 14 inputs were then used for the harmonic regression before being used in the random forest models.

| Band | Type |
|---|---|
| B2 | Blue |
| B3 | Green |
| B4 | Red |
| B5 | Red Edge 1 |
| B6 | Red Edge 2 |
| B7 | Red Edge 3 |
| B8 | Near-IR (narrow) |
| B8A | Near-IR (wide) |
| B11 | Shortwave-IR1 |
| B12 | Shortwave-IR2 |

**Table 1: Bands and indices used**

| Index | Equation |
|---|---|
| NDVI | $\frac{B8 - B4}{B8 + B4}$ |
| RENDVI | $\frac{B6 - B5}{B6 + B5}$ |
| SAVI | $\frac{1.5(B8 - B4)}{B8 + B4 + 0.5}$ |
| NBR | $\frac{B12 - B8}{B12 + B8}$ |

# Harmonic Regression

To extract the characteristics from each band/index, the following model was fit (DOY=Day of Year):

$$band = a + b * cos\left(\frac{2\pi}{365} DOY\right) + c * sin\left(\frac{2\pi}{365} DOY\right) + * cos\left(2 * \frac{2\pi}{365} DOY\right) + e * sin\left(2 * \frac{2\pi}{365} DOY\right)$$
$$+ f * cos\left(3 * \frac{2\pi}{365} DOY\right) + g * sin\left(3 * \frac{2\pi}{365} DOY\right) + h * cos\left(4 * \frac{2\pi}{365} DOY\right) + i * sin\left(4 * \frac{2\pi}{365} DOY\right)$$

Of the 14 bands/indices, only clear separation between native and invasive iNaturalist data was seen for B3, B5, and RENDVI (highlighted in green circles below).

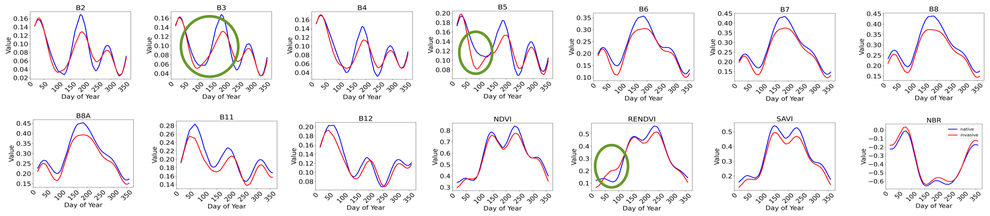



**Figure 1: Average results of harmonic regression for iNaturalist data**

# Models Investigated

These 4 following random forest (RF) models were investigated for this classification task:

- **Basic RF**: no spatial information is used in the model
- **Geographic RF**: developed by Georganos et al. to account for spatial autocorrelation and heterogeneity by training subRFs on spatial subsets of data[2]
- **Spatial RF**: a more efficient method of geographic RF developed by Weidemann et al. that instead trains individual decision trees on subsets of the data and weights data by its distance from the center of the tree[3]
- **Spatial RF, tuned**: same as above but tuned for the number of spatial neighbors generating the least error

All models used the 9 regression coefficients and each pixel's regression's root mean square error (RMSE) for each of the 10 bands and 4 indices resulting in 140 explanatory variables for the basic model. The other 3 models also took in the coordinates of the pixels.

# Model Performance

Model performance was assessed by calculating overall accuracy and Cohen's Kappa using the remaining test data (30% of the iNaturalist data):

- **Overall accuracy** is the percent of correctly classified sites out of total sites
- **Cohen's Kappa** is a statistical measure of how well the true and predicted labels agree on a scale of 0 to 1 with 0 meaning perfect disagreement and 1 meaning perfect agreement

**Table 2: Model performance metrics**

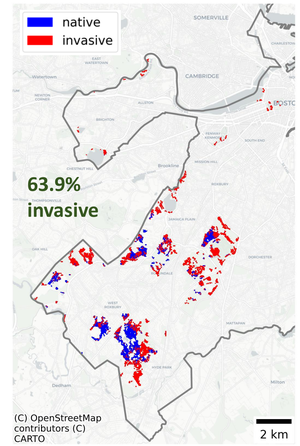
| Model | Runtime (s) | Overall Accuracy | Cohen's Kappa |
|---|---|---|---|
| Basic | 0.09 | 91.1% | 0.82 |
| Geographic | 151 | 93.3% | 0.86 |
| Spatial | 1.69 | 84.4% | 0.68 |
| Spatial, tuned | 1.58 | 86.7% | 0.73 |

All the models performed well with the **basic and geographic being clearly better at discriminating the two categories**. However, there was a clear trade-off between the increased precision of geographic RF and the runtime it requires.

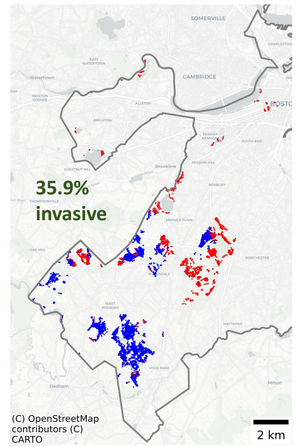
# Prediction Results

The most accurate basic and geographic models were used to predict the rest of the pixels in the study area that were not in the iNaturalist observations, and the predicted area covered by invasives was determined:

Basic Random Forest




Geographic Random Forest




**Map 2: Final prediction results**

# Main Takeaways

1. **The workflow is accurate**: It can pick up on small differences between the classes. This could be a major help to the Boston Parks Department, Mass Audubon, and others managing invasives in our area by **better targeting their resources**.
2. **Spatial machine learning is needed for spatial questions**: Though they take longer to run, the increased classification precision and accuracy **is essential to valid predictions on larger datasets.** The small differences between the basic and geographic model became much more apparent once applied to larger predictions because of **higher levels of spatial autocorrelation** than was viewed in the sparse training data.
3. **Geographic was better than spatial**: With the rise of high-performance computing, the difference in runtime of the two different spatial model implementations may not be as important as the increased accuracy and precision.

# Future Work

- ❑ **Add more training data** by conducting field sampling with ecologists and arborists
  - ❑ This would improve the representativeness of the sample
  - ❑ This may allow for differentiating species (which could not be done in this work because of too few sample locations)
- ❑ **Extend** to the rest of the Charles River Watershed and/or Massachusetts to test scalability
  - ❑ But will the runtime required become too large?
- ❑ Investigate **changes over time**
  - ❑ This could help municipalities and other land stewards determine if their actions are sufficient and/or effective

Chad Fisher | Advanced Remote Sensing | UEP 294

References:
1. Yang X, Qiu S, Zhu Z, Rittenhouse C, Riordan D, Cullerton M (2023). Mapping understory plant communities in deciduous forests from Sentinel-2 time series. Remote Sensing of Environment, 293. https://doi.org/10.1016/j.rse.2023.113601
2. Georganos S, Grippa T, Gadiaga AN, Linard C, Lennert M, Vanhuysse S, Mboga N, Wolff E, Kalogirou S (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International, 36(2). https://doi.org/10.1080/10106049.2019.1595177
3. Wiedemann N, Martin H, Westerholt R (2023). Benchmarking regression models under spatial heterogeneity. GIScience 2023. http://dx.doi.org/10.4230/LIPIcs.GISCience.2023.11

Images: PowerPoint Stock Images - Web Images and The Cultural Landscape Foundation