PROJECT ID: CAFS.20.79

YEAR: 1 of 3

PROJECT TITLE: Multi-regional evaluation of new machine learning algorithms for mapping tree species distribution and abundance

INVESTIGATOR(S): Kasey Legaard, University of Maine Center for Research on Sustainable Forests, and School of Forest Resources, kasey.legaard@maine.edu; Aaron Weiskittel, University of Maine Center for Research on Sustainable Forests; Larry Whitsel, University of Maine Advanced Computing Group; Erin Simons-Legaard, University of Maine School of Forest Resources

PROJECT DESCRIPTION:

Satellite remote sensing has the potential to satisfy many of the information needs of forest management through the provision of high-resolution maps of forest attributes. From the Sentinel program, for example, new 10-20 m resolution imagery is freely available each week for a given area, providing the potential to map forest conditions in near real-time across large, dynamic landscapes. Yet the integration of satellite-derived data into forest management planning has generally been slower than might be expected. Arguably, satellite remote sensing not yet delivered sufficient value over traditional sources of spatial information.

Satellite-derived maps are obtained from empirical models relating imagery to reference data, most commonly field measurements. Uncertainty in both field and remotely sensed data can cause biased models and severely detrimental patterns of map error. Bias originates from fundamental aspects of the remote sensing problem – comparing image pixels to a limited number of field measurements – and typically cannot be reduced or eliminated with traditional statistical approaches. A common pattern of error in estimates of continuous forest attributes (e.g., biomass, species relative abundance) is the over-estimation of low values and under-estimation of high values, caused by attenuation bias in predictive models (also referred to as regression dilution bias or regression to the mean). Systematic error in classification products (e.g., change detection maps) often results in excessive over- or under-estimation of mapped classes. Systematic error degrades map value as users are required to integrate additional data or take additional steps to mitigate excessive spatial uncertainty.

For the past several decades, machine learning (ML) algorithms have been adopted and refined to improve map accuracy. However, several decades of data and algorithm development in satellite remote sensing have not yielded robust solutions for eliminating systematic map error. Our research team (K. Legaard, E. Simons-Legaard, A. Weiskittel, L. Whitsel) has specifically targeted this problem and has developed a ML method that is capable of minimizing both total and systematic error in satellite-derived maps. Our mapping approach combines the strength of Support Vector Machines (SVMs) to model complex, nonlinear relationships based on limited training data, a common condition in forestry applications, with the adaptability of a multi-objective Genetic Algorithm (GA). The GA drives the evolution of SVMs to simultaneously increase accuracy and reduce or eliminate systematic error. Whereas other ML methods typically ignore prediction bias or provide limited and indirect means of influencing bias, control of bias is built into our ML framework through the direct minimization of systematic error.

We recently published a detailed description of our first-generation multi-objective ML algorithm, applied to the prediction of tree species relative abundance (percent of aboveground live biomass) from 30 m Landsat imagery at USFS FIA plot locations in northern Maine. Our approach compared favorably to a number of widely used alternatives including Random Forest and nearest neighbor methods, producing the least systematic error and yielding predictions that more closely matched FIA measurements. Here we propose to extend our multi-objective ML algorithms and applications in a number of important ways, including the incorporation of 10 m Sentinel 2 imagery and LiDAR data. Algorithm performance will be rigorously evaluated in four separate regions in an effort to provide a set of transparent benchmarks for further development and future comparison to alternative products. Successful workflows will be fully integrated into an automated image processing and multi-objective ML software system recently developed at UMaine to enable low-cost, high-volume data production. **Our goal is to support adaptive, data-driven forest management through the provision of high-quality, affordable spatial data products.**

OBJECTIVES:

1) Develop and validate maps of tree species relative abundance at 10 m pixel resolution;

2) Implement and validate alternative approaches to pooling species-level predictions into a satellite-derived canopy composition map;

3) Develop and validate annual time series of forest disturbance, 1985-present;

4) Implement and validate alternative approaches to aboveground biomass estimation:

5) Apply and validate species relative abundance, canopy composition, disturbance history, and biomass estimation methods across four large study areas in the northwest, northcentral, northeast, and southeast regions.

METHODS:

For Objective 1 and 2, species maps will be produced using newly developed automated image processing and multi-objective machine learning workflows (minimizing total and systematic error) developed at UMaine to support high-throughput processing in contrast to published work. We will specifically evaluate a number of methodological improvements over our previously published work, including: (a) use of Sentinel 2 imagery processed at 10 m pixel resolution; (b) integrated use of Sentinel 2 and Landsat 8 imagery, processed at 10 m resolution, as a possible means of improving predictions through improved resolution of phenological effects; (c) use of two-stage species modeling, chaining models of species relative abundance to models of species occurrence/non-occurrence to improve outcomes for rare species; and (d) evaluation of alternative approaches to combining species maps into a single canopy composition product.

For the remaining objectives, key methodological steps would include (a) production of Landsat forest disturbance time series (1985-present) using multi-objective ML to eliminate year-to-year prediction bias; (b) estimation of aboveground live biomass by multi-objective ML using Sentinel/Landsat imagery plus disturbance time series variables (e.g., time since last disturbance, metrics of disturbance intensity and recovery); (c) biomass estimation by kNN imputation of plot measurements based on multi-objective ML predictions of species relative abundance and disturbance history; and (d) biomass estimation by multi-objective ML integrating commercial LiDAR with satellite imagery. Reference data for the analysis would be provided by the US Forest Service, Forest Inventory and Analysis Program, while independent verification data and LiDAR data is available at selected NASA Carbon Monitoring System study sites in Maine, South Carolina, Minnesota, and Oregon.

MAJOR FINDINGS:

- Fully and semi-automated machine learning and mapping workflows for tree species relative abundance (% AGB) at 10 m pixel resolution
- Progress on species map validation in Maine suggests that mean absolute error typically ranges from 5-15% for Maine tree species, but we are currently comparing alternative map production workflows and alternative validation strategies
- Sentinel + Landsat integration, including automated warping/coregistration of imagery plus improved Landsat haze correction routines, with preliminary results suggesting limited benefits for species mapping
- Full implementation of a 2-stage species modeling approach, where species occurrence is mapped during stage 1 and relative abundance is mapped within areas of predicted occurrence during stage 2
- 2-stage species modeling reduces computation time and appears to improve outcomes for certain species by reducing the influence of low abundance samples on predictions at higher abundance, but validation is more complicated, and good outcomes requires control of error patterns during both stages
- New workflows and new software to integrate species maps into forest type and canopy composition maps, including both raster and vector output
- Polygon stand delineation from Sentinel-based species maps output, consistent with industry stand typing practices based on overstory composition
- New software for automated Sentinel image retrieval and integration
- New software for ML prediction post-processing, including automated and seamless mosaicking for large area applications
- New approach and new software to detect and delineate forest roads using Sentinel imagery, required to eliminate or suppress linear artifacts on species maps caused by intermixing of road/forest reflectance
- Workflows and software to produce high accuracy forest change detection maps from Sentinel and Landsat imagery (10 m and 30 m resolution), with balanced omission and commission error rates (accuracies generally 90-94%)
- Progress on new workflows and software to facilitate change detection over large areas and long time series, using semi-automated, semi-supervised techniques that maintain accuracy and control error across space and time
- New methods and software to improve disturbance and harvest delineation through multi-model or multi-map fusion

DELIVERABLES:

- New workflows and software:
    - Sentinel-based species modeling and mapping at 10 m pixel resolution
    - 2-stage species modeling using multi-objective machine learning to control error patterns in both stages
    - Automated and seamless map mosaicking using an efficient least-cost path algorithm to identify cutlines in areas of map overlap

- Automated Sentinel image retrieval from the European Space Agency's Copernicus API Hub
- Improved Landsat haze correction
- Forest roads detection and delineation, based on image morphology, adaptive thresholding, and object analysis
- Forest type and canopy composition mapping, including stand delineation, based on satellite-derived species maps
- Improved Sentinel- and Landsat-based disturbance and harvest mapping based on multi-model and multi-map fusion techniques

MEMBER COMPANY BENEFITS:

Key member benefits from year 1 project activities include the continued development, exploration, and evaluation of novel machine learning and geospatial integration methods developed at the University of Maine. Year 1 outcomes included the identification of multiple areas worthy of continued R&D, including control of error patterns during 2-stage species modeling, roads delineation, and multi-model/multi-map fusion to improve disturbance detection and harvest delineation. Progress has included the integration of all methods and workflows within high-throughput software developed to support low-cost data production across large areas. Validation of workflows and map products is underway in Maine, and validation at additional study sites will follow. Our intension is to establish the generalizability and value of the methods across conditions managed by forest industry in multiple regions. Members will ultimately benefit from the development and implementation of these methods as they should reduce the time and costs associated with current forest inventory and mapping methods, providing new opportunities to manage forests more efficiently.